

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**AN INVESTIGATION OF FEATURE SELECTION
METHODS FOR MACHINE LEARNING**

by

Buse Nur BALTACIOĞLU

September, 2024

İZMİR

AN INVESTIGATION OF FEATURE SELECTION METHODS FOR MACHINE LEARNING

**A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science in
Department of Statistics, Statistics Program**

**by
Buse Nur BALTACIOĞLU**

September, 2024

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**AN INVESTIGATION OF FEATURE SELECTION METHODS FOR MACHINE LEARNING**" completed by **Buse Nur BALTACIOĞLU** under supervision of and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....
Assoc. Prof. Dr. İdil YAVUZ

Supervisor

.....
Assoc. Prof. Dr. Özgül VUPA

Jury Member

.....
Asst. Dr. Ümit KUVVETLİ

Jury Member

.....
Prof.Dr. Okan FISTIKOĞLU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I am deeply grateful to Assoc. Prof. Dr. İdil YAVUZ, my esteemed supervisor, for her invaluable guidance, unwavering support and insightful perspectives at every stage of this project. Her mentorship has been instrumental in shaping my academic journey, providing not only valuable ideas and experiences, but also unwavering encouragement to push me forward.

I would like to express my sincere gratitude to the dedicated academics in the Department of Statistics, whose shared knowledge and experience have contributed greatly to my growth during both my undergraduate and postgraduate studies. Their guidance has been instrumental in deepening my understanding of the field.

I would also like to express my infinite gratitude to my precious mother, Gülayşe DEMİRCİ, and my sister, Hande BALTACIOĞLU. Their unwavering encouragement and support have been a constant source of enthusiasm and motivation in every situation.

Finally, I would like to pay tribute to my grandmother, Selatin DEMİRCİ, who sadly passed away during the process of writing this thesis. Her memory and spirit have been a guiding light, inspiring me to persevere and honour her legacy through my work.

Buse Nur BALTACIOĞLU

AN INVESTIGATION OF FEATURE SELECTION METHODS FOR MACHINE LEARNING

ABSTRACT

In the current era of big data, a multitude of fields, including medical, geological, financial, and social media, are regarded as sources of data. With the growing significance of data, an increasing number of organisations are generating a vast amount of data on a daily basis. One of the pivotal challenges in statistical science is to comprehend the interrelationships between multivariate data sets in the context of big data analysis. These relationships have been developed through the utilisation of applied statistical methods in a broad sense. The objective is to analyse the current situation by utilising the existing data, predict the features of future data and their relations with each other.

Due to the nature of high-dimensional data obtained from real-life scenarios, issues such as the lack of correlation between the features and the target variable, high correlation between the features, the number of observations being less than the number of features, or the number of features being so high that interpretation is hindered, may arise. These circumstances give rise to a number of negative consequences, including bias, a lack of interpretability, an increase in workload and a loss of time. In light of this, shrinkage and feature selection methods have been developed.

Feature selection methods relate to the identification of which features will be included in the model, while shrinkage is concerned with the reduction of coefficients (punishment) within the model. These methods have the effect of reducing the computation time, improving the learning performance, providing a better understanding of the learning model or data, and offering an effective solution to the over-fitting problem.

Feature selection is categorized as Supervised, Unsupervised, Semi-supervised, and

Reinforcement Learning based. Each category is subdivided into four subcategories: Filter, Wrapper, Embedded, and Hybrid. In this thesis, Filter, Wrapper and Embedded approaches created as Supervised Learning scenarios are discussed.

This thesis applies feature selection methods appropriate for the type of features and target variable to four different real medical data sets. The 10-fold cross validation prediction performances were compared by applying the most commonly used machine learning methods. The analysis was conducted using the R programming language.

Keywords: Machine learning, feature selection, filter methods, wrapper methods, embedded methods

MAKİNE ÖĞRENMESİ İÇİN ÖZNİTELİK SEÇİM YÖNTEMLERİNİN İNCELENMESİ

ÖZ

İçinde bulunduğumuz büyük veri çağında, tıbbi, jeolojik, finansal ve sosyal medya dahil olmak üzere çok sayıda alan veri kaynağı olarak kabul edilmektedir. Verinin artan önemiyle birlikte, giderek artan sayıda kuruluş günlük olarak çok büyük miktarda veri üretmektedir. İstatistik bilimindeki en önemli zorluklardan biri, büyük veri analizi bağlamında çok değişkenli veri setleri arasındaki karşılıklı ilişkileri anlamaktır. Bu ilişkiler, geniş anlamda uygulamalı istatistiksel yöntemlerin kullanılmasıyla geliştirilmiştir. Amaç, mevcut verileri kullanarak mevcut durumu analiz etmek, gelecekteki verilerin özelliklerini ve birbirleriyle olan ilişkilerini tahmin etmektir.

Gerçek hayat senaryolarından elde edilen yüksek boyutlu verilerin doğası gereği, özellikler ile hedef değişken arasında korelasyon olmaması, özellikler arasında yüksek korelasyon olması, gözlem sayısının özellik sayısından az olması veya özellik sayısının yorumlamayı engelleyecek kadar fazla olması gibi sorunlar ortaya çıkabilmektedir. Bu durumlar, yanlışlık, yorumlanabilirlik eksikliği, iş yükünde artış ve zaman kaybı gibi bir dizi olumsuz sonuca yol açmaktadır. Bunun ışığında, küçültme ve özellik seçme yöntemleri geliştirilmiştir.

Özellik seçimi yöntemleri, hangi özelliklerin modele dahil edileceğinin belirlenmesiyle ilgiliyken, küçültme ise model içindeki katsayıların azaltılmasıyla (cezalandırılmasıyla) ilgilidir. Bu yöntemler, hesaplama süresini azaltma, öğrenme performansını iyileştirme, öğrenme modelinin veya verinin daha iyi anlaşılmasını sağlama ve aşırı uyum sorununa etkili bir çözüm sunma etkisine sahiptir.

Özellik seçimi; Denetimli, Denetimsiz, Yarı denetimli ve Pekiştirmeli Öğrenme tabanlı olarak kategorize edilir. Her kategori kendi içinde dört alt kategoriye ayrılır: Filtre, Sarmalayıcı, Gömülü ve Hibrit. Bu tezde, Denetimli Öğrenme senaryoları

olarak oluşturulan Filtre, Sarmalayıcı ve G  m  l   yaklaşımlar ele alınmıştır.

Bu tezde,   zellik t  r  ne ve hedef deęiřkене uygun   zellik seim y  ntemleri d  rt farklı gerek tıbbi veri setine uygulanmıştır. En yaygın kullanılan makine   ğrenimi y  ntemleri uygulanarak 10 kat apraz doęrulama tahmin performansları karřılařtırılmıştır. Analiz R programlama dili kullanılarak gerekleřtirilmiştir.

Anahtar kelimeler: Makine   ğrenmesi,   znitelik seimi, filtre y  ntemleri, sarmalayıcı y  ntemler, g  m  l   y  ntemler



CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZ	vi
LIST OF FIGURES	xiv
LIST OF TABLES	xv
CHAPTER ONE – INTRODUCTION	1
1.1 Purpose, scope and objectives	1
1.2 Data sets	2
1.2.1 Obsessive Compulsive Disorder	2
1.2.2 Breast Cancer	5
1.2.3 Sleep Duration	9
1.2.4 SENIC.....	12
CHAPTER TWO – MACHINE LEARNING.....	16
2.1 Supervised Learning	16
2.1.1 Linear Regression.....	18
2.1.2 Logistic Regression	22
2.1.3 Multinomial Logistic Regression.....	25
2.1.4 Ordinal Logistic Regression	26
2.2 Unsupervised Learning.....	29
2.3 Semi-supervised and Reinforcement Learning	30
CHAPTER THREE – DATA PRE-PROCESSING AND PERFORMANCE METRICS.....	31
3.1 Standardization	31

3.2	K-Fold Cross Validation Technique	31
3.3	Performance Metrics	32
3.3.1	Accuracy	33
3.3.2	RMSE	34
CHAPTER FOUR – FEATURE SELECTION		35
CHAPTER FIVE – FILTER METHODS		37
5.1	Feature Ranking Selection Algorithms	38
5.1.1	Information Gain.....	38
5.1.2	Gain Ratio.....	39
5.1.3	Symmetrical Uncertainty	40
5.1.4	Gini Index	41
5.1.5	Chi-squared Test	42
5.1.6	Cramer’s V.....	44
5.1.7	OneR Algorithm	46
5.1.8	ANOVA.....	47
5.1.9	Kruskal Wallis Test.....	49
5.1.10	Pearson’s Correlation	51
5.1.11	Spearman’s Correlation	52
5.1.12	Relief Family of Algorithms.....	54
5.1.12.1	Relief Algorithm	54
5.1.12.2	ReliefF Algorithm.....	56
5.1.12.3	RReliefF Algorithm	58
5.2	Feature Subset Selection Algorithms.....	59
5.2.1	Correlation-based Feature Selection.....	59
5.2.2	Consistency Feature Subset Selection	61
CHAPTER SIX – WRAPPER METHODS.....		63
6.1	Subset Selection Methods	63
6.1.1	Best Subset Selection	63

6.1.2 Stepwise Selection.....	65
6.1.2.1 Sequential Forward Selection.....	65
6.1.2.2 Sequential Backward Elimination.....	67
6.2 Recursive Feature Elimination	69
CHAPTER SEVEN – EMBEDDED METHODS	71
7.1 Shrinkage Methods	71
7.1.1 Ridge Regression	72
7.1.2 The Lasso	74
7.1.3 Elastic Net	76
7.2 Tree-based Methods.....	78
7.2.1 Regression Trees	79
7.2.2 Classification Trees.....	81
7.2.3 Bagging	82
7.2.4 Random Forest.....	85
7.2.5 Boosting	86
7.3 Support Vector Machines	88
7.3.1 Linear Support Vector Machine.....	88
7.3.2 Radial Support Vector Machine.....	91
7.3.3 Polynomial Support Vector Machine.....	92
CHAPTER EIGHT – APPLICATION	95
8.1 Obsessive Compulsive Disorder Data Set Results.....	98
8.2 Breast Cancer Data Set Application Results	103
8.3 Sleep Duration Data Set Application Results.....	108
8.4 SENIC Data Set Application Results.....	112
CHAPTER NINE – CONCLUSION	117
REFERENCES.....	120

LIST OF FIGURES

	Page
Figure 1.2.1 Description of features for OCD data set.....	3
Figure 1.2.2 Descriptive statistics of the features for OCD data set.....	4
Figure 1.2.3 Graphical displays for the OCD data set.....	5
Figure 1.2.4 Description of features for the Breast Cancer data set.....	6
Figure 1.2.5 Descriptive statistics of features for the Breast Cancer data set.....	7
Figure 1.2.6 Histogram of features for the Breast Cancer data set	8
Figure 1.2.7 Description of features for the Sleep Duration data set	9
Figure 1.2.8 Descriptive statistics of features for the Sleep Duration data set	10
Figure 1.2.9 Bar chart of features for the Sleep Duration data set and histogram of the sleep duration target variable	11
Figure 1.2.10 Description of features for the SENIC data set	12
Figure 1.2.11 Descriptive statistics of features for the SENIC data set.....	13
Figure 1.2.12 Distribution and correlation of features for the SENIC data set	14
Figure 2.0.1 Taxonomy of machine learning	16
Figure 2.1.2 Supervised learning.....	17
Figure 2.1.3 Full model of the Sleep Duration data set.....	21
Figure 2.1.4 Full model of the Breast Cancer data set	24
Figure 2.1.5 Full model of the OCD data set	28
Figure 2.2.6 Unsupervised learning	30
Figure 3.2.1 K-fold cross validation technique.....	32
Figure 4.0.1 Taxonomy of feature selection.....	36
Figure 5.0.1 Taxonomy of filter methods	37

Figure 5.1.2	Information Gain with OCD data set	38
Figure 5.1.3	Gain Ratio with OCD data set	39
Figure 5.1.4	Symmetrical Uncertainty with OCD data set.....	40
Figure 5.1.5	Gini Index with Sleep Duration data set	42
Figure 5.1.6	Chi-squared Test with OCD data set.....	43
Figure 5.1.7	Cramer's V with OCD data set	45
Figure 5.1.8	OneR Algorithm with OCD data set	46
Figure 5.1.9	ANOVA with Sleep Duration data set	49
Figure 5.1.10	Kruskal Wallis Test with Sleep Duration data set	50
Figure 5.1.11	Pearson's Correlation with SENIC data set.....	52
Figure 5.1.12	Spearman's Correlation with SENIC data set	53
Figure 5.1.13	Relief Algorithm with Breast Cancer data set.....	56
Figure 5.1.14	ReliefF Algorithm with OCD data set.....	57
Figure 5.1.15	ReliefF Algorithm with Sleep Duration data set	59
Figure 5.2.16	Correlation-based Feature Selection with SENIC data set	60
Figure 5.2.17	Consistency Feature Subset Selection with OCD data set.....	62
Figure 6.0.1	Taxonomy of wrapper methods	63
Figure 6.1.2	Best Subset Selection with Breast Cancer data set	64
Figure 6.1.3	Sequential Forward Selection with Breast Cancer data set.....	66
Figure 6.1.4	Sequential Backward Elimination with Breast Cancer data set	68
Figure 6.2.5	Recursive Feature Elimination with SENIC data set.....	70
Figure 7.0.1	Taxonomy of embedded methods.....	71
Figure 7.1.2	Parameter tuning for Ridge Regression	73
Figure 7.1.3	Ridge Regression with Breast Cancer data set	73
Figure 7.1.4	Parameter tuning for The Lasso	75

Figure 7.1.5The Lasso with Breast Cancer data set	75
Figure 7.1.6Parameter tuning for Elastic Net	77
Figure 7.1.7Elastic Net with Breast Cancer data set	77
Figure 7.2.8Regression Tree with Sleep Duration data set.....	80
Figure 7.2.9Classification Tree with Breast Cancer data set	82
Figure 7.2.10Bagging with Breast Cancer data set	84
Figure 7.2.11Random Forest with Breast Cancer data set	86
Figure 7.2.12Parameter tuning for Boosting.....	87
Figure 7.3.11Linear Support Vector Machine with Sleep Duration data set.....	90
Figure 7.3.12Radial Support Vector Machine with Sleep Duration data set.....	92
Figure 7.3.13Parameter tuning for Polynomial Support Vector Machine	93
Figure 8.1.1Cross-validation accuracy values for the OCD data set using filter methods	99
Figure 8.1.2The optimal model chosen for the OCD data set.....	102
Figure 8.2.3Cross-validation accuracy values for the Breast Cancer data set using filter methods	104
Figure 8.2.4The optimal model chosen for the Breast Cancer data set.....	106
Figure 8.3.5Cross-validation RMSE values for the Sleep Duration data set using filter methods	109
Figure 8.3.6The optimal model chosen for the Sleep Duration data set	111
Figure 8.3.7Ranking the importance of features in the optimal model for the Sleep Duration data set.....	112
Figure 8.4.8Cross-validation RMSE values for the SENIC data set using filter methods	113
Figure 8.4.9The optimal model chosen for the SENIC data set	115

Figure 8.4.1 Ranking the importance for features in the optimal model for the SENIC data set.....	116
---	-----



LIST OF TABLES

	Page
Table 3.3.1 Confusion matrix	33
Table 8.0.1 Grouping of feature selection methods according to data type.....	96
Table 8.1.2 Model performances for the Obsessive Compulsive Disorder data set	101
Table 8.2.3 Model performances for the Breast Cancer data set.....	105
Table 8.3.4 Model performances for the Sleep Duration data set	110
Table 8.4.5 Model performances for the SENIC data set	114



CHAPTER ONE

INTRODUCTION

1.1 Purpose, scope and objectives

Feature selection or variable selection is an important research topic in knowledge discovery. There are major studies in Machine Learning feature selection areas. Most of these studies are Supervised Learning algorithms. This is because the Supervised Learning algorithms are regression-based. In the 1920s, R. A. Fisher created modern regression analysis (Aldrich, 2005) and later works on dimension reduction for regression were developed (Cook, 2007). About 20 years later, Hotelling conducted a study on the selection of variables to be used in the regression model (Hotelling, 1940). With the development of technology, the more advanced statistical works began to be carried out in the early 1960s. Since then, regression variable selection studies have evolved into supervised feature selection studies (Huang, 2015).

With the increasing importance of information obtained from data, the age of big data has begun. Data sizes keep increasing exponentially from day to day. High-dimensional data causes many issues such as bias, non-interpretation, increased workload, and loss of time. For this reason, feature selection methods have been developed.

Recently, feature selection studies have been conducted on topics such as bioinformatics (Saeys et al., 2007), cancer diagnosis (Abeel et al., 2010), neuroimaging (Mwangi et al., 2014), gene expression (Omara et al., 2018), fraud detection (Singh & Jain, 2019), credit rating (Yang et al., 2020), mood classification in music (Saari et al., 2010). When these studies in different fields are examined, it has been observed that explanatory models employed in these works generally achieve better performance when feature selection methods are applied.

This thesis is a comprehensive review of feature selection methods in Supervised Learning algorithms and aims to help determine and apply suitable feature selection

algorithms on four real medical data sets. By doing so, the results in this thesis will provide useful insights to researchers dealing with classification and regression problems.

1.2 Data sets

The analysis of the feature selection methods initially concentrated on the types of input and output (target) variables. Input and output variables were classified into four categories: categorical-categorical, numerical-categorical, categorical-numerical and numerical-numerical. In accordance with these categories, the OCD, Breast Cancer, Sleep Duration and SENIC data sets were selected, respectively.

1.2.1 Obsessive Compulsive Disorder

This data set examines the correlation between music habits and psychological well-being. It was created by Catherine Rasgaitis and obtained from Kaggle. The data set presents empirical research aimed at comprehending the interplay between music and mental health. The data was gathered from various sources, including Reddit forums and community libraries, with a focus on participant representation and inclusivity.

The participants were questioned about their musical background and listening habits. Additionally, the frequency of their listening to 16 different types of music was determined. Additionally, participants were asked to rate their experience of psychological health problems, such as anxiety, depression, insomnia, and obsessive-compulsive disorder (OCD), on a scale of 0 to 10. This study was conducted by taking OCD as the target variable and transforming it into three ordinal categories. The study also explored the potential impact of music therapy on patients' quality of life and psychological well-being.

OCD is characterised by recurring obsessions and compulsions, and interventions

such as music therapy may have a positive effect on relieving symptoms. This data set could prove to be a valuable resource for researchers investigating the correlation between musical habits and psychological well-being. It can aid in the development of novel approaches in both music therapy and psychological health. Figure 1.2.1 below shows the data set edited for this study.

OCD	Features	Interpretation
X1	Age	Age $\leq$ 25 : 1, Age $\leq$ 50 : 2, Age > 50 : 3.
X2	Primary streaming service	"Apple Music", "I do not use a streaming service.", "Spotify", "YouTube Music", "Other streaming service"
X3	Hours per day	Hours $\leq$ 3 : 1, Hours $\leq$ 6 : 2, Hours $\leq$ 9 : 3, Hours > 9 : 4.
X4	While working	Yes:1, No:2
X5	Instrumentalist	Yes:1, No:2
X6	Composer	Yes:1, No:2
X7	Exploratory	Yes:1, No:2
X8	Foreign languages	Yes:1, No:2
X9-24	Classical Country EDM Folk Gospel Hip-hop Jazz Kpop Latin Lofi Metal Pop RB Rap Rock Video game music	"Rarely", "Sometimes", "Never", "Very frequently"
X25	OCD	"low", "middle", "high"

Figure 1.2.1 Description of the features for OCD data set

The above Figure 1.2.1 provides descriptions of target variable with 24 features from 714 individuals in the Obsessive Compulsive Disorder data set. Appropriate methods were selected for this data set, where input and output features are categorical. The performance of appropriate Supervised Machine Learning methods was evaluated using the widely used 10-fold cross-validation.

Age			Streaming	Hours	While_working	Instrumentalist	Composer	Exploratory	Foreign_languages
1:500	Apple Music		: 51	1:447	Yes:563	Yes:228	Yes:124	Yes:509	Yes:396
2:174	I do not use a streaming service.:	70	2:182	No :151	No :486	No :590	No :205	No :318	
3: 40	Spotify		:452	3: 47					
	YouTube Music		: 91	4: 38					
	Other streaming service		: 50						
	Classical	Country		EDM		Folk	Gospel		Hiphop
Rarely	:252	Rarely	:231	Rarely	:188	Rarely	:214	Rarely	:133
Sometimes	:193	Sometimes	:105	Sometimes	:143	Sometimes	:140	Sometimes	:49
Never	:166	Never	:331	Never	:294	Never	:283	Never	:521
Very frequently:	103	Very frequently:	47	Very frequently:	89	Very frequently:	77	Very frequently:	11
	Jazz	Kpop		Latin		Lofi	Metal		Pop
Rarely	:244	Rarely	:169	Rarely	:168	Rarely	:204	Rarely	:189
Sometimes	:170	Sometimes	: 65	Sometimes	: 84	Sometimes	:160	Sometimes	:129
Never	:250	Never	:406	Never	:432	Never	:267	Never	:253
Very frequently:	50	Very frequently:	74	Very frequently:	30	Very frequently:	83	Very frequently:	143
	RB	Rap		Rock		Video_game_music	OCD		
Rarely	:206	Rarely	:211	Rarely	: 93	Rarely	:192	low	:488
Sometimes	:174	Sometimes	:187	Sometimes	:213	Sometimes	:182	middle:	132
Never	:220	Never	:190	Never	: 87	Never	:225	high	: 94
Very frequently:	114	Very frequently:	126	Very frequently:	321	Very frequently:	115		

Figure 1.2.2 Descriptive statistics of the features for OCD data set

The Figure 1.2.2 in summarises the features of OCD. The features 'Foreign_languages' reflects participants' listening habits for foreign language songs. The data shows that the majority of participants (59.4%) prefer foreign language songs, suggesting an interest in music from different cultures and openness to exploring various genres. Approximately 40.6% of individuals do not prefer listening to songs in a foreign language. This indicates that musical preferences may be influenced by personal and cultural backgrounds. This data can be utilised in various fields such as the music industry and radio stations to assess the popularity of songs in different languages and comprehend music consumption habits. Additionally, it may aid in the development of music education programs that promote language acquisition and cultural appreciation.

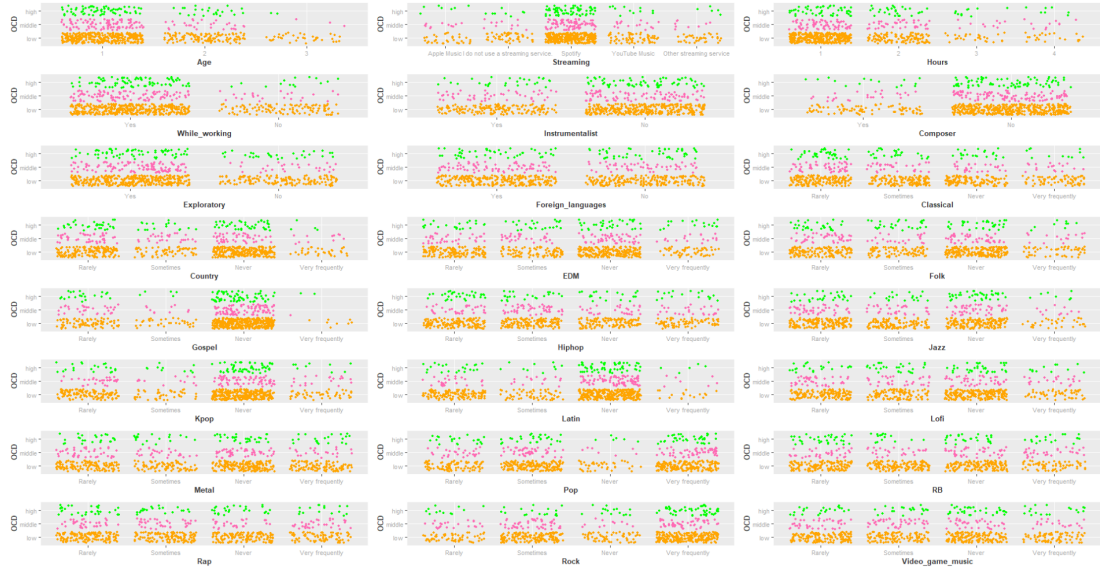


Figure 1.2.3 Graphical displays for the OCD dataset

Figure 1.2.3 presents graphs of OCD data set features based on OCD level. The orange, pink, and green dots represent the low, middle, and high categories of the OCD level, respectively. The data points corresponding to the categories are randomly distributed to show their density in the graph in Figure 1.2.3. This graph provides general information about the individuals who participated in the OCD survey. For instance, people aged 25 and under showed more interest in this survey. The most commonly used platform for listening to music is Spotify. Additionally, most of the participants prefer to listen to music for less than 6 hours and while working. They are not necessarily instrumentalists or composers, but they enjoy exploring new music. Pop and rock are the very frequently listened to genres, while country, gospel, Kpop, and Latin music are never listened to.

1.2.2 Breast Cancer

The Breast Cancer Wisconsin data set was obtained from hospitals at the University of Wisconsin. The data set is obtained by digitizing thin aspiration images of a Breast mass from patients. It consists of the features of the cell nucleus of 194 patients. The purpose of the data set is to estimate the recurrence of Breast Cancer in patients whose

cancer has been excreted (Mangasarian et al., 1995). The data includes 35 columns with ID and output. The $X_4 - X_{35}$ features define the measurements related to the shape properties of each cell nucleus, tumour size and lymph node status.

According to the World Health Organization, there are more than 2.3 million Breast Cancer patients worldwide each year. According to this statistic, Breast Cancer is the most common type of cancer among adults. Breast Cancer is the first or second leading cause of cancer death in women in 95% of countries (World Health Organization Regional Office for Africa, 2022).

Breast Cancer Wisconsin (Diagnostic)	Features	Interpretation
X1	ID number	
X2	Outcome	R = recur, N = nonrecur
X3	Time	
X4 - X33	Ten real-valued features are computed for each cell nucleus	
		radius (mean of distances from center to points on the perimeter)
		texture (standard deviation of gray-scale values)
		perimeter
		area
		smoothness (local variation in radius lengths)
		compactness ($\text{perimeter}^2 / \text{area} - 1.0$)
		concavity (severity of concave portions of the contour)
		concave points (number of concave portions of the contour)
		symmetry
		fractal dimension ("coastline approximation" - 1)
X34	Tumor size	diameter of the excised tumor in centimeters
X35	Lymph node status	number of positive axillary lymph nodes observed at time of surgery

Figure 1.2.4 Description of features for the Breast Cancer data set

Figure 1.2.4 above presents the features and explanations of the Breast Cancer data

set. Methods have been selected for this data set with numeric input and categorical output. The feature ID that represents a unique number for each patient has been deleted from the data before analysis was conducted. To estimate the performance of the methods, appropriate supervised Machine Learning methods have been applied and the commonly used 10-fold cross-validation was utilized for performance evaluation.

	mean <dbl>	sd <dbl>	median <dbl>	min <dbl>	max <dbl>	range <dbl>	skew <dbl>	kurtosis <dbl>
X3	46.94	34.52	39.50	1.00	125.00	124.00	0.51	-0.84
X4	17.40	3.17	17.29	10.95	27.22	16.27	0.32	-0.31
X5	22.30	4.34	21.80	10.38	39.28	28.90	0.54	0.83
X6	114.78	21.43	113.70	71.90	182.10	110.20	0.39	-0.16
X7	969.09	353.16	929.10	361.60	2250.00	1888.40	0.72	0.39
X8	0.10	0.01	0.10	0.07	0.14	0.07	0.42	0.35
X9	0.14	0.05	0.13	0.05	0.31	0.27	0.60	0.38
X10	0.16	0.07	0.15	0.02	0.43	0.40	0.67	0.65
X11	0.09	0.03	0.09	0.02	0.20	0.18	0.69	0.66
X12	0.19	0.03	0.19	0.13	0.30	0.17	0.75	0.91
X13	0.06	0.01	0.06	0.05	0.10	0.05	0.97	1.71
X14	0.60	0.31	0.54	0.19	1.82	1.63	1.25	1.58
X15	1.27	0.53	1.18	0.36	3.50	3.14	1.27	2.12
X16	4.26	2.19	3.78	1.15	13.28	12.13	1.26	1.54
X17	70.29	48.02	58.45	13.99	316.00	302.01	1.80	4.50
X18	0.01	0.00	0.01	0.00	0.03	0.03	3.84	25.30
X19	0.03	0.02	0.03	0.01	0.14	0.13	2.03	7.04
X20	0.04	0.02	0.04	0.01	0.14	0.13	1.71	4.37
X21	0.02	0.01	0.01	0.01	0.04	0.03	1.26	2.63
X22	0.02	0.01	0.02	0.01	0.06	0.05	2.09	4.85
X23	0.00	0.00	0.00	0.00	0.01	0.01	1.64	3.83
X24	20.99	4.25	20.52	12.84	35.13	22.29	0.61	0.23
X25	30.18	6.07	30.30	16.67	49.54	32.87	0.38	0.26
X26	140.14	28.83	136.50	85.10	232.20	147.10	0.65	0.23
X27	1401.76	587.04	1295.00	508.10	3903.00	3394.90	1.20	1.99
X28	0.14	0.02	0.14	0.08	0.22	0.14	0.47	1.01
X29	0.36	0.17	0.35	0.05	1.06	1.01	1.15	2.15
X30	0.44	0.17	0.40	0.02	1.17	1.15	0.80	1.65
X31	0.18	0.05	0.18	0.03	0.29	0.26	-0.13	0.10
X32	0.32	0.07	0.31	0.16	0.66	0.51	1.13	2.45
X33	0.09	0.02	0.09	0.06	0.21	0.15	1.49	4.51
X34	2.87	1.95	2.50	0.40	10.00	9.60	1.70	2.96
X35	3.21	5.48	1.00	0.00	27.00	27.00	2.22	4.75

Figure 1.2.5 Descriptive statistics of features for the Breast Cancer data set

The Figure 1.2.5 presents descriptive statistics on the features of Breast Cancer. For instance, when examining the X_{27} feature, the mean value is 1401.76 and the standard deviation is 587.04, with a median value of 1295.00. These values indicate that the X_{27} feature typically has high values. The feature has a wide data range, with a minimum

value of 508.10 and a maximum value of 3903.00.

The distribution is right-skewed (positive skewness), indicating that there are more values at the right end. The distribution is right-skewed, indicating that there are more values at the right end. The skewness and kurtosis coefficients were taken into account to determine this. The distribution is right-skewed, indicating that there are more values at the right end. Furthermore, the kurtosis coefficient has a high value of 1.99, indicating a significant deviation from the mean. To summarise, feature X_{27} exhibits a wide, right-skewed distribution with high variance. This information enhances our understanding of the feature's properties.

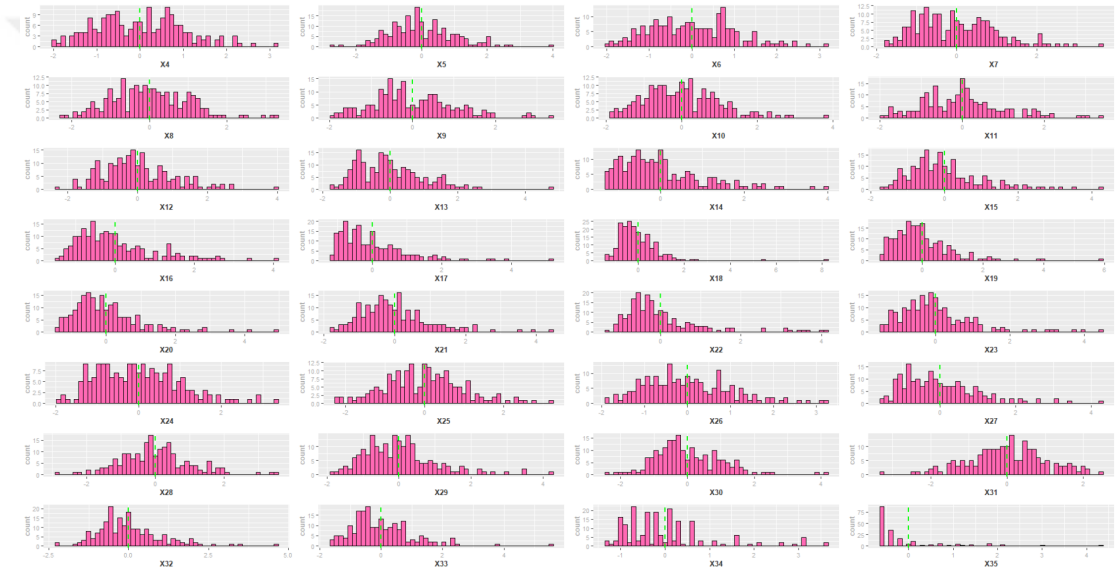


Figure 1.2.6 Histogram of features for the Breast Cancer data set

Figure 1.2.6 displays histograms of standardized measurements of the shape features, tumour size and lymph node status of each nucleus of the $X_4 - X_{35}$ Breast Cancer data set. The histograms reveal that the second measurements of the cell nucleus, namely $X_{14} - X_{23}$, are right-skewed and contain extreme values. Additionally, the symmetric and fractal dimension features also exhibit extreme values and have a right-skewed structure. Furthermore, the study participants tumour sizes (X_{35}) ranged from 0.4 cm to 10 cm, with an average of 2 cm and a standard deviation of approximately 2.9 cm. The tumour-size histogram has been standardized, and it displays extreme values when observed.

1.2.3 Sleep Duration

A subset of the Pima Indian Diabetes database (Tigga & Garg, 2020) contains the findings from a survey conducted with 849 participants over the age of 18. The purpose of the subset of this data set that we consider deals with determining the features that affect sleep duration and estimating sleep duration. Sleep duration and habits are closely related to health. Several studies have shown that short (generally less than 7 hours) or long (over 8 hours) sleep duration increases the risk of loss of life by 17-24% (Hublin et al., 2007).

Sleep	Features	Interpretation
X1	Age	18 or above
X2	Gender	Male - Female
X3	Family history with diabetes	Yes - No
X4	Diagnosed with high blood pressure	Yes - No
X5	PhysicallyActive	"None"
		"Less than half an hour"
		"More than half an hour"
		"One hour or more"
X6	BMI	"Underweight"
		"Normal weight"
		"Overweight"
		"Obese"
X7	Smoking	Yes - No
X8	Alcohol consumption	Yes - No
X9	Hours of sleep	Numeric
X10	SoundSleep	SoundSleep ≤ 4 : 0,
		SoundSleep ≤ 5.5 : 1,
		SoundSleep ≤ 7 : 2,
		SoundSleep > 7 : 3.
X11	Regular intake of medicine?	Yes - No
X12	Junk food consumption	Yes - No
X13	Stress	"Not at all"
		"Sometimes"
		"Often"
		"Always"
X14	Blood pressure level	"Low"
		"Normal"
		"High"
X15	Number of pregnancies	0 - 1 - 2 - 3
X16	Gestation diabetes	Yes - No
X17	Frequency of urination	"Not much"
		"Quite much"
X18	Diabetic	"Diabetic"
		"Non diabetic"

Figure 1.2.7 Description of features for the Sleep Duration data set

Figure 1.2.7 above presents the features and explanations of the Sleep Duration data set. Methods have been selected for the Sleep Duration data set which consists of categorical input and numerical output. In order to estimate the performance of several supervised Machine Learning methods, 10-fold cross-validation was used.

Age	Gender	Family_Diabetes	highBP	PhysicallyActive	BMI	Smoking	Alcohol	Sleep
less than 40:417	Male :519	no :441	no :640	one hr or more :191	Normal weight:386	no :764	no :680	Min. : 4.000
40-49 :147	Female:330	yes:408	yes:209	less than half an hr:291	Obese :181	yes: 85	yes:169	1st Qu.: 6.000
50-59 :148				none :129	Overweight :230			Median : 7.000
60 or older :137				more than half an hr:238	Underweight : 52			Mean : 6.939
								3rd Qu.: 8.000
								Max. :11.000
SoundSleep	RegularMedicine	JunkFood	Stress	BPLevel	Pregancies	Pdiabetes	UriationFreq	Diabetic
0:258	no :541	often :174	always : 61	low : 25	0:698	0 :836	not much :611	no :601
1:141	yes:308	very often : 52	sometimes :511	normal:631	1: 28	yes: 13	quite often:238	yes:248
2:304		occasionally:623	very often:150	high :193	2: 63			
3:146			not at all:127		3: 60			

Figure 1.2.8 Descriptive statistics of features for the Sleep Duration data set

Figure 1.2.8 presents a summary of the features of the Sleep Duration data set. The participants were divided into four age groups: under 40 years old, 40-49 years old, 50-59 years old, and 60 years old and above. The age group with the highest frequency is those under 40 years of age, while the age group with the lowest frequency is those aged 60 and over. The data shows that 59.8% of the participants are male. Participants without a family history of diabetes outnumber those with such a history. More than half of the participants engage in less than half an hour of physical activity. The majority of participants have a normal BMI. Although the proportion of participants who smoke is low, the proportion of participants who consume alcohol is even lower.



Figure 1.2.9 Bar chart of features for the Sleep Duration data set and histogram of the sleep duration target variable

Based on the data presented in Figure 1.2.9, it can be concluded that the majority of participants were younger than 40 years of age. Gender distribution shows that male participants are higher. Additionally, almost half of the participants reported a family history of diabetes. Regarding hypertension (high blood pressure), it was observed that the majority of participants did not have it and the blood pressure level was generally normal. Body mass index (BMI) distribution showed that the majority of participants were normal or slightly overweight. When smoking and alcohol consumption habits were examined, it was determined that the majority of participants almost did not use either of them. When deep sleep levels are examined, the majority of people have deep sleep between 5.5 and 7 hours and less than 4 hours. Almost 60% of the participants do not use medication regularly. It was observed that the majority of the participants consumed junk food occasionally. Analysis of stress levels revealed that a significant proportion of participants sometimes experienced stress. Additionally, after examining the skewness and kurtosis values for the target variable Sleep, it is seen that the skewness value is close to zero, which shows that the Sleep variable is closer to normal distribution. Based on the results of the Shapiro-Wilk normality test, it can be concluded that the Sleep variable follows a normal distribution because $p - value = 0.0782$ is greater than $\alpha = 0.05$.

1.2.4 SENIC

The data set considers 9 exploratory features to explain the risk of infection of 113 hospitals participating in the SENIC project. According to the US, hospital infection causes at least 30.000 deaths a year, increases the length of hospitalization by 7-10 days, and thus costs an additional 5-10 billion USD annually (Gül et al., 2022).

Senic	Features	Interpretation
X1	ID	Unique hospital identifier
X2	Length of Stay	Average length of stay of all patients in hospital (days).
X3	Age	Average age of patients (in years).
X4	Infection Risk	Estimated probability of acquiring infection in hospital (percent).
X5	Routine Culturing Ratio	Ratio of number of cultures performed to the number of patients without signs or symptoms of hospital-acquired infection, times 100.
X6	Routine Xray Ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100.
X7	Number of Beds	Number of beds in hospital during study period.
X8	Medical School	Indicator of whether the hospital is associated with a medical school (1 = Yes, 2 = No).
X9	Region	Indicator of the geographic region for hospital (1 = NE, 2 = NC, 3 = S, 4 = W).
X10	Average Census	Number of patients per day in hospital during study period.
X11	Number Nurses	number of full-time equivalent registered and licensed practical nurses during study period (number of full time plus one half the number of part time).
X12	Available Facilities	Percent of 35 potential facilities and services that are provided by the hospital.

Figure 1.2.10 Description of features for the SENIC data set

Figure 1.2.10 above presents the features and explanations of the SENIC data set. The SENIC data set that consists of numeric input and numeric output, appropriate

supervised Machine Learning methods have been selected. For illustration purposes the two categorical features, X_8 and X_9 , have not been used as candidate features for these models so that the numerical-numerical setting could be studied. To compare the performance of the methods, again 10-fold cross-validation was used.

	Mean	Std.Dev	Min	Median	Max	IQR	Skewness	Kurtosis
X10	191.371681	153.759564	20.0	143.00	791.00	184.00	1.342984491	1.52102685
X11	173.247788	139.265390	14.0	132.00	656.00	152.00	1.342382454	1.35491643
X12	43.159292	15.200861	5.7	42.90	80.00	22.90	0.072223045	-0.49752105
X2	9.648319	1.911456	6.7	9.42	19.56	2.13	2.014314486	7.48380118
X3	53.231858	4.461607	38.8	53.20	65.90	5.30	-0.101237983	0.89725776
X4	4.354867	1.340908	1.3	4.40	7.80	1.50	-0.116597463	0.06674647
X5	15.792920	10.234707	1.6	14.10	60.50	11.90	1.567681306	3.62178044
X6	81.628319	19.363826	39.6	82.30	133.50	24.60	0.007669835	-0.32915721
X7	252.168142	192.842687	29.0	186.00	835.00	206.00	1.342231775	1.09930587

Figure 1.2.11 Descriptive statistics of features for the SENIC data set

Figure 1.2.11 shows important findings. The X_2 feature, which represents the length of hospital stay, has an average value of approximately 10 days with a standard deviation of approximately 2, indicating a spread around the mean. The patients' average age in the X_3 feature is approximately 53 years, with a standard deviation of 4, indicating a certain spread in the age distribution.

The feature X_4 expresses the risk of infection in the hospital as a percentage. On average, the risk of infection is 4.35%, with a standard deviation of 1.34%. This suggests that the risk of infection in the hospital may vary in certain situations. The routine culturing rate (X_5) is the proportion of cultures performed among patients who do not exhibit specific symptoms. The mean culturing rate was 15.79, with a standard deviation of 10.23, indicating a wide distribution of measures across hospitals.

X_6 refers to the routine x-ray rate, with a mean ratio of 81.63 and a standard deviation of 19.36, suggesting varying procedures and policies among hospitals. These findings offer a significant viewpoint for comprehending the variety and range of hospital services.

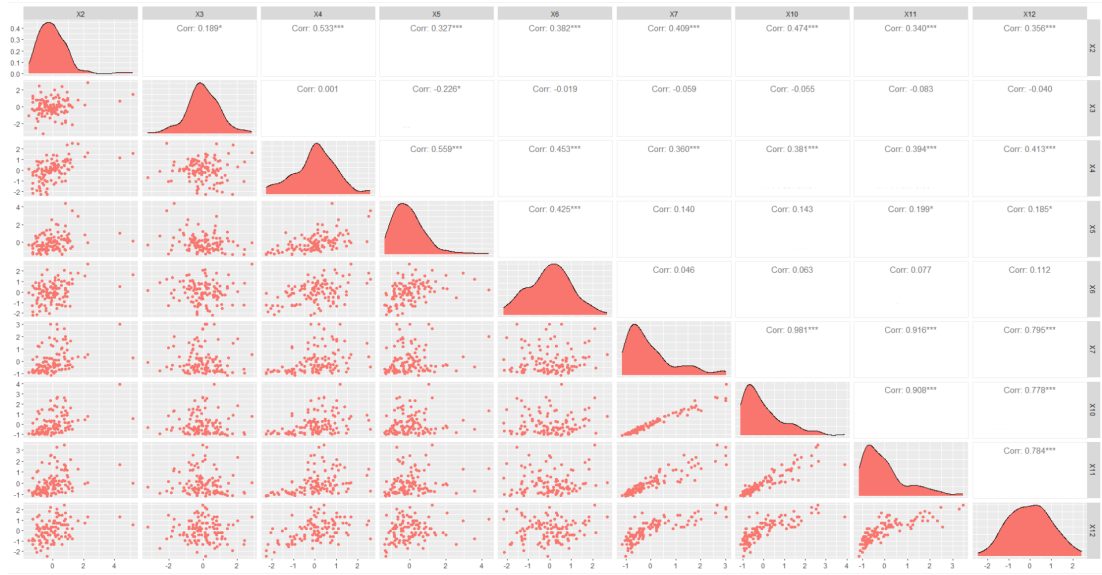


Figure 1.2.12 Distribution and correlation of features for the SENIC data set

After examining Figure 1.2.12, for instance, it is clear that the distributions of the X_{10} (Average Census) and X_{11} (Number Nurses) features are skewed to the right and have higher kurtosis than a normal distribution. These skewness and kurtosis values indicate that these features do not follow a normal distribution. The right skewness of these features suggests that higher-than-average values occur more frequently than below-average values.

However, after examining the skewness and kurtosis values for the target variable X_4 (Infection Risk), it is evident that the skewness value is almost zero, indicating that the variable X_4 is closer to a normal distribution. Based on the results of the Shapiro-Wilk normality test, it can be concluded that the X_4 variable follows a normal distribution, as the $p - value = 0.1339$ is greater than $\alpha = 0.05$.

Figure 1.2.12 presents correlation analyses, which are important for understanding relationships between features. For instance, the high correlation (0.9809) between X_7 (Number of Beds) and X_{10} (Average Population) indicates a high positive relationship between the density of a large hospital and the number of beds. Similarly, the correlation between X_{10} and X_{11} (0.9079) indicates that busier hospitals employ more nurses. However, the data shows a negative correlation (-0.2258) between X_5

(Routine Culture Rate) and X_3 (Age), indicating that the routine culture rate decreases as the average age of the hospital increases.

These findings aid in understanding the relationships and distributions among the features in the data set. This information can be used to analyze important issues such as hospital operations or service quality.



CHAPTER TWO

MACHINE LEARNING

The Machine Learning literature developed as a multidisciplinary field. Particularly, the main fields that contribute are mathematics, statistics, cybernetics and computer science. Machine Learning is often confused with artificial intelligence in terms of learning and decision-making, but it is a sub-topic of artificial intelligence. They share a common development history until 1957. In 1957, with the invention of the perceptron in a study led by psychologist Frank Rosenblatt at Cornell University (Fradkov, 2020) the two fields separated. Then the historical process of Machine Learning started and it was divided into Supervised, Unsupervised, Semi-supervised and Reinforcement branches (Géron, 2022). The Machine Learning algorithms are presented in Figure 2.0.1 in the form of a general taxonomy.

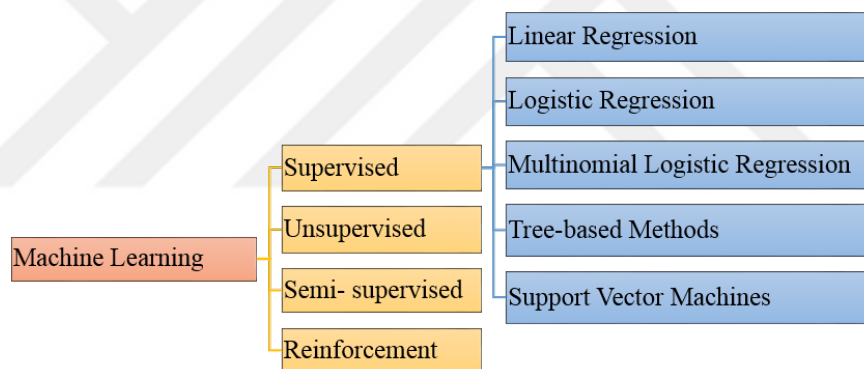


Figure 2.0.1 Taxonomy of machine learning

2.1 Supervised Learning

Supervised Learning is a paradigm that is widely used in the field of Machine Learning. In this approach, a Machine Learning model is trained using labelled data. Labeled data refers to a data set where each data sample is characterised by features and is also assigned a target variable (label). Subfields such as regression and classification define the core algorithms that Supervised Learning employs.

Classification in the field of Supervised Learning involves predicting category labels

or classes for a given input. The input data can be any type of data, such as images, text, or numerical data, and the output is a discrete set of possible labels or classes. Common classification algorithms include Logistic Regression, Decision Trees, and Support Vector Machines (Hastie et al., 2009).

In contrast, regression is a Supervised Learning problem that aims to predict a numerical value for a given input. In regression problems, the output variable is numerical, and the goal is to learn the relationship between the input and output variables. Linear Regression, Polynomial Regression, and Neural Networks are all examples of regression algorithms that are commonly used (Bishop & Nasrabadi, 2006).

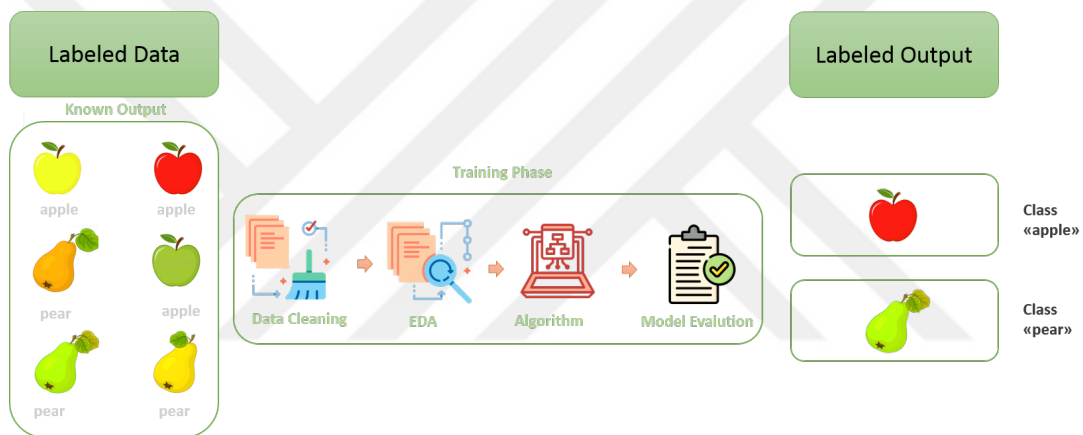


Figure 2.1.2 Supervised learning

Figure 2.1.2 presents a flow diagram illustrating the process of training a Supervised Learning model using a data set comprising images of apples and pears. The initial step, namely the collection and labelling of apple and pear images, is then highlighted. Subsequently, the data cleaning step is employed to identify and rectify any missing or anomalous values within the data set. Subsequently, in the exploratory data analysis (EDA) phase, the most significant features are identified by examining the structure, distributions and relationships of the data set. Subsequently, the algorithm selection step illustrates the process of selecting the most suitable algorithm for the project and training the model. The model is trained on the training set, where it learns the features of apple and pear images. Its performance is then evaluated on the test set. Finally, it is

stated that the test set is employed to assess the efficacy of the model on genuine data after it has been trained. This step is crucial for assessing the model's performance in real-world scenarios.

In the following sub-sections brief introductions to Linear Regression and Logistic Regression models will be given and the shrinkage models, tree-based approaches and support-vector machines will be discussed in Chapter 7.

2.1.1 Linear Regression

Linear Regression is a statistical model used for explaining the relationship between a target variable and one or more independent features. The goal of Linear Regression is to find the most appropriate straight line that defines the relationship between features and the target variable (James et al., 2013). The regression line is defined by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon,$$

where Y is the target variable, $X_1, X_2, \dots, X_p$ are the features, β_0 is the intercept, and β_i are the coefficients of the features for $i = 1, \dots, p$. The term ϵ represents the discrepancies between the model predicted values of $\hat{Y}$ and the actual observed values of Y .

In the context of Linear Regression, it is important to consider some basic assumptions to ensure the reliability and validity of model results. These basic assumptions cover many different aspects. First, the linear hypothesis posits that the relationship between the features and target variable must be linear. Second, error independence asserts that the errors, that is, the difference between the actual and predicted values, must be uncorrelated, ensuring that the error of each data point is not influenced by other factors. The homogeneity assumption states that the variance of these errors will be constant across different levels of the features. Additionally, normality of errors assumes that errors follow a normal distribution, which can affect

the performance of some statistical tests. Finally, the multicollinearity assumption suggests that features should exhibit minimal correlation with each other to easily distinguish their individual effects on the target variable (James et al., 2013).

To calculate the coefficients for prediction in Linear Regression, one commonly used method is Ordinary Least Squares (OLS). OLS aims to minimize the Sum of Squared Errors (SSE), as mentioned earlier. This minimization process leads to the estimation of the coefficients, such as $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$. Once the model is trained, these coefficients are used in the regression equation to make predictions (James et al., 2013). Predictions are calculated by substituting the values of the features into the equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p,$$

where $\hat{Y}$ represents the predicted value of the target variable. These coefficients, derived from the training process, enable the model to predict the outcome based on new data, making Linear Regression a valuable tool in predictive analytics.

The SSE (Sum of Squared errors) measurement is used to determine the difference between the actual and predicted values of the target variable in Linear Regression. It is also known as the sum of squared residual (SSR). Mathematically, the SSE is calculated as follows:

$$SSE = \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j X_{ij} \right)^2. \quad (2.1)$$

The SSE is an important measure of the fit of the Linear Regression model. A lower value of SSE indicates a better fit of the model, meaning that the model is able to explain more of the variance in the target variable.

However, R-squared (R^2) is a crucial metric that gauges the efficacy of a regression model. This metric indicates the extent to which the independent features can be used to predict the target variable. The R^2 value is defined as a ratio between 0 and 1, with a value closer to 1 indicating a greater fit between the model and the data. A high R^2

value indicates that the model explains a significant proportion of the variance of the target variable, whereas a low value may suggest that the model does not fit the data adequately.

The R^2 value is calculated by subtracting the ratio of the sum of the squares of the differences between the model predicted values and the actual values (SSE) and the ratio of the sum of the squares of the differences between the mean value of the target variable and the actual values (SST) from 1. This ratio is used to assess the degree of fit between the model and the data.

Consequently, as the sum of squared errors (SSE) diminishes, the R^2 value increases. This indicates that the model is able to explain a greater proportion of the variance in the data set with a smaller number of errors

For illustration purposes a multiple Linear Regression model is applied to the Sleep Duration data set where the outcome is sleep duration and all the features where some are categorical and some are numerical are included in the model. The results can be seen in Figure 2.1.3.

```

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-1.9167 -0.6273 -0.0013  0.6037  3.3141

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.23295    0.27533   22.638 < 2e-16 ***
`Age40-49`     0.05318    0.10225    0.520 0.603101
`Age50-59`     0.41294    0.11668    3.539 0.000424 ***
`Age60 or older` -0.28712    0.14040   -2.045 0.041169 *
GenderFemale    0.09227    0.08752    1.054 0.292083
Family_Diabetesyes 0.11563    0.07167    1.613 0.107038
highBPyes      0.34699    0.11841    2.930 0.003479 **
`PhysicallyActiveless than half an hr` -0.19311    0.09699   -1.991 0.046814 *
PhysicallyActivenone -0.01507    0.11794   -0.128 0.898339
`PhysicallyActivemore than half an hr` 0.20905    0.09786    2.136 0.032961 *
BMIObese       0.24344    0.09634    2.527 0.011701 *
BMIOverweight  0.09365    0.08631    1.085 0.278190
BMIUnderweight -0.26837    0.15076   -1.780 0.075432 .
Smokingyes     -0.31374    0.13270   -2.364 0.018296 *
Alcoholyes     0.10277    0.10165    1.011 0.312295
SoundSleep1    0.19656    0.10860    1.810 0.070667 .
SoundSleep2    0.70480    0.09334    7.551 1.16e-13 ***
SoundSleep3    2.30969    0.11179   20.661 < 2e-16 ***
RegularMedicineyes -0.03897    0.09671   -0.403 0.687117
`JunkFoodvery often` 0.20955    0.16482    1.271 0.203954
JunkFoodoccasionally -0.07985    0.09060   -0.881 0.378378
Stresssometimes 0.23107    0.15080    1.532 0.125835
`Stressvery often` 0.11769    0.16208    0.726 0.467961
`Stressnot at all` 0.11345    0.16973    0.668 0.504050
BPLevelnormal  -0.27814    0.20709   -1.343 0.179613
BPLevelhigh    -0.80627    0.24643   -3.272 0.001114 **
Pregancies1    -0.02510    0.20369   -0.123 0.901961
Pregancies2     0.47164    0.14831    3.180 0.001528 **
Pregancies3    -0.38269    0.15695   -2.438 0.014969 *
Pdiabetesyes   -0.05847    0.30585   -0.191 0.848449
`UriationFreququite often` 0.16555    0.08497    1.948 0.051705 .
Diabeticyes    -0.02635    0.10688   -0.246 0.805372
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9452 on 817 degrees of freedom
Multiple R-squared:  0.4604,    Adjusted R-squared:  0.44
F-statistic: 22.49 on 31 and 817 DF,  p-value: < 2.2e-16

```

Figure 2.1.3 Full model of the Sleep Duration data set

As it can be seen from Figure 2.1.3, age, blood pressure physical activity level, BMI, smoking, sleeping sound and number of pregnancies are the features that have statistically significant effects on sleep duration and collectively these features explain 44% of variation amongst the durations ($\alpha = 0.05$).

2.1.2 Logistic Regression

Logistic Regression is a statistical method used to model the relationship between a binary target variable and one or more independent features. Logistic Regression aims to predict the probability distribution of a binary outcome based on the features (Smith & Johnson, 2020).

To model the relationship between the success probability of the binary outcome and the features, the Logistic Regression model utilizes a logistic function (also known as a sigmoid function). The logistic (or logit) function takes any input value and outputs a value between 0 and 1, which represents the predicted success probability of the binary outcome.

The Logistic Regression model estimates the coefficients of the features that maximize the likelihood of observing the observed binary outcome. These estimates can be obtained via the Newton-Raphson algorithm. These coefficients can be used for calculating the odds ratio of the binary outcome associated with each features (Kleinbaum & Klein, 2002). The model has the following mathematical form:

$$P(Y = 1|X_1, \dots, X_p) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}},$$

where $P(Y = 1|X_1, \dots, X_p)$ is the probability of the target variable (Y) taking on the value 1 given the values of the features $(X_1, X_2, \dots, X_p)$, e the Euler number is the base of the natural logarithm, β_0 is the intercept and $\beta_1, \beta_2, \dots, \beta_p$ are the coefficients of the features.

For a model with feature p, the log odds of category Y=1 compared to the reference category Y=0 are modeled as follows, since $P(Y = 0) = 1 - P(Y = 1)$:

$$\ln\left(\frac{P(Y = 1|X_1, \dots, X_p)}{1 - P(Y = 1|X_1, \dots, X_p)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p,$$

where, represents the logit link function of the Logistic regression model. The odds ratio comparing category to the reference category can be calculated by exponentiating

the coefficient (e^{β}).

The coefficients estimated from the model provide information about the association between the features and each category of the target variable, relative to the reference category.

For illustration purposes a Logistic Regression model is applied to the Breast Cancer data where the outcome is whether the cancer is recurrent and all the features that are numerical are included in the model. The results can be seen in Figure 2.1.4.



```

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.92002  -0.40158  -0.09114  -0.00037   2.89865

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3098     0.6020  -5.498 3.84e-08 ***
X3            -2.8166     0.5737  -4.909 9.15e-07 ***
X4           -13.6711    15.6298  -0.875  0.38175
X5            -2.4101     1.0565  -2.281  0.02253 *
X6             1.7253    14.8754   0.116  0.90767
X7            12.2293     6.6484   1.839  0.06585 .
X8             3.7998     1.2590   3.018  0.00254 **
X9            -0.9104     2.1860  -0.416  0.67709
X10           -1.3074     2.2104  -0.591  0.55419
X11           -2.0072     2.0660  -0.972  0.33127
X12            1.2672     0.8133   1.558  0.11924
X13           -1.6644     1.4031  -1.186  0.23555
X14           -8.2564     3.8106  -2.167  0.03026 *
X15           -3.4195     1.0040  -3.406  0.00066 ***
X16           11.8417     3.8509   3.075  0.00210 **
X17           -2.2933     2.4704  -0.928  0.35325
X18            1.7076     1.0119   1.688  0.09151 .
X19            4.8214     1.8865   2.556  0.01060 *
X20           -5.6640     2.2156  -2.556  0.01057 *
X21           -2.2548     1.1457  -1.968  0.04905 *
X22            2.1371     1.2142   1.760  0.07839 .
X23            1.2544     1.6863   0.744  0.45694
X24           19.7507     8.0544   2.452  0.01420 *
X25            2.9254     1.3159   2.223  0.02620 *
X26           -8.7904     4.6889  -1.875  0.06083 .
X27           -9.1683     5.9883  -1.531  0.12576
X28           -0.6180     1.0041  -0.616  0.53821
X29           -5.2798     2.1202  -2.490  0.01276 *
X30            5.5372     2.0293   2.729  0.00636 **
X31           -0.4443     1.0871  -0.409  0.68275
X32           -2.3460     1.4363  -1.633  0.10239
X33            1.0511     1.8067   0.582  0.56070
X34           -0.2856     0.3179  -0.898  0.36901
X35            0.9318     0.3662   2.544  0.01095 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212.519  on 193  degrees of freedom
Residual deviance:  96.438  on 160  degrees of freedom
AIC: 164.44

Number of Fisher Scoring iterations: 7

```

Figure 2.1.4 Full model of the Breast Cancer data set

The Logistic Regression model including all the features of the Breast Cancer data set is given in Figure 2.1.4. 14 features were statistically significant at $\alpha = 0.05$ for explaining the variation in the probability of having recurrent Breast Cancer. For example the feature X_{15} : The estimated texture2 coefficient of the cancer cell is approximately -3.42. The odds ratio for this feature is then 0.03 meaning that a 1 unit increase in X_{15} results in a 97% decrease in the odds of cancer recurrence.

2.1.3 Multinomial Logistic Regression

A statistical method that models the relationship between a categorical target variable with three or more categories and one or more features is called Multinomial Logistic Regression. This method extends Binary Logistic Regression and can be used when there are more than two categories in the target variable (Kleinbaum & Klein, 2002).

In Multinomial Logistic Regression, the target variable is modeled using a set of Binary Logistic Regression equations, where each equation compares one category to a reference category. The equations are estimated using maximum likelihood estimation, and the coefficients of the features are estimated separately for each equation.

The probability of each category is modeled as a function of the features using the softmax function, which maps the input features to a probability distribution over the categories. In the case where the probability $P(Y = j|X_1, \dots, X_p)$ is given when Y has more than two possible levels, namely 1,2,...,J, it can be expressed as follows:

$$P(Y = j|X_1, \dots, X_p) = \frac{e^{\beta_{j0} + \beta_{j1}X_1 + \dots + \beta_{jp}X_p}}{1 + \sum_{k=1}^{J-1} e^{\beta_{0k} + \beta_{1k}X_1 + \dots + \beta_{pk}X_p}}.$$

The equation for calculating is the probability of the target variable getting categories up to $j=1,2,\dots,J-1$, while the last category (J) is the reference, is given by $P(Y = j|X_1, \dots, X_p)$, where the values of the features ($X_1, X_2, \dots, X_p$) are taken into account. Technical notations for this category are $\beta_{j0}, \beta_{j1}, \beta_{j2}, \dots$, the coefficients of

the features for the j -th category are represented by β_{jp} .

The logit form for $P(Y = j|X_1, \dots, X_p)$ is:

$$\ln\left(\frac{P(Y = j|X_1, \dots, X_p)}{P(Y = J|X_1, \dots, X_p)}\right) = \beta_{j0} + \beta_{j1}X_1 + \dots + \beta_{jp}X_p.$$

The logit form is useful for understanding the linear relationship between the features and the log-odds of the j^{th} category relative to the reference category.

2.1.4 Ordinal Logistic Regression

Ordinal Logistic Regression is a regression analysis technique used when the target variable has ordered categories. It extends the concept of Logistic Regression to handle situations where the target variable has more than two categories and there is a natural ordering among these categories. Ordinal scales are particularly useful when the categories have a meaningful order, but the spacing or magnitude between them may not be uniform. Therefore, it is important to acknowledge that the intervals between adjacent categories on the ordinal scale may not represent equal or consistent changes in the underlying feature being measured.

In Ordinal Logistic Regression, the goal is to model the cumulative odds of an observation falling into or below a certain category versus the odds of it falling into a higher category (Kleinbaum & Klein, 2002). The model assumes that the cumulative odds follow a logistic function.

Let's denote $P(Y \leq j|X_1, \dots, X_p)$ as the probability of the target variable being in category j or below given the values of the features $X_1, X_2, \dots, X_p$. The cumulative probability can be expressed as:

$$P(Y \leq j|X_1, \dots, X_p) = \frac{e^{\beta_{j0} + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_{j0} + \beta_1 X_1 + \dots + \beta_p X_p}}.$$

Here, β_{j0} represents the intercept parameter for category j , and $\beta_1, \beta_2, \dots, \beta_p$ are the

coefficients associated with the features.

Ordinal Logistic Regression estimates parameters using maximum likelihood estimation, similar to Logistic Regression. However, instead of estimating separate models for each category, it directly models cumulative probabilities, taking into account the ordinal nature of the target variable. Proportional odds or cumulative logit model is commonly used for such scenarios. In this context, the logit function is expressed as follows:

$$\ln\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \beta_{j0} + \beta_1 X_1 + \cdots + \beta_p X_p,$$

allows estimating the log-odds of the cumulative probability of an outcome being at or below a given level compared to being above that level. Ordinal Logistic Regression models the cumulative probabilities of multiple ordinal outcomes, as opposed to Binary Logistic Regression, which focuses on the probability of a single outcome.

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
Age2	"-0.07"	"0.126"	"-0.557"	"0.577"	" "	"0.932"
Age3	"-0.77"	"0.294"	"-2.618"	"0.009"	" * "	"0.463"
`StreamingI do not use a streaming service.`	"-0.272"	"0.254"	"-1.07"	"0.285"	" "	"0.762"
StreamingSpotify	"-0.112"	"0.192"	"-0.585"	"0.559"	" "	"0.894"
`StreamingYouTube Music`	"-0.147"	"0.233"	"-0.632"	"0.527"	" "	"0.863"
`StreamingOther streaming service`	"-0.092"	"0.271"	"-0.341"	"0.733"	" "	"0.912"
Hours2	"0.202"	"0.122"	"1.665"	"0.096"	" "	"1.224"
Hours3	"0.438"	"0.202"	"2.169"	"0.03"	" * "	"1.55"
Hours4	"0.414"	"0.223"	"1.861"	"0.063"	" "	"1.513"
While_workingNo	"-0.162"	"0.137"	"-1.185"	"0.236"	" "	"0.851"
InstrumentalistNo	"-0.045"	"0.128"	"-0.355"	"0.723"	" "	"0.956"
ComposerNo	"0.153"	"0.147"	"1.036"	"0.3"	" "	"1.165"
ExploratoryNo	"0.038"	"0.123"	"0.312"	"0.755"	" "	"1.039"
Foreign_languagesNo	"0.138"	"0.113"	"1.218"	"0.223"	" "	"1.148"
ClassicalSometimes	"-0.088"	"0.129"	"-0.684"	"0.494"	" "	"0.915"
ClassicalNever	"-0.172"	"0.142"	"-1.21"	"0.226"	" "	"0.842"
`ClassicalVery frequently`	"-0.161"	"0.168"	"-0.96"	"0.337"	" "	"0.851"
CountrySometimes	"0.205"	"0.155"	"1.321"	"0.186"	" "	"1.227"
CountryNever	"-0.185"	"0.124"	"-1.496"	"0.135"	" "	"0.831"
`CountryVery frequently`	"-0.263"	"0.228"	"-1.156"	"0.248"	" "	"0.769"
EDMSometimes	"0.223"	"0.146"	"1.533"	"0.125"	" "	"1.25"
EDMNever	"0.023"	"0.134"	"0.174"	"0.862"	" "	"1.024"
`EDMVery frequently`	"0.208"	"0.17"	"1.225"	"0.221"	" "	"1.231"
FolkSometimes	"0.076"	"0.143"	"0.533"	"0.594"	" "	"1.079"
FolkNever	"-0.168"	"0.132"	"-1.275"	"0.202"	" "	"0.845"
`FolkVery frequently`	"-0.213"	"0.181"	"-1.178"	"0.239"	" "	"0.808"
GospelSometimes	"-0.223"	"0.228"	"-0.978"	"0.328"	" "	"0.8"
GospelNever	"-0.227"	"0.137"	"-1.657"	"0.097"	" "	"0.797"
`GospelVery frequently`	"-0.182"	"0.438"	"-0.415"	"0.678"	" "	"0.834"
HiphopSometimes	"-0.002"	"0.152"	"-0.011"	"0.991"	" "	"0.998"
HiphopNever	"-0.048"	"0.172"	"-0.279"	"0.78"	" "	"0.953"
`HiphopVery frequently`	"-0.029"	"0.221"	"-0.13"	"0.897"	" "	"0.972"
JazzSometimes	"0.07"	"0.14"	"0.5"	"0.617"	" "	"1.073"
JazzNever	"0.209"	"0.129"	"1.621"	"0.105"	" "	"1.232"
`JazzVery frequently`	"0.219"	"0.22"	"0.996"	"0.319"	" "	"1.245"
KpopSometimes	"0.06"	"0.196"	"0.305"	"0.76"	" "	"1.062"
KpopNever	"0.092"	"0.134"	"0.687"	"0.492"	" "	"1.096"
`KpopVery frequently`	"0.084"	"0.198"	"0.424"	"0.671"	" "	"1.088"
LatinSometimes	"0.033"	"0.189"	"0.173"	"0.862"	" "	"1.033"
LatinNever	"0.189"	"0.136"	"1.386"	"0.166"	" "	"1.208"
`LatinVery frequently`	"0.299"	"0.273"	"1.096"	"0.273"	" "	"1.349"
LofiSometimes	"0.088"	"0.142"	"0.621"	"0.535"	" "	"1.092"
LofiNever	"-0.022"	"0.136"	"-0.16"	"0.873"	" "	"0.978"
`LofiVery frequently`	"-0.024"	"0.182"	"-0.13"	"0.897"	" "	"0.977"
MetalSometimes	"0.089"	"0.156"	"0.571"	"0.568"	" "	"1.093"
MetalNever	"-0.014"	"0.141"	"-0.101"	"0.92"	" "	"0.986"
`MetalVery frequently`	"-0.034"	"0.16"	"-0.211"	"0.833"	" "	"0.967"
PopSometimes	"-0.216"	"0.148"	"-1.456"	"0.145"	" "	"0.806"
PopNever	"-0.187"	"0.226"	"-0.827"	"0.408"	" "	"0.83"
`PopVery frequently`	"-0.159"	"0.155"	"-1.028"	"0.304"	" "	"0.853"
RBSometimes	"-0.32"	"0.148"	"-2.166"	"0.03"	" * "	"0.726"
RBNever	"-0.119"	"0.145"	"-0.823"	"0.411"	" "	"0.888"
`RBVery frequently`	"0.164"	"0.176"	"0.934"	"0.35"	" "	"1.179"
RapSometimes	"-0.118"	"0.153"	"-0.772"	"0.44"	" "	"0.889"
RapNever	"-0.039"	"0.168"	"-0.233"	"0.816"	" "	"0.962"
`RapVery frequently`	"-0.17"	"0.207"	"-0.82"	"0.412"	" "	"0.844"
RockSometimes	"-0.226"	"0.176"	"-1.286"	"0.198"	" "	"0.798"
RockNever	"-0.019"	"0.207"	"-0.093"	"0.926"	" "	"0.981"
`RockVery frequently`	"-0.201"	"0.18"	"-1.115"	"0.265"	" "	"0.818"
Video_game_musicSometimes	"0.079"	"0.141"	"0.558"	"0.577"	" "	"1.082"
Video_game_musicNever	"0.196"	"0.139"	"1.411"	"0.158"	" "	"1.217"
`Video_game_musicVery frequently`	"0.109"	"0.169"	"0.645"	"0.519"	" "	"1.115"
low middle	"0.157"	"0.386"	"0.407"	"0.684"	" "	"1.17"
middle high	"0.85"	"0.387"	"2.196"	"0.028"	" * "	"2.34"

Figure 2.1.5 Full model of the OCD data set

The study employed an Ordinal Logistic Regression model that included all the features of the OCD data set. This data set aims to classify a person as whether they have Low, Middle, or High levels of obsessive compulsive disorder using 24 features. Statistical values for the full model can be found in Figure 2.1.5.

Upon examination of the model in terms of its features, three features were identified as statistically significant at the $\alpha = 0.05$ level. For example, when other features are in the model and held constant, individuals aged 50 years and older are approximately 54% less likely to have low rather than high levels of OCD compared to individuals aged under 25 years.

Similarly, the coefficient estimate of individuals who listen to music between 6 and 9 is statistically significant at $\alpha = 0.05$ level. Individuals who listened to 6 to 9 hours of music were approximately 1.5 times more likely to have low rather than high levels of OCD compared to those who listened to less than 3 hours of music.

Another statistically significant feature at the $\alpha = 0.05$ level is R&B. Compared to those who listen to R&B rarely, those who listen to R&B sometimes exhibit a 28% reduction in the odds of having low rather than high levels of OCD.

2.2 Unsupervised Learning

Unsupervised Learning problems are different from the Supervised Learning problems by nature. The data sets for these problems do not include an outcome variable which provides labelling for the experimental units. Unsupervised Learning is generally carried out using clustering based algorithms. Therefore, we train the machine without labeled data in Unsupervised Learning.

The process of Unsupervised Learning is visualized in Figure 2.2.6. In this process the machine is provided by images of apples and pears of different colors. The features are extracted from these images and data cleaning, exploratory data analysis, fitting of algorithms and model checking is carried out in an order. After the model is validated, it can be used for assigning the initial images to appropriate clusters of apple and pear.



Figure 2.2.6 Unsupervised learning

Unsupervised Machine Learning can be used to answer problems of labelling by itself or more commonly the outputs obtained by these algorithms are fed into Supervised Learning machines. For further information in some of the popular Unsupervised Learning algorithms there is significant research in the literature (Géron, 2022).

2.3 Semi-supervised and Reinforcement Learning

Most Semi-supervised Learning algorithms are a combination of supervised and Unsupervised Learning algorithms. It allows us to train the machine using labeled and unlabeled data. For example, it is used to tag people in our photos in advanced apps like Google Photos (Géron, 2022).

Reinforcement Learning has a different algorithm compared to other methods. This learning method can observe the environment and perform tasks accordingly. After performing the task, if the task is correct, it will be rewarded, otherwise it will be punished. This process is repeated many times and it is aimed to develop a Machine Learning strategy. Reinforcement Learning is often used in advanced applications such as driverless cars like AlphaGo (Géron, 2022).

CHAPTER THREE

DATA PRE-PROCESSING AND PERFORMANCE METRICS

Data pre-processing is a crucial step in the data analysis process. It refers to the transformation of raw data into a form that is suitable for analysis. This involves a series of steps to clean, transform, and format data, ensuring that it is accurate, complete, and relevant for the analysis that will be performed.

3.1 Standardization

Standardization, also known as feature scaling, is a data preprocessing technique used to rescale the data values so that they have a mean (μ_x) of 0 and a standard deviation (σ_x) of 1. This technique involves subtracting the mean value of each feature and dividing by its standard deviation (Hastie et al., 2009). Mathematically, the standardization of a feature x can be expressed as:

$$Z = \frac{x - \mu_x}{\sigma_x}.$$

Standardization is often used in Machine Learning algorithms to bring all the features to a similar scale, which can improve the performance and accuracy of the model. If the features are not scaled properly, some features may have a larger impact on the model than others. For example, if one feature has a range of values from 0 to 1000 and another feature has a range of values from 0 to 1, the first feature may dominate the model's predictions, even if it is not actually more important.

3.2 K-Fold Cross Validation Technique

K-fold cross-validation is a method in Machine Learning used to assess a model's performance on a data set. It involves partitioning the data set into k equal-sized subsets or folds. The model is trained on $k-1$ folds and assessed on the remaining fold. This

process is repeated k times, ensuring that each fold is used as the test set exactly once (Hastie et al., 2009). The process can be visualized as in Figure 3.2.1.

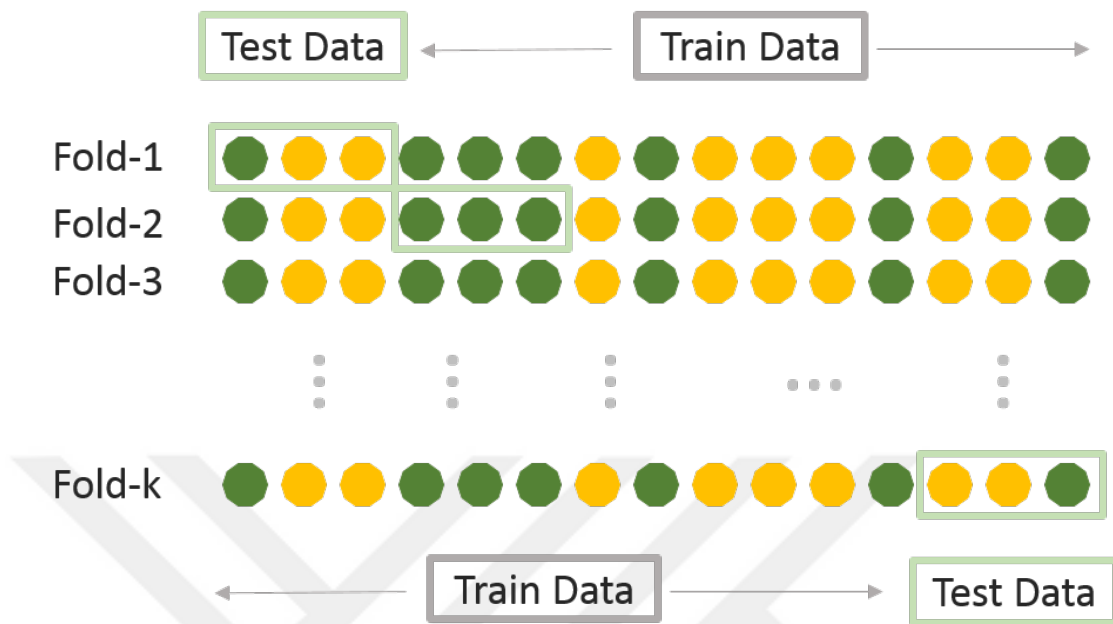


Figure 3.2.1 K-fold cross validation technique

The advantage of k -fold is that it provides a more accurate assessment of model performance than a simple train/test division. This is due to the fact that all data points within the data set are used for both training and testing. This ultimately results in a more meaningful assessment of model performance. In addition, k -fold cross-validation can help to reduce the variance in the model's performance estimate. This is because it averages the results of k iterations with different test sets.

3.3 Performance Metrics

Performance metrics, also known as evaluation metrics, are essential for the assessment of the efficiency and quality of Machine Learning models. Accuracy, precision, recall, F1 score, and AUC-ROC are commonly used metrics in classification tasks. Accuracy measures the correctness of predictions overall, while precision focuses on the variance of predictions. Recall is a measure of the model's ability to identify positive instances. A balanced assessment of precision and recall is

provided by the F1 score. AUC-ROC evaluates the trade-off between true positive and false positive rates over different classification thresholds (Hastie et al., 2009). In regression tasks, metrics like mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R-squared, and mean absolute percentage error (MAPE) are commonly used. These metrics, capture different aspects of the model's performance, including the magnitude of errors, goodness of fit, and percentage-based errors (Bishop & Nasrabadi, 2006) .

3.3.1 Accuracy

For a binary classification problem the possible scenarios that can unfold are summarized in Table 3.3.1. The rows of this table represent the predicted outcomes from the model and the columns represent reality. The diagonal therefore represents the correct decisions achieved from the model.

	Condition Positive	Condition Negative	Total
Test Positive	True positive T_p	False positive F_p	$N_1 = T_p + F_p$
Test Negative	False negative F_n	True negative T_n	$N_2 = F_n + T_n$
Total	$N_3 = T_p + F_n$	$N_4 = F_p + T_n$	$N = T_p + F_p + F_n + T_n$

Table 3.3.1 Confusion matrix

Here T_p stands for it represents the cases where the model correctly identifies a positive instance as positive. T_n is the instances that are actually negative and are correctly identified as negative by the model. F_p means that the model has incorrectly identified a negative instance as positive. F_n represents the number of instances that are incorrectly predicted as negative by a model, but are actually positive.

Accuracy is a performance metric that is commonly used in classification problems to measure the proportion of correctly classified samples out of the total number of samples (Hastie et al., 2009). It is defined as follows:

$$Accuracy = \frac{T_p + T_n}{N}.$$

3.3.2 *RMSE*

RMSE stands for Root Mean Squared Error, which is a commonly used performance metric in regression problems to measure the difference between the predicted ($\hat{y}$) and actual values (y). It is a measure of the average distance between the predicted values and the actual values, expressed in the same units as the target variable (Hastie et al., 2009). It is defined as follows:

$$RMSE = \sqrt{\frac{\sum(\hat{y} - y)^2}{n}}.$$

RMSE is calculated by taking the square root of the average of the squared differences between the predicted and actual values. The square of the differences is used to ensure that the negative and positive errors do not cancel each other out.

CHAPTER FOUR

FEATURE SELECTION

The purpose of Feature Selection methods is to model the most relevant features out of all the features in the data set, while removing the redundant ones from the model (Kumar & Minz, 2014). This can be achieved by ranking, subset selection, or shrinkage. Thus, the dimension reduction is done. When Feature Selection is applied, multicollinearity, over-fitting, training time, prediction time, and model complexity are reduced. In general, Feature Selection methods are divided into four categories: Filter, Wrapper, Embedded, and Hybrid (Wah et al., 2018).

A Hybrid method in Feature Selection is a combination of multiple Feature Selection techniques to obtain a final subset of features that are most relevant for building a predictive model. It involves integrating different Feature Selection methods, each with its own strengths and weaknesses, to produce a more robust set of features for the model (Taheri & Nezamabadi-pour, 2014).

For example, one hybrid method could be to use a filter method to remove highly correlated features, then apply a wrapper method like Recursive Feature Elimination (RFE) to further narrow down the feature set, and finally use an Embedded method like Lasso regularization to perform final feature selection (Lu et al., 2017).

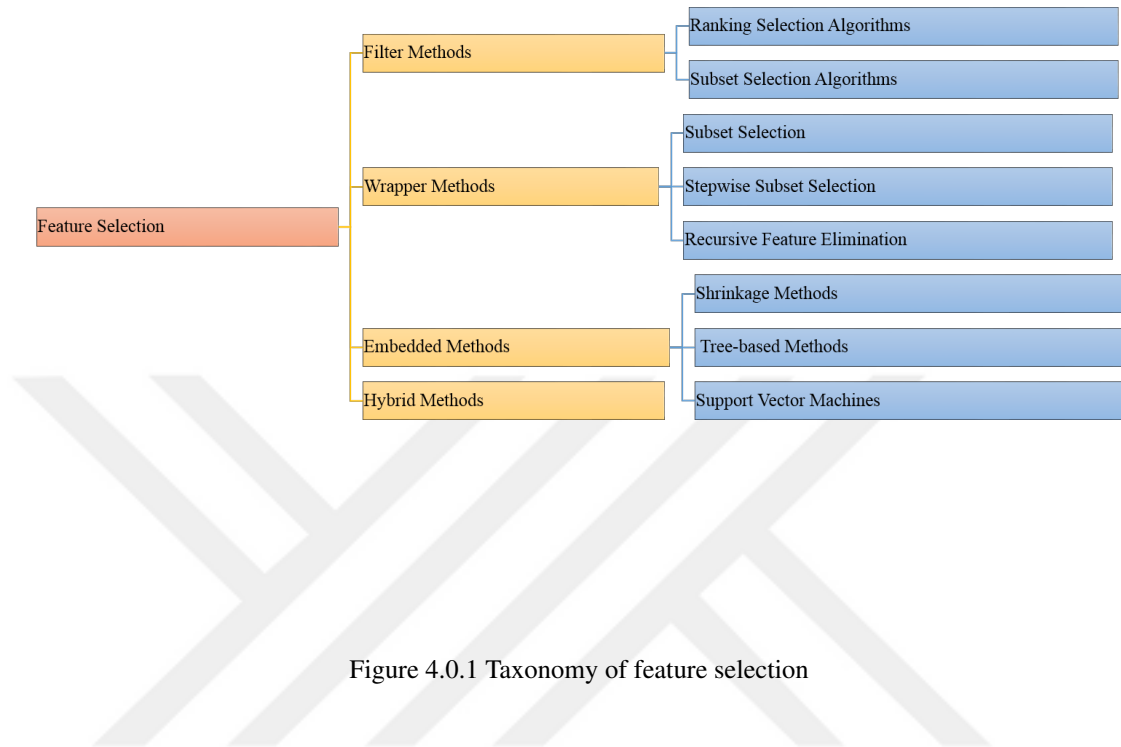


Figure 4.0.1 Taxonomy of feature selection

A taxonomy of Feature Selection models is given in Figure 4.0.1 with emphasis on the approaches discussed in this thesis.

CHAPTER FIVE

FILTER METHODS

The Filter methods can be seen as data pre-processing techniques because these methods select features independent of the model under consideration (Kumar & Minz, 2014). Because Filter methods are fast, scalable, and model-independent, they ignore correlation between features and possible interactions. Filter methods have been categorized according to whether the input features and output features are categorical or continuous (Saeys et al., 2007). The two approaches Filter methods use are evaluating the suitability of features using a ranking procedure before selecting high scoring features and the other one is to compare subsets of features (Bommert et al., 2022). The Filter method is presented in Figure 5.0.1 in the form of a general taxonomy.

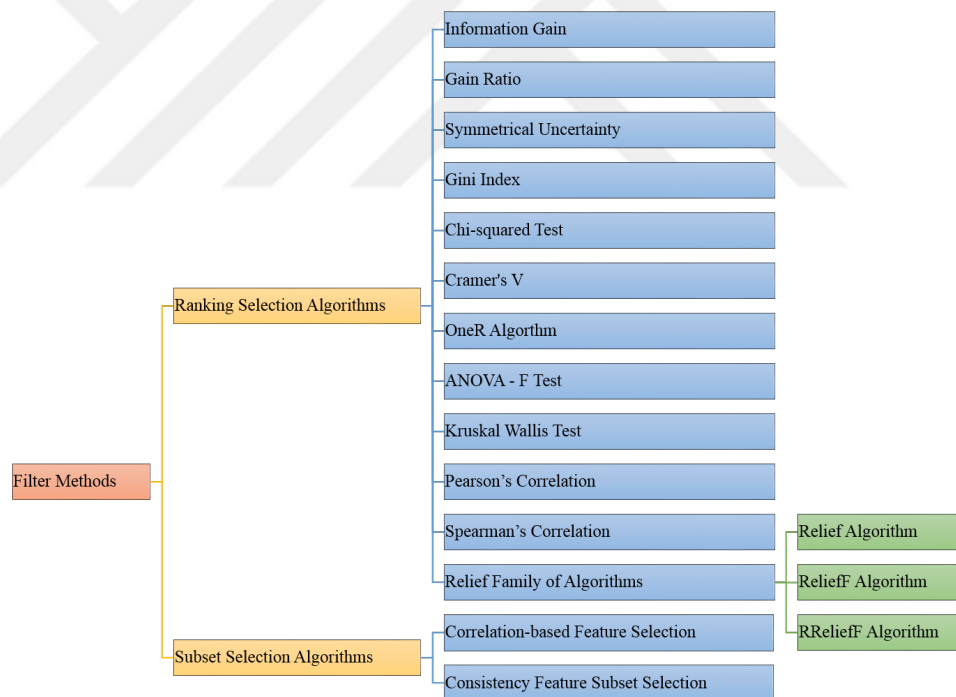


Figure 5.0.1 Taxonomy of filter methods

5.1 Feature Ranking Selection Algorithms

5.1.1 Information Gain

Entropy-based searches have been carried out for feature ranking selection methods. Entropy is a measure of the uncertainty or disorder associated with a feature (Wah et al., 2018). Information Gain is measured by the reduction in entropy with or without the feature. Each feature is evaluated independently, and the greater the reduction in entropy, the more important that feature is for prediction (Ladha & Deepa, 2011). It is easy to calculate and interpret. Information Gain can only be used when both input and output are categorical. Entropy has the following mathematical form for a feature:

$$H(X) = - \sum_i P(x_i) \log_2 P(x_i), \quad (5.1)$$

The entropy of the target variable given one of the features can be calculated by:

$$H(Y|X) = - \sum_i P(x_i) \sum_j P(y_j|x_i) \log_2 P(y_j|x_i).$$

The unconditional entropy of the target variable is then:

$$H(Y) = - \sum_i P(y_i) \log_2 P(y_i). \quad (5.2)$$

The Information Gain can then be defined as the decrease in entropy of Y caused by the inclusion of feature X. Mathematically the Information Gain can be written as:

$$IG(Y, X) = H(Y) - H(Y|X). \quad (5.3)$$

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
RBSometimes	"-0.294394701616024"	"0.132475651631566"	"-2.22225516908405"	"0.0262660659918821"	" * "	"0.745"
RBNever	"-0.101075681498226"	"0.120953543452937"	"-0.835657051565045"	"0.403347863282063"	" "	"0.904"
`RBVery frequently`	"0.18642798806107"	"0.14013635406756"	"1.33033279837717"	"0.183408645345934"	" "	"1.205"
low middle	"0.410637972758656"	"0.0875796212986966"	"4.68873884894005"	"2.74893932231478e-06"	" * "	"1.508"
middle high	"1.05947327471442"	"0.0933300925756434"	"11.3518935369718"	"7.25737714968418e-30"	" * "	"2.885"

Figure 5.1.2 Information Gain used to model OCD data set

The OCD data set was analyzed using Ordinal Logistic Regression with feature selection based on the Information Gain method. This method is used with statistical significances given as shown in Figure 5.1.2. In the feature selection made by the

Information Gain method, it suggested building a model with 1 feature out of 24 features. Approximately 68.3% cross-validation accuracy was achieved in the model established with 1 feature, and the categories of the target variable and the individual feature of sometimes listening to R&B were found to be statistically significant at the $\alpha = 0.05$ level. In comparison to those who rarely listen to R&B, individuals who sometimes listen to R&B are 25% less likely to exhibit a low level of OCD than a high level of OCD.

5.1.2 Gain Ratio

The Gain Ratio is an entropy based method. It is an unsymmetrical method developed to eliminate Information Gain bias. As with the Information Gain, each feature is evaluated and ranked on its own. The Gain Ratio is calculated by dividing the Information Gain value by the entropy of the feature of interest. This calculation can be seen as a kind of normalization. Because the values convert to a range of 0 and 1 (Chelvan & Perumal, 2017). As with Information Gain, input and output feature are categorical. It has the following mathematical form:

$$GR(X) = \frac{IG(Y,X)}{H(X)}.$$

The Gain Ratio is calculated by dividing the Information Gain in Equation 5.3 by the feature entropy in Equation 5.1.

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
Age2	"-0.0758978494370764"	"0.110424695631607"	"-0.687326770546719"	"0.491876832645073"	" "	"0.927"
Age3	"-0.696183256803506"	"0.253660192832719"	"-2.74455068818235"	"0.00605937960427298"	" * "	"0.498"
low middle	"0.427888254650798"	"0.0571137965399336"	"7.49185451805189"	"6.79071710746009e-14"	" * "	"1.534"
middle high	"1.07414392216888"	"0.066005121832493"	"16.2736450194703"	"1.51843486122993e-59"	" * "	"2.927"

Figure 5.1.3 Gain Ratio used to model OCD data set

The OCD data set was analyzed using Ordinal Logistic Regression with feature selection based on the Gain Ratio method. This method is used with statistical significance as shown in Figure 5.1.3. In feature selection using the Gain Ratio method, it is recommended to create a model with 1 feature out of 24 features. In the

model created with 1 feature, approximately 68.3% cross-validation accuracy was achieved with the target variable and target variable categories. The categories of the target variable and the OCD level in individuals over 50 years of age were statistically significant at the $\alpha = 0.05$ level. The likelihood of exhibiting low rather than high levels of OCD is approximately 50% lower in individuals aged 50 years and above compared to those below 25 years of age.

5.1.3 Symmetrical Uncertainty

Another entropy-based method is Symmetrical Uncertainty. As the name suggests, this method provides a symmetrical normalization. It is obtained by dividing the Information Gain value by the sum of the entropies of the feature and target variable. The Symmetrical Uncertainty is very similar to Gain Ratio except it is symmetrical (Singh et al., 2014). Symmetrical Uncertainty can be used on categorical input and output features. It has the following mathematical form:

$$SU(X) = \frac{2 * IG(Y, X)}{H(X) + H(Y)}.$$

Symmetrical Uncertainty is calculated by dividing twice the Information Gain in Equation 5.3 by the sum of the feature entropy in Equation 5.1 and the target variable entropy in Equation 5.2.

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
Hours2	"0.235215582228635"	"0.108768427941878"	"2.16253545885876"	"0.030576926985522"	" * "	"1.265"
Hours3	"0.419346232633134"	"0.183089467453706"	"2.29038971200878"	"0.0219987353864846"	" * "	"1.521"
Hours4	"0.419022052554239"	"0.199744833964862"	"2.09778668232266"	"0.035923993487024"	" * "	"1.52"
low middle	"0.592756972736366"	"0.0622442735543699"	"9.5230764034638"	"1.68126667423635e-21"	" * "	"1.809"
middle high	"1.24158768218411"	"0.0718739527323624"	"17.2745151057354"	"7.31859246544822e-67"	" * "	"3.461"

Figure 5.1.4 Symmetrical Uncertainty used to model OCD data set

According to the information provided, OCD data set was analysed using Ordinal Logistic Regression with feature selection based on the Symmetric Uncertainty method. This method is used with statistical significances given as shown in Figure 5.1.4. The model selecting only 1 feature achieved 68.3% cross-validation

accuracy. The categories of the target variable and all categories of the hours feature were statistically significant at the $\alpha = 0.05$ level at the OCD level.

The findings indicate that individuals who listen to music for a duration of between 3 and 6 hours per day are 1.2 times more likely to exhibit low levels of OCD compared to those who listen to music for a duration of less than 3 hours per day. Similarly, individuals who listen to music for more than six hours per day are approximately 1.5 times more likely to exhibit low levels of OCD than those who listen to music for less than three hours per day.

5.1.4 *Gini Index*

Gini index, also known as Gini impurity or Gini coefficient, is a measure of the impurity or randomness of a set of examples in a classification problem. The Gini Index quantifies the likelihood of misclassifying a randomly selected example from a set of examples, taking into account the distribution of classes within the set. The index ranges from 0 to 1, with 0 indicating perfect purity (all examples belong to the same class) and 1 indicating maximum impurity (the examples are evenly distributed across all classes) (Hastie et al., 2009). It has the following mathematical form:

$$Gini = \sum_{k=1}^K \hat{P}_{mk}(1 - \hat{P}_{mk}). \quad (5.4)$$

where k is the number of classes and p_{mk} represents the proportion of training observations in the m^{th} region that belong to the k^{th} class.

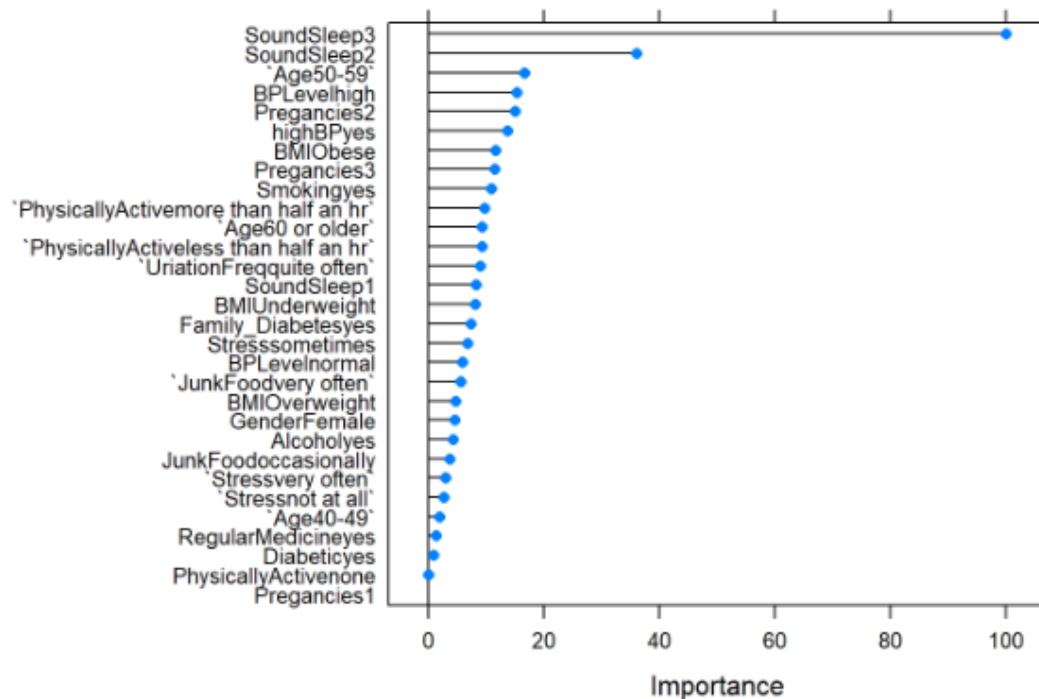


Figure 5.1.5 Ranking the importance of features in the Sleep Duration data set modelled by the Gini Index

Based on the information provided, a Linear Regression was applied to the Sleep Duration data set using feature selection based on the Gini Index method. The data set initially had 17 features. By applying the Gini Index method, the features were ranked according to their relevance to the target variable in Figure 5.1.5. After ranking the features, the top 16 features with the highest Gini Index scores were selected for the model. Using these 16 selected features, the Linear Regression model achieved the lowest cross-validation RMSE of 0.96.

5.1.5 Chi-squared Test

The feature selection method using Chi-square statistics is used to test the relationship or independence of two categorical features. If the features are numeric, they can be converted into categories and this method can be applied. As a result, the larger the calculated Chi-square value, the stronger the relationship of the distributions of the features and their classes (Wah et al., 2018). The features that

score the highest are selected by sorting all features (Bahassine et al., 2020). The expected counts have the following mathematical form:

$$E_{ij} = \frac{R_i C_j}{N}.$$

The statistic can then be calculated as:

$$X^2 = \sum_{i=1}^a \sum_{j=1}^b \left(\frac{O_{ij} - E_{ij}}{E_{ij}} \right). \quad (5.5)$$

where a denotes the number of intervals, b refers to the number of classes, N represents the total number of samples, R_i signifies the number of samples in the i^{th} interval, C_j signifies the number of samples in the j^{th} class and O_{ij} represents the number of samples in the i^{th} interval j^{th} class.

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
RBSometimes	-.0557701202560763"	"0.252153409233046"	"-2.21175356802463"	"0.0269836969557159"	" *	"0.573"
RBNever	-.01593237702041"	"0.237475931437541"	"-0.670904917562155"	"0.502281103439448"	" "	"0.853"
`RBVery frequently`	"0.2149869083359"	"0.292624199168558"	"0.734686020318036"	"0.46230736151088"	" "	"1.24"
RockSometimes	-.0357149774632864"	"0.286980223581668"	"-1.24451005778531"	"0.213311900006367"	" "	"0.7"
RockNever	-.0709002165963803"	"0.342755009368999"	"-0.206853917983301"	"0.836123946202716"	" "	"0.932"
`RockVery frequently`	-.0333351890441741"	"0.280042638866955"	"-1.19036119567532"	"0.233904457846169"	" "	"0.717"
Hours2	"0.327763749051667"	"0.202742946242626"	"1.61664686799721"	"0.10595452893962"	" "	"1.388"
Hours3	"0.673104454084312"	"0.33498013083218"	"2.00938620571955"	"0.0444961919831762"	" *	"1.96"
Hours4	"0.656044165499313"	"0.356257750522879"	"1.84148741897246"	"0.0655501625824517"	" "	"1.927"
CountrySometimes	"0.40368809632512"	"0.256118153138922"	"1.57617916331824"	"0.114984519120225"	" "	"1.497"
CountryNever	-.0308571890956046"	"0.209094422353047"	"-1.47575381248112"	"0.14009994022778"	" "	"0.734"
`CountryVery frequently`	-.0300128393688743"	"0.372557402925017"	"-0.805589665733064"	"0.420479486284682"	" "	"0.741"
ExploratoryNo	-.0192159016282242"	"0.207629847314059"	"-0.0925488405294561"	"0.926261988563972"	" "	"0.981"
While_workingNo	-.025621487102988"	"0.234465546270263"	"-1.09023048874029"	"0.275611626174291"	" "	"0.774"
RapSometimes	-.010249178280659"	"0.255581020011122"	"-0.401014843755338"	"0.688409195296485"	" "	"0.903"
RapNever	-.0088692484060616"	"0.289262801042726"	"-0.3066155880333"	"0.759135986688818"	" "	"0.915"
`RapVery frequently`	-.00921634102967689"	"0.343470193068233"	"-0.268330155445133"	"0.788445196349308"	" "	"0.912"
Age2	-.0107112448586424"	"0.210261100008336"	"-0.509425892769404"	"0.610453730861827"	" "	"0.898"
Age3	-1.41646744875292"	"0.544096254087079"	"-2.60333982840878"	"0.00923203943177892"	" *	"0.243"
EDMSometimes	"0.344995308772741"	"0.243874919026662"	"1.41464038265872"	"0.157173962294072"	" "	"1.412"
EDMNever	"0.10023160901053"	"0.227131522415586"	"0.44129325575133"	"0.659000707989479"	" "	"1.105"
`EDMVery frequently`	"0.37247459633203"	"0.28041073874665"	"1.32831787397615"	"0.184073117390901"	" "	"1.451"
LofiSometimes	"0.180285149437798"	"0.236939631798729"	"0.760890645727615"	"0.44672386223452"	" "	"1.198"
LofiNever	-.00966027401448802"	"0.23028429440324"	"-0.41943393569096"	"0.674855581894068"	" "	"0.908"
`LofiVery frequently`	-.00481724661788775"	"0.297191235654953"	"-0.162092485913034"	"0.871233023320813"	" "	"0.953"
HiphopSometimes	-.00014563324407891"	"0.255021908181763"	"-0.000571062027888238"	"0.999544358449452"	" "	"1"
HiphopNever	-.01878258549310056"	"0.29639850807702"	"-0.0601415136893116"	"0.952042926640862"	" "	"0.982"
`HiphopVery frequently`	-.00779684725553828"	"0.367561701597219"	"-0.212123494413523"	"0.832010690066312"	" "	"0.925"
GospelSometimes	-.036346544167851"	"0.381215332449617"	"-0.953438675572018"	"0.34036785316897"	" "	"0.695"
GospelNever	-.0348697396082162"	"0.223936572690079"	"-1.55712571597114"	"0.119440640587662"	" "	"0.706"
`GospelVery frequently`	-.0275699061560434"	"0.741926914600581"	"-0.371598679243033"	"0.710191672601713"	" "	"0.759"
Video_game_musicSometimes	"0.121547886825222"	"0.232187567363879"	"0.523490074017335"	"0.600633256686678"	" "	"1.129"
Video_game_musicNever	"0.285996293130118"	"0.235369130995363"	"1.21509686474456"	"0.224329149241291"	" "	"1.331"
`Video_game_musicVery frequently`	"0.123405155658299"	"0.27897264602006"	"0.44235575730348"	"0.658231790211094"	" "	"1.131"
FolkSometimes	"0.139028458466198"	"0.241819864945044"	"0.57492571380682"	"0.565341538556426"	" "	"1.149"
FolkNever	-.0293259814074971"	"0.21992777696071"	"-1.33343690427682"	"0.182380468441107"	" "	"0.746"
`FolkVery frequently`	-.0364168002914654"	"0.30645307731245"	"-1.1883201515989"	"0.234702629235606"	" "	"0.695"
LatinSometimes	"0.172266490216906"	"0.322244260301341"	"0.534583579722456"	"0.592937840600076"	" "	"1.188"
LatinNever	"0.403181831570112"	"0.228008319389699"	"1.7682768446752"	"0.0770146329783754"	" "	"1.497"
`LatinVery frequently`	"0.5814583727275423"	"0.448533766806214"	"1.29635361096422"	"0.194853683194752"	" "	"1.789"
Foreign_languagesNo	"0.208348086335745"	"0.190102414125692"	"1.09597811944666"	"0.27308635000434"	" "	"1.232"
PopSometimes	-.0391456733094513"	"0.24473039697608"	"-1.59954275370531"	"0.109700056791042"	" "	"0.676"
PopNever	-.0482685972648224"	"0.396493227745356"	"-1.21738768501293"	"0.22345674635327"	" "	"0.617"
`PopVery frequently`	-.0295467305849802"	"0.250761551975593"	"-1.17827993774165"	"0.238685023198874"	" "	"0.744"
KpopSometimes	-.02050106575948245"	"0.329061692965242"	"-0.0760059834660437"	"0.939414337853147"	" "	"1.025"
KpopNever	"0.146240612729612"	"0.225954898928771"	"0.647211516204887"	"0.51749508999969"	" "	"1.157"
`KpopVery frequently`	"0.19828408694357"	"0.331484718880671"	"0.598169615821563"	"0.549726762988396"	" "	"1.219"
low middle	"0.332676251117686"	"0.495695076673737"	"0.671130835815526"	"0.502137184621859"	" "	"1.395"
middle high	"1.53689394867591"	"0.500508848651772"	"3.07066289200653"	"0.00213584146663968"	" *	"4.65"

Figure 5.1.6 Chi-squared Test used to model OCD data set

Based on the information provided, Ordinal Logistic Regression was applied to the OCD data set using feature selection based on Chi-square Test. Chi-square Test statistical significances are given as shown in Figure 5.1.6. Chi-square Test selected

18 features out of 24 features. Of these 18 features, 3 features were statistically significant at the $\alpha = 0.05$ level. The model has a cross validation accuracy of 68%. For instance, individuals over the age of 50 have low OCD level rather than high OCD level decreases by approximately 75% compared to individuals under the age of 25.

5.1.6 Cramer's V

Cramer's V is a measure of association based on the Chi-squared statistic. It can be used when all the features and the target variable are categorical. Cramer's V measures how strongly the feature and target variable are correlated (Martens et al., 2011). It has the following mathematical form:

$$CV = \sqrt{\frac{X^2}{n}}.$$

Cramer's V statistic is obtained by dividing the Chi-square test statistic in Equation 5.5 value by the sample size and taking the square root. Cramer's V statistic takes values between 0 and 1. If the value is 0, there is no relationship, and 1 means that there is a very strong relationship (Akoglu, 2018).

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
RBSometimes	"-0.307"	"0.146"	"-2.107"	"0.035"	" * "	"0.736"
RBNever	"-0.1"	"0.138"	"-0.722"	"0.47"	" "	"0.905"
`RBVery frequently`	"0.149"	"0.174"	"0.857"	"0.392"	" "	"1.161"
RockSometimes	"-0.197"	"0.169"	"-1.166"	"0.243"	" "	"0.821"
RockNever	"-0.009"	"0.2"	"-0.047"	"0.962"	" "	"0.991"
`RockVery frequently`	"-0.146"	"0.165"	"-0.886"	"0.375"	" "	"0.864"
Hours2	"0.185"	"0.119"	"1.549"	"0.121"	" "	"1.203"
Hours3	"0.414"	"0.199"	"2.08"	"0.038"	" * "	"1.513"
Hours4	"0.359"	"0.218"	"1.643"	"0.1"	" "	"1.432"
CountrySometimes	"0.217"	"0.153"	"1.421"	"0.155"	" "	"1.242"
CountryNever	"-0.158"	"0.121"	"-1.299"	"0.194"	" "	"0.854"
`CountryVery frequently`	"-0.209"	"0.221"	"-0.947"	"0.344"	" "	"0.811"
ExploratoryNo	"0.034"	"0.119"	"0.288"	"0.773"	" "	"1.035"
While_workingNo	"-0.16"	"0.134"	"-1.196"	"0.232"	" "	"0.852"
RapSometimes	"-0.091"	"0.149"	"-0.612"	"0.54"	" "	"0.913"
RapNever	"-0.04"	"0.163"	"-0.243"	"0.808"	" "	"0.961"
`RapVery frequently`	"-0.12"	"0.205"	"-0.586"	"0.558"	" "	"0.887"
Age2	"-0.069"	"0.122"	"-0.563"	"0.573"	" "	"0.934"
Age3	"-0.772"	"0.282"	"-2.738"	"0.006"	" * "	"0.462"
EDMSometimes	"0.232"	"0.143"	"1.621"	"0.105"	" "	"1.261"
EDMNever	"0.042"	"0.132"	"0.317"	"0.751"	" "	"1.043"
`EDMVery frequently`	"0.249"	"0.166"	"1.503"	"0.133"	" "	"1.283"
LofiSometimes	"0.105"	"0.14"	"0.748"	"0.454"	" "	"1.11"
LofiNever	"-0.031"	"0.133"	"-0.231"	"0.817"	" "	"0.97"
`LofiVery frequently`	"-0.043"	"0.178"	"-0.239"	"0.811"	" "	"0.958"
HiphopSometimes	"0"	"0.149"	"-0.001"	"0.999"	" "	"1"
HiphopNever	"-0.04"	"0.168"	"-0.237"	"0.813"	" "	"0.961"
`HiphopVery frequently`	"-0.029"	"0.217"	"-0.132"	"0.895"	" "	"0.972"
GospelSometimes	"-0.214"	"0.225"	"-0.949"	"0.343"	" "	"0.808"
GospelNever	"-0.215"	"0.133"	"-1.613"	"0.107"	" "	"0.807"
`GospelVery frequently`	"-0.187"	"0.435"	"-0.43"	"0.667"	" "	"0.829"
Video_game_musicSometimes	"0.062"	"0.138"	"0.454"	"0.65"	" "	"1.064"
Video_game_musicNever	"0.189"	"0.137"	"1.381"	"0.167"	" "	"1.208"
`Video_game_musicVery frequently`	"0.071"	"0.165"	"0.433"	"0.665"	" "	"1.074"
FolkSometimes	"0.092"	"0.142"	"0.648"	"0.517"	" "	"1.096"
FolkNever	"-0.157"	"0.128"	"-1.227"	"0.22"	" "	"0.855"
`FolkVery frequently`	"-0.197"	"0.179"	"-1.1"	"0.271"	" "	"0.821"
LatinSometimes	"0.063"	"0.186"	"0.34"	"0.734"	" "	"1.065"
LatinNever	"0.216"	"0.131"	"1.641"	"0.101"	" "	"1.241"
`LatinVery frequently`	"0.347"	"0.268"	"1.293"	"0.196"	" "	"1.415"
Foreign_languagesNo	"0.126"	"0.11"	"1.147"	"0.251"	" "	"1.135"
PopSometimes	"-0.224"	"0.143"	"-1.568"	"0.117"	" "	"0.799"
PopNever	"-0.201"	"0.222"	"-0.908"	"0.364"	" "	"0.818"
`PopVery frequently`	"-0.166"	"0.148"	"-1.124"	"0.261"	" "	"0.847"
KpopSometimes	"0.036"	"0.193"	"0.185"	"0.853"	" "	"1.036"
KpopNever	"0.08"	"0.132"	"0.606"	"0.544"	" "	"1.083"
`KpopVery frequently`	"0.099"	"0.196"	"0.507"	"0.612"	" "	"1.105"
low middle	"0.247"	"0.289"	"0.854"	"0.393"	" "	"1.281"
middle high	"0.935"	"0.291"	"3.21"	"0.001"	" * "	"2.547"

Figure 5.1.7 Cramer's V used to model OCD data set

Ordinal Logistic Regression was applied to the OCD data set using feature selection based on Cramer's V. Statistical significances are given as shown in Figure 5.1.7. Cramer's V selected 18 features out of 24 features. The cross-validation accuracy of the model is approximately 68.5%. Of these 18 features, 3 were statistically significant at the $\alpha = 0.05$ level. To illustrate, individuals who sometimes listen to R&B are 36% less likely to exhibit a low level of OCD than a high level of OCD, in comparison to those who rarely listen to R&B.

5.1.7 OneR Algorithm

This algorithm can be used when the features and the target variable are both categorical. The OneR Algorithm is a classification algorithm that creates a score for each feature and ranks the features according to these scores. The features with smaller scores are considered as better candidates for the model (Holte, 1993).

The algorithm starts by creating a cross tabulation between the feature and the target variable. For each category of the feature the algorithm selects the class of Y with the smallest frequency. The score for a feature is then calculated by summing these smallest frequencies for each level of the feature (Morariu et al., 2013). These scores are considered as errors this feature would have made if the target variable was categorized naively using the features classes only by assigning all the Y's into the dominant category for that class. As a result, these scores can be interpreted as the errors that the feature would make if the target variable was categorized naively, using only the classes of the feature and assigning all instances to the dominant class for each level of the feature.

By calculating and sorting the features based on these error scores, the algorithm prioritizes the features with smaller errors. This implies that features with smaller scores are considered better candidates for the model, as they are expected to contribute more accurate or informative predictions.

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
Age2	"-0.0758978494370764"	"0.110424695631607"	"-0.687326770546719"	"0.491876832645073"	" "	"0.927"
Age3	"-0.696183256803506"	"0.253660192832719"	"-2.74455068818235"	"0.00605937960427298"	" * "	"0.498"
low middle	"0.427888254650798"	"0.0571137965399336"	"7.49185451805189"	"6.79071710746009e-14"	" * "	"1.534"
middle high	"1.07414392216888"	"0.066005121832493"	"16.2736450194703"	"1.51843486122993e-59"	" * "	"2.927"

Figure 5.1.8 OneR Algorithm used to model OCD data set

Based on the information provided, Ordinal Logistic Regression was applied to the OCD data set using feature selection based on the OneR algorithm. The statistical significances of the OneR algorithm are shown in Figure 5.1.8. The OneR algorithm selected 1 feature out of 24. The cross-validation accuracy of the model is approximately 68.3%. For instance, the chance of exhibiting a low level of OCD

rather than a high level of OCD is approximately 50% lower in individuals aged 50 and above in comparison to those below the age of 25.

5.1.8 ANOVA

ANOVA, which means Analysis of Variance, is a statistical method used to determine whether two or more categories of the feature come from the same population. Its primary purpose is to assess whether there are statistically significant differences between the means of different categories, while considering both within-category and between-category variation (Shakeela et al., 2021).

In the context of feature selection, ANOVA is utilized to identify the most relevant features within a given set for a predictive model. By applying ANOVA, we can evaluate the relationship between each feature and the target variable, and determine whether the variation in the target variable can be significantly explained by the variation in a specific feature (Nasiri & Alavi, 2022).

During the feature selection process, the F-values of different features are compared. A feature with a high F value indicates a strong relationship with the target variable. This allows prioritization and selection of the most effective features, facilitates dimensionality reduction, and improves the performance and interpretability of predictive models. This assesses whether there are significant differences in means between different categories of feature on a continuous target variable.

The ANOVA F-statistic is calculated using the formula:

$$SSR = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2,$$

where n_i is the number of observations in category i .

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

where x_{ij} represents the j th observation in category i .

$$F = \frac{SSR/(k-1)}{SSE/(n-k)}$$

where SSR (sum of squares regression) represents the sum of squares between categories (variance explained by the feature), SSE (sum of squares error) represents the sum of squares within categories (variance within each category), k is the number of categories, and n is the total number of observations.

ANOVA has important assumptions such as independence, normality, and homogeneity of variances. Violating these assumptions can affect the results of the ANOVA. If the assumptions are not met, alternative approaches such as nonparametric tests, data transformations, Welch's ANOVA, robust ANOVA can be considered. It is important to consider and select these techniques to maintain the validity of the statistical analysis depending on the nature and severity of the violation.

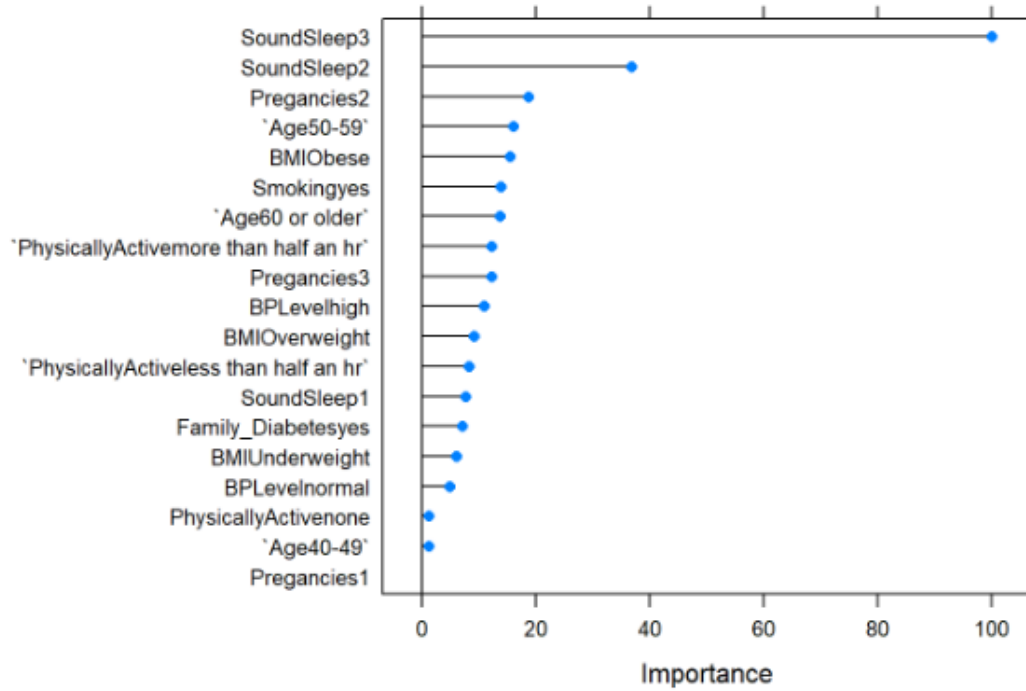


Figure 5.1.9 Ranking the importance of features in the Sleep Duration data set modelled by the ANOVA

Based on the information provided, a Linear Regression was applied to the Sleep Duration data set using feature selection based on the ANOVA. The data set initially had 17 features. By applying the ANOVA, the features were ranked according to their relevance to the target variable in Figure 5.1.9. After ranking the features, the top 8 features with the highest ANOVA scores were selected for the model. Using these 8 selected features, the Linear Regression model achieved the lowest cross-validation RMSE of 0.96.

5.1.9 Kruskal Wallis Test

The Kruskal Wallis Test is used if there are categorical features and a numerical target variable in the data set and these features do not meet the normality assumption. The Kruskal Wallis Test is one of the most widely used non-parametric analysis of variance (McDonald, 2009). Calculation takes place by ranking the number values in the features (Ni, 2012). The Kruskal Wallis Test rank sum test statistic is used as a score for feature selection (Bommert et al., 2020). The test works by ranking the

combined data from all categories, calculating the sum of ranks for each category, and then applying a statistical formula to determine if the differences in ranks between the categories are statistically significant. It has the following mathematical form:

$$R_i = \sum_{j=1}^{n_i} \text{rank}(x_{ij}),$$

$$H = \frac{12}{n(n+1)} \left(\sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3(n+1).$$

n_i : denotes the number of observations in the category i , $i \in \{1, \dots, k\}$

R_i is the ordinal sum of the observations. H gives the Kruskal Wallis Test total rank statistics.

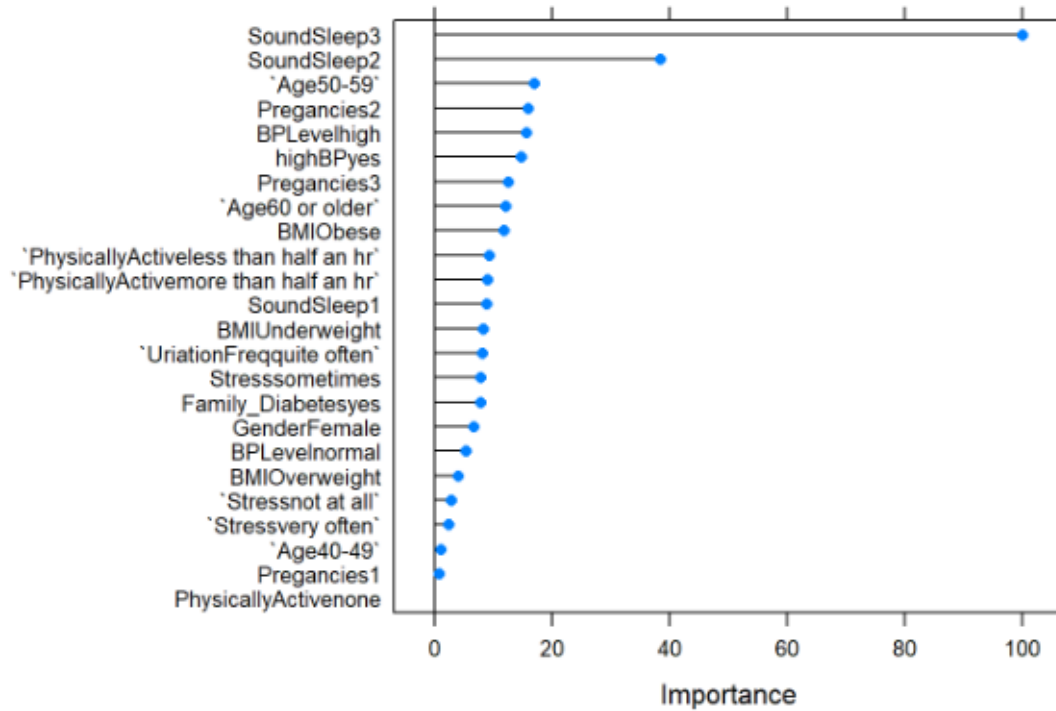


Figure 5.1.10 Ranking the importance of features in the Sleep Duration data set modelled by the Kruskal Wallis Test

Based on the information provided, a Linear Regression was applied to the Sleep Duration data set using feature selection based on the Kruskal Wallis Test method.

The data set initially had 17 features. By applying the Kruskal Wallis Test, the features were ranked according to their relevance to the target variable in Figure 5.1.10. After ranking the features, the top 11 features with the highest Kruskal Wallis Test scores were selected for the model. Using these 11 selected features, the Linear Regression model achieved the lowest cross-validation RMSE of 0.96.

5.1.10 Pearson's Correlation

Pearson's Correlation coefficient is the most used method to understand the relationship between the features themselves and the target variable. The Pearson's Correlation coefficient helps us to understand the linear relationship between two normal continuous variables (Jebli et al., 2021). It has the following mathematical form:

$$Cov(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1},$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}},$$

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}},$$

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y}.$$

The Pearson's Correlation coefficient is obtained by dividing the covariance of the feature and target variable by the standard deviations. The Pearson's Correlation coefficient takes values between -1 and 1. This coefficient gives information about the linear relationship between X and Y. For example; If $r_{xy} = -1$, it is understood that there is a strong negative relationship, if $r_{xy} = 0$, they are linearly unrelated, and if $r_{xy} = 1$, there is a strong positive relationship (Zhou et al., 2016).

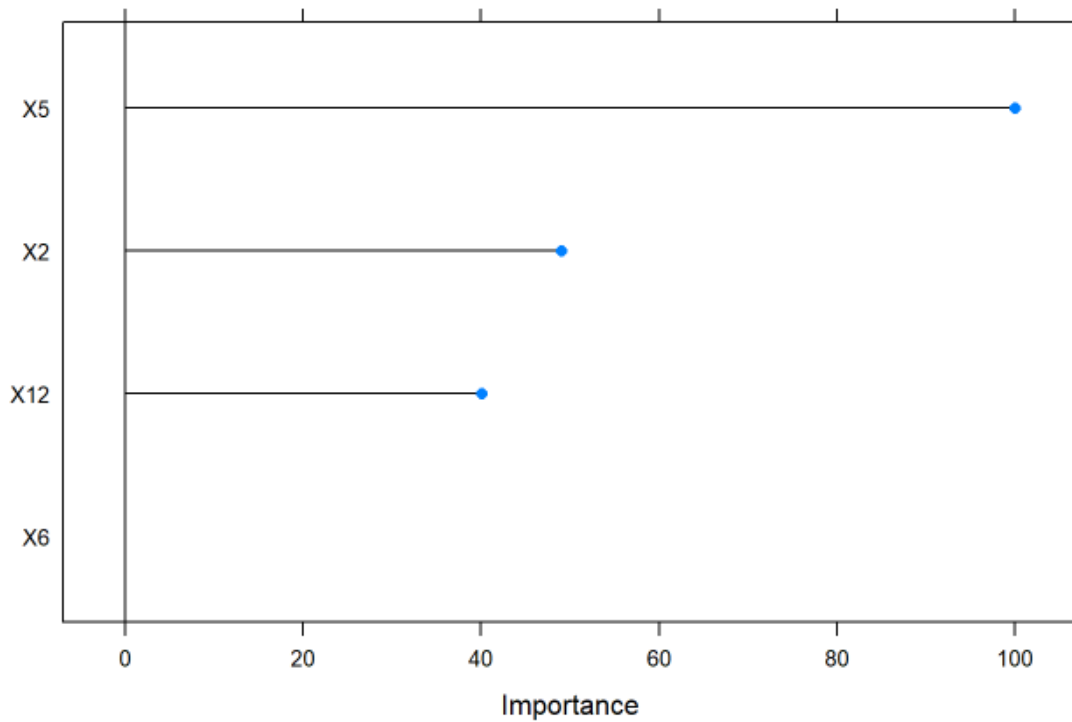


Figure 5.1.11 Ranking the importance of features in the SENIC data set modelled by the Pearson's Correlation

Based on the information provided, a Linear Regression was applied to the SENIC data set using feature selection based on the Pearson's Correlation. The data set initially had 8 features. By applying the Pearson's Correlation, the features were ranked according to their relevance to the target variable in Figure 5.1.11. After ranking the features, the top 4 features with the highest Pearson's Correlation scores were selected for the model. Using these 4 selected features, the Linear Regression model achieved the lowest cross-validation RMSE of 0.72.

5.1.11 Spearman's Correlation

Spearman's Correlation coefficient is a non-parametric measure of the statistical relationship between the feature and the target variable (Liu et al., 2010). Spearman's Correlation coefficient is calculated using the ranks of the feature and target variable. For this reason, it is also called the Spearman rank correlation coefficient (Zhu, 2020).

It has the following mathematical form:

$$D_i = \text{rank}(X_i) - \text{rank}(Y_i),$$

$$\rho_{\text{spearman}} = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)}.$$

D_i denote the difference between the ranks of X_i and Y_i for $i = 1, \dots, n$. Spearman's Correlation coefficient is calculated by the formulas.

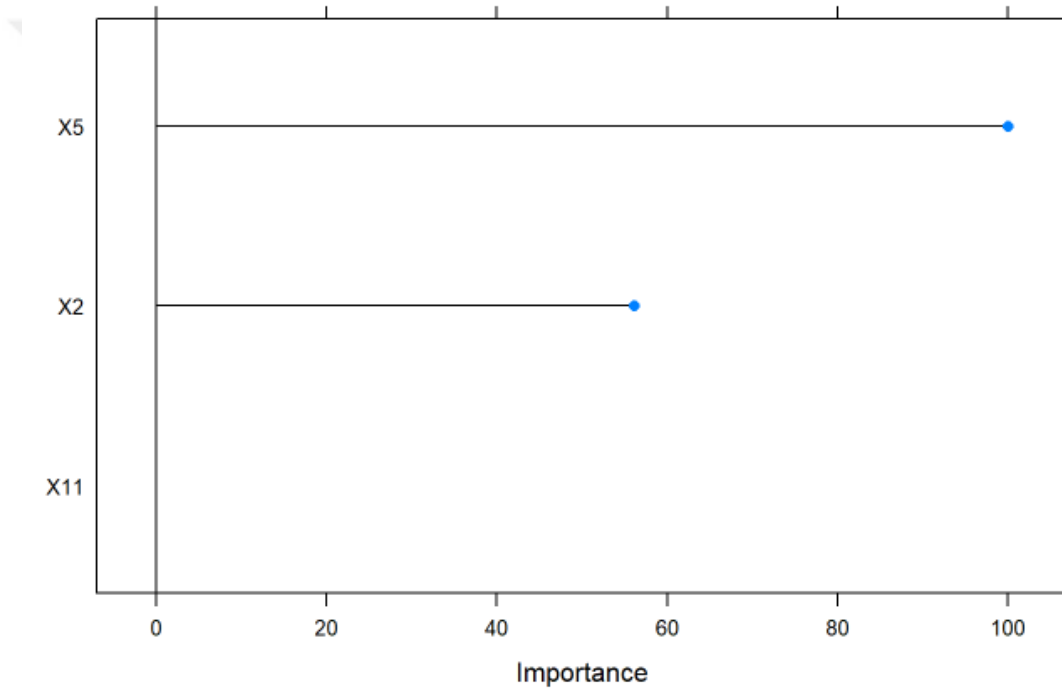


Figure 5.1.12 Ranking the importance of features in the SENIC data set modelled by the Spearman's Correlation

Based on the information provided, a multiple Linear Regression model was applied to the SENIC data set using feature selection based on the Spearman's Correlation. The data set initially had 8 features. By applying the Spearman's Correlation, the features were ranked according to their relevance to the target variable in Figure 5.1.12. After ranking the features, the top 3 features with the highest Spearman's Correlation scores

were selected for the model. Using these 3 selected features, the Linear Regression model achieved the lowest cross-validation RMSE of 0.73.

5.1.12 Relief Family of Algorithms

The Relief Family of Algorithms encompasses various variants that share the core concept of evaluating feature relevance based on the differences between nearest hit and miss instances. While the original Relief algorithm was initially designed for binary classification problems, several extensions and modifications have been proposed to handle multi-class classification and address other challenges, such as imbalanced data sets (Kira & Rendell, 1992).

m : iteration number

R_i : randomly selected observation

H : nearest hit

M : nearest miss

X : all features

$W[X]$: weight of all features

5.1.12.1 Relief Algorithm

The Relief Algorithm is a feature selection method that aims to identify the most relevant features in a data set for binary classification tasks. In the Relief Algorithm, the number of iterations m is determined first. Then the weights of all the features are initialized from zero and updated at each iteration. An iteration process begins by choosing a random set R_i of observations. The nearest hit H for the observations belonging to the same class of the target variable and the nearest miss M for the other class are calculated with the Manhattan distance measure (Robnik-Šikonja &

Kononenko, 2003). The weights of each feature in the Relief Algorithm are updated using the following mathematical form:

$$W[X] := W[X] - \frac{|X_{R_i} - H|}{m(\max(X_i) - \min(X_i))} - \frac{|X_{R_i} - M|}{m(\max(X_i) - \min(X_i))}.$$

In this equation, $W[X]$ represents the weight of feature X , X_{R_i} represents the value of feature X for the randomly selected observation R_i , H represents the nearest hit observation, M represents the nearest miss observation, $\max(X_i)$ represents the maximum value of feature X across all observations, $\min(X_i)$ represents the minimum value of feature X across all observations, and m represents the total number of iterations.

Using this weight update formula, the Relief Algorithm adjusts the weights of each feature based on the differences between the randomly selected instance and its nearest hit and miss instances. The formula takes into account the relative differences by normalizing the differences using the range of feature values. By iteratively updating the weights and repeating the process, the algorithm identifies the most relevant features that contribute significantly to the classification task. Sorting the features based on their weights provides a ranking of feature importance, with higher weights indicating stronger relevance.

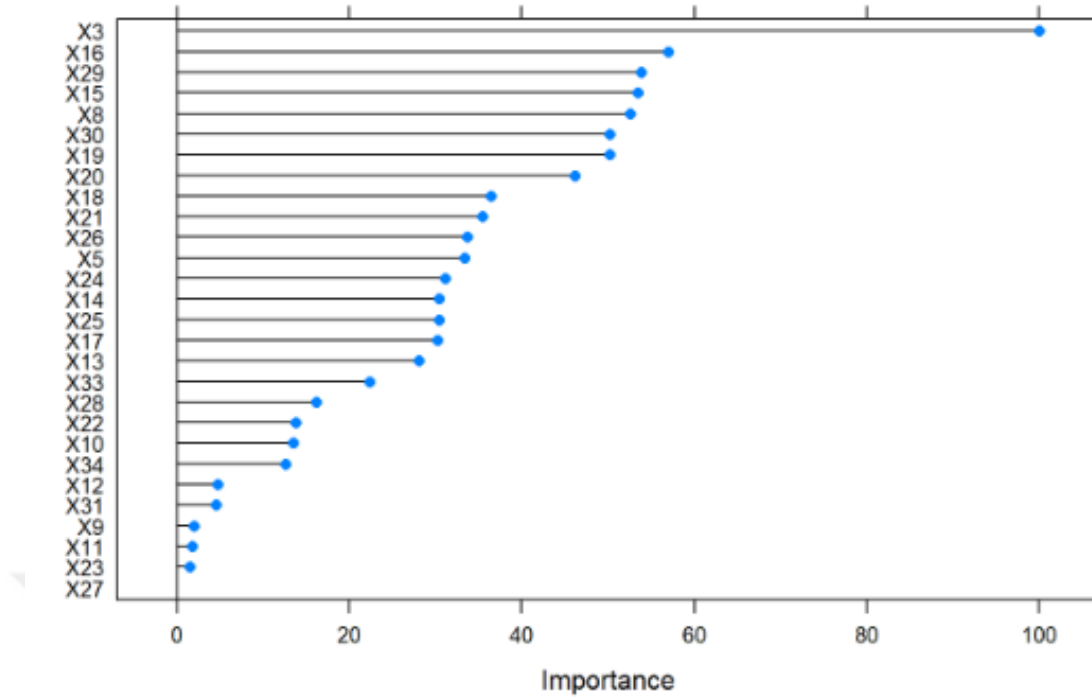


Figure 5.1.13 Ranking the importance of features in the Breast Cancer data set modelled by the Relief Algorithm

Based on the information provided, a Logistic Regression was applied to the Breast Cancer data set using feature selection based on the Relief Algorithm. The data set initially had 33 features. By applying the Relief Algorithm, the features were ranked according to their relevance to the target variable in Figure 5.1.13. After ranking the features, the top 28 features with the highest Relief Algorithm scores were selected for the model. Using these 28 selected features, the Logistic Regression model achieved the highest cross-validation accuracy of 82%.

5.1.12.2 *ReliefF Algorithm*

The Relief Algorithm can handle both nominal and numerical features but has limitations regarding incomplete data and its applicability to only two-class problems. To address these limitations, an extension called ReliefF was introduced. ReliefF is specifically designed for multi-class problems and behaves more robustly in noisy data sets (Robnik-Šikonja & Kononenko, 2003). Unlike Relief, ReliefF calculates

nearest hits and misses for each class and updates feature weights using the average values across all classes. This allows ReliefF to handle incomplete data and provide solutions for multi-class classification.

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
FolkSometimes	"0.098"	"0.141"	"0.696"	"0.486"	"	"1.103"
FolkNever	"-0.17"	"0.129"	"-1.32"	"0.187"	"	"0.844"
`FolkVery frequently`	"-0.147"	"0.176"	"-0.833"	"0.405"	"	"0.863"
PopSometimes	"-0.191"	"0.146"	"-1.307"	"0.191"	"	"0.826"
PopNever	"-0.181"	"0.222"	"-0.812"	"0.417"	"	"0.835"
`PopVery frequently`	"-0.089"	"0.15"	"-0.596"	"0.551"	"	"0.914"
Video_game_musicSometimes	"0.093"	"0.138"	"0.676"	"0.499"	"	"1.098"
Video_game_musicNever	"0.12"	"0.135"	"0.884"	"0.377"	"	"1.127"
`Video_game_musicVery frequently`	"0.084"	"0.167"	"0.503"	"0.615"	"	"1.087"
JazzSometimes	"0.066"	"0.137"	"0.485"	"0.627"	"	"1.069"
JazzNever	"0.214"	"0.127"	"1.694"	"0.09"	"	"1.239"
`JazzVery frequently`	"0.181"	"0.215"	"0.841"	"0.4"	"	"1.198"
MetalSometimes	"0.07"	"0.154"	"0.457"	"0.648"	"	"1.073"
MetalNever	"-0.027"	"0.139"	"-0.196"	"0.845"	"	"0.973"
`MetalVery frequently`	"-0.014"	"0.158"	"-0.092"	"0.927"	"	"0.986"
GospelSometimes	"-0.213"	"0.223"	"-0.955"	"0.34"	"	"0.808"
GospelNever	"-0.238"	"0.131"	"-1.809"	"0.07"	"	"0.788"
`GospelVery frequently`	"-0.281"	"0.428"	"-0.656"	"0.512"	"	"0.755"
RBSometimes	"-0.344"	"0.144"	"-2.38"	"0.017"	"	"0.709"
RBNever	"-0.106"	"0.142"	"-0.751"	"0.452"	"	"0.899"
`RBVery frequently`	"0.141"	"0.171"	"0.822"	"0.411"	"	"1.151"
`StreamingI do not use a streaming service.`	"-0.286"	"0.249"	"-1.152"	"0.249"	"	"0.751"
StreamingSpotify	"-0.12"	"0.19"	"-0.634"	"0.526"	"	"0.887"
`StreamingYouTube Music`	"-0.201"	"0.23"	"-0.873"	"0.383"	"	"0.818"
`StreamingOther streaming service`	"-0.231"	"0.266"	"-0.867"	"0.386"	"	"0.794"
ExploratoryNo	"0.024"	"0.12"	"0.198"	"0.843"	"	"1.024"
LatinSometimes	"0.008"	"0.185"	"0.042"	"0.966"	"	"1.008"
LatinNever	"0.19"	"0.134"	"1.416"	"0.157"	"	"1.209"
`LatinVery frequently`	"0.219"	"0.265"	"0.828"	"0.408"	"	"1.245"
RockSometimes	"-0.242"	"0.173"	"-1.397"	"0.162"	"	"0.785"
RockNever	"-0.056"	"0.203"	"-0.274"	"0.784"	"	"0.946"
`RockVery frequently`	"-0.19"	"0.178"	"-1.068"	"0.285"	"	"0.827"
ClassicalSometimes	"-0.072"	"0.126"	"-0.571"	"0.568"	"	"0.931"
ClassicalNever	"-0.161"	"0.14"	"-1.146"	"0.252"	"	"0.852"
`ClassicalVery frequently`	"-0.117"	"0.162"	"-0.719"	"0.472"	"	"0.89"
Foreign_languagesNo	"0.119"	"0.11"	"1.083"	"0.279"	"	"1.127"
KpopSometimes	"0.053"	"0.193"	"0.274"	"0.784"	"	"1.054"
KpopNever	"0.054"	"0.13"	"0.415"	"0.678"	"	"1.055"
`KpopVery frequently`	"0.083"	"0.194"	"0.426"	"0.67"	"	"1.086"
EDMSometimes	"0.214"	"0.145"	"1.473"	"0.141"	"	"1.238"
EDMNever	"-0.002"	"0.133"	"-0.012"	"0.99"	"	"0.998"
`EDMVery frequently`	"0.208"	"0.168"	"1.235"	"0.217"	"	"1.231"
HiphopSometimes	"-0.045"	"0.132"	"-0.339"	"0.735"	"	"0.956"
HiphopNever	"-0.067"	"0.149"	"-0.449"	"0.654"	"	"0.935"
`HiphopVery frequently`	"-0.106"	"0.169"	"-0.623"	"0.533"	"	"0.9"
LofiSometimes	"0.111"	"0.14"	"0.795"	"0.427"	"	"1.118"
LofiNever	"-0.037"	"0.132"	"-0.278"	"0.781"	"	"0.964"
`LofiVery frequently`	"0.021"	"0.179"	"0.117"	"0.907"	"	"1.021"
Hours2	"0.248"	"0.117"	"2.114"	"0.035"	"	"1.281"
Hours3	"0.433"	"0.197"	"2.2"	"0.028"	"	"1.542"
Hours4	"0.402"	"0.215"	"1.867"	"0.062"	"	"1.494"
low middle	"0.209"	"0.355"	"0.588"	"0.556"	"	"1.232"
middle high	"0.89"	"0.356"	"2.498"	"0.012"	"	"2.435"

Figure 5.1.14 ReliefF Algorithm used to model OCD data set

Based on the information provided, the OCD data set was analyzed using Ordinal Logistic Regression with feature selection based on the ReliefF Algorithm. ReliefF Algorithm, as shown in Figure 5.1.14, their statistical significance is shown. The model selected 18 features based on the ReliefF Algorithm. Using these selected features, the Ordinal Logistic Regression model achieved a cross-validation accuracy

of approximately 68.5%. 2 features are statistically significant at the $\alpha = 0.05$ level. Example, in comparison to those who rarely listen to R&B, individuals who sometimes listen to R&B are 30% less likely to exhibit a low level of OCD than a high level of OCD.

5.1.12.3 *RReliefF Algorithm*

RReliefF is an algorithm that extends the ReliefF feature selection method to regression problems. It is designed to handle data sets where the target variable is continuous, rather than categorical as in classification problems.

In regression problems, the goal is to predict a continuous target variable, such as a numerical value or a real number. RReliefF addresses the challenge of feature selection in regression tasks by considering the relative distances between the predicted target variable values of different observations.

Unlike the original ReliefF Algorithm, which calculates the nearest hits and misses for each class in multi-class classification problems, RReliefF focuses on determining the relative distances between predicted target variable values. This approach allows RReliefF to assess the relevance of features in explaining the variations in the continuous target variable.

By evaluating the influence of each feature on the predicted target variable values, RReliefF identifies the most important features for regression tasks. It helps in selecting a subset of features that have a significant impact on predicting the continuous target variable (Robnik-Šikonja & Kononenko, 2003).

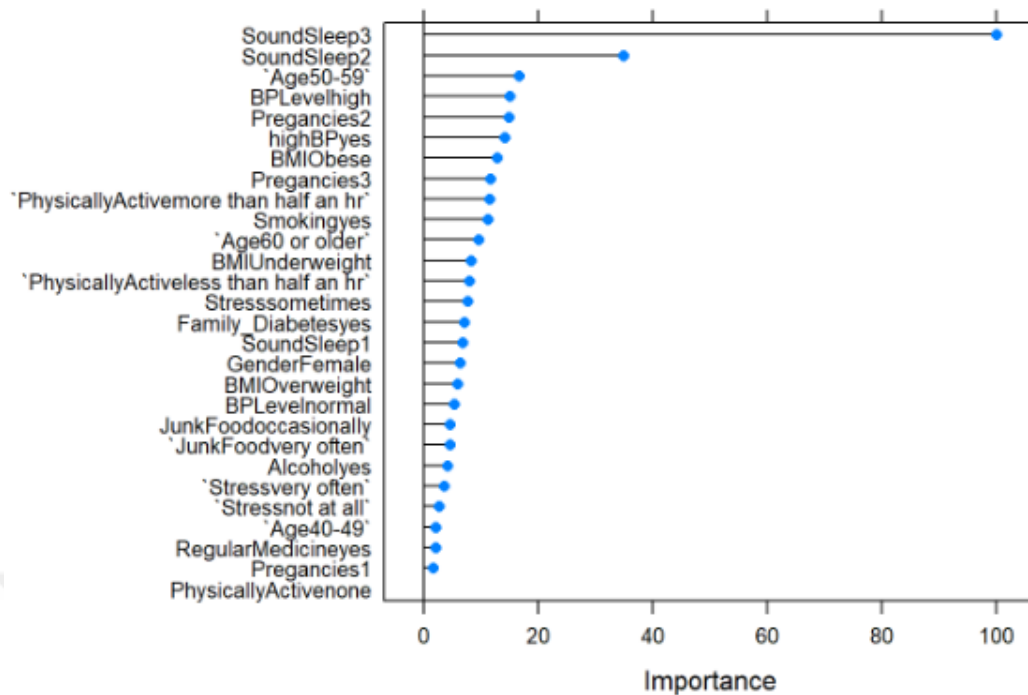


Figure 5.1.15 Ranking the importance of features in the Sleep Duration data set modelled by the RReliefF Algorithm

Based on the information provided, a Linear Regression was applied to the Sleep Duration data set using feature selection based on the RReliefF Algorithm. The data set initially had 17 features. By applying the RReliefF Algorithm, the features were ranked according to their relevance to the target variable in Figure 5.1.15. After ranking the features, the top 14 features with the highest RReliefF Algorithm scores were selected for the model. Using these 14 selected features, the Logistic Regression model achieved the highest cross-validation RMSE of 0.96.

5.2 Feature Subset Selection Algorithms

5.2.1 Correlation-based Feature Selection

Correlation-based Feature Selection (CFS) is a method that selects features by selecting a subset, not by ranking the features with a criterion value. This method is concerned with the correlations among the features themselves and between the

feature subset and the target variable. CFS focuses on finding subsets of features that are individually highly relevant to the target variable but have low correlation between them (Ladha & Deepa, 2011). It is often used with heuristics because it needs to perform a search among all candidate subsets (Palma-Mendoza et al., 2019). It has the following mathematical form:

$$M_s = \frac{k\bar{r}_{yf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}}.$$

M_s : Value assigned to a subset of s with k features determined heuristically

$\bar{r}_{yf}$: The mean of the correlations between each feature in the s subset and the target variable and of correlations between features in subset s

$\bar{r}_{ff}$: The mean of the correlations between features and features in subset s and the mean of the correlations between features in subset s

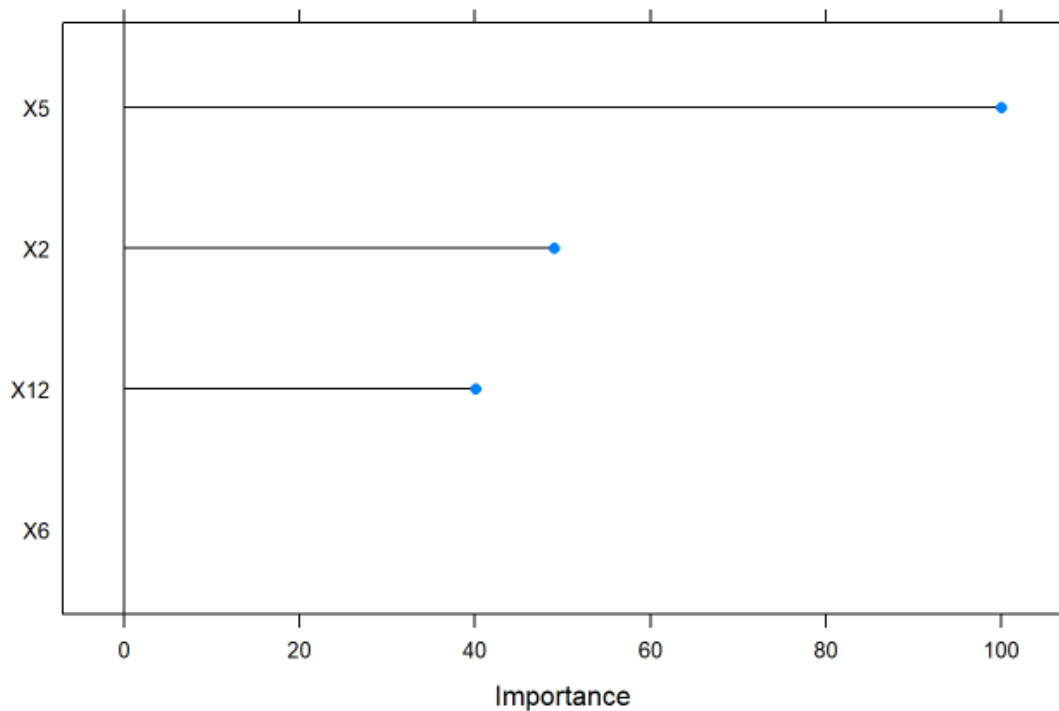


Figure 5.2.16 Subset determined by the Correlation-based Feature Selection of SENIC data set

Based on the information provided, a Linear Regression was applied to the SENIC

data set using feature selection based on the Correlation-based Feature Selection method. The data set initially had 8 features. By applying the Correlation-based Feature Selection method, the features were determined according to their relevance to the target variable in Figure 5.2.16. After determining the features, the top 4 features with the highest Correlation-based Feature Selection method scores were selected for the model. Using these 4 selected features, the Linear Regression model achieved the lowest cross-validation RMSE of 0.72.

5.2.2 Consistency Feature Subset Selection

The Consistency-based feature selection method evaluates the stability or Consistency of features across different subsets of the data set. It assumes that relevant features will consistently produce similar results when the data is randomly sampled or partitioned (Olasehinde et al., 2018). This process requires the identification of all potential observation combinations for a given feature set. For example, let X_1 and X_2 be binary features, the observation combinations of these features are as follows: $\{(0,0), (0,1), (1,0), (1,1)\}$. For each of the above combinations, the number of occurrences of each combination in the target variable, D_i , is determined. The most common combination is determined as the majority class M_i . The sum of the differences between the number of occurrences and the majority numbers provides an indication of the number of inconsistencies. The consistency ratio is calculated by comparing the number of inconsistencies with the total number of observations. The method mentioned above can be expressed mathematically as follows:

$$Consistency = 1 - \frac{\sum_{i=1}^k (|D_i| - |M_i|)}{N}.$$

k : Number of combinations

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
Age2	"-0.038"	"0.118"	"-0.323"	"0.747"	"	"0.963"
Age3	"-0.643"	"0.268"	"-2.401"	"0.016"	" *	"0.526"
Hours2	"0.234"	"0.113"	"2.073"	"0.038"	" *	"1.264"
Hours3	"0.435"	"0.19"	"2.285"	"0.022"	" *	"1.544"
Hours4	"0.46"	"0.21"	"2.196"	"0.028"	" *	"1.584"
Foreign_languagesNo	"0.145"	"0.103"	"1.407"	"0.16"	"	"1.156"
ClassicalSometimes	"-0.035"	"0.122"	"-0.282"	"0.778"	"	"0.966"
ClassicalNever	"-0.173"	"0.133"	"-1.295"	"0.195"	"	"0.841"
`ClassicalVery frequently`	"-0.178"	"0.156"	"-1.143"	"0.253"	"	"0.837"
LatinSometimes	"0.059"	"0.181"	"0.325"	"0.745"	"	"1.061"
LatinNever	"0.147"	"0.126"	"1.165"	"0.244"	"	"1.159"
`LatinVery frequently`	"0.219"	"0.254"	"0.863"	"0.388"	"	"1.245"
LofiSometimes	"0.142"	"0.135"	"1.051"	"0.293"	"	"1.152"
LofiNever	"0.02"	"0.126"	"0.155"	"0.877"	"	"1.02"
`LofiVery frequently`	"0.05"	"0.168"	"0.299"	"0.765"	"	"1.052"
MetalSometimes	"0.026"	"0.149"	"0.174"	"0.862"	"	"1.026"
MetalNever	"-0.082"	"0.133"	"-0.614"	"0.539"	"	"0.921"
`MetalVery frequently`	"-0.064"	"0.154"	"-0.417"	"0.677"	"	"0.938"
PopSometimes	"-0.202"	"0.14"	"-1.436"	"0.151"	"	"0.817"
PopNever	"-0.219"	"0.219"	"-1"	"0.317"	"	"0.803"
`PopVery frequently`	"-0.128"	"0.144"	"-0.893"	"0.372"	"	"0.879"
RBSometimes	"-0.304"	"0.138"	"-2.207"	"0.027"	" *	"0.738"
RBNever	"-0.121"	"0.132"	"-0.918"	"0.358"	"	"0.886"
`RBVery frequently`	"0.154"	"0.154"	"0.996"	"0.319"	"	"1.166"
RockSometimes	"-0.175"	"0.165"	"-1.061"	"0.289"	"	"0.84"
RockNever	"-0.058"	"0.198"	"-0.294"	"0.768"	"	"0.943"
`RockVery frequently`	"-0.124"	"0.166"	"-0.748"	"0.454"	"	"0.883"
low middle	"0.333"	"0.235"	"1.415"	"0.157"	"	"1.395"
middle high	"1.004"	"0.237"	"4.226"	"0"	" *	"2.728"

Figure 5.2.17 Consistency Feature Subset Selection used to model OCD data set

Ordinal Logistic Regression was applied to the OCD data set using feature selection based on Consistency Feature Subset Selection. Statistical significances are given as shown in Figure 5.2.17. Consistency Feature Subset Selection selected 10 features out of 24 features. The cross-validation accuracy of the model is approximately 68.2%. Of these 10 features, 3 were statistically significant at the $\alpha = 0.05$ level. For example, in comparison to those who rarely listen to R&B, individuals who sometimes listen to R&B are 26% less likely to exhibit a low level of OCD than a high level of OCD.

CHAPTER SIX

WRAPPER METHODS

Wrapper methods are placed on a predetermined model. A search procedure is defined for all possible subsets, subsets are created and evaluated. Evaluation is obtained by model performance. In short, Wrapper methods “wrap” a search algorithm around the model to search the feature subset space (Saeys et al., 2007). In the subset search procedure, the number of features grows exponentially, so heuristics are also used to find an optimal subset. The Wrapper method is presented in Figure 6.0.1 in the form of a general taxonomy.

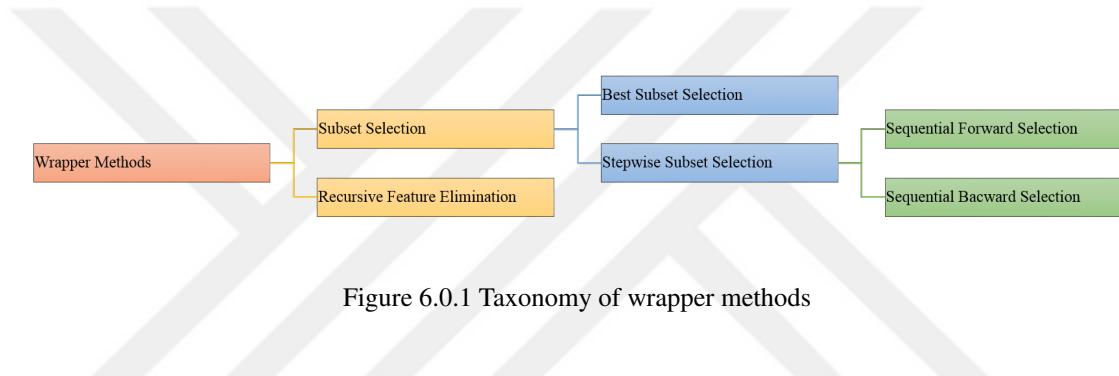


Figure 6.0.1 Taxonomy of wrapper methods

6.1 Subset Selection Methods

6.1.1 *Best Subset Selection*

The Best Subset Selection wraps around the least squares regression model (Gareth et al., 2013). Iterates the model for each possible combination of features. The optimal model is decided by comparing the results of the models. The algorithm searches through all possible feature subsets and evaluates them based on a specified criterion, such as minimizing root mean squared error or maximizing R-squared. Best Subset Selection algorithm: A performance metric or criterion is first selected to evaluate subsets. Creates models using all possible feature combinations in the given feature set. Then, performance metrics are compared for each model. According to the evaluation criteria, the subset of features that provide the highest performance is selected (Kuhn et al., 2013) .

```

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.09271  -0.39987  -0.12960  -0.00084   2.78859

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9438     0.4961  -5.934 2.96e-09 ***
X3            -2.5081     0.4921  -5.097 3.45e-07 ***
X4           -11.8873     5.5769   -2.132  0.03304 *
X5            -1.8474     0.9222   -2.003  0.04514 *
X7             11.0974     5.1309    2.163  0.03055 *
X8              3.1035     0.9456    3.282  0.00103 **
X11           -2.5824     1.2528   -2.061  0.03928 *
X13           -1.8705     0.9622   -1.944  0.05189 .
X14           -8.1772     3.1393   -2.605  0.00919 **
X15           -2.8476     0.8467   -3.363  0.00077 ***
X16             9.7251     3.3255    2.924  0.00345 **
X18             1.0116     0.6201    1.631  0.10281
X19             4.0036     1.4576    2.747  0.00602 **
X20           -5.0706     1.8051   -2.809  0.00497 **
X21           -1.7175     0.6555   -2.620  0.00878 **
X22             1.7506     0.8656    2.022  0.04313 *
X23             2.0027     0.8688    2.305  0.02115 *
X24           19.1078     5.9718    3.200  0.00138 **
X25             2.1879     1.1533    1.897  0.05781 .
X26           -7.2291     4.0599   -1.781  0.07498 .
X27          -10.5186     4.2581   -2.470  0.01350 *
X29           -4.6725     1.5868   -2.945  0.00323 **
X30             4.3713     1.5774    2.771  0.00559 **
X32           -1.1546     0.8290   -1.393  0.16370
X35             0.8608     0.2896    2.972  0.00296 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212.52  on 193  degrees of freedom
Residual deviance: 102.13  on 169  degrees of freedom
AIC: 152.13

Number of Fisher Scoring iterations: 7

```

Figure 6.1.2 Best Subset Selection used to model Breast Cancer data set

In Figure 6.1.2, 24 features were selected for the Breast Cancer data set by the Best Subset Selection method. 19 features were statistically significant at $\alpha = 0.05$ for explaining the variation in the probability of having recurrent Breast Cancer. For

example the feature X_{35} : Lymph status which reports the number of positive axillary lymph nodes during surgery has a coefficient estimate of 0.8608. The odds ratio for this feature is then about 2.37 meaning that a 1 unit increase in X_{35} results in a 137% increase in the odds of the cancer recurring.

6.1.2 Stepwise Selection

The Best Subset Selection are that the search space grows when there are many features. Therefore, computational difficulty can lead to increased computation time, overfitting of coefficient estimates, and high variance (Gareth et al., 2013). For this reason, Stepwise Selections have been developed as an alternative to the Best Subset Selection. Stepwise Selection is a feature selection technique used in statistical modeling to systematically select a subset of features from a larger set of potential predictors. It involves iteratively adding or removing features based on their statistical significance or predictive power. There are two common variants of Stepwise Selection: Sequential Forward Selection and Sequential Backward Elimination (Hastie et al., 2009).

6.1.2.1 Sequential Forward Selection

Sequential Forward Selection (SFS) is a variant of the Stepwise Selection algorithm, which starts with an empty model and iteratively adds features to the model until no additional features meet the predefined criteria for inclusion. The difference between SFS and Forward Selection is that SFS evaluates each feature independently and selects the best feature at each step, whereas Forward Selection considers all possible combinations of features at each step (Gareth et al., 2013). Sequential Forward Selection algorithm: It starts with an empty model. In each iteration, a feature that maximizes the measure of performance is added. After adding the new feature, the performance of the model is evaluated and additional features added, the process is repeated until it does not meet the predefined criteria for

inclusion. After the SFS process is completed, the performance of the final selected model is evaluated using appropriate measurements.

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.78506  -0.60481  -0.25630  -0.03079   2.58339

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0565     0.3154  -6.520 7.04e-11 ***
X3            -1.7288     0.3461  -4.994 5.90e-07 ***
X15           -0.5058     0.3308  -1.529 0.126167
X27           -8.1319     3.0397  -2.675 0.007467 **
X35            0.5201     0.2031   2.561 0.010437 *
X4           -12.2055     4.0178  -3.038 0.002383 **
X24           10.5289     3.2406   3.249 0.001158 **
X7             9.3576     3.7845   2.473 0.013412 *
X5            -0.3816     0.2503  -1.525 0.127307
X21           -0.9289     0.3796  -2.447 0.014396 *
X23            1.5719     0.4759   3.303 0.000957 ***
X13           -2.1071     0.6409  -3.287 0.001011 **
X8             1.0178     0.3949   2.577 0.009952 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212.52  on 193  degrees of freedom
Residual deviance: 136.94  on 181  degrees of freedom
AIC: 162.94

Number of Fisher Scoring iterations: 6
```

Figure 6.1.3 Sequential Forward Selection used to model Breast Cancer data set

In Figure 6.1.3, 12 features are selected for the Breast Cancer data set using a Sequential Forward Selection method. 10 features were statistically significant at $\alpha = 0.05$ and explained variation in Breast Cancer recurrence probability. For example, the coefficient estimate for the feature X_{35} : lymph status, which indicates the number of positive axillary lymph nodes during surgery, is 0.5201. The odds ratio for this feature is approximately 1.68. This means that when X_{35} increases by 1 unit, the probability of cancer recurrence increases by 68%.

6.1.2.2 *Sequential Backward Elimination*

Sequential Backward Elimination (SBE) is a variant of the Stepwise Selection algorithm, where the feature selection process starts with a model that includes all potential predictors and iteratively removes features until no further elimination improves the predefined criteria (Gareth et al., 2013). Sequential backward selection algorithm: It starts with a model that contains all potential predictors. In each iteration, a feature that minimizes the performance criterion or has the least impact on the performance of the model is removed. After the feature is removed, the performance of the model is evaluated and the process is repeated until the removal of additional features does not significantly affect the performance of the model. After the SBE process is completed, the performance of the final selected model is evaluated using appropriate measurements.

```

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.09271  -0.39987  -0.12960  -0.00084   2.78859

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9438     0.4961  -5.934 2.96e-09 ***
X3            -2.5081     0.4921  -5.097 3.45e-07 ***
X4           -11.8873     5.5769   -2.132  0.03304 *
X5            -1.8474     0.9222   -2.003  0.04514 *
X7             11.0974     5.1309    2.163  0.03055 *
X8              3.1035     0.9456    3.282  0.00103 **
X11           -2.5824     1.2528   -2.061  0.03928 *
X13           -1.8705     0.9622   -1.944  0.05189 .
X14           -8.1772     3.1393   -2.605  0.00919 **
X15           -2.8476     0.8467   -3.363  0.00077 ***
X16             9.7251     3.3255    2.924  0.00345 **
X18             1.0116     0.6201    1.631  0.10281
X19             4.0036     1.4576    2.747  0.00602 **
X20           -5.0706     1.8051   -2.809  0.00497 **
X21           -1.7175     0.6555   -2.620  0.00878 **
X22             1.7506     0.8656    2.022  0.04313 *
X23             2.0027     0.8688    2.305  0.02115 *
X24            19.1078     5.9718    3.200  0.00138 **
X25             2.1879     1.1533    1.897  0.05781 .
X26           -7.2291     4.0599   -1.781  0.07498 .
X27          -10.5186     4.2581   -2.470  0.01350 *
X29           -4.6725     1.5868   -2.945  0.00323 **
X30             4.3713     1.5774    2.771  0.00559 **
X32           -1.1546     0.8290   -1.393  0.16370
X35             0.8608     0.2896    2.972  0.00296 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212.52  on 193  degrees of freedom
Residual deviance: 102.13  on 169  degrees of freedom
AIC: 152.13

Number of Fisher Scoring iterations: 7

```

Figure 6.1.4 Sequential Backward Elimination used to model Breast Cancer data set

In Figure 6.1.4, 24 features are selected for the Breast Cancer data set using a sequential backward selection method. 19 features were statistically significant at the $\alpha = 0.05$ level in explaining variation in Breast Cancer recurrence probability. For

example, the coefficient estimate for feature X_{22} , which is one of the features that defines the measured shape properties of each nucleus, is 1.75. The odds ratio for this variant is approximately 5.76. This means that a one unit increase in X_{22} increases the probability of cancer recurrence by 476%.

6.2 Recursive Feature Elimination

RFE stands for Recursive Feature Elimination. It was developed as an alternative to the backward elimination method. While the backward elimination method uses combinatorial search to calculate feature ranking, RFE enhances the process of selecting features based on their ranking by eliminating the assumption of orthogonality inherent to correlation-based methods. The underlying concept of RFE is to iteratively remove features from the dataset and fit the remaining features into a model, thus determining which features have the greatest impact on the model's performance. This process is summarized in detail in the following quote (Guyon et al., 2002).

The RFE process can be summarized as follows: Initially, a prediction model is trained using all input features. The importance of each feature is then evaluated using a specific metric, such as accuracy or mean squared error. The least important features are then removed from the dataset and the model is retrained on the surviving features. These steps are repeated until the desired number of features is reached. Finally, the final feature set is used to train the final model.

RFE can be performed using a variety of machine learning algorithms, including linear regression, logistic regression, or support vector machines (Zhang et al., 2020).

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables  RMSE Rsquared  MAE  RMSESD RsquaredSD  MAESD Selected
1 0.8295    0.3273 0.6755 0.14852    0.2215 0.11376
2 0.7845    0.4264 0.6426 0.10442    0.2045 0.08979
3 0.7658    0.4348 0.6302 0.09754    0.2179 0.07577
4 0.7257    0.4831 0.5854 0.10171    0.2185 0.08128      *
5 0.7461    0.4552 0.6016 0.14191    0.2231 0.11246
8 0.7591    0.4538 0.6026 0.15412    0.2102 0.11170

The top 4 variables (out of 4):
X5, X2, X6, X12
```

Figure 6.2.5 Ranking the importance of features in the SENIC data set modelled by the Recursive Feature Elimination

The RFE approach with Linear Regression modelling selected 4 out of 9 features for the SENIC dataset in Figure 6.2.5. The RFE model built with the features Routine culturing, Length, Routine Xray and Percentage of facilities available explains the probability of infection with a low cross-validation RMSE value of 0.7257.

CHAPTER SEVEN

EMBEDDED METHODS

Embedded methods are Machine Learning algorithms that make feature selection and model training at the same time. In other words, the feature selection method is Embedded in the algorithms. The Embedded method is presented in Figure 7.0.1 in the form of a general taxonomy.

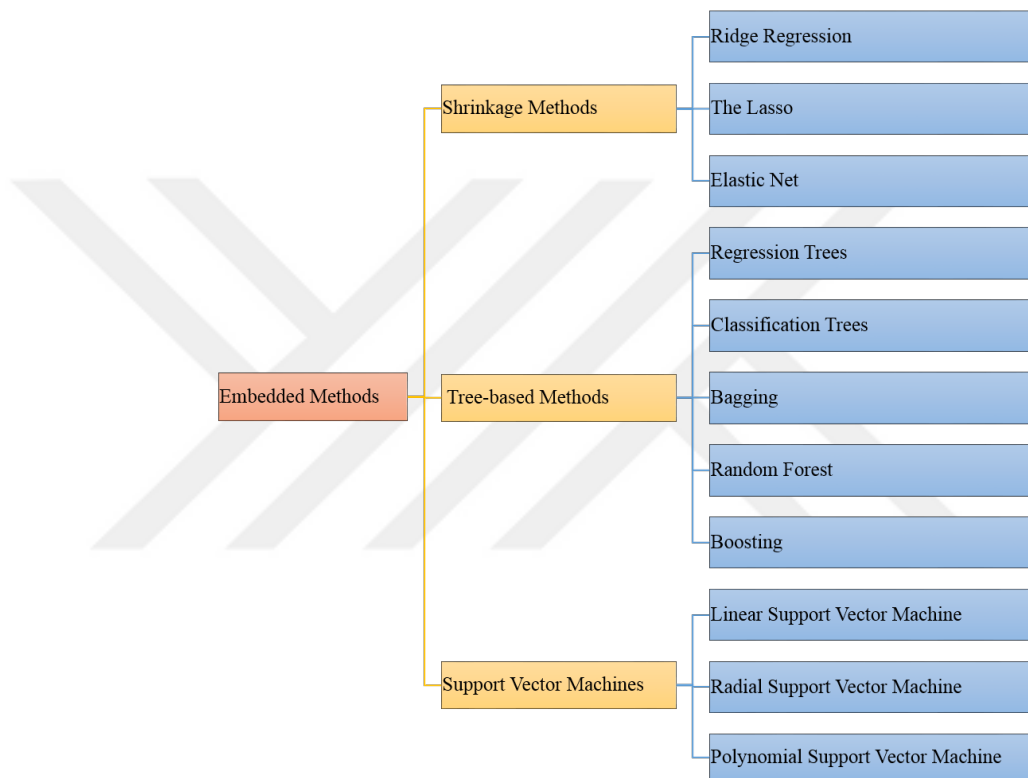


Figure 7.0.1 Taxonomy of embedded methods

7.1 Shrinkage Methods

Shrinking methods have been developed as a solution to multicollinearity problems and overfitting problems. It works by adding a penalty term that aims to reduce the size of coefficients in the model and reduces variance.

λ : Hyperparameter controlling the strength of the penalty term

When $\lambda = 0$, there is no penalized parameter and the model is standard Linear

Regression. When $\lambda = 1$, all coefficients disappear and a null model is formed. In short, as the lambda increases, more coefficients are set to zero and are removed from the model. However, as the lambda increases, the variance decreases but the bias increases.

α : Hyperparameter controlling a linear combination of L1 and L2 penalty terms

When $\alpha = 0$, the elastic network is reduced to Ridge Regression, and when $\alpha = 1$, it is reduced to Lasso regression. When alpha is between 0 and 1, the Elastic Net balances Lasso's sparsity with the improved numerical stability of the Ridge Regression.

7.1.1 Ridge Regression

Ridge Regression is a Linear Regression algorithm that adds L2 editing terms to the sum of squares error (Hastie et al., 2009). The Ridge Regression adds a L2 regulating term that encourages coefficients to be small, but does not necessarily have to be zero. The L2 regulation term acts as a penalty for the magnitude of coefficients, reducing the deviation in the model and providing interpretable coefficient estimates. It has the following mathematical form:

$$L2_{penalty} = \lambda \sum_{j=1}^p \beta_j^2,$$

$$\sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2.$$

Ridge Regression model is modeled with the penalty term L2 added to the sum of squares error 2.1.

For instance, when examining the Ridge Regression application of the Breast Cancer data set, Figure 7.1.2 displays the cross-validation accuracies of the λ hyperparameter between 0 and 1.

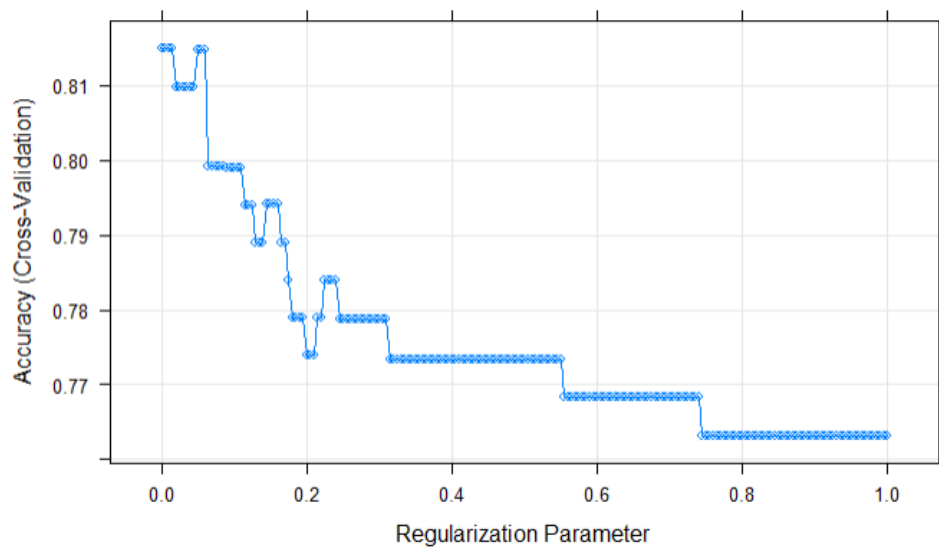


Figure 7.1.2 Parameter tuning for Ridge Regression

The optimal hyperparameter value of the model was determined to be $\lambda = 0.015$, resulting in a cross-validation accuracy of 81.5%.

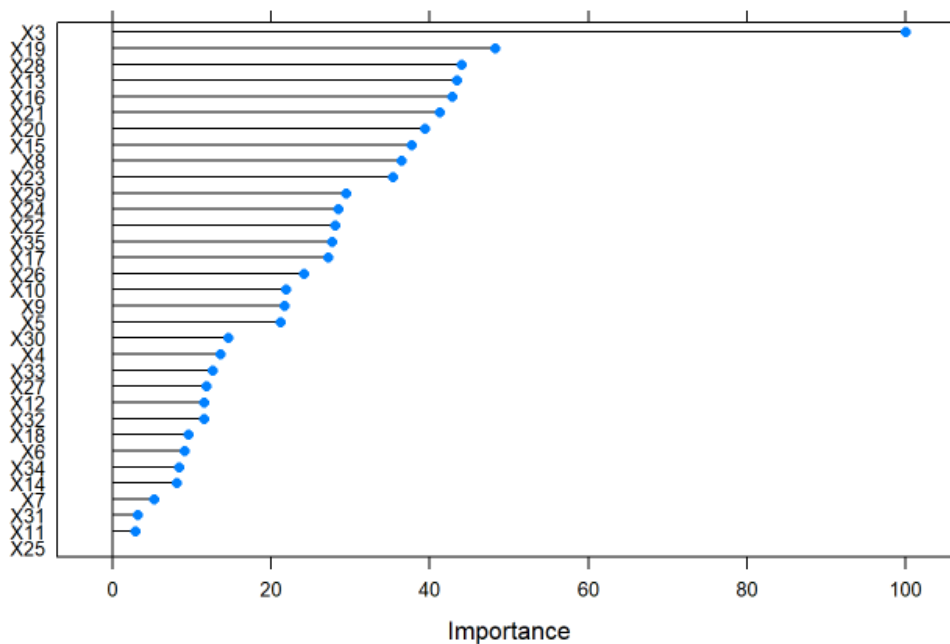


Figure 7.1.3 Ranking the importance of features in the Breast Cancer data set modelled by the Ridge Regression

Additionally, Figure 7.1.3 provides the features importance for Ridge Regression.

The time feature is the most important, followed by the structural features of each cell nucleus.

7.1.2 The Lasso

Least Absolute Shrinkage and Selection Operator (LASSO) is a Linear Regression algorithm that adds L1 editing terms to the sum of squares error (Hastie et al., 2009). The LASSO adds a L1 regulating term that encourages some of the coefficients to be exactly zero. The L1 regulation term acts as a penalty for the magnitude of coefficients, resulting in feature selection. It has the following mathematical form:

$$L1_{penalty} = \lambda \sum_{j=1}^p |\beta_j|,$$

$$\sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|.$$

The Lasso model is modeled with the penalty term L1 added to the sum of squares error 2.1.

When examining the Lasso application of the Breast Cancer data set, Figure 7.1.4 displays the cross-validation accuracies of the λ hyperparameter between 0 and 1.

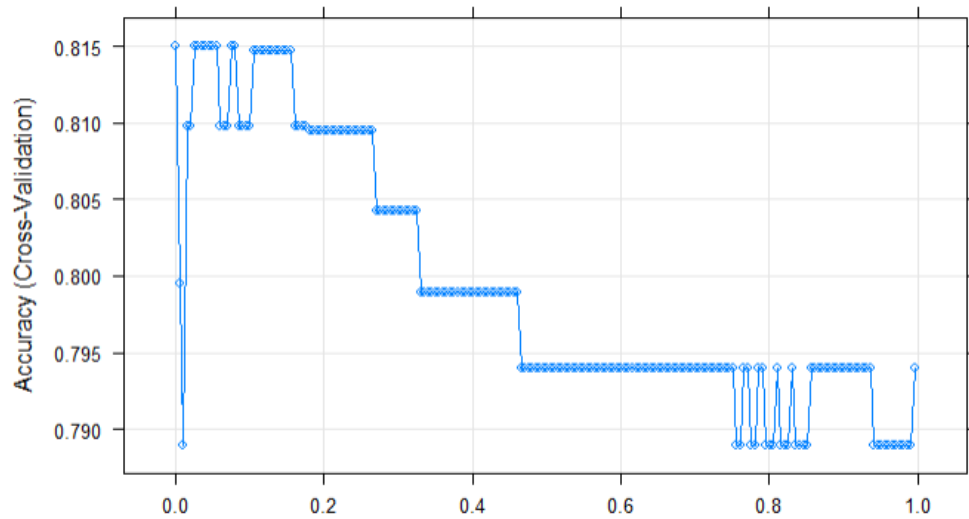


Figure 7.1.4 Parameter tuning for The Lasso

The optimal hyperparameter value of the model was determined to be $\lambda = 0.081$, resulting in a cross-validation accuracy of 81.5%.

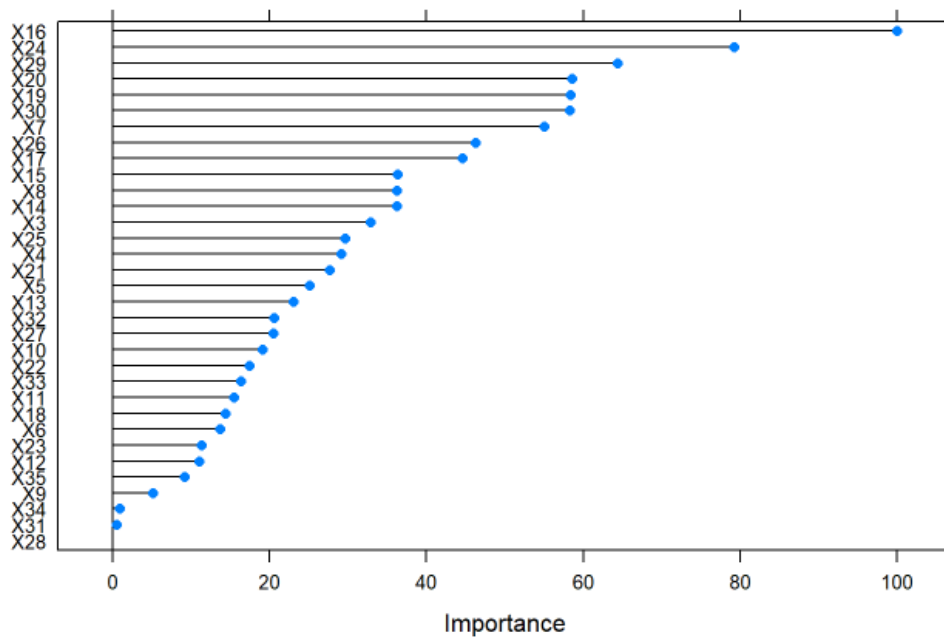


Figure 7.1.5 Ranking the importance of features in the Breast Cancer data set modelled by the Lasso

Furthermore, Figure 7.1.5 provides the feature importance for the Lasso. The structural features of cell nucleus are the most important.

7.1.3 Elastic Net

Elastic Net is a hybrid regularization technique that combines L1 and L2 regularization methods. It now adds both L1 and L2 regularization terms to the sum of squares error, providing a balance between Lasso's feature selection capabilities and Ridge Regression's stability (Hastie et al., 2009). Elastic Net is extensively utilized in various Machine Learning applications, including time series estimation, Logistic Regression, and Linear Regression.

$$Elasticnet_{penalty} = \sum_{j=1}^p (\alpha |\beta_j| + (1 - \alpha) \beta_j^2),$$

$$\sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p (\alpha |\beta_j| + (1 - \alpha) \beta_j^2).$$

Elastic Net model is modeled with the penalty term L1 and L2 added to the sum of squares error Equation 2.1.

For instance, Figure 7.1.6 displays the cross-validation accuracies of λ and α hyperparameters between 0 and 1 when examining the Elastic Net application of the Breast Cancer data set.

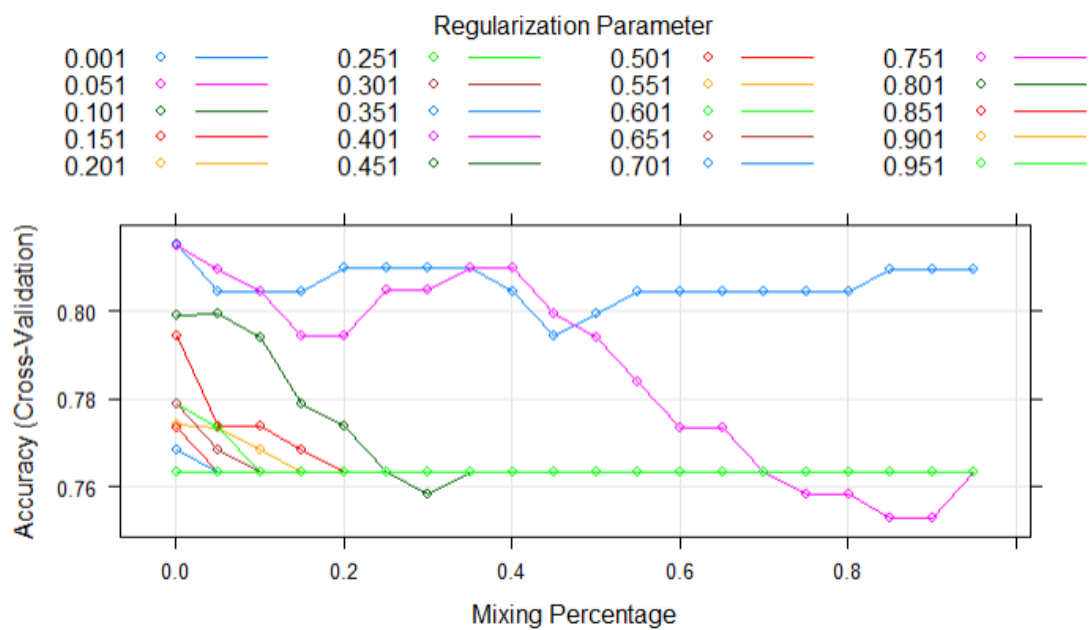


Figure 7.1.6 Parameter tuning for Elastic Net

The optimal hyperparameter values for the model were determined as $\alpha = 0.001$ and $\lambda = 0.001$, resulting in a cross-validation accuracy of 81.5%.

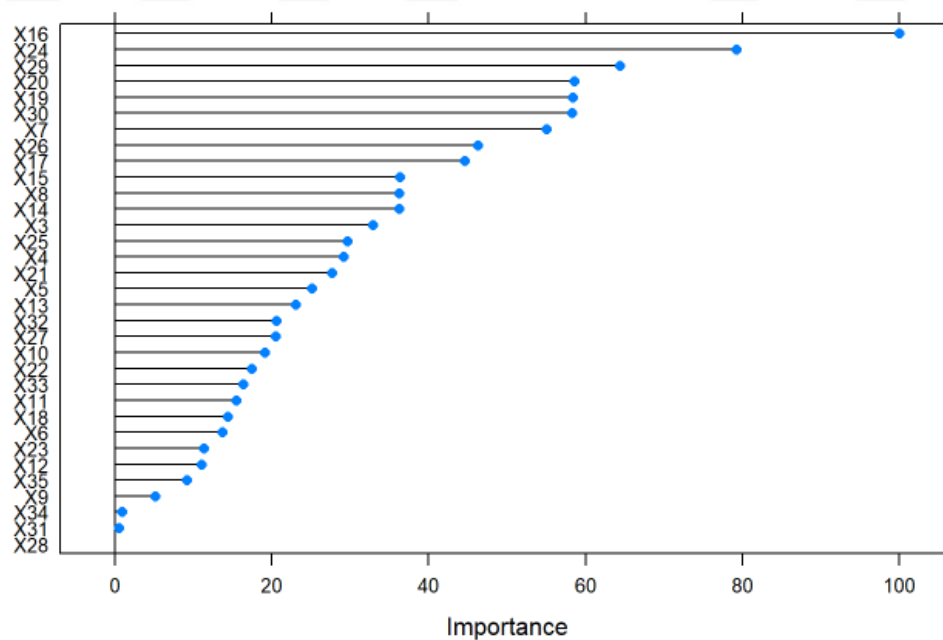


Figure 7.1.7 Ranking the importance of features in the Breast Cancer data set modelled by the Elastic Net

Additionally, Figure 7.1.7 provides information on the feature importance for Elastic Net. The time feature is the most important, followed by the structural features of each cell nucleus.

7.2 Tree-based Methods

Tree-based methods are a family of Machine Learning algorithms that utilize decision trees as a fundamental building block for making predictions or classifying data points. Decision trees are hierarchical structures composed of nodes and branches. Each node in the tree symbolises a feature whilst the branches illustrate the potential results or values that stem from that feature. The leaves of the tree indicate the final predictions or classifications (Hastie et al., 2009).

The process of Tree-based methods consists of several essential steps. Initially, the data set is divided into various subsets by partitioning it based on different feature values. The objective is to identify the optimal splits that maximise Information Gain or minimise impurity. The partitioning occurs recursively, giving rise to a tree-like structure. Each node in the decision tree corresponds to a feature and split condition, and the branches reflect the potential outcomes. The decision tree is built to establish subsets that are homogeneous in relation to the target variable. The final predictions or classifications are made at the leaf nodes. In regression tasks, the mean or median of the target variable is usually the predicted value in that specific leaf, while in classification tasks, the majority class label is assigned to the leaf node. Ensemble methods such as Random Forest and Boosting entail the creation of several decision trees that combine their predictions to enhance performance. Each tree in the ensemble undergoes training on a random subset of the data or features, and the final forecast is attained by aggregating the predictions from all the trees. This method improves predictive accuracy and robustness (Hastie et al., 2009).

7.2.1 Regression Trees

A Regression Tree is a decision tree-based algorithm used for regression analysis, where the goal is to predict a continuous target variable. It partitions the feature space into regions based on feature values and learns a piecewise constant function within each region. The splitting of the data is performed recursively based on selected features and split points that optimize certain criteria, such as minimizing the variance of the target variable within each partition (Hastie et al., 2009).

The objective is to partition the feature space, which includes the possible values of $X_1, X_2, \dots, X_p$, into j distinct and non-overlapping regions, $R_1, R_2, \dots, R_j$. The regions are formed as high-dimensional boxes or rectangles to simplify the model and aid interpretation. A same prediction is made for each observation that falls into a particular region R_j . This predictive value is an average of target values for training observations within that region. The regions $R_1, R_2, \dots, R_j$ are constructed via the recursive binary splitting method, which follows a top-down, greedy approach. The aim is to identify the feature X_j and the cutpoint s that yield the largest decrease in the sum of squared errors in each region. This iterative process entails choosing the optimal feature and cutpoint that minimizes the SSE. However, overfitting can be a potential issue with this method. Pruning the tree using the cost-complexity pruning technique is one solution. This process involves considering a sequence of subtrees, indexed by a tuning parameter α , and selecting the subtree that minimises a trade-off between complexity and fit to the training data. By incrementing α , the branches are pruned from the tree, resulting in a smaller subtree. The most optimal α value can be determined using cross-validation. Applying cost complexity pruning aids in regulating the complexity of the tree, reducing overfitting, and improving generalisation to unseen data (Hastie et al., 2009). The pruning method is characterized by the following mathematical formula:

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|.$$

There is a subtree T , which is a subset of the original tree T_0 , for every value of the tuning parameter α . This sub-tree is selected to minimize the sum of squared errors within each region R_m , in addition to α times the number of terminal nodes present within the sub-tree $|T|$. The objective is to achieve an equilibrium between the fit of the subtree to the training data and its complexity, which is controlled by α . As α increases, the penalty for having more terminal nodes in the subtree increases, leading to a smaller and potentially simpler tree. This trade-off is important for avoiding overfitting and finding an optimal subtree that balances prediction accuracy and model complexity.

The presented model is an example of a decision tree created using the Regression Tree method to predict sleep duration based on features such as age, physical activity, and age group.

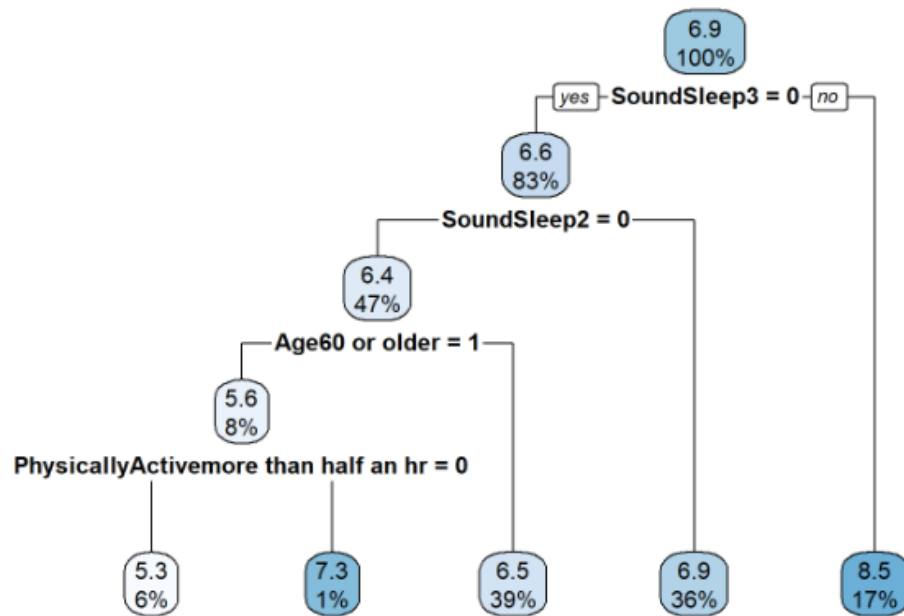


Figure 7.2.8 Regression Tree used to model Sleep Duration data set

Figure 7.2.8 shows the decision tree, which provides insight into the importance of these features in predicting sleep duration. The data set is split at the root node based on the SoundSleep feature, demonstrating the significance of sleep deep in predicting sleep duration. Subsequent divisions are based on other features such as age and physical activity. For instance, individuals aged 60 or above who engage in less

physical activity tend to have lower sleep duration, while younger individuals who engage in more physical activity tend to have better sleep duration.

7.2.2 *Classification Trees*

In Classification Trees, the aim is to predict a categorical response, rather than a numerical one. The predicted category for an observation in a Classification Tree is determined by the most common category among the training observations in the region to which it belongs. The fraction of training observations in a region that do not belong to the most common class is measured by the classification error rate. It is a measure employed to assess the quality of splits in Classification Trees. However, it is not sensitive enough, hence the Gini Index is often used as an alternative measure. The Gini Index assesses the purity of nodes in the tree by examining the distribution of classes. The Gini coefficient is similarly efficient in capturing node purity. In constructing a Classification Tree, either the Gini Index is employed to ascertain the optimal splits. Even with qualitative features, decision trees can still be created by assigning qualitative values to different branches. Node purity assumes significance, as it increases the certainty in predicting the class for test observations. Even if a split does not decrease the classification error, it can enhance the Gini Index by augmenting the purity of the node. Gini Index formula 5.4 is described.

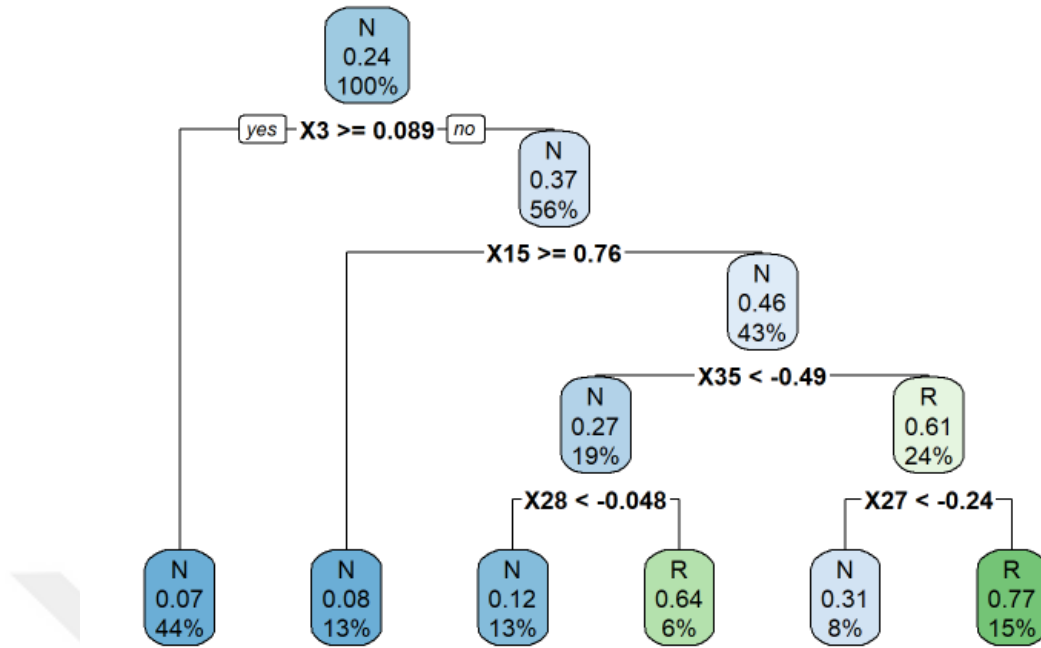


Figure 7.2.9 Classification Tree used to model Breast Cancer data set

The provided model is a decision tree created using the Classification Trees algorithm to predict the recurrence or non-recurrence of Breast Cancer based on various features obtained from cell nuclei and clinical data. Figure 7.2.9 displays the decision tree of the model. The root node of the tree splits the data set based on feature X_3 , indicating that time plays a crucial role in predicting cancer recurrence. Further divisions were based on additional features, such as X_{15} and X_{35} , which represent the structural features of the cell nucleus, particularly the cancer cell's texture2 and the number of lymph nodes, respectively. The model demonstrated approximately 73% accuracy in cross-validation when predicting results, with an optimal complexity parameter (C_p is a parameter used to determine the depth of trees and the extent of their branching.) of 0.018 selected through cross-validation.

7.2.3 Bagging

The creation of multiple bootstrapped training sets from the original data set is the basic concept of Bagging. In bootstrapping, each training set is used to train a

different decision tree. This results in a collection of individual trees with high variance but low bias. The predictions from these trees are then averaged to produce a single low-variance model (Hastie et al., 2009). This has the following mathematical form:

$$\hat{f}_{bag}(X) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(X).$$

B : the number of prediction models generated through bootstrapping

b : the b -th model among the B models generated. In the context of bagging, each model indexed by the value of 'b' is trained on a distinct bootstrapped sample of the original data.

$\hat{f}^b(X)$: the prediction generated by the b -th individual model (in this case, a decision tree) for a given inputs X .

By randomly selecting samples from the original training set with replacement, bootstrapped training sets are generated. Each bootstrapped training set is used to build a separate Regression Tree. These trees are typically grown deep, meaning they are allowed to have many levels and capture complex patterns in the data. Each individual Regression Tree has high variance because of its depth and the different training data used for its construction. However, these trees tend to have low bias, meaning they can closely fit the training data. The predictions of the individual trees are directly averaged for regression problems. However, for classification problems, a majority vote is used to determine the final prediction. Each tree predicts the class for a given test observation, and the overall prediction is the most commonly occurring class among the predictions of all the trees.

While interpreting the assortment of bagged trees may pose challenges, an overall feature importance summary can be attained by evaluating the reduction in SSE (for Regression Trees) or Gini Index (for Classification Trees) achieved through splits over each feature. Larger reductions indicate more important features, offering insights into their impact on explaining variability or improving classification accuracy within the

ensemble (Hastie et al., 2009).

Bagging is a technique that involves training multiple decision trees on bootstrapped samples of the original data set and then combining their predictions to increase overall accuracy and robustness. In the case of the Breast Cancer data set, this ensemble consists of 194 samples and 33 features, including measurements such as time, specific properties of the cell nucleus, tumour size, and lymph node status.

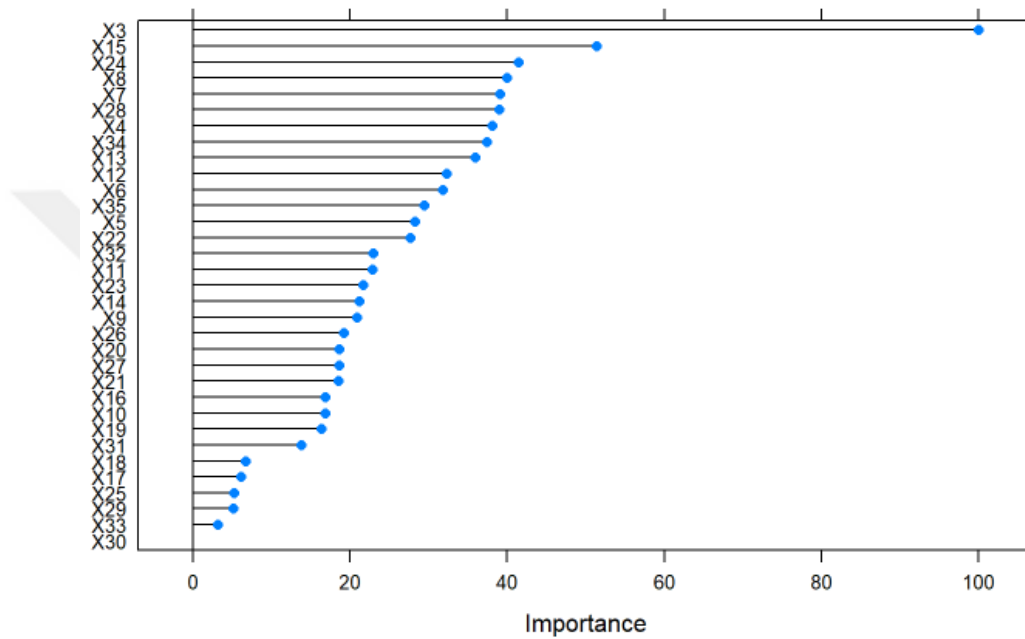


Figure 7.2.10 Ranking the importance of features in the Breast Cancer data set modelled by the Bagging

Resampling results indicate an overall accuracy of 77.87%. Furthermore, Figure 7.2.10 displays the feature importance scores, indicating the relative contribution of each feature to the model. Notably, features such as X_3 (time), X_{15} (texture2 of the cancer cell), X_{24} (radius3 of the cancer cell) and X_8 (smoothness1 of the cancer cell) exhibit higher importance scores, demonstrating their significant role in predicting cancer outcomes.

7.2.4 *Random Forest*

Random Forests are an improvement on bagged trees, as they aim to reduce correlation among the trees and improve the accuracy of predictions. By randomly selecting a subset of features at each split during the construction of decision trees, Random Forests decorrelate the trees and provide more reliable predictions. This approach prevents all trees from relying heavily on a dominant feature, which can occur in bagged trees and lead to high correlation among the predictions.

The primary differentiation between Bagging and Random Forests is in the determination of the feature subset size. Random Forests select a random subset of features, typically of size $\sqrt{p}$, at each split, whereas Bagging considers all features. This selective process in Random Forests helps overcome the issue of high correlation and improves prediction accuracy (Hastie et al., 2009). In this study, high-accuracy models were examined by creating Random Forest models with subsets at all feature numbers.

The Random Forest algorithm utilises a combination of multiple decision trees on the Breast Cancer data set. Each tree is trained on a random subset of the data and a random subset of the features.

Resampling results indicate that the Random Forest model has an overall accuracy of approximately 80.97%, demonstrating its effectiveness in predicting cancer outcomes. The optimal number of features considered in each bin (mtry) was determined to be 33. The estimated error rate of the model on unseen data, known as the out-of-bag (OOB) estimate, is approximately 20.1%.

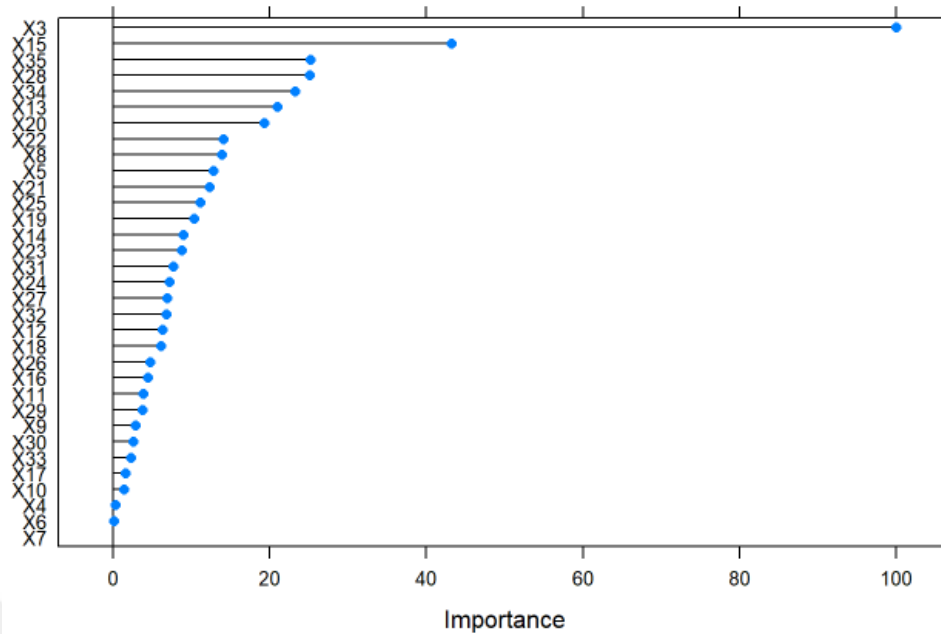


Figure 7.2.11 Ranking the importance of features in the Breast Cancer data set modelled by the Random Forest

Furthermore, Figure 7.2.11 presents feature importance scores that demonstrate the relative contribution of each feature in explaining the model. Notably, features such as X_3 (time), X_{15} (cancer cell texture2), X_{35} (lymph node status), and X_{28} (cancer cell area3) exhibit higher importance scores, indicating their significant role in predicting Breast Cancer outcomes.

7.2.5 Boosting

Boosting is an alternative technique for enhancing predictions with decision trees. The procedure is sequential, constructing trees using information from previous trees. In contrast to Bagging, Boosting does not entail bootstrap sampling. Instead, an adapted version of the initial data set is fitted to each tree (Hastie et al., 2009).

For regression, Boosting entails combining a large number of decision trees. The process starts with an original model. A tree is fitted to the errors in the current model. This new tree is added to the current model, updating the errors. The whole procedure continues sequentially, with each tree targeting regions where the existing

model performs weakly. The shrinkage parameter regulates the learning speed, and smaller trees are generally preferred for Boosting. The mathematical representation is as follows:

$$\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^b(x).$$

Boosting involves various tuning parameters, including the number of trees in the ensemble B and the shrinkage parameter λ . Careful consideration of the number of trees is necessary to avoid overfitting. Cross-validation serves as a method to help determine the optimal value. The shrinkage parameter controls the learning rate, with smaller values requiring the use of more trees to achieve optimal performance. Finally, the complexity of the boosted ensemble is determined by the number of splits in each tree.

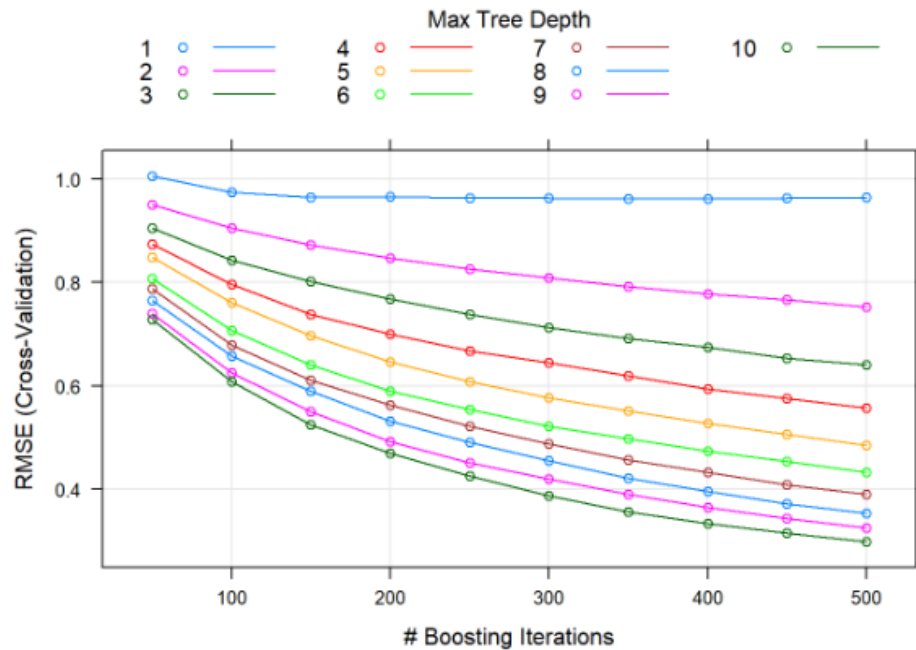


Figure 7.2.12 Parameter tuning for Boosting

The presented model is a Boosting model designed to predict sleep duration based on various features. The model's performance is presented in Figure 7.2.12, including the parameters used to minimize RMSE and maximize R-squared value. Results are

presented using resampling. The final model was selected based on the smallest RMSE value with a minimization of 0.1, and the minimum number of observations at each terminal node was set to 10. The model consists of 500 iterations and 10 interaction depths.

7.3 Support Vector Machines

The primary aim of SVM is to identify the best optimal hyperplane that can distinguish various classes of data points in a high-dimensional feature space. The hyperplane in question maximises the space between both classes, ensuring optimal separation of the nearest data points from each class. These data points, also called support vectors, represent the decision boundary.

To obtain the optimal hyperplane, the Support Vector Machine (SVM) formulates an optimisation problem that aims to maximise the margin while minimising the classification errors. The solution to this problem involves finding the support vectors and determining the appropriate weights (coefficients) for each support vector to define the hyperplane.

7.3.1 *Linear Support Vector Machine*

The Linear Support Vector Machine is a classification and regression-based machine learning method. It is a version of the Support Vector Machine algorithm that employs a linear kernel function. The objective is to identify a hyperplane that optimally separates the data points in the feature space, with the greatest possible margin (Bishop & Nasrabadi, 2006). For instance, when employed for classification, the SVM algorithm is configured to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class (referred to as support vectors). During training, the SVM algorithm aims to identify the optimal hyperplane by minimizing the classification error and maximizing the margin.

In regression, a linear SVM operates in a manner analogous to classification and seeks to identify the hyperplane that best aligns with the data points. Regression tasks involve predicting continuous target variables rather than class labels. In linear SVM regression, the objective is to identify the hyperplane that minimises the discrepancy between the predicted and actual target values. As with classification, the algorithm optimises the location of the hyperplane during training to minimise the error.

In linear SVM, the decision boundary is a hyperplane defined by the following equation:

$$q'(x, w, b) = w^T \phi(x) + b = \sum_{i=1}^n w_i \phi_i(x_i) + b.$$

The function ϕ maps inputs x from space X to corresponding feature vectors $\phi(x)$ in the higher-dimensional F space. The resulting feature vector $\phi(x)$ is then used to predict a linear algorithm called $q'(x, w, b)$; w and b represent the weight vector and bias term, respectively.

As an example, the Sleep data set model is a Support Vector Machine with a linear kernel used for regression purposes to predict sleep duration based on various features. SVM aims to find the optimal hyperplane that best separates data points into different classes.

```

Support Vector Machines with Linear Kernel

849 samples
17 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 764, 765, 764, 763, 765, 764, ...
Resampling results across tuning parameters:

  C      RMSE      Rsquared    MAE
0.1  1.011275  0.3673369  0.7594089
0.2  1.011616  0.3673036  0.7595865
0.3  1.010878  0.3681665  0.7590180
0.4  1.010889  0.3683187  0.7590378
0.5  1.011889  0.3677512  0.7597016
0.6  1.012285  0.3673680  0.7599648
0.7  1.012031  0.3677055  0.7598639
0.8  1.012039  0.3677116  0.7598673
0.9  1.011922  0.3677794  0.7596435
1.0  1.011796  0.3678762  0.7597087
1.1  1.011992  0.3678548  0.7599382
1.2  1.011897  0.3678744  0.7598230
1.3  1.012021  0.3677370  0.7598765
1.4  1.011956  0.3678841  0.7599891
1.5  1.011946  0.3678406  0.7598160
1.6  1.011802  0.3680120  0.7598221
1.7  1.011987  0.3678979  0.7599132
1.8  1.012015  0.3678156  0.7599203
1.9  1.011982  0.3678484  0.7598898
2.0  1.012046  0.3678274  0.7599394

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was C = 0.3.

```

Figure 7.3.13 Linear Support Vector Machine used to model Sleep Duration data set

The model's performance was assessed through cross-validated resampling at various tuning parameters. The optimal parameters were selected, resulting in the smallest RMSE of approximately 1.01, indicating the average deviation of the predictions from the actual values.

The model depicted in Figure 7.3.13 employs a linear kernel and an epsilon support vector regression algorithm with an epsilon parameter of 0.1 and a cost parameter of 0.3.

7.3.2 Radial Support Vector Machine

Radial SVM, also identified as the Radial Basis Function (RBF) SVM, is a modification of the Support Vector Machine algorithm that utilises a radial basis function as the kernel function. The RBF kernel allows for the modeling of non-linear decision boundaries in a higher-dimensional feature space. It is widely used in various applications, including pattern recognition, classification, and regression tasks (Bishop & Nasrabadi, 2006). The RBF kernel function is defined as:

$$K(x,y) = \exp(-\frac{1}{2\sigma^2}\|x-y\|^2),$$

where x and y denote the input feature vectors, the squared Euclidean distance between them is given by $\|x-y\|^2$. The hyperparameter σ regulates the width of the Gaussian function.

The sigma parameter controls the influence of each training example on the decision boundary. A reduction in sigma creates a smoother decision boundary which can preclude overfitting, whereas an increase in sigma generates a more intricate boundary that results in a closer match with the training data.

```

Support Vector Machines with Radial Basis Function Kernel

849 samples
17 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 764, 765, 764, 763, 765, 764, ...
Resampling results across tuning parameters:

   C      RMSE      Rsquared    MAE
0.25  0.8864673  0.5712176  0.6345350
0.50  0.7574881  0.6807643  0.5254728
1.00  0.5972373  0.7953913  0.3950689
2.00  0.4659514  0.8701947  0.2847131
4.00  0.3576518  0.9218078  0.2180683
8.00  0.2680054  0.9542570  0.1772445
16.00 0.2181348  0.9673621  0.1565786
32.00 0.2107519  0.9683306  0.1530348
64.00 0.2151596  0.9671818  0.1543951
128.00 0.2151597  0.9671820  0.1543985

Tuning parameter 'sigma' was held constant at a value of 0.0210606
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were sigma = 0.0210606 and C = 32.

```

Figure 7.3.14 Radial Support Vector Machine used to model Sleep Duration data set

Figure 7.3.14 shows a Support Vector Machine (SVM) regression model with a Radial Basis Function kernel. The model predicts sleep duration based on various determinants.

To evaluate the model's performance, resampling with cross-validation was used on different tuning parameters, particularly the regularization parameter. The SVM regression with RBF kernel uses an epsilon support vector regression algorithm with an epsilon parameter of 0.1 and a cost parameter of 32. 320 support vectors were used to train data points based on Euclidean distances in the transformed feature space. The training error was approximately 1.38%. The optimal RMSE value chosen resulted in a small value of about 0.2107519.

7.3.3 Polynomial Support Vector Machine

Polynomial SVM is a variant of the Support Vector Machine algorithm that enables non-linear decision boundaries by using polynomial kernels. The polynomial

kernel function maps the original feature space to a high-dimensional feature space where linear separation may be possible. In this high-dimensional space, the polynomial SVM learns a hyperplane to classify the data (Bishop & Nasrabadi, 2006). The definition of the polynomial kernel function is as follows:

$$K(x,y) = (xy + c)^d.$$

The input feature vectors are denoted as x and y , with the constant c and degree of the polynomial, d . The function of the kernel involves calculating the dot product of the input vectors, raising it to the d -th power, and adding the constant c .

The degree parameter (d) determines the degree of complexity of the decision boundary. An increased degree enables the model to capture more complex patterns in the data, but also increases the risk of overfitting. The constant term (c) affects the importance of higher order terms in the polynomial.

The provided model is a Support Vector Machine regression with a polynomial kernel used to predict outcomes based on various features in the healthcare setting.

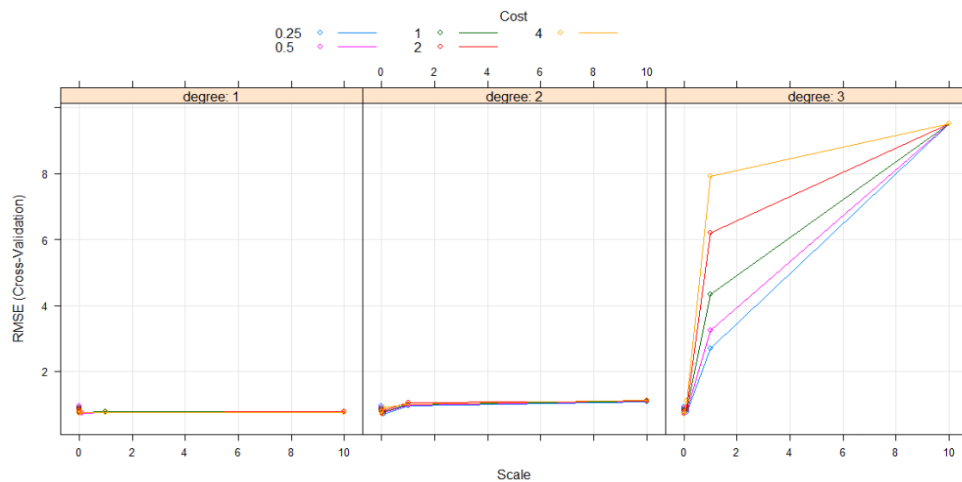


Figure 7.3.15 Parameter tuning for Polynomial Support Vector Machine

Figure 7.3.15 shows the model created using 10-fold cross-validation resampling on different hyperparameters. The optimal model was selected based on the smallest

RMSE obtained with these parameters. The SVM regression with polynomial kernel function and these hyperparameters has 100 support vectors, and the RMSE value is 0.70.



CHAPTER EIGHT

APPLICATION

Although various approaches have been proposed in the literature, no study has yet examined which methods are advantageous in specific respects for researchers. This study serves as an application and guide to Feature Selection methods in Machine Learning problems.

The analysis of Feature Selection methods is presented first, with a focus on their input and output (target) variables. The input and output variables are grouped into four categories: categorical - categorical, categorical- -numerical, numerical - categorical, and numerical - numerical. These categories are presented in Table 8.0.1.

	Input		Output	
Method	Categorical	Numerical	Categorical	Numerical
Information Gain	+		+	
Gain Ratio	+		+	
Symmetrical Uncertainty	+		+	
Gini Index	+		+	+
Chi-squared Test	+		+	
Cramer's V	+		+	
OneR Algorithm	+		+	
ANOVA - F Test	+			+
Kruskal Wallis Test	+			+
Pearson's Correlation		+		+
Spearman's Correlation		+		+
Relief Algorithm	+		+	
ReliefF Algorithm		+	+	
RReliefF Algorithm	+	+		+
Correlation-based Feature Selection		+		+
Consistency Feature Subset Selection	+		+	
Best Subset Selection	+	+	+	+
Sequential Forward Selection	+	+	+	+
Sequential Backward Elimination	+	+	+	+
Recursive Feature Elimination	+	+	+	+
Ridge Regression	+	+	+	+
The Lasso	+	+	+	+
Elastic Net	+	+	+	+
Regression Trees	+	+		+
Classification Trees	+	+	+	
Bagging	+	+	+	+
Random Forest	+	+	+	+
Boosting	+	+	+	+
Linear Support Vector Machines	+	+	+	+
Radial Support Vector Machines	+	+	+	+
Polynomial Support Vector Machines	+	+	+	+

Table 8.0.1 Grouping of feature selection methods according to data type

Filter methods are classified into two categories: Ranking and Subset Feature Selection methods. It seems in Table 8.0.1 that Filter subset methods are only applicable when the data set consists of either categorical-categorical or numerical-numerical variables. In Filter ranking methods, when the data set is categorical, Information Gain, Gain Ratio, Symmetric Uncertainty, Gini Index, Chi-square Test, Cramer's V, OneR Algorithm, and Relief Algorithm are used. When working with data sets that combine numerical and categorical variables, different algorithms are used depending on the type of data. The ReliefF Algorithm is used for numerical-categorical data sets, while the Gini Index and ANOVA are employed for categorical-numerical data sets. The Kruskal Wallis Test and RReliefF Algorithm are used for numerical-categorical data sets. Finally, for numerical data sets, the Pearson Correlation, Spearman Correlation, and RReliefF Algorithm are used.

Wrapper algorithms are applicable to all types of data structures. The process involves iterating through various feature combinations and evaluating their performance using a specified learning algorithm, making them suitable for any data format.

Embedded algorithms, similar to Wrapper methods, can handle various data structures except for tree-based classification and Regression Trees. These algorithms include feature selection in the model building process, either during training or as a regularization component.

The study compares the 10-fold cross-validation prediction performances on four distinct real medical data sets Chapter 1, using selected feature methods suitable for each feature's type and target variable, along with commonly used Machine Learning methods. The study employed various packages, including caret, funModeling, MASS, RWeka, FSelector, ggplot2, rpart.plot, nnet, FSinR, ineq in R software. Details regarding the methods and algorithms applied to each data set are described below. The study evaluates and compares the prediction accuracies and feature selection successes for each data set.

8.1 Obsessive Compulsive Disorder Data Set Results

Descriptive data analysis was performed on the data set for Obsessive Compulsive Disorder. Filter and Wrapper Feature Selection methods were applied and the data set was modeled using the Ordinal Logistic Regression model. Additionally, embedded methods suitable for the data set were applied. The effectiveness of the methods was evaluated using 10-fold cross-validation, a technique commonly used in such cases. A model using Ordinal Logistic Regression with 24 features was developed and a cross-validation accuracy of 67.64% was achieved.

Filter methods were applied to the OCD data set and the performance of the model was tested on features 1 to 24. The performance of the model with each method tested on features 1 through 24 is shown in Figure 8.1.1. A subset with 10 features were created using Consistency Feature Subset Selection, resulting in a cross-validation accuracy of 68.21%. Among the Filter methods, information theory-based methods such as Information Gain, Gain Ratio, Symmetric Uncertainty, and Gini Index, as well as the rule-based OneR Algorithm, were used to select one feature out of 24. The cross-validation accuracy was 68.35%. The Chi-squared Test, Cramer's V, and ReliefF Algorithm have an approximate 68% cross-validation accuracy with 18 features. Upon examining the Ordinal logistic regression models applied to all these Filter methods in terms of feature significance, the R&B, Age, and Hours features stand out.

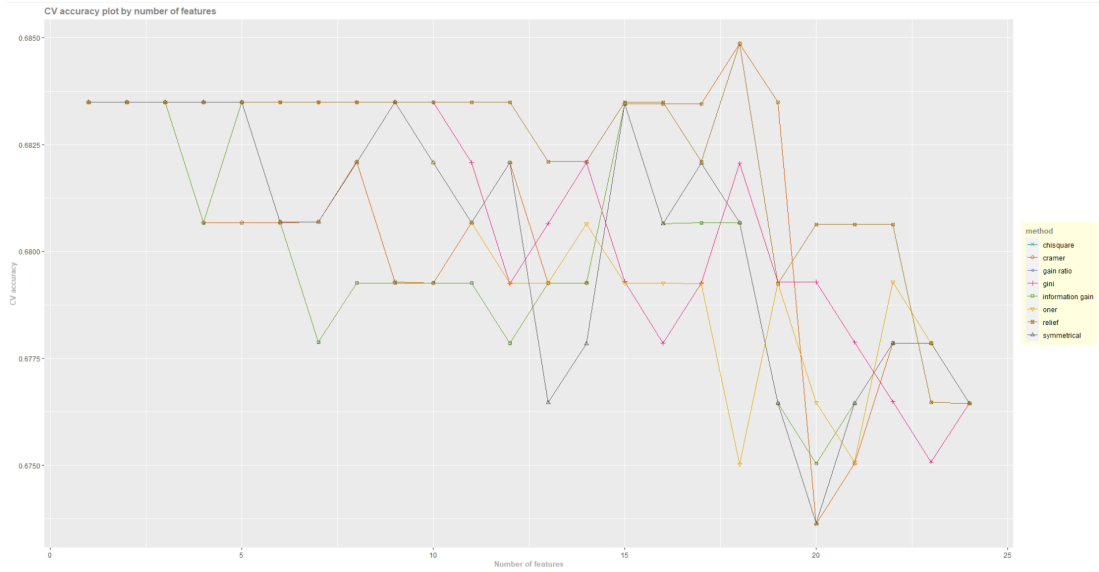


Figure 8.1.1 Cross-validation accuracy values for the OCD data set using filter methods

When the wrapping algorithms were examined, it was discovered that the forward selection method had a cross-validation accuracy of 67.64% with 24 features, which was the same as the full model. Best Subset Selection and Sequential Backward Elimination methods showed 68.21% accuracy using 4 features. Recursive Feature Elimination stood out by achieving the highest cross-validation accuracy of 68.35% when only 1 feature was selected. A potentially simpler and more interpretable model should be chosen compared to methods that select a larger number of features, highlighting its effectiveness in identifying the most effective features for predicting OCD levels. When the feature significances were examined in the background, Recursive Feature Elimination selected the Country feature, but it was not significant at the $\alpha = 0.05$ level. Sequential Forward Selection is not an optimal model because it is the same as the full model. Finally, Best Subset Selection and Sequential Backward Elimination methods selected all 4 features the same, and 3 of these features were significant at the $\alpha = 0.05$ level.

During the model-building stage, various Embedded methods were used for estimating the level of OCD in the data set. Among the tested methods, Ridge Regression and Radial Support Vector Machine performed the best, achieving the highest accuracy scores of 68.49%. This demonstrates their effectiveness in capturing

underlying patterns in the data while minimizing overfitting. Additionally, Elastic Net, Boosting, and Polynomial Support Vector Machine performed competitively, with accuracy rates ranging from 68.35% to 68.36%. In contrast, Lasso, Classification Trees, Bagging, and Linear Support Vector Machine yielded slightly lower accuracies, ranging from 64.85% to 67.23%. Random Forest achieved an accuracy of 68.35%, which is comparable to many other methods in terms of prediction performance. Table 8.1.2 illustrates the performance of the model with each method.



OCD	# feature	Accuracy
Information Gain	1	68.35%
Gain Ratio	1	68.35%
Symmetrical Uncertainty	1	68.35%
Gini Index	1	68.35%
Chi-squared Test	18	68.07%
Cramer's V	18	68.49%
OneR Algorithm	1	68.35%
ReliefF Algorithm	18	68.49%
Consistency Feature Subset	10	68.21%
Best Subset Selection	4	68.21%
Sequential Forward Selection	24	67.64%
Sequential Backward Elimination	4	68.21%
Recursive Feature Elimination	1	68.35%
Ridge Regression	2	68.49%
The Lasso		64.85%
Elastic Net		68.35%
Classification Trees		67.23%
Bagging		66.10%
Random Forest		68.35%
Boosting		68.36%
Linear Support Vector Machines		68.35%
Radial Support Vector Machines		68.49%
Polynomial Support Vector Machines		68.35%
Full model	24	67.64%

Table 8.1.2 Model performances for the Obsessive Compulsive Disorder data set

The analysis revealed that 3 out of the 24 traits could be significantly associated with obsessive compulsive disorder. Knowledge-based techniques achieved an accuracy of

approximately 68.35%, while some statistical tests and the Relief Algorithm showed around 68%. Recursive Feature Elimination stood out with an accuracy of 68.35%. In modelling, Ridge Regression and Radial Support Vector Machine achieved the highest accuracy of 68.49%, while the others ranged from 64.85% to 68.36%. These findings offer valuable insight into effective prediction approaches for OCD levels.

The selection of the most appropriate model based on the results may vary depending on the researcher's focus on specific contexts. In order to create the most suitable model for this application, some constraints need to be taken into account. Since there are no significant differences in accuracy, it is recommended to choose a model based on the number of features. There are 6 models with 1 feature for this data set. R&B, Age and Hours features attract attention. In line with the purpose of the data set, it was found that Best Subset and Sequential Backward Elimination had meaningful features by selecting the same 4 features. These are models chosen to achieve high accuracy using a limited number of features. As an example, let's look at this model selection model in more detail.

	Coefficients	Std. Error	Z Stat	p-value	significant	odds
Age2	"-0.084"	"0.113"	"-0.744"	"0.457"	" "	"0.92"
Age3	"-0.696"	"0.257"	"-2.713"	"0.007"	" *	"0.499"
Hours2	"0.2"	"0.111"	"1.795"	"0.073"	" "	"1.221"
Hours3	"0.418"	"0.186"	"2.246"	"0.025"	" *	"1.518"
Hours4	"0.44"	"0.205"	"2.152"	"0.031"	" *	"1.553"
CountrySometimes	"0.171"	"0.145"	"1.177"	"0.239"	" "	"1.186"
CountryNever	"-0.181"	"0.111"	"-1.627"	"0.104"	" "	"0.835"
`CountryVery frequently`	"-0.167"	"0.204"	"-0.819"	"0.413"	" "	"0.846"
RBSometimes	"-0.33"	"0.135"	"-2.436"	"0.015"	" *	"0.719"
RBNever	"-0.078"	"0.123"	"-0.636"	"0.525"	" "	"0.925"
`RBVery frequently`	"0.14"	"0.144"	"0.973"	"0.331"	" "	"1.15"
low middle	"0.391"	"0.119"	"3.281"	"0.001"	" *	"1.479"
middle high	"1.058"	"0.124"	"8.549"	"0"	" *	"2.879"

Figure 8.1.2 The optimal model chosen for the OCD data set

The optimal model chosen for the OCD data set is Best Subset Selection and Sequential Backward Elimination models, which are the same as this model. It aims to classify a person according to whether they have Low, Moderate or High levels of obsessive-compulsive disorder using 4 features out of 24. Statistical values for this model can be found in Figure 8.1.2.

When the significance of the model in terms of features was examined, it was seen that 3 features were statistically significant at the $\alpha = 0.05$ level. When the significance of the model in terms of features was examined, it was seen that 3 features were statistically significant at the $\alpha = 0.05$ level. The likelihood of exhibiting low rather than high levels of OCD is approximately 50% lower in individuals aged 50 years and above compared to those below 25 years of age.

Similarly, individuals who listen to music for more than six hours per day are approximately 1.5 times more likely to exhibit low levels of OCD than those who listen to music for less than three hours per day.

In another instance, individuals who sometimes listen to R&B are 30% less likely to exhibit a low level of OCD than a high level of OCD, in comparison to those who rarely listen to R&B.

8.2 Breast Cancer Data Set Application Results

Exploratory data analysis was carried out on the Breast Cancer data set, which consists of 33 numerical inputs and 1 categorical output. The data was preprocessed by standardisation. After employing Filter and Wrapper Feature Selection methods, Logistic Regression was used to model the data set, which includes both numerical inputs and categorical outputs. Additionally, Embedded methods suitable for the data set were applied. To assess the efficacy of the methods, 10-fold cross-validation was utilised, which is commonly used in such cases. A Logistic Regression model was developed with 33 features and achieved a cross-validation accuracy of 77%.

The ReliefF Algorithm was determined to be the only appropriate Filter method for the Breast Cancer data set. Figure 8.2.3 displays the performance of the model with each method tested from 1 to 33 features. Upon examination, we observed an 80% cross-validation accuracy with 13 features and an 82% cross-validation accuracy with 28 features.

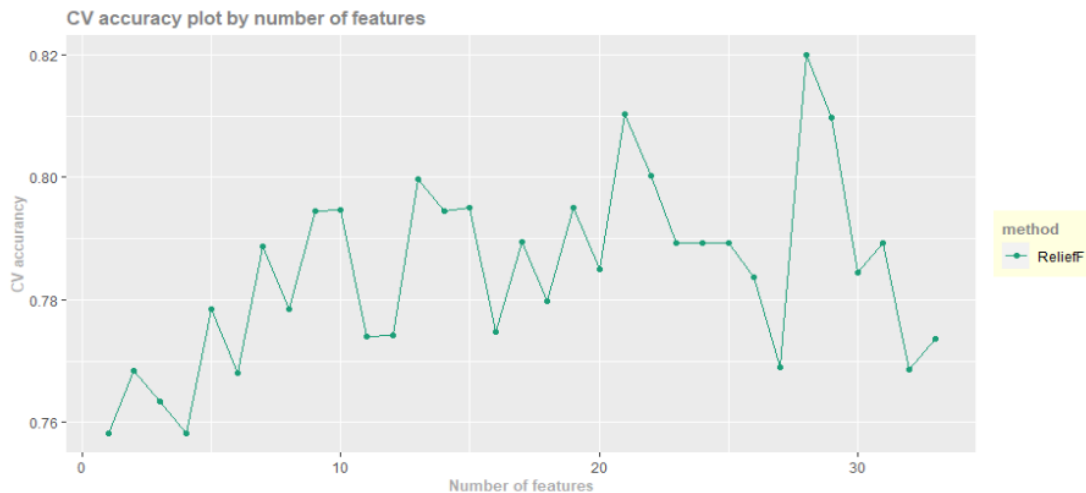


Figure 8.2.3 Cross-validation accuracy values for the Breast Cancer data set using filter methods

When examining the Wrapper algorithms, it was found that the Best Subset and Backward methods possess a cross-validation accuracy of 77%, with 24 features used. Among the Wrapper methods, the Forward method exhibited superior performance with an accuracy of 80% using 12 features. Recursive Feature Elimination resulted in 76% cross-validation accuracy with 3 features employed.

This data set utilized 10 Embedded methods, excluding Regression Tree. Classification Tree, Boosting, and Bagging have cross-validation accuracies of 73%, 77%, and 78%, respectively. The remaining methods all achieve the highest accuracy among the Embedded methods, with a cross-validation accuracy of 81%. Table 8.2.3 displays the performance of the model with each methods.

Breast Cancer	# Features	Accuracy
ReliefF Algorithm	28	82%
Best Subset Selection	24	77%
Sequential Forward Selection	12	80%
Sequential Backward Elimination	24	77%
Recursive Feature Elimination	3	76%
Ridge Regression	4	81%
The Lasso		81%
Elastic Net		81%
Classification Trees		73%
Bagging		78%
Random Forest		81%
Boosting		77%
Linear Support Vector Machines		81%
Radial Support Vector Machines		81%
Polynomial Support Vector Machines		81%
Full model	33	77%

Table 8.2.3 Model performances for the Breast Cancer data set

Based on the analysis, out of the 33 features, 17 of them can be considered to be meaningfully associated with cancer recurrence. The recommended method for this data set is not explicitly mentioned, but it can be inferred that the Filter feature selection method using ReliefF Algorithm 28 achieved the highest cross-validated accuracy of 77%. Among the Wrapper algorithms, the Forward method 12 achieved an accuracy of 80% in cross-validation. Embedded shrinkage methods, as well as Random Forest, Support Vector Machine Linear, Radial, and Polynomial, all achieved an accuracy of 81% in cross-validation.

Based on the results, the optimal model selection may vary depending on the researcher's focus within specific contexts. To establish an optimal model for this

application, certain constraints must be considered. Given the medical nature of the data, a minimum accuracy threshold of 80% has been set. Furthermore, a model with a concise set of features is preferred. The chosen model for achieving high accuracy while using a limited number of features is Sequential Forward Selection (SFS). Sequential Forward Selection model has a sensitivity of 89.9% for accurately detecting recurrent cancer cases. The specificity rate, i.e. the rate of correctly predicting cases where cancer will not recur, is 59%. Accuracy rate is 80%. Let's examine the Sequential Forward Selection model in more detail.

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.78506  -0.60481  -0.25630  -0.03079   2.58339

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.099e+01  8.154e+00   2.574 0.010055 *
X3           -5.008e-02  1.003e-02  -4.994  5.9e-07 ***
X15          -9.586e-01  6.268e-01  -1.529 0.126167
X27          -1.385e-02  5.178e-03  -2.675 0.007467 **
X35           9.493e-02  3.707e-02   2.561 0.010437 *
X4           -3.848e+00  1.267e+00  -3.038 0.002383 **
X24           2.480e+00  7.633e-01   3.249 0.001158 **
X7            2.650e-02  1.072e-02   2.473 0.013412 *
X5           -8.802e-02  5.773e-02  -1.525 0.127307
X21          -1.682e+02  6.872e+01  -2.447 0.014396 *
X23           8.061e+02  2.441e+02   3.303 0.000957 ***
X13          -2.891e+02  8.793e+01  -3.287 0.001011 **
X8            8.074e+01  3.132e+01   2.577 0.009952 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212.52  on 193  degrees of freedom
Residual deviance: 136.94  on 181  degrees of freedom
AIC: 162.94

Number of Fisher Scoring iterations: 6
```

Figure 8.2.4 The optimal model chosen for the Breast Cancer data set

Figure 8.2.4 shows the model of the original Breast Cancer data set generated by Sequential Forward Selection. Sequential Forward Selection selected 12 features out of 33. Of these 12 features, 10 are statistically significant at the $\alpha = 0.05$ level in explaining variation in the likelihood of Breast Cancer recurrence.

Based on the information provided in Figure 8.2.4, here is an example summary of the relationships between cancer recurrence and various features:

X_3 : The coefficient estimate for duration of previous cancer recurrence is -0.05. The odds ratio for this trait is 0.95, meaning that a unit increase in X_3 results in a 5% decrease in the odds of cancer recurrence.

X_{27} : The estimated area3 coefficient of the cancer cell is approximately -0.96. The odds ratio for this trait is 0.99, meaning that a one-unit increase in X_{27} results in a 1% decrease in the odds of cancer recurrence.

X_{35} : The coefficient estimate for the lymph status feature, which reports the number of positive axillary lymph nodes at the time of surgery, is approximately 0.095. The odds ratio for this feature is 1.09, meaning that a 1 unit increase in X_{35} results in a 9% increase in the odds of cancer recurrence.

X_4 : The estimated radius1 coefficient of the cancer cell is approximately -3.848. The odds ratio for this trait is 0.02, meaning that a 1 unit increase in X_4 results in a 98% decrease in the odds of cancer recurrence.

X_{24} : The estimated radius3 coefficient of the cancer cell is approximately 2.48. The odds ratio for this trait is approximately 12, meaning that a one-unit increase in X_{24} results in a 12-fold increase in the odds of cancer recurrence.

X_7 : The estimated area1 coefficient of the cancer cell is approximately -0.026. The odds ratio for this trait is 1.02, meaning that a one-unit increase in X_7 results in a 2% increase in the odds of cancer recurrence.

The Breast Cancer data set revealed that Sequential Forward Selection identified key features that influence cancer recurrence. Specifically, the duration of previous

recurrence and certain features of cancer cells, such as area and radius, were associated with decreased or increased odds of recurrence. Additionally, the number of positive axillary lymph nodes during surgery and specific features of cancer cell morphology also played significant roles. These findings provide valuable insights into the features that affect the likelihood of Breast Cancer recurrence.

8.3 Sleep Duration Data Set Application Results

The Sleep Duration data set, comprising of 17 categorical inputs and 1 numerical output, underwent exploratory data analysis. Filter and Wrapper Feature Selection methods were employed, followed by Linear Regression to model the data set, which included both numerical inputs and categorical output. Additionally, Embedded methods suitable for the data set were applied. To evaluate the effectiveness of the methods, 10-fold cross-validation was used, as is common in such cases. A Linear Regression model was developed using 17 features and achieved a cross-validation RMSE of 0.96.

Among the tested Filter methods, ANOVA, Gini Index, Kruskal Wallis Test, and RReliefF Algorithm were found to be suitable for this data set. Figure 8.3.5 displays the performance of these methods for features 1 through 17. The full RMSE value obtained is 0.96, which is the lowest value among these methods. In particular, ANOVA 8 achieved this value, while Kruskal Wallis Test and RReliefF Algorithm achieved a cross-validation RMSE values of 11 and 14, respectively, and Gini Index achieved a cross-validation RMSE value of 16.

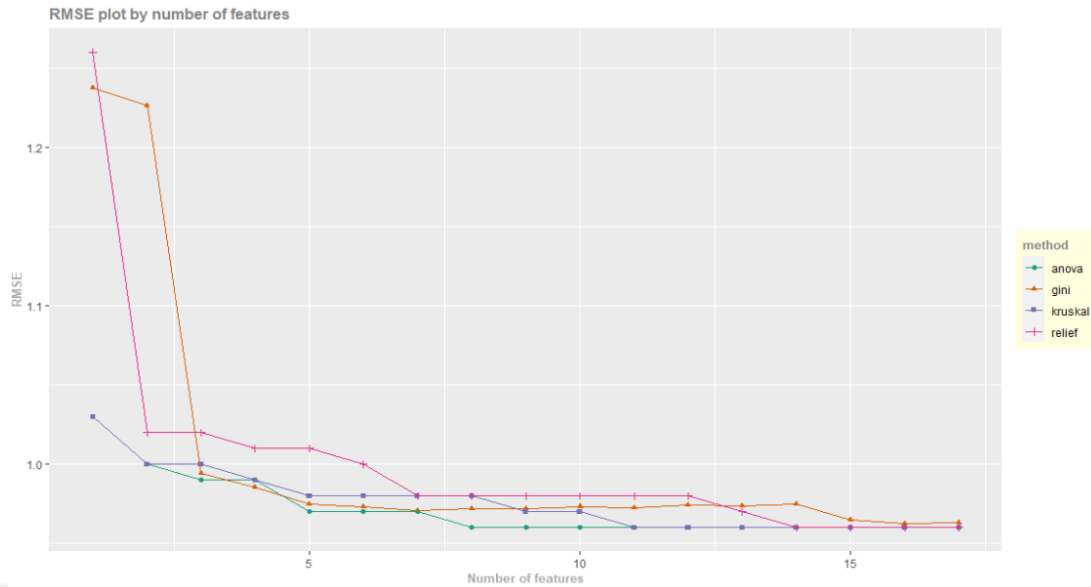


Figure 8.3.5 Cross-validation RMSE values for the Sleep Duration data set using filter methods

Upon examining Wrapper methods, it is evident that 4 features are selected across 4 methods. The cross-validation RMSE value of Best Subset and Sequential Forward Selection is 0.99, whereas that of Sequential Backward and Recursive Feature Elimination methods is 1.01.

Ten Embedded methods were used, excluding the Classification Tree. The Ridge Regression, The Lasso, and Elastic Net methods produced the same RMSE value as the full model. Among the Tree-based algorithms, the Random Forest model had a cross-validation RMSE value of 0.23. Among the Support Vector Machines, the Radial and Polynomial models had the lowest cross-validation RMSE value for this data set. The performance of the Sleep Duration data set with each method is displayed in this Table 8.3.4.

Sleep	# Features	RMSE
Gini Index	16	0.96
ANOVA - F Test	8	0.96
Kruskal Wallis Test	11	0.96
RReliefF Algorithm	14	0.96
Best Subset Selection	4	0.99
Sequential Forward Selection	4	0.99
Sequential Backward Elimination	4	1.01
Recursive Feature Elimination	4	1.01
Ridge Regression	3	0.96
The Lasso		0.96
Elastic Net		0.96
Regression Trees		1.01
Bagging		0.82
Random Forest		0.23
Boosting		0.30
Linear Support Vector Machines		1.01
Radial Support Vector Machines		0.21
Polynomial Support Vector Machines		0.21
Full model	17	0.96

Table 8.3.4 Model performances for the Sleep Duration data set

The analysis conducted using Linear Regression on 17 features established relationships, resulting in an RMSE of 0.96. It is worth noting that the ANOVA algorithm 8, used as a Filter feature selection method, produced a similar RMSE value. However, no significant differences in RMSE values were observed between Wrapper and Embedded shrinkage methods. The Embedded method was found to be most effective when used with Support Vector Machines, Radial and Polynomial, achieving an impressive RMSE of 0.21. These findings suggest the potential efficacy

of the Embedded method, particularly in conjunction with Support Vector Machines, for accurate predictions in the context of the examined relationships.

The selection of the best model depends on the researcher's specific focus. For this application, the SVM Polynomial model was chosen due to its low RMSE value and high R^2 value, as it is medical data. The cross-validation RMSE value of the SVM Polynomial model is 0.21, and the R^2 value is 0.97. Now, let's examine the SVM polynomial model in detail.

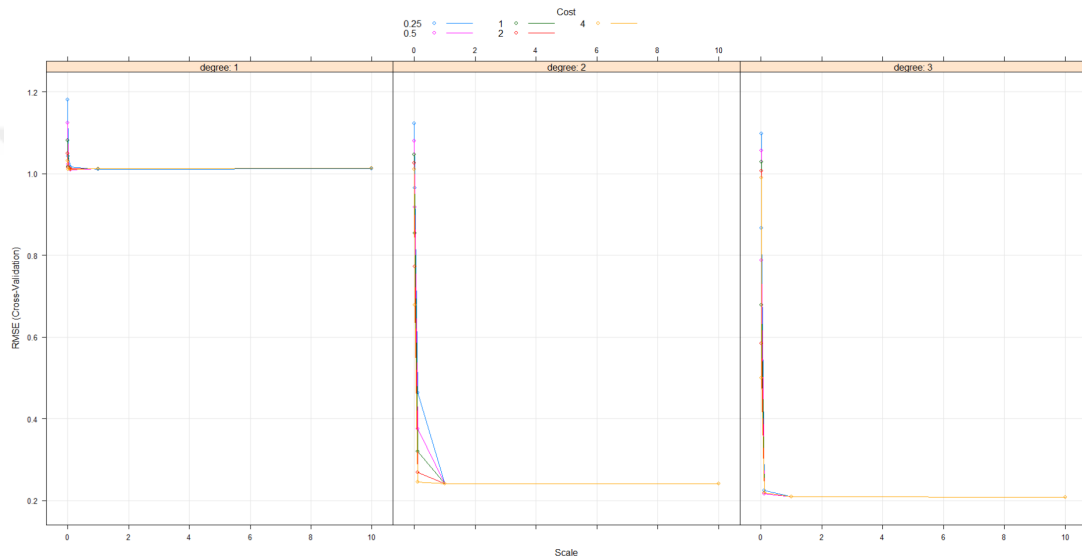


Figure 8.3.6 The optimal model chosen for the Sleep Duration data set

The analysis of Support Vector Machines with a Polynomial Kernel on the data set consisting of 849 samples and 17 features was conducted using a 10-fold cross-validated resampling approach. Results for tuning parameters, including degree, scale, and C, are systematically in Figure 8.3.6. The degree parameter represents the degree of the polynomial kernel, which in this case is 3, indicating the inclusion of cubic terms in the kernel function. The scale parameter determines the impact of input features on the decision boundary. A higher scale value usually leads to a smoother decision boundary. In this case, the scale is set to 10. The offset parameter is an extra constant in the kernel function that enables the user to shift the kernel function, thereby influencing the location of the decision boundary. In this case, the offset is set to 1. These parameter values were selected during the model training and

tuning process, most likely based on their ability to minimize the RMSE or maximize another evaluation metric on your data set. The specific values for these parameters may vary depending on the characteristics of the data and the problem you are trying to solve. The final model configuration represents a deliberate balance of polynomial degree, scaling, and regularization parameter to attain the lowest RMSE within the data set.

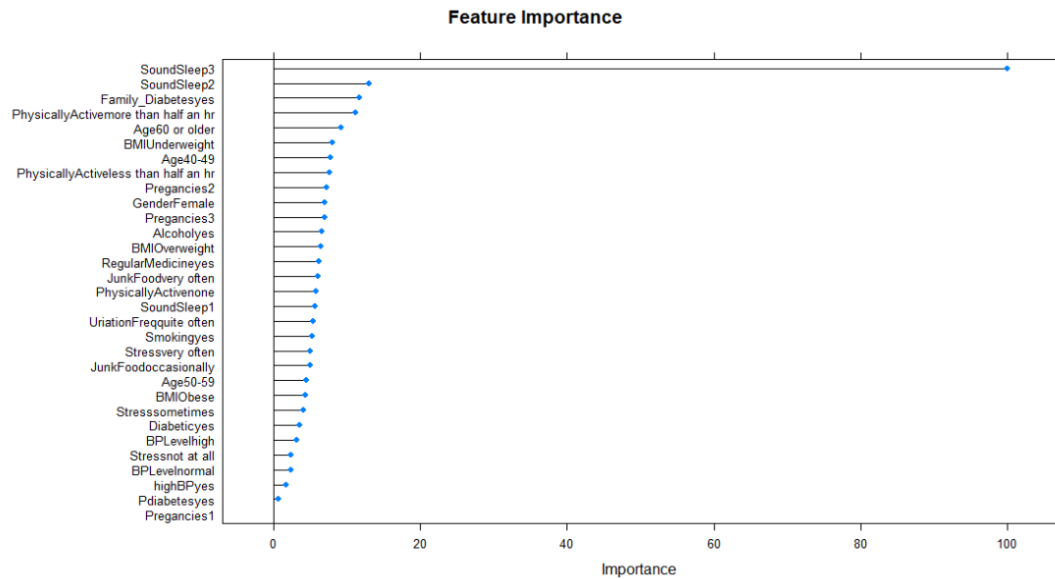


Figure 8.3.7 Ranking the importance of features in the optimal model for the Sleep Duration data set

Figure 8.3.7 shows the feature importance of the Polynomial Kernel Support Vector Machine model with sleep duration. For example, deep sleep duration, having a family history of diabetes, and doing more than an hour and a half of physical activity are features of high importance to this model.

8.4 SENIC Data Set Application Results

Exploratory data analysis was performed on the SENIC data set, which comprises 8 numerical inputs and 1 numerical output. The data set was modelled using Linear Regression for Filter and Wrapper Feature Selection methods. Additionally, Embedded methods suitable for the data set were applied. To evaluate the effectiveness of the methods, 10-fold cross-validation was employed, as is common in

such cases. A Linear Regression model was developed using 8 features, and cross-validation achieved an RMSE of 0.77.

When assessing Filter methods for the data set, Pearson's Correlation, Spearman's Correlation, RReliefF Algorithm, and Correlation-based Feature Selection were found to be suitable. Correlation-based Feature Selection, which operates through direct subset selection, produced a model with a cross-validation RMSE value of 0.72, using 4 features. Figure 8.4.8 shows the comparative performance of these methods for features 1 to 8. The RReliefF Algorithm performed similarly to the full model with 8 features. Pearson's Correlation, which focused on 4 features, produced a model with a cross-validation RMSE value of 0.72, while Spearman's Correlation, which emphasized 3 features, resulted in a model with an RMSE value of 0.73.

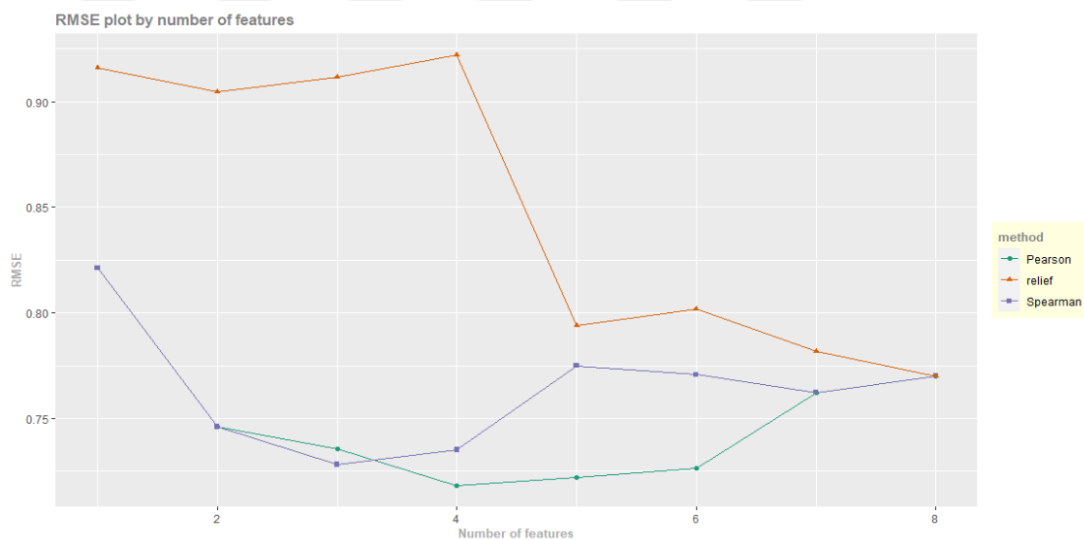


Figure 8.4.8 Cross-validation RMSE values for the SENIC data set using filter methods

Upon examining the Wrapper methods, it was found that the cross-validation RMSE value for the Best Subset and Sequential Forward Selection with 3 features is 0.72. In contrast, the cross-validation RMSE value for the Sequential Backward and Iterative feature selection elimination methods with 4 features is 0.73.

Ten Embedded methods were used, excluding the Classification Tree. The Ridge Regression and Elastic Net methods produced an RMSE value of 0.73, while the Lasso produced an RMSE value of 0.74. Among the tree-based algorithms, the

Random Forest and Boosting models produced the best value with an RMSE of 0.7. Similarly, the Polynomial between Support Vector Machines had the lowest cross-validation RMSE value for this data set. The model performance for the SENIC data set with each method is shown in Table 8.4.5.

Senic	# Features	RMSE
Pearson's Correlation	4	0.72
Spearman's Correlation	3	0.73
RReliefF Algorithm	8	0.77
Correlation-based Feature Selection	4	0.72
Best Subset Selection	3	0.72
Sequential Forward Selection	3	0.72
Sequential Backward Elimination	4	0.73
Recursive Feature Elimination	4	0.73
Ridge Regression	3	0.73
The Lasso		0.74
Elastic Net		0.73
Regression Trees		0.77
Bagging		0.71
Random Forest		0.70
Boosting		0.70
Linear Support Vector Machines		0.74
Radial Support Vector Machines		0.76
Polynomial Support Vector Machines		0.70
Full model	8	0.77

Table 8.4.5 Model performances for the SENIC data set

The data set showed an RMSE value of 0.72 when using the Filter method and Pearson's Correlation and Correlation-based Feature Selection method with 4 features. The Wrapper algorithms Best and Forward with 3 features also showed an RMSE of

0.72. Among the Embedded methods, Random Forest, Boosting, and Support Vector Machine Polynomial were identified as the most effective, all achieving an RMSE of 0.70. Let's take a closer look at the Boosting model, which had one of the lowest RMSE values.

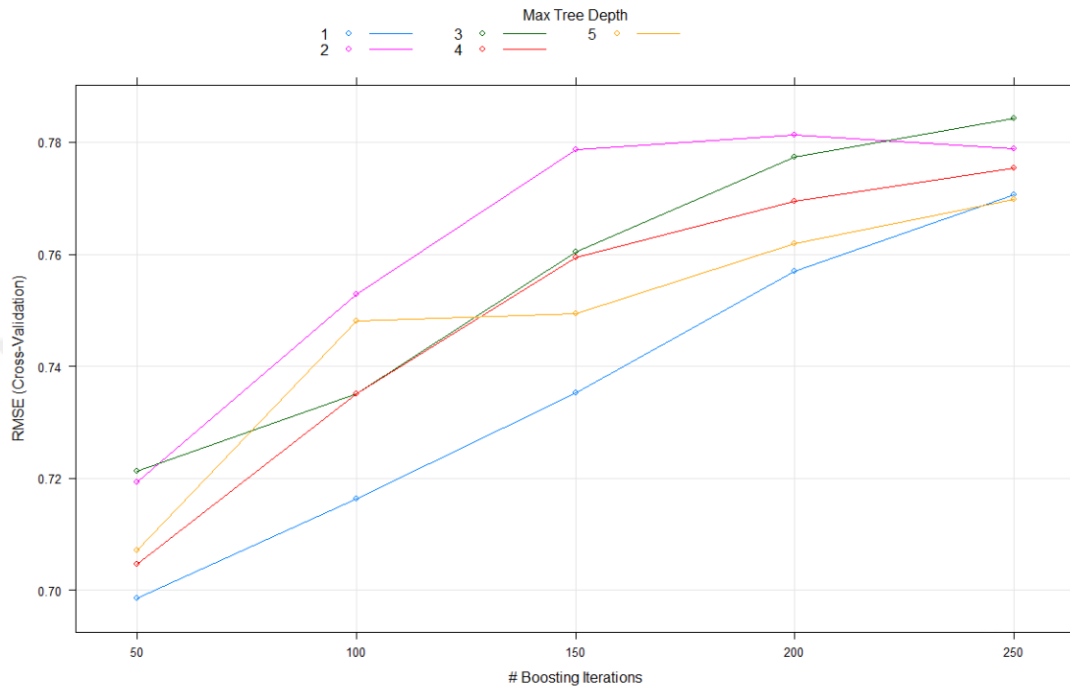


Figure 8.4.9 The optimal model chosen for the SENIC data set

The Boosting model was analysed on a data set comprising of 113 samples and 8 features. The model's complexity is affected by the interaction depth, which determines the maximum interaction depth between trees. Higher values capture more complex data relationships but may lead to overfitting. The results are shown in Figure 8.4.9. The performance of the model can be improved by using a larger number of trees to represent sequentially constructed decision trees. The contribution of each tree can be regulated by adjusting the shrinkage or learning rate. Lower values of shrinkage require more trees to achieve a similar effect, but they can improve generalization. However, this improvement comes at a computational cost. The final model, which was chosen due to its minimal RMSE value, included 50 trees, 1 interaction depth, 0.1 shrinkage, and a minimum of 10 observations at a node. The optimized configuration resulted in an RMSE of 0.6985 and an R-squared value of 0.5425. These values were achieved through rigorous parameter calibration to ensure

optimal model performance within the data set.

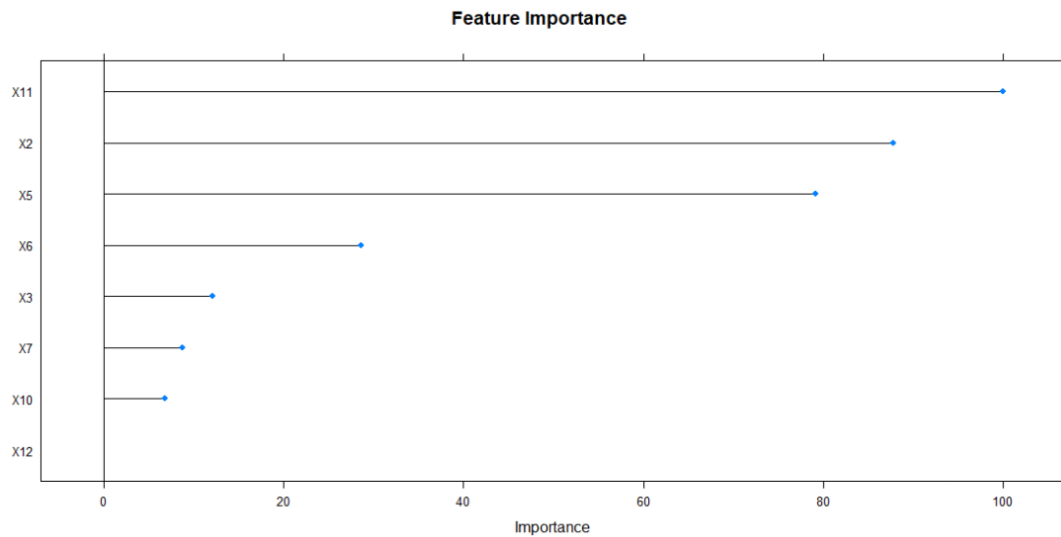


Figure 8.4.10 Ranking the importance for features in the optimal model for the SENIC data set

In Figure 8.4.10, it can be seen that the likelihood of infection in hospital is influenced by various features, each with different degrees of importance. The feature with the greatest impact is X_{11} , which represents the number of registered and licensed practical nurses during the study period and is assigned a score of 100 for its significance. The features with the highest importance scores are X_2 (Length of Stay), X_5 (Routine Culturing Ratio), and X_6 (Routine Xray Ratio), with scores of 87.77, 79.12, and 28.62, respectively. The estimation of the likelihood of infection in the hospital is influenced by several features, including X_3 (Age), X_7 (Number of Beds), X_{10} (Average Census), and X_{12} (Available Facilities). However, these features have comparatively lower importance scores ranging from 12.14 to 0. It is important to note that these scores indicate the relative impact of each feature on the estimated likelihood of infection.

CHAPTER NINE

CONCLUSION

This thesis presents a comprehensive review and application of Feature Selection methods in the context of Supervised Machine Learning algorithms. The aim of applying Feature Selection methods in Machine Learning problems is to guide researchers in selecting appropriate methods in the field of life sciences. The analysis of four different data sets is included. The study on Obsessive Compulsive Disorder (OCD), Breast Cancer, Sleep Duration and SENIC (Hospital Infection) provided valuable insights into the comparison of various feature selection techniques and their impact on predictive model performance across these data sets.

The implementation of Feature Selection methods involved a systematic process of data preprocessing, model development, and evaluation using 10-fold cross-validation. The study assessed the efficacy of these methods in enhancing feature counts, model accuracy, interpretability, and prediction performance. In conclusion, the use of different feature selection techniques in medical data sets has provided a detailed comprehension of their effectiveness and limitations, guiding researchers in selecting suitable methods for their specific data analysis requirements.

The initial findings present an overview of Feature Selection methods, which include Filter, Wrapper, and Embedded algorithms. Each method is tailored to different data structures and modeling needs, as demonstrated in Chapter 8. Filter methods, such as sorting and subset selection, are classified based on the type of data they process, including categorical, numeric, and mixed data sets. However, it has been observed that Wrapper and Embedded algorithms can work regardless of data structures.

The findings are presented across the data sets. For instance, in the OCD data set, all techniques achieved approximately 68% accuracy. Out of the 24 traits, three were significantly associated with obsessive-compulsive disorder. This led to the adoption of Best Subset Selection and Backward Elimination, which focus on four out of the 24 features, including fewer and more meaningful ones. The Breast Cancer data set

was analysed using the Sequential Forward Selection method, which selected 12 out of 33 features. This revealed the potential of the model created with fewer features, as it achieved higher accuracy (80%) than the full model (77%). In the sleep data set, the Embedded methods, particularly Radial and Polynomial Support Vector Machines, achieved an impressive 0.21 RMSE compared to the full model. The SENIC data set yielded a Boosting model with an RMSE of 0.70 after rigorous parameter tuning. These findings suggest that Wrapper algorithms are more effective for categorical output variables, while Embedded methods are more effective for numerical output variables.

In addition to these findings, it was noted that applications in R software required a "for" loop to evaluate model performance from 0 to the number of features in the data set for Filter methods. If no specific cut-off number in the number of features was specified at the beginning, it was found that the workload was higher than for other Feature Selection methods. As Embedded methods involve hyperparameterisation, they were found to take slightly longer than Wrapper methods. The study serves as a valuable resource for researchers and practitioners who wish to apply Feature Selection methods to the analysis of high-dimensional data, enabling them to understand and predict the application process.

This thesis is based on an experimental design where the input and output features consist of categorical or numerical data types. In addition to this experimental design, it is among the new areas of study that feature selection methods should be developed for mixed data types. In this context, feature selection methods using likelihood ratio-based generalized fitting and conditional independence tests have been proposed, extending the Markov blanket algorithm (Lee et al., 2020). In addition, methodologies such as a new correlation criterion (Liu et al., 2013) and density-based feature clustering (Yan et al., 2021) have been developed for mixed-type data. Furthermore, filter feature selection based on spectral gap score and supervised kernel-based (Solorio-Fernández et al., 2019) approaches have been used to create more accurate and reliable models for mixed data types. Future research could aim to improve existing methodologies, investigate new approaches and develop automated

frameworks to facilitate the feature selection process. Efforts to enhance the interpretability and explicability of feature selection methodologies, particularly in the context of medical data analysis, have the potential to advance healthcare analytics and improve patient outcomes.

The thesis enhances the comprehension of Feature Selection methods in the context of Machine Learning and offers practical guidance to researchers and practitioners in selecting suitable methods for various types of data sets with confidence. The insights gained from this work will serve as a catalyst for further innovation and advancement in the search for effective feature selection strategies for Machine Learning in healthcare. This will provide a roadmap to improve patient care and treatment outcomes.

REFERENCES

- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3), pp. 392–398.
- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3), pp. 91–93.
- Aldrich, J. (2005). Fisher and regression. *Statistical Science*, 20(4), pp. 401–417.
- Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved chi-square for arabic text classification. *Journal of King Saud University-Computer and Information Sciences*, 32(2), pp. 225–231.
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern Recognition and Machine Learning*, vol. 4 of *Information Science and Statistics*. New York: Springer.
- Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143, 106839.
- Bommert, A., Welchowski, T., Schmid, M., & Rahnenführer, J. (2022). Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics*, 23(1), bbab354.
- Chelvan, P. M., & Perumal, K. (2017). On feature selection algorithms and feature selection stability measures: A comparative analysis. *AIRCC's International Journal of Computer Science and Information Technology*, pp. 159–169.
- Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1), pp. 1–26.
- Fradkov, A. L. (2020). Early history of machine learning. *IFAC-PapersOnLine*, 53(2), pp. 1385–1390.
- Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. New York: Springer.

- Gül, H. C., Artuk, C., & Güven, G. Ö. (2022). *Sağlık Bakım Hizmetlerinde Enfeksiyonların Kontrolü ve Önlenmesi*. İstanbul: Akademisyen Kitabevi.
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46, pp. 389–422.
- Géron, A. (2022). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol: O'Reilly Media, Inc.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. New York: Springer.
- Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1), pp. 63–90.
- Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *The Annals of Mathematical Statistics*, 11(3), pp. 271–283.
- Huang, S. H. (2015). Supervised feature selection: A tutorial. *Artif. Intell. Res.*, 4(2), pp. 22–37.
- Hublin, C., Partinen, M., Koskenvuo, M., & Kaprio, J. (2007). Sleep and mortality: a population-based 22-year follow-up study. *Sleep*, 30(10), pp. 1245–1253.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
- Jebli, I., Belouadha, F.-Z., Kabbaj, M. I., & Tilioua, A. (2021). Prediction of solar energy guided by pearson correlation using machine learning. *Energy*, 224, 120109.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992*, pp.249–256. Morgan Kaufmann.
- Kleinbaum, D., & Klein, M. (2002). *Statistics for biology and health: logistic regression. A self-learning text. 2nd ed.* Atlanta: Springer Verlag NY.

- Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, vol. 26. New York: Springer.
- Kumar, V., & Minz, S. (2014). Feature selection: a literature review. *SmartCR*, 4(3), pp. 211–229.
- Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International journal on computer science and engineering*, 3(5), pp. 1787–1797.
- Lee, J., Jeong, J.-Y., & Jun, C.-H. (2020). Markov blanket-based universal feature selection for classification and regression of mixed-type data. *Expert Systems with Applications*, 158, 113398.
- Liu, D., Cho, S.-Y., Sun, D.-m., & Qiu, Z.-D. (2010). A spearman correlation coefficient ranking for matching-score fusion on speaker recognition. In *TENCON 2010-2010 IEEE Region 10 Conference, IEEE*, pp. 736–741.
- Liu, H., Wei, R., & Jiang, G. (2013). A hybrid feature selection scheme for mixed attributes data. *Computational and Applied Mathematics*, 32, pp. 145–161.
- Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., & Gao, Z. (2017). A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*, 256, pp. 56–62.
- Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations research*, 43(4), pp. 570–577.
- Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems*, 51(4), pp. 782–793.
- McDonald, J. H. (2009). *Handbook of Biological Statistics*, vol. 2. Baltimore, MD: Sparky House Publishing.
- Morariu, D., Cretulescu, R., & Breazu, M. (2013). Feature selection in document classification. In *The fourth international conference in romania of information science and information literacy, ISSN-L*, pp. 2247–0255.

- Mwangi, B., Tian, T. S., & Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2), pp. 229–244.
- Nasiri, H., & Alavi, S. A. (2022). A novel framework based on deep learning and anova feature selection method for diagnosis of covid-19 cases from chest x-ray images. *Computational intelligence and neuroscience*, 2022(1), 4694567.
- Ni, W. (2012). A review and comparative study on univariate feature selection techniques.
- Olasehinde, O. O., Alese, B., & Adetunmbi, A. (2018). Performance evaluation of bayesian classifier on filter-based feature selection techniques. *Connections*, 82332(100.00), 175341.
- Omara, H., Lazaar, M., & Tabii, Y. (2018). Effect of feature selection on gene expression datasets classification accuracy. *International Journal of Electrical and Computer Engineering*, 8(5), pp. 3194–3203.
- Palma-Mendoza, R.-J., de Marcos, L., Rodriguez, D., & Alonso-Betanzos, A. (2019). Distributed correlation-based feature selection in spark. *Information Sciences*, 496, pp. 287–299.
- Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1), pp. 23–69.
- Saari, P., Eerola, T., & Lartillot, O. (2010). Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6), pp. 1802–1812.
- Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19), pp. 2507–2517.
- Shakeela, S., Shankar, N. S., Reddy, P. M., Tulasi, T. K., & Koneru, M. M. (2021). Optimal ensemble learning based on distinctive feature selection by univariate anova-f statistics for ids. *International Journal of Electronics and Telecommunications*, pp. 267–275.

- Singh, A., & Jain, A. (2019). Adaptive credit card fraud detection techniques based on feature selection method. In *Advances in computer communication and computational sciences*, pp. 167–178.
- Singh, B., Kushwaha, N., Vyas, O. P., et al. (2014). A feature subset selection technique for high dimensional data using symmetric uncertainty. *Journal of Data Analysis and Information Processing*, 2(04), 95.
- Smith, J., & Johnson, J. (2020). Logistic regression analysis. *Journal of Statistical Methods*, 10, pp. 100–120.
- Solorio-Fernández, S., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2019). A supervised filter feature selection method for mixed data based on the spectral gap score. In *Pattern Recognition: 11th Mexican Conference, MCPR 2019, Querétaro, Mexico, June 26–29, 2019, Proceedings 11*, Springer, pp. 3–13.
- Taheri, N., & Nezamabadi-pour, H. (2014). A hybrid feature selection method for high-dimensional data. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, IEEE, pp. 141–145.
- Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, 167, pp. 706–716.
- Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, 26(1).
- World Health Organization Regional Office for Africa (2022). *Title of the article or page*. from <https://www.afro.who.int/>.
- Yan, X., Sarkar, M., Gebru, B., Nazmi, S., & Homaifar, A. (2021). A supervised feature selection method for mixed-type data using density-based feature clustering. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, pp. 1900–1905.
- Yang, S., Islam, M. T., et al. (2020). Principal component analysis and factor analysis for feature selection in credit rating. *arXiv preprint arXiv:2011.09137*.

Zhang, Y.-H., Pan, X., Zeng, T., Chen, L., Huang, T., & Cai, Y.-D. (2020). Identifying the rna signatures of coronary artery disease from combined lncrna and mrna expression profiles. *Genomics*, 112(6), pp. 4945–4958.

Zhou, H., Deng, Z., Xia, Y., & Fu, M. (2016). A new sampling method in particle filter based on pearson correlation coefficient. *Neurocomputing*, 216, pp. 208–215.

Zhu, L. (2020). Selection of multi-level deep features via spearman rank correlation for synthetic aperture radar target recognition using decision fusion. *IEEE Access*, 8, pp. 133914–133927.

