DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DIAGNOSIS OF THYROID DISEASE USING MACHINE LEARNING TECHNIQUES

By
Ege SAVCI

October, 2021
İZMİR

DIAGNOSIS OF THYROID DISEASE USING MACHINE LEARNING TECHNIQUES

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 Computer Science

By
Ege SAVCI

October, 2021
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DIAGNOSIS OF THYROID DISEASE USING MACHINE LEARNING TECHNIQUES"** completed by **EGE SAVCI** under supervision of **ASSOC. PROF. DR. FIDAN NURİYEVA** 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. Fidan NURİYEVA

_____

Supervisor

Prof. Dr. Efendi NASİBOĞLU

_____

Jury Member

Prof. Dr. Burak ORDİN

_____

Jury Member

_____

Prof. Dr. Özgür ÖZÇELİK

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank my supervisor Assoc. Prof. Dr. Fidan NURİYEVA for her guidance and efforts to support this study.

I would also like to thank to my family and my friends.

Ege SAVCI

DIAGNOSIS OF THYROID DISEASE USING MACHINE LEARNING TECHNIQUES

ABSTRACT

Thyroid disease is quite common in our country. In this respect, it is very important that the diagnosis of this disease is fast and accurate. It is possible and a good option to detect this disease with machine learning techniques, which is usually diagnosed with various laboratory tests and imaging tests.

There are two types of thyroid disease (hyperthyroid and hypothyroid) in the target class, as well as healthy patients, on the dataset containing a total of 7000 lines of analysisand information about the people undergoing the tests. In this multi-classification problem, it is aimed to determine whether the patients are hyperthyroid, hypothyroid or healthy.

Since there were unbalanced distributions on the dataset and overfitting occurred, various operations were required. These included feature extraction by correlation, undersampling and oversampling. In addition, the parameters affecting the result were found and the unimportant ones were removed. After these processes, the algorithms ran on these datasets and it was seen that the best result was feature extraction with correlation.

Support vector machines, artificial neural networks, logistic regression, k-nearest neighbors and decision trees (random forest algorithm) were used for prediction. The most successful among them were artificial neural networks, k-nearest neighbors and support vector machines.

Keywords: Thyroid disease, multiclassification, neural network, random forest, support vector machines, k-nearest neighbors, feature extraction, oversampling, undersampling

TİROİD HASTALIĞININ MAKİNE ÖĞRENİMİ TEKNİKLERİ İLE TEŞHİSİ

ÖZ

Tiroit hastalığı ülkemizde oldukça yaygındır. Bu doğrultuda, bu hastalığın teşhisinin de hızlı ve doğru olmasının önemi yüksektir. Genellikle çeşitli laboratuvar tahlilleri ve görüntüleme testleri ile teşhisi koyulan bu hastalığı makine öğrenmesi teknikleri ile de belirlenmesi mümkün ve iyi bir seçenektir.

Toplam 7000 satırlık tahlil yapılan insanlara ilişkin bilgilerin yer aldığı veri seti üzerinde, hedef sınıfında tiroit hastalığının iki çeşidi (hipertroit ve hipotiroit) ve sağlıklı insanlar yer almaktadır. Bu çoklu sınıflandırma probleminde, hastaların bu üç gruptan hangisine ait olduğunun bulunması hedeflenmiştir.

Veri seti üzerinde dengesiz dağılımlar olduğundan ve fazla uyum gerçekleştiğinden çeşitli işlemler yapılması gerekmiştir. Bunlar arasında korelasyon ile öznitelik çıkarımı, düşük ve yüksek hızda örnekleme yer almıştır. Ayrıca sonuca etki eden parametreler bulunup önemsiz olanlar çıkarılmıştır. Bu işlemler sonrasında algoritmalar uygulanmıştır ve en iyi sonuç veren yaklaşımın korelasyon ile öznitelik çıkarımı olduğu görülmüştür.

Tahmin doğrultusunda destek vektör makineleri, yapay sinir ağları, lojistik regresyon, k-yakın komşu ve karar ağaçları (rastgele orman algoritması) kullanılmıştır. Bunlar arasında en başarılı olanlar ise yapay sinir ağları, k-en yakın komşu ve destek vektör makineleri olmuştur.

Anahtar kelimeler: Tiroit hastalığı, çoklu sınıflandırma, yapay sinir ağı, rastgele orman, destek vektör makineleri, k-en yakın komşu, öznitelik çıkarımı, yüksek hızda örnekleme, düşük hızda örnekleme

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER 1
INTRODUCTION

In this thesis we are going to discuss different machine learning algorithms for detecting thyroid diseases. Thyroid disease is widely common in the world, especially on women. It is estimated that approximately worldwide about 200 million people have a thyroid disorder (Jonklaas, Bianco, Bauer, 2014).

The thyroid gland is a tiny organ in thefront of the neck that produces thyroid hormones. When your thyroid isn't functioning properly, it can have a negative impact on your entire body. Hyperthyroidism is a condition that occurs when your body produces too much thyroid hormone. Hypothyroidism is a condition in which your body produces too little thyroid hormone. Both illnesses are dangerous and require medical attention. This is why the false diagnosesmight be severe, and need to be perfectly examined by experts.

With recent development on machine learning and data mining, it is not unusual for experts to get help by these algorithms and machines. Our aim is to help experts through their examination process with provided data. Hence, we need to create models for specific goal and make the data meaningful for them. Since it is matter of health, sensitivity of the results is very important. We aim to predict whether the patient has thyroid disease or not, therefore we value every little development on our outputs.

The purpose of this study is also comparing the algorithms' accuracy while predictingthe class of diagnose. In order to increase the efficiency of the corresponding classifier, data preparation and reducing features will be important, also we aimed that showing themost essential part of creating models, preparing the data, for the best result. Accordingly,we've implemented feature selection with correlation, mutual information, andoversampling/undersampling over data to tune our model for the better results. From outputs of these techniques, we've reached the optimal tunes for

our different algorithms.

All algorithms and prediction models has been implemented with Python. The reason behind this is Python is one of the most used and wanted programming languages for machine learning. With its packages such as Numpy, Sci-kit Learn,Pandas and Keras makes it easier to deploy a model and tune in however you need.

CHAPTER 2
LITERATURE

Generally, diagnosis of diseases and cancers are very popular in machine learning, because of potential data amount in healthcare makes it suitable for prediction algorithms.There are several important researches about this topic.

One of the most related work is “Diagnosis of Thyroid Disease Via Support Vector Machines” (Korhan, 2016). In this work, they have used the same data within this work. However, they approached the problem only with Support Vector Machines (SVM). Theyhave prepared the data with feature selection and optimization. The results are given withtheir accuracy, sensitivity and specifity.

The other work on thyroid diagnosis is “Diagnosis of Thyroid by Hybrid Particle Swarm Optimization with Artificial Neural Network” (Adak & Yumusak, 2016). They‘ve used Neural Networks with PSO (Particle Swarm Optimization) and trained neural networks with it. Also, results are optimized with genetic algorithms. For the results, they’ve compared them with back propagation method. Again, only one approach used inthis work for the diagnosis, which was Artificial Neural Network.

Another paper who used same dataset within this thesis, they’ve managed to used NaïveBayes, Decision Trees, Neural Networks and Radial Basis Function Network with the helpof a KNIME software. They also compared the results adding and removing specific attributes. And concluded that the best approach was Decision Trees among the others (Irina Ionită & Liviu Ionită, 2016).

Decision Tree used in this work for diagnosis Thyoid Disease, “MLTDD : Use Of Machine Learning Techniques For Diagnosis Of Thyroid Gland Disorder” (Al-muwaffaq& Bozkus, 2016). PCA (Principle Component Analysis) used for feature selection, which be used in this work too. They’ve also developed an app for real life use cases, takinginputs of new patients as parameters. The accuracy is very high with their model.

In this paper they have compared three algorithms, logistic regression, decision trees and k-nearest neighbor. They've taken the data from the Graven Institute in Australia. Thedataset contains 5 attributes and 215 instances. They used information gaining for decisiontree and evaluate the best model among many. For kNN, they've used Euclidean distanceas they evaluate the model. As result, they've achieved 81% accuracy for logisticregression, 87% for decision tree and 96% for kNN algorithm. In general, the paper describes in detail the data preparation, training, and testing, as well as a step-by-step discussion of each of the strategies employed and a comparison of the methods' accuracyin prediction (Chaubey, Bisen, Arjaria, & Yadav, 2021).

Another paper called "An experimental comparative study on thyroid disease diagnosisbased on feature subset selection and classification" (Kousarrizi, Seiti, & Teshnehlab, 2012), which gathered their dataset from UCI machine learning repository with 215 instances and 5 attributes. Addition to this dataset, they worked on a real dataset gatheredby the Intelligent System Laboratory of K.N.Toosi University of Technology from ImamKhomeini hospital. Pattern recognition was used to extract or choose a feature set for usage in the pre-processing stage. For feature selection, it's been used Sequential forwardselection and sequential backward selection. Diagnosis has been predicted by Support Vector Machines with 3- and 10-fold cross validation with the accuracy of 98%.

CHAPTER 3
THYROID DISEASE AND DIAGNOSIS

Thyroid disease is quite common in the world, especially in women. It is estimated thatin the United States more than 12 percent of inhabitants have developed in some way some condition Thyroid gland hormone, in addition, it is estimated that approximately 20 million Americans have some form of Thyroid disease (Pearce, Farwell, Braverman, 2003). In the case of India, it is estimated that there are approximately more than 42 million people who suffer a similar situation (Patil, Rehman & Jialal, 2020). In Turkey, it is the second most common type of cancer. For the national case, it is calculated that approximately 4% of the population suffers from some disorder hormonal (Gültekin & Boztaş, 2009) and it is estimated that approximately worldwide about 200 million people have a thyroid disorder (Jonklaas, Bianco, Bauer, 2014).

It is important to say that worldwide pathologies of this type they affect mostly women,especially if they are in advanced ages, according to information obtained from thyroid aware women are four to seven times more possibility of being affected by a thyroid disorder that men, according to the International Thyroid Federation the effects that can cause such a disorder in people range from excessive tiredness, depression, anxiety, irritability, difficulty concentration, headaches and migraines, lethargy, intolerance to hot or cold places, menstruation disturbance and even alteration in the autoimmune system (American Thyroid Association, 2019).

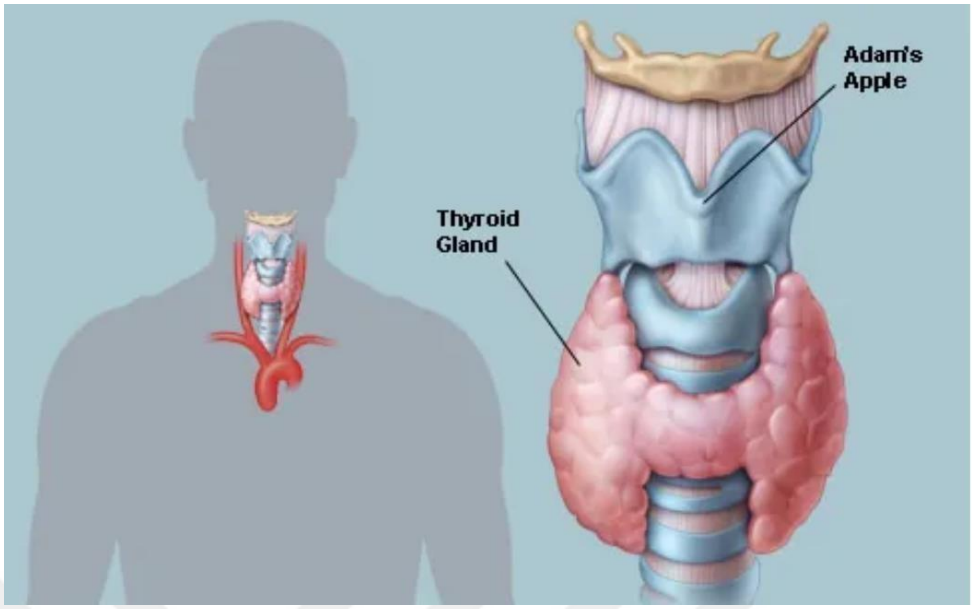

Figure 3.1 Thyroid Gland (Hoffman, 2021)

The thyroid is an organ located in the front of our neck (Figure 3.1). The function of the thyroid gland is to produce, store and, when necessary, transport thyroid hormones into the blood, thus regulating our metabolism. The thyroid is a small gland, weighing 1,520 grams and the size of a walnut. It is located under the skin in the front of the neck and is shaped like a butterfly. The wings of a butterfly are called the left and right leaves,and the part that connects these two leaves in the middle is called the isthmus. Each leaf is 4 cm long and 12 cm wide. The thyroid gland is just behind the bulge of the trachea (larynx) called the Adam's apple and moves up and down when you swallow. For this reason, your doctor will ask you to swallow during the test. The thyroid gland is an organthat uses the mineral iodine extracted from food and water to produce thyroid hormones.After iodine enters the bloodstream from the intestines together with water and food, it reaches the thyroid gland in our neck and is used to produce thyroid hormones. The iodinethat enters the thyroid combines with the amino acid tyrosine here to form thyroid hormones called T3 and T4. T4 is called T4 because there are four iodine molecules in thehormone structure, and T3 is because there are 3 iodine molecules in the T3 hormone structure. The amino acid tyrosine is provided by the protein foods we eat. Adequate intake of protein and iodine through food and water is necessary for the normal productionof

thyroid hormones. The hormones T3 and T4 produced in the glands are then released into the bloodstream, where they enter all organs and cells of the body and play a role in the production of two active thyroid hormones, levothyroxine (T4) and triiodothyronine (T3), which helps the body to adjust body temperature, blood pressure and heart rate. (Yıldırmkaya et. al., 1996).

Hyperthyroidism and hypothyroidism, two of the most prevalent thyroid disorders, arecaused by aberrant thyroid activity (Hershman, 2020). Hypothyroidism (sometimes calledunderactive thyroid or low thyroid) is a disorder in which the thyroid gland does not generate enough hormones. Hypothyroidism can lead to a variety of health issues,including obesity, joint discomfort, infertility, and heart disease if not treated properly. Hyperthyroidism (overactive thyroid) is a disorder in which the thyroid gland produces too much of the hormone thyroxin. In this scenario, the body's metabolism is rapidly increasing, resulting in rapid weight loss, a quick or erratic heartbeat, and a brain(Broughton & Ahmad, 2019).

3.1 Diagnosis and Treatment

The most sensitive inspection method is ultrasound. Even in healthy people over 50 years old, about 50% of nodules can be detected by ultrasound. In fact, more than 90% ofthese nodules are benign, and less than 10% are malignant (they contain cancer cells). It is very important that the radiologist performing the ultrasound examination is experienced and able to provide detailed information about the structure of the nodule (Cleveland Clinic, 2020).

After the TSH hormone level is confirmed to be high, the low sT4 result confirms the diagnosis. In most cases, along with commencing treatment without further testing, anti- thyroid antibody measurement is used to identify whether the condition is autoimmune or not (Yıldırımkaya et. al., 1996).

Hyperthyroid is much clear disease and easier to diagnose. The general examinations performed in light of the symptoms are sufficient to make a diagnosis.

Thyroid hormonesand TSH levels in the blood are measured for a precise diagnosis. T3 and T4 levels are usually high in both or just one of the patients. Thyroid scintigraphy provides informationon nodules. Thyroid uptake testing, on the other hand, provides information on all of the thyroid's functions.

However, if early diagnosis is not possible, then nodules grow without patient's knowledge and it causes side effects. In these cases where they can not treat with medicineonly, it can be applied direct surgery or nuclear treatment on glands. To determine this there are several ways. Fine-needle aspiration biopsy (FNAB) is the easiest way to tell if nodules larger than 11.5 cm in diameter are benign or cancerous by ultrasound. Interventional radiologists with experience in thyroid biopsies use small needles to remove cells from the nodules under ultrasound guidance, examine these cells under a pathology microscope, and makea diagnosis. The biopsy result can be benign or malignant and, in some cases, it can be suspicious.

When the biopsy result is malignant, that is, compatible with cancer, part or all of the thyroid gland is surgically removed. Radioactive iodine treatment can then be done to destroy the remaining cells. With radioactive iodine, all tissues that contain the thyroid gland are exposed to intense radiation. After treatment, the patient was kept in a room where all the walls were covered with lead for 23 days.

90% of cases with suspicious biopsy results can be diagnosed as benign by repeated fine needle biopsy and / or core needle biopsy. Therefore, it is not the correct way to act immediately when the result is suspicious. In centers that recommend immediate surgeryrather than a second biopsy, 90% of unnecessary thyroid surgeries are performed on benign thyroid nodules. Although these procedures have many risks associated with permanent hoarseness or anesthesia, many patients have to rely on synthetic thyroid hormones for a lifetime after the operation. In these patients, a second needle biopsy usinga technique (hollow needle biopsy) can make a clear diagnosis to a large extent and prevent unnecessary operations. The reason for this is that compared with FNA, the use of skill biopsy can remove larger tissue masses and can diagnose pathology more accurately.

Research conducted in recent years has shown that aerobatic biopsy can makea clear diagnosis 80-90% of patients with suspicious results in the first FNA. Therefore, for these patients, aerobatic biopsy should be performed first, and if the aerobatic biopsy cannot get a clear result, then surgery should be considered.

If the biopsy result is benign, if the patient does not have any symptoms, ultrasound examinations of the nodules are performed every 6 months. However, if benign nodules are larger than a certain diameter, cause discomfort, or grow rapidly, treatment is generallyrecommended. Complaints about benign nodules may be due to the huge influence of thyroid nodules or the production of hormones. Depending on the mass effect of the nodule, complaints such as cosmetics, difficulty swallowing, dyspnea, voice changes and neck pain may appear. In addition, depending on the hormones produced by the nodules,symptoms such as palpitations, irritability, hand tremors, insomnia and sweating may appear.

Classic treatment is surgical removal of part or all of the thyroid gland. However, thyroid surgery has some risks and drawbacks. It leaves a permanent incision mark on theneck, the patient must use the drug for life, and there are risks associated with anesthesia.Carrying out all these risks on benign nodules is not suitable for modern medicine today and it is a method that has been questioned.

New treatment methods have been developed in the treatment of benign nodules and have been applied with success in recent years. Percutaneous ablation is a good alternativeto surgical procedures and has been widely used around the world in the last 10 years. Inmy country, there are interventional radiologists with experience in the treatment of non-surgical thyroid nodules. This method uses multiple needles to enter the nodules under theguidance of local anesthesia and ultrasound, and destroys the nodules by heating (thermalablation) with laser, radiofrequency, or microwave energy. The preferred method for cystic (liquid) nodules is alcohol injection therapy (chemical ablation). After treatment ofthyroid nodules, the patient can return to normal life in a few hours without pain and without a scalpel, using modern methods. Plus, you don't have to use synthetic thyroid hormones for your entire life. Therefore,

the preferred method for benign nodules shouldbe modern non-surgical methods. If necessary, classical surgical treatment should be applied in the second stage.

CHAPTER 4
MATERIALS AND METHODS

4.1 Dataset

The data set used for this thesis downloaded from university of California of Irvin (UCI) repository site. It consists of 3772 are training instances, 3428 testing instances and 22 attributes with total 7200 instances and 3 classes. Dataset describes the main characteristics of the thyroid data set and its attributes: Each measurement consists of 21 values – 15 binary and 6 are continuous. Each of the measurement vectors is assigned oneof three classes, which correspond to hyperthyroidism, hypothyroidism, or normal thyroidfunction (Quinlan, 1987).
Ranges of raw data's binary and continuous values like below.

Table 3.1 Range of dataset variables

	Age	Sex	On thyroxine	Query on_thyroxine	On antithyroid medication	Sick	Pregnant	Thyroid surgery	I131 treatment	Query hypothyroid	Query hyperthyroid	Lithium	Goitre	Tumor	Hypopituit	Psych	TSH	T3	TT4	T4U	FTI	outputs Cl
min	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.002757	0.009727	0.043946	0.007782	1
max	0.97	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.97744	0.500215	0.77178	0.988283	0.974955	3

4.1.1 Data Explanation

Age: Age of patients, continuous.

Sex: Sex of patients, categorical (0,1).

On thyroxine: Usage of hormone when thyroid gland does not make enough thyroxine, categorical (0,1).
On anti thyroid medication: Usage of medication when thyroid gland make too much thyroxine, categorical (0,1).
Sick: Sickness, categorical (0,1)

Pregnant: Pregnancy, categorical (0,1)

Thyroid Surgery: Have a thyroid operation, categorical (0,1)

IT3T Treatment: Have IT3T treatment, categorical (0,1)

Lithium: Lithium usage, categorical (0,1)

Goiter: Has goiter, categorical (0,1)

Tumor: Has tumor, categorical (0,1)

Hypopituitary: Has hypopituitary, categorical (0,1)

Psych: Has psychological symptoms, categorical (0,1)

TSH: Thyroid-stimulating hormone, a blood test that measures this hormone, continuous

T3: Triiodothyronine hormone, plays a crucial part in the body's metabolic control, continuous.

TT4: Thyroxine Test, a type of thyroid hormone. Too much or too little can indicate thyroid disease, continuous.

T4U: Thyroxine utilization rate, continuous.

FTI: The free T4 index (FTI) is a measure for the free hormone level, continuous.

Output Class: Target class, Hypothyroid, hyperthyroid and normal, categorical (1,2,3).

4.1.2 Frequency and Range Table

Table 4.1 Frequency and Range Table of Dataset

Age	**Range**	**Average**
Number	1-97	52
Sex	**Count of Sex**	**Frequency**
Female	5009	70%
Male	2191	30%

Table 4.1 Continues

Row Labels	Count of Onthyroxine	Frequency
F	6260	87%
T	940	13%
RowLabels	**Count of On antithyroidmedication**	**Frequency**
F	7108	99%
T	92	1%
RowLabels	**Count of Sick**	**Frequency**
F	6924	96%
T	276	4%
RowLabels	**Count of Pregnant**	**Frequency**
F	7122	99%
T	78	1%
RowLabels	**Count of Thyroidsurgery**	**Frequency**
F	7099	99%
T	101	1%
RowLabels	**Count of IT3T treatment**	**Frequency**
F	7079	98%
T	121	2%
Row Labels	**Count of Query hypothyroid**	**Frequency**
F	6728	93%
T	472	7%
RowLabels	**Count of Query hyperthyroid**	**Frequency**
F	6705	93%
T	495	7%

Table 4.1 Continues

RowLabels	Count of Lithium	Frequency
F	7109	99%
T	91	1%
RowLabels	**Count of Goiter**	**Frequency**
F	7141	99%
T	59	1%
RowLabels	**Count of Tumor**	**Frequency**
F	7016	97%
T	184	3%
RowLabels	**Count of Hypopituitary**	**Frequency**
F	7199	100%
T	1	0%
Row Labels	**Count of Psych**	**Frequency**
F	6848	95%
T	352	5%

4.2 Materials and Methodology

In this research, all implementations have been made with Python 3.6.5. The reason behind this is Python is a very strong and trending programming language among machinelearning. It is easy to process data and creating models. With packages you can tune yourmodels with high detail.

Among the machine learning techniques, several classification algorithms have been used. Logistic Regression (LR), Random Forest (RF), Support Vector Machine

(SVM), Artificial Neural Network (ANN) and k-nearest neighbors algorithms are mostly used in classification problems, also developed for such classification problems.

One of the essential things in our thesis is the target has 3 different classes. That meansbinary classification algorithms would not fit into this problem. For this, we've implemented multi-class classification since we have 3 different target classes (non-sick,hyperthyroid, hypothyroid).

Since the dataset is very wide, it needs to be processed to reduce redundancy and possible overfitting problems. Also, target class distribution is highly unbalanced. There are 6666 non-sick, 368 hypothyroid and 166 hyperthyroid instances.

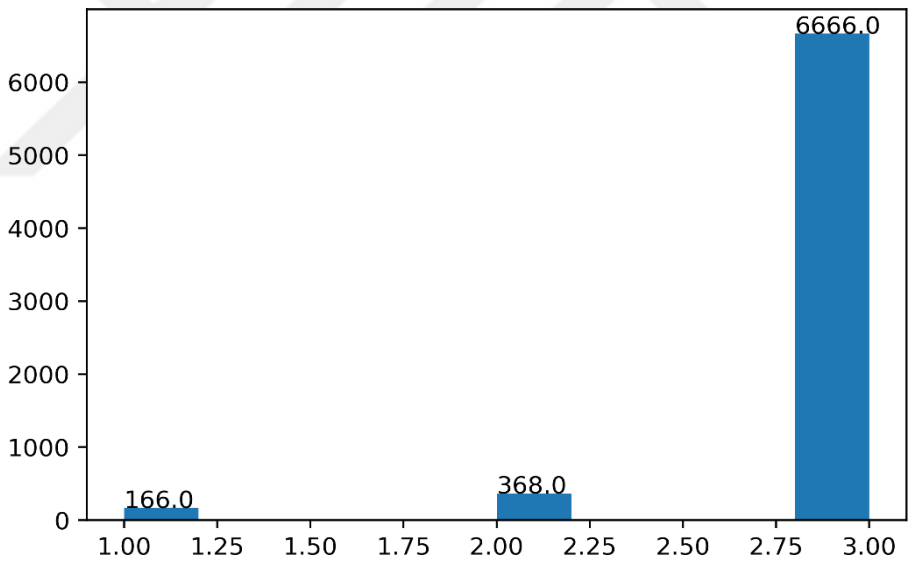

Figure 4.1 Frequencies of output class

As it can be seen in Figure 4.1, total number of instances for target class of 3 (non- sick), is very high amongst the other classes. Since their distribution is unbalanced, training and testing sub datasets would be much smaller and it can lead wrong results sinceit is very hard to train the dataset with so little data.

4.3 Pre-Processing

The "curse of dimensionality", which refers to a number of issues that arise while examining data in a high-dimensional domain, has greatly increased the computational burden. This computational burden can also lead to false results by running algorithms ormodels. So, analyzing data in high-dimensional space requires good amount of time for pre-processing. Despite it is costly for both time and resource, it is a must for any high- dimensional data. Because that computational burden's cost is much higher than pre- processing's.

Data processing becomes significantly more difficult in high-dimensional space because data becomes sparser and a large number of samples are required to train models.One of the most successful tools for reducing dimensionality and addressing the problem mentioned above is feature selection. It has been widely used as an effective preprocessing tool in a variety of applications, including data mining, machine learning and pattern recognition.

A vast number of attributes and limited available samples are common characteristics of biological data learning. Filtering out insignificant features prior to model fitting, for example, by discarding the ones that are least associated with the result, is a typical method. Mutual information (MI) is a commonly used measure of association in this context (Guyon & Elisseeff, 2003). It has also been shown to favor variables with more categories. It is also closely related to the Gini index and it has been proven also to be biased in favor of variables with more categories (Achard et al., 2005).

Feature selection, in general, yields a feature subset, with the purpose of obtaining themost informative subset by selecting crucial traits and eliminating unnecessary aspects from the original feature set. During the selection process, candidate traits and target classes are invariant, but the relationship between them is not. It will be a constantly changing value as more characteristics are added to the subset. As a

result, determining the relationships between candidate features, selected features, and categories during the selection process is difficult. Classifier independent Filter, classifier-dependent Wrapper, and Embedded techniques are three types of supervised algorithms that deal with diverse interactions between data and classifiers.

4.4 Scaling Data

Feature scaling is a technique for normalizing a set of independent variables or data components. It is also known as data normalization in data processing and is usually done during the data preprocessing step. This strategy is necessary for accurate prediction and outcomes. When one of the columns has a very high value in comparison to the others, the impact of that column will be significantly greater than the impact of the other low- valued columns. Not every dataset requires normalization. It is required only when features have different ranges.

Some examples of algorithms where feature scaling matters are,

- k-nearest neighbours with a Euclidean distance measure
- k-means
- Logistic Regression, SVMs, Neural Networks
- Tree based models are not distance based models and can handle varying ranges of features. Hence, Scaling is not required while modelling trees.

The data has been normalized by min-max normalization. One of the most popular methods of data normalization is min-max normalization. This technique converts its minimum value to 0, and maximum to 1. All between values transforming to decimal values between 0 and 1. In this way all data values are more balanced for the model.

The table of data before and after normalized below.

Table 4.2 Compare of raw and normalized dataset

TSH		T3		TT4	
Before	After	Before	After	Before	After
0.0006	0.00290	0.015	0.072	0.120	0.580
0.00025	0.00111	0.0300	0.1332	0.1430	0.6350
0.00190	0.0102	0.024	0.129	0.102	0.551
:	:	:	:	:	:

T4U		FTI	
Before	After	Before	After
0.082	0.396	0.1460	0.7068
0.133	0.590	0.1080	0.4796
0.131	0.708	0.0780	0.4215
:	:	:	:

4.5 Correlation Feature Selection

In this case, we've aimed to find the specific columns that has correlated with target or with each other. It is better to remove high correlated features within columns, because it causes overfit or underfit with low correlation. However, if specific column has high correlation with the target, it needs to be included in the training data, since it is the significant column effecting the target column.

Correlation is a measure of the linear association between two or more variables. The higher the correlation, the more linearly associated the variables are. Correlation is often a useful property. If two variables are correlated, we can predict one from the other. Therefore, we generally look for features or predictor variables that are highly correlated with the target, particularly for linear machine learning models. However, if two predictor variables are highly correlated among themselves, they provide in essence, redundant information about the target because with just one of them we can make accurate predictions on the target. The second predictor does not add additional information. Therefore, to make good machine learning models, in general will look for variables that are highly correlated with the target yet uncorrelated among themselves. In other words, we want the predictors to be correlated with the target but the predictors should not be correlated among themselves. In fact, the center hypothesis of correlation feature selection is that a good set of features, this is, a good set of predictive variables, contains variables that are highly correlated with the class or the target but uncorrelated with each other. Correlated features do not affect classification accuracy per se. The problem arises when we have datasets with loads of features and in extreme cases more features than examples. This is commonly known as the cause of dimensionality. If 2 numerical features are correlated, then one doesn't add much additional information over the other feature. So, if the numbers of features are high, removing correlated features is a good approach to reduce the feature space without losing information (Wei, Zhao, Feng, He, & Yu, 2020). One thing though, is that correlated features do affect more than interpretability and in particular, in linear models. If two variables are correlated, the model will fit coefficients to both variables that somehow capture the

correlation. Therefore, these will be misleading on the true importance of each of the individual features. And this is also true for ensembletree models, if two features are correlated, tree methods will assign roughly the same importance to both but half of the importance they would assign if we had only one of thecorrelated features in the dataset. The data's correlation heat maps are below.

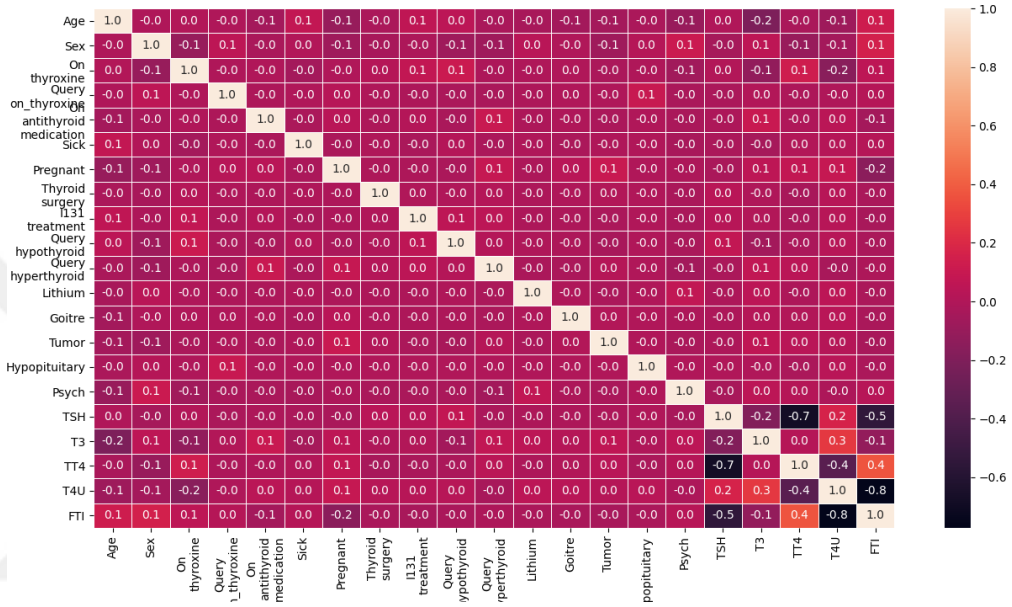

Figure 4.2 Correlation heatmap of dataset

Overall, the columns we're going to use mainly for our model are TSH, T3, TT4, T4U, and FTI, along with categorical variables sex and On-thyroxine.

In Figure 4.3 below, there is the FTI and TSH values' correlation graph. There canbe seen negative correlation and the line between two variables. As it has been mentionedbefore, the correlation between two variable is -0.5.

That means there is an inverse proportion between variables, when one is high the other one is lower.

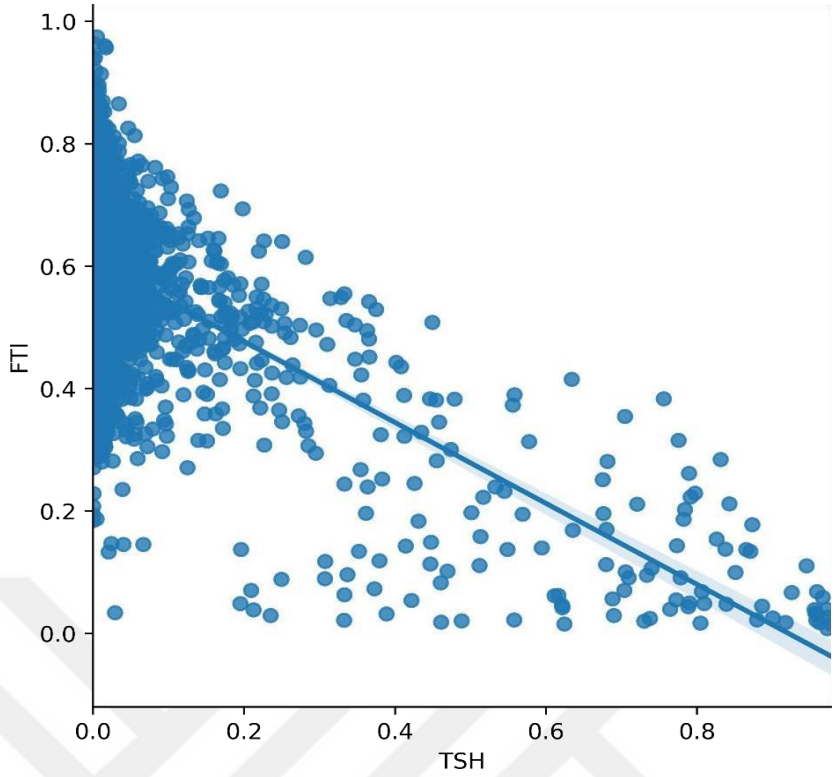

Figure 4.3 Correlation graph between FTI and TSH values.

4.6 Mutual Information

Mutual Information measures non-linear relationships between two random variables. It also shows how much knowledge can be gleaned from one random variable by viewinganother random variable.

It is intertwined with the idea of entropy. This is due to the fact that it can also be referred to as the reduction of uncertainty of a random variable when another is known.

As a result, a high mutual information value denotes a significant reduction in uncertainty, whereas a low value denotes a minor reduction. If the mutual information is 0, the two random variables are considered independent.

Mutual information is especially guides us to select important features. As you

can seebelow (Figure 4.4), we've managed to find mutual information with target class. We've aimed the both hypothyroid and hyperthyroid as a comparing class.

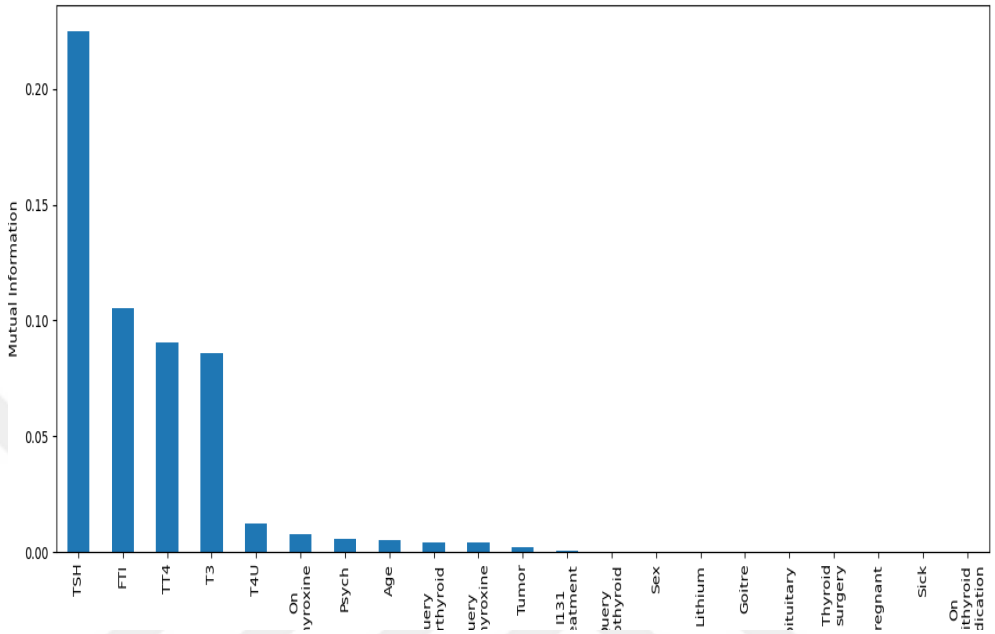

Figure 4.4 Mutual Information Graph

4.7 Feature Importance

One of the most extensively used data analysis approaches is feature importance, whichis utilized for two main goals. One of them is to choose features for predictive models, removing the least predictive features to simplify and potentially increase the model's generality, and to apply business, medical, or other insights, such as customer-valued product characteristics or treatment contributions, to business, medical, or other insights.The likelihood of reaching that node to compute feature importance is weighted by the decrease in node impurity.

The node probability is calculated by dividing the number of samples that reach the node by the total number of samples (Breiman & Friedman, 1984).

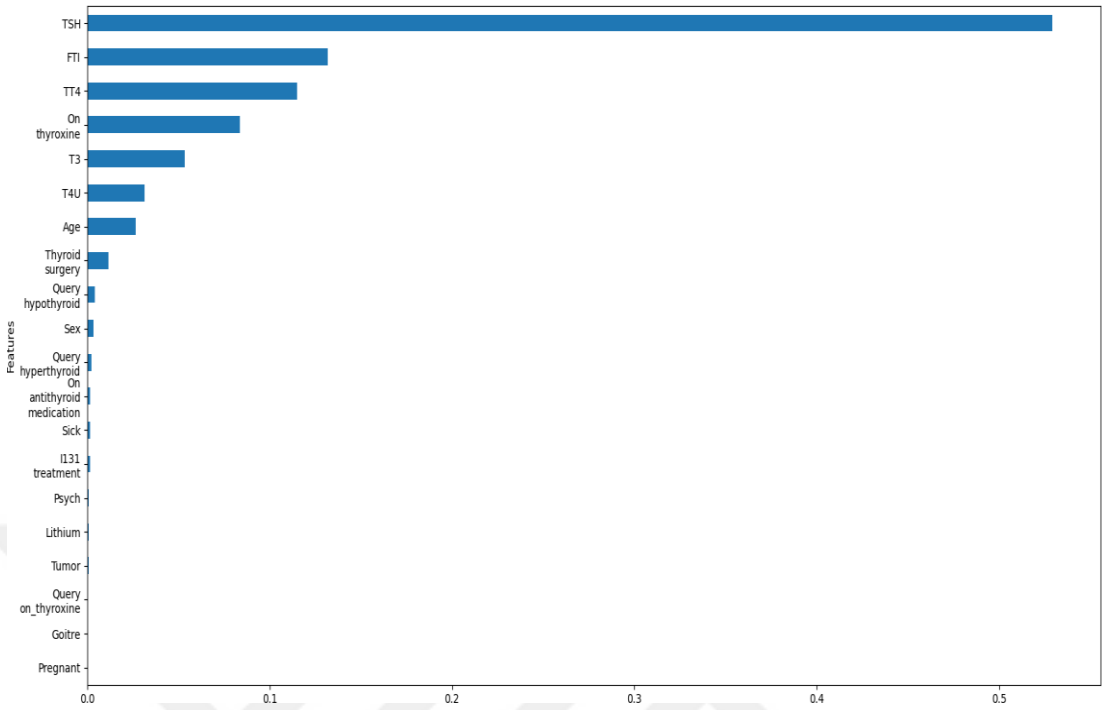

Figure 4.5 Feature Importance Order

In our data, we've found the most characteristically important features with random classifier technique. Results are also similar with correlated features that we've discovered. Our data with most important features as above (Figure 4.5).

Importance of features above. Most important features by random classifier tree are TSH, FTI, TT4, On-thyroxine, T3 and T4U. Again, these are the most correlated features.

In machine learning, outliers are as important as correlated features. In this case, we can remove those outliers as well as we use featured variables. According to our feature importance list, we can remove the least significant values. These are the values that closeto 0 feature importance or the least correlated features.

According to these outputs, we can remove the most unrelated field for our target classto maintain the accuracy for our algorithms. In this case, we've removed Goitre,

Pregnant,Query On-Thyroxine fields.

4.8 Over and Under Sampling

In machine learning, predicted values depends on model's ability to capture characteristics of different classes. In some cases, data has been skewed towards positiveor negative class. If this skewness has not been normalized or optimized, it may result to overfitting or underfitting while predicting classes. The issues that come from data imbalance are handled by two resampling procedures, namely oversampling and undersampling, from a data-centric approach. Both resampling strategies add data preprocessing expenses to the equation, which might be overpowering in the case of verylarge-scale training data (Dey, 2019).

Practically, this process is taking skewed part and optimize it for the other values', means that number of values in that target class equalized with each other.

Oversampling does not result in the loss of information because all samples from the minority and majority classes are kept. However, because to the higher number of trainingexamples, oversampling generates more training time throughout the learning process. Furthermore, using a complicated oversampling strategy results in substantial computational expenses during data preprocessing.

Undersampling has been presented as a useful way to improve a classifier's sensitivity.However, this strategy may result in the omission of potentially useful data that is criticalto the learning process (Ertekin, 2013).

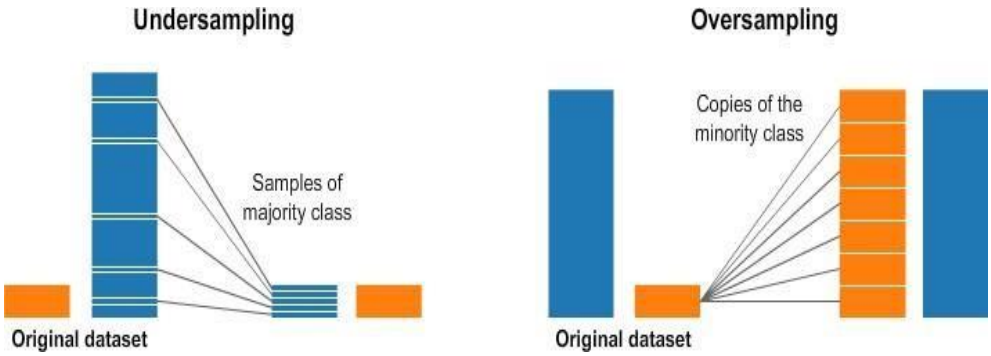

Figure 4.6 Explaining the concept of undersampling-oversampling

In our data most target values consist of one class, which it corresponds as the "not sick". Other two classes are 2 and 3, hypo and hyper thyroid respectfully.

Here is the data before and after sampling.

Table 4.3 Compare of Sampled Data

Class	RawData	Undersampled Data	Oversampled Data
1	6666	368	6666
2	368	350	5000
3	166	166	3000

Oversampling can be defined as adding more copies to the minority class. Oversampling can be a good choice when there is not many data to work with. However,tend to result in data that is non-smooth, has boundaries or small features. One way to avoid this pitfall is to combine undersampling and boosting. You might also want to manually resample or repair any holes in the data algorithmically

CHAPTER 5
ALGORITHMS

5.1 Support Vector Machine

SVM is a widely used in classification of high-dimensional data. It is applied by training on data. Any method used in convex optimization can be chosen for training. Basically, it allows separating a data set that cannot be linearly separated in low dimensions by moving it to a higher dimension with the help of a plane.

Support vector machines (SVMs) were created with binary classification in mind. It's still a work in progress to figure out how to make it work for multiclass classification. Numerous methods have been developed, in which a multiclass classifier is often built byintegrating several binary classifiers (Hsu & Lin, 2002).

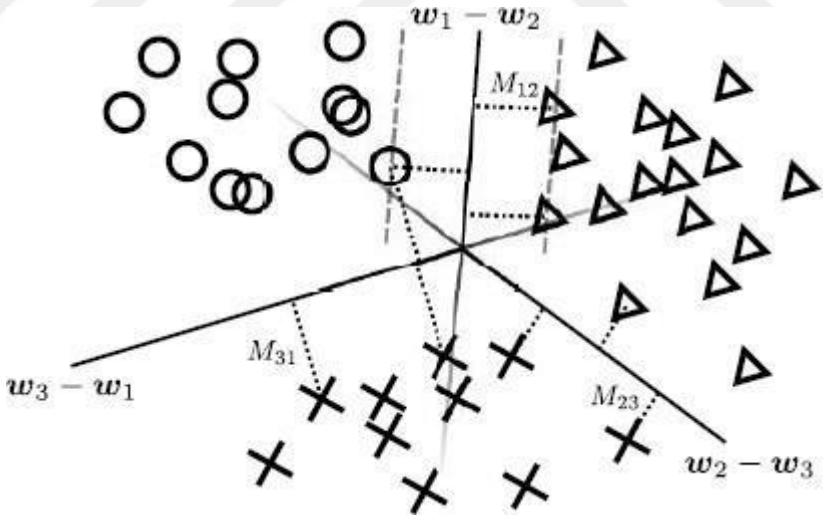

Figure 5.1 Margins of multiclass SVM (Aggarwal,2014)

The purpose is to transfer data points to a high-dimensional space so that classes can be separated mutually linearly. This is known as a One-to-One technique, in which the multiclass problem is broken down into several binary classification

problems. Each pair of classes has a binary classifier (Mircia, Imre, & Balaci, 2010). Another approach one can use is One-to-Rest. The breakdown is set to a binary classifier per class in that manner (Chen, Yang & Wang, et al, 2012).

We've implemented one-to-one approach with Python in our thesis. 10 folds been usedin SVM algorithm with 30% training and 70% test data separation.

5.2 Multinomial Logistic Regression

Multinomial Logistic Regression is a subset of logistic regression that can handle multiclass classification problems. To support multi-class classification issues natively, this process requires updating the loss function to cross-entropy loss and to a multinomialprobability distribution, predict the probability distribution (Sultana & Jilani, 2015). Thisapproach is using in different fields such as image classification, mail classification etc.

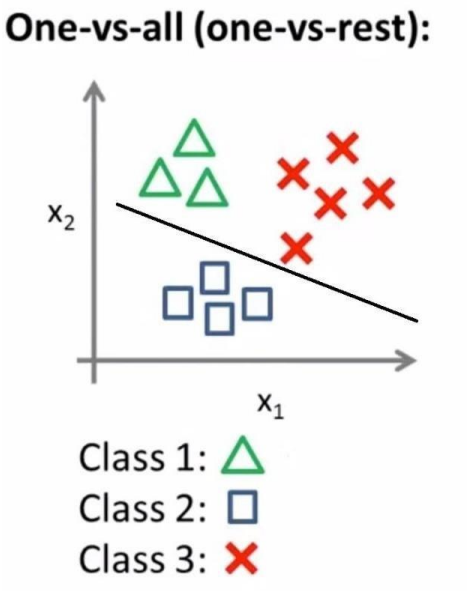

Figure 5.2 Multinomial Logistic Regression with 3 class representation (Debajyoti Saha)

When modifying model weights during training, cross-entropy loss with one-vs-allmethod is used (Figure 5.2). The objective is to minimize the loss; the lower the loss, thebetter the model. A perfect model has zero cross-entropy loss. Most of the time, it's used for multi-class and multi-label classifications.

In our thesis, we used one-vs-rest scheme for multi-class classification and cross-entropy loss. The number of bits necessary to represent or transmit the average event fromone distribution compared to another is calculated using cross entropy, which is based onthe entropy theory.

5.3 Random Forest

Random Forest is a decision tree algorithm that uses a divide-and-conquer strategy to construct decision trees from a randomly partitioned dataset. A group of decision tree classifiers is referred to as the forest. An attribute selection indicator such as informationgain, gain ratio, or Gini index is used to generate individual decision trees for each attribute. Each tree is built using a different random sample. The most popular class is chosen as the final result when each tree votes on a categorization challenge.

Each forest tree votes with a single unit, allocating each input to the most likely class label. It's a fast, noise-resistant algorithm with a successful ensemble for detecting non- linear patterns in data. It is capable of handling both numerical and category data with ease (Titapiccolo et al., 2013). One of the most significant advantages of Random Forest is that it does not suffer from overfitting, even as the forest grows larger.

The principle of random forest algorithms can be shown as below (Figure 5.3).

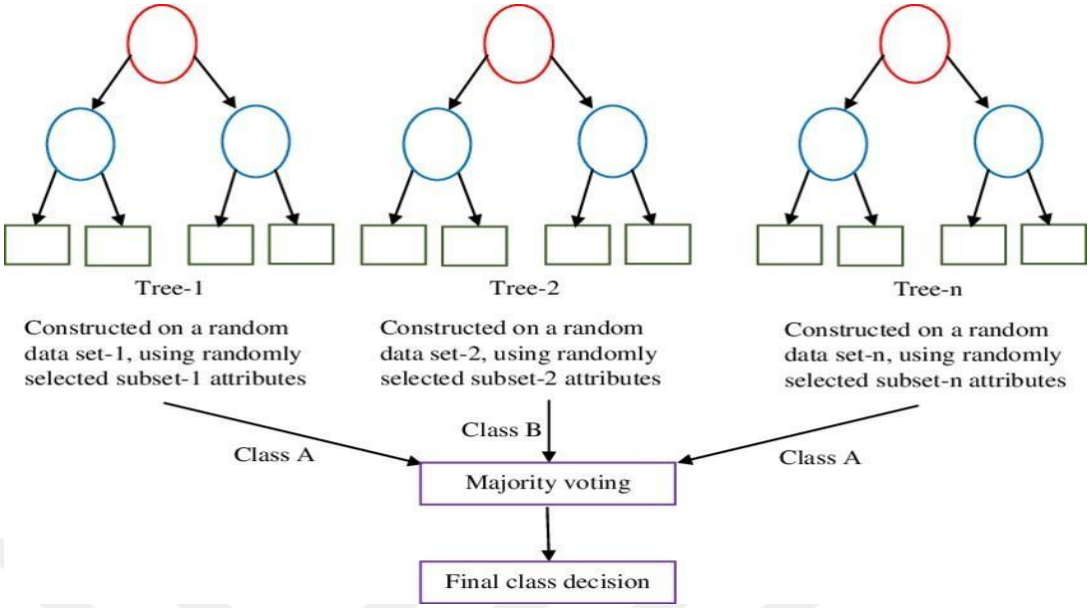

Figure 5.3 Random Forest Principle (Kiranmai & Ahuja, 2018)

In the case of regression, the final result is the average of all tree outputs, and it is bothsimpler and more powerful than other non-linear classification techniques. Data has beentrained with 20 trees, entropy as the loss function method. It's been tried with 5,10 and 20 folds.

5.4 Artificial Neural Network

Artificial neural networks are systems that can learn events through examples, derive and discover new information using the information they have learned, and thus respond to the effects of the environment similar to those of humans with the knowledge,experience and experiences they have gained (Haciefendioğlu, 2012).

Two hidden layers and an output layer make up the artificial neural network model.The neurons in the buried layers are 50 and 20 respectively. Rectified Linear Unit is selected for activation in these layers' functions. Since the problem is multiclass

classification, the output layer consists of 3 neurons and the activation function is chosenas softmax. The dataset is trained for 200 epochs.

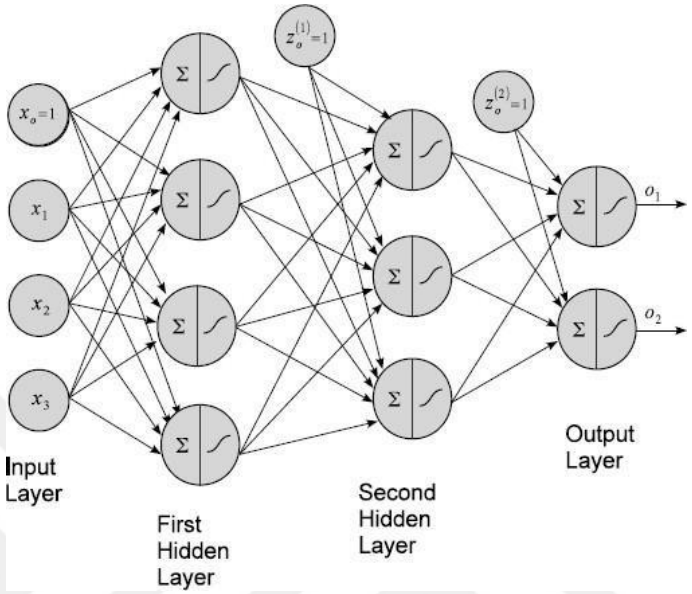

Figure 5.4 Neural Network with two hidden layers (Aggarwal, 2014)

5.5 K-Nearest Neighbors

The k-nearest neighbors (k-NN) algorithm is a data categorization method that uses thedata points closest to it to estimate the likelihood that a data point belongs to one of two categories. To solve classification and regression problems, the supervised machine learning algorithm k-nearest neighbor is applied. The k-NN algorithm assigns a class to anew observation and is one of the most extensively used classification algorithms in a range of disciplines (Güney & Atasoy, 2012).

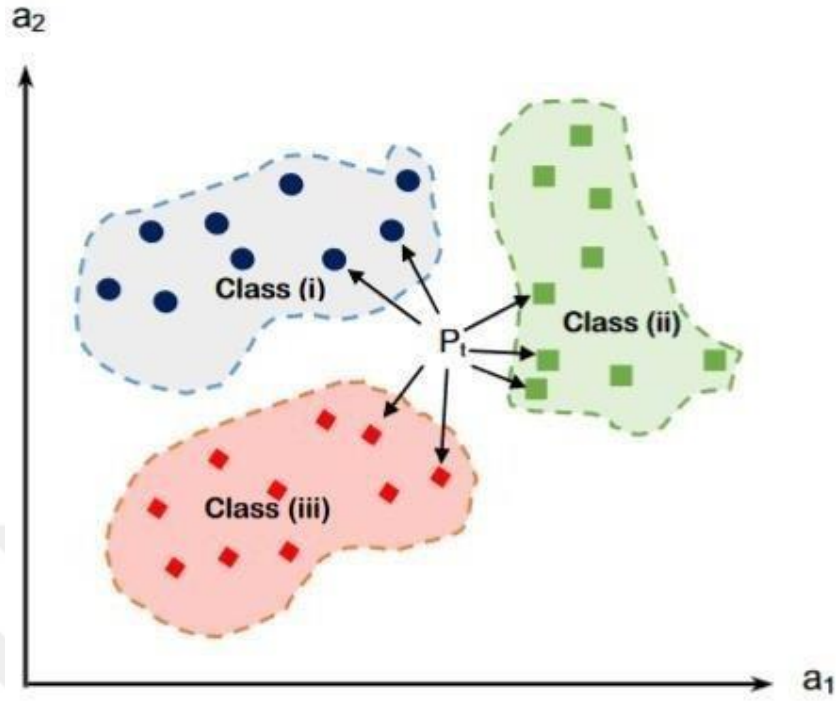

Figure 5.5 K-nearest Neighbor Classifier Example (Atallah, Badawy, & El-Sayed,2019)

There are four ways to calculate the distance measure between the data point and its nearest neighbor: Hamming distance, Manhattan distance, Euclidean distance, and Minkowski distance. Out of the three, Euclidean distance is the most commonly used distance function or metric. Which is used in this thesis too.

KNN doesn't work well if there are too many features. That is why, feature selection isa must for this problem. Because the dataset itself has high dimension. For this algorithmdataset reduced to 5 columns which were the most correlated with the target class.

KNN has been applied with k=2 to k=8, with step of 1. The result graph has been shown below (Figure 5.6). As you can see below since k increases the accuracy has dropped as well.

Graph also explains that the best k value is k=4 since it has the best accuracy on the test dataset. In this way we found the optimal k value for our model.

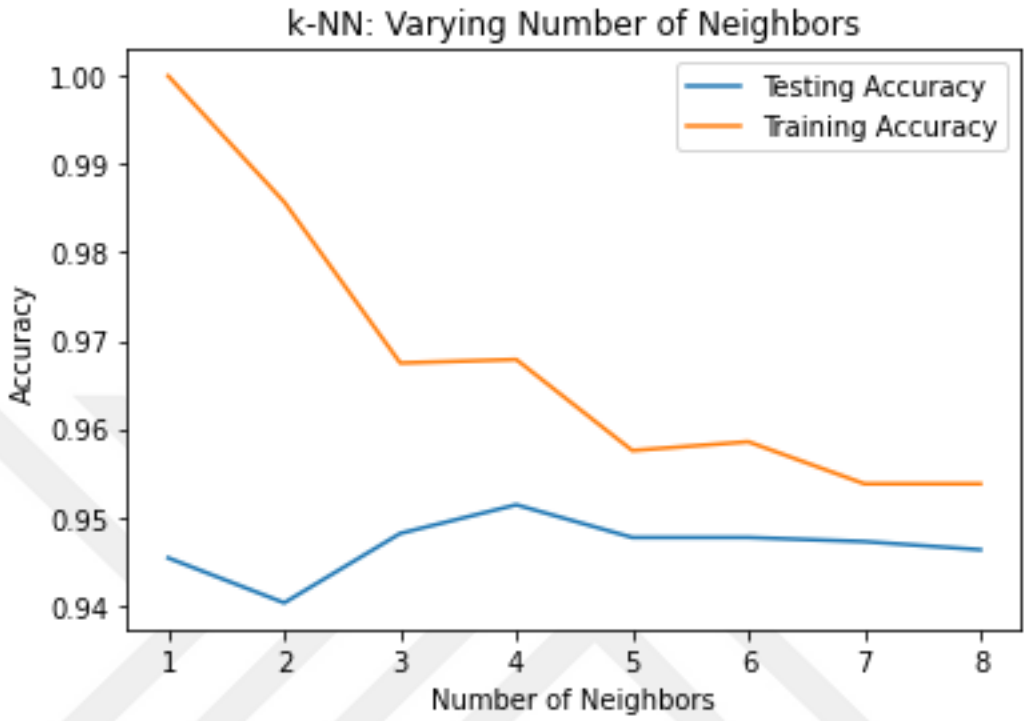

Figure 5.6 Accuracy graph of k-NN

CHAPTER 6
RESULTS AND METRICS

With the best result of accuracy is the ANN, SVM and K-nearest neighbors algorithms. K-nearest neighbors and random forest algorithms highly effected by the undersampled data in a negative way. Especially for random forest tree, there must have been more suitable variables since the categorical variables are imbalanced highly. The fact that these algorithms are widely used for multi-class classification, there has been many tune and parameters in these algorithms. Finding the best one depends on experienceand tests on various data.

The "best" choice really depends on the underlying problem and what you are willing to tolerate more, false positives (FP) or false negatives (FN). If it less concerned about producing FN predictions and want to minimize FP predictions, it would probably focus more on achieving high precision scores (compare with the formula). For the opposite case, it would focus more on recall respectively. If it cannot afford such a trade-off between these two, the F1-score is a good measure because it is the harmonic mean ofprecision and recall.

Accuracy calculated by sum of True Positives and True Negatives divided by sum of True Positive, False Positive, False Negative and True Negative values (6.1).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6.1}$$

Results table with algorithms we implemented is below (Table 6.1).

Table 6.1 Result Table by Accuracy

Algorithm	**With oversampling**	**With undersampling**	**With feature extraction**
SVM	% 89	% 70	% 94
Multinomial Logistic Regression	% 85	% 68	% 88
ANN	% 90	% 86	% 98
Random Forest	% 62	% 68	% 79
K-nearest neighbors	% 71	% 85	% 95

CHAPTER 7
CONCLUSION

For most of healthcare diagnosis problem, it is important that the data is reliable and has enough dimension for the needed model. Thyroid disease can be diagnosed with specific hormone tests, however being a multiclass classification problem making this diagnose with machine learning algorithms tricky since it is challenging to find which algorithms or techniques will be used. For this purpose, this paper had compared the techniques and algorithms for multi-class classification problem.

The most challenging part was lack of dimension in dataset. It is a highly imbalanced and not very big. This made the algorithms overfit easily hence it's been applied some techniques to prevent that.

To overcome to that challenge, we aimed to select the most possible contribute to our prediction class. We've achieved this by finding correlation between variables, extracting mutual information between target class and our variables, and use over/under sampling to gather maximum efficiency. We've been tried these in multiple ways such as with or without over/undersampling or by not normalize the values. All our predictions calculatedwith the best approach by these methods.

Results are showing that when there is an imbalanced data, undersampling is not alwaysa good technique with such small dataset. For same reason, it was not very effective to apply oversampling as well. Since training and test datasets are imbalanced, the best way to prevent this was using oversampling and feature importance together, thus it would reduce the overfit possibility. For prevent this we excluded the non-important features aswell as detect the important and not important features. Specific algorithms needed scalingsince their structure needs such dataset. Overall, the best way to diagnoses thyroid diseasewas artificial neural networks, however SVM and K-nearest neighbors performed well enough with over 95% accuracy. These algorithms are very effective on multi-classclassification if they tuned well.

For future work, dataset can be expanded vertically, it is important to make sample wider in machine learning. Visual tests also can be included, such as computerized tomography and ultrasound image of gland. With the information excluded from the images it’ll help the outcome along with numeric or categorical variables. Visual classifications are also for detecting malignant thyroid nodule as autoimmune, micro- nodular, nodular, and normal, which could help diagnosis of cancer.

REFERENCES

Achard, S., Pham, D. T., & Jutten, C. (2005). Criteria based on mutual information minimization for blind source separation in post nonlinearmixtures. *Signal Processing*, *85*(5), 965-974.

Adak, M.F., & Yumusak, N. (2016). Diagnosis of Thyroid by Hybrid Particle Swarm Optimization with Artificial Neural Network. *4th International Symposium on Innovative Technologies in Engineering and Science, Antalya, Turkey,* 25-30.

Aggarwal, C. (2014). Support Vector Machines. *In Data Classification: Algorithms and Applications (1st ed.)* (223-225). Boca Raton: Chapman & Hall/CRC.

Al-muwaffaq, I., & Bozkus, Z. (2016). MLTDD : Use of Machine Learning Techniques for Diagnosis of Thyroid Gland Disorder. *The Fourth International Conference on Database and Data Mining, Dubai, UAE, 10-78.*

Atallah, D.M., Badawy, M. & El-Sayed, A. (2019). Intelligent feature selection with modified K- nearest neighbor for kidney transplantation prediction. *SN Applied Sciences 1, 1297.*

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees* (1st Edition). Boca Raton: Routledge.

Broughton, C., & Ahmad, B. (2019). Thyroid Anatomy and Physiology. In *Advanced Practice in Endocrinology Nursing* (pp. 497-503). Cham:Springer.

Chaubey, Bisen, Arjaria, & Yadav (2021). Thyroid disease prediction usingmachine learning approaches. *National Academy Science Letters, 44*(3),233-238.

Chen, H. L., Yang, B., Wang, G., Liu, J., Chen, Y. D., & Liu, D. Y. (2012). A three-stage expert system based on support vector machines for thyroiddisease diagnosis. *Journal Of medical Systems, 36*(3), 1953-1963.

Cleveland Clinic, (2020). *Thyroid disease: Causes, symptoms, risk factors,testing & treatment*. Retrieved May 25, 2021, from https://my.clevelandclinic.org/health/diseases/8541-thyroid-disease.

Debajyoti Saha, (2021). *Multinomial Logistic Regression in Python*. Retrieved September 2, 2021 from https://www.codespeedy.com/multinomial-logistic-regression-in-python.

Dey, T. K., Giesen, J., Goswami, S., Hudson, J., Wenger, R., & Zhao, W.(2001). Undersampling and oversampling in sample-based shape modeling. In *Proceedings Visualization*, Ohio: IEEE, 83-545.

Ertekin, Ş. (2013). Adaptive Oversampling For Imbalanced Data Classification In *Information Sciences And Systems.* (261-269). Cham: Springer.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and featureselection. *Journal Of Machine Learning Research, 3*, 1157-1182.

Güney, S., & Atasoy, A. (2012). Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm and support vector machine in an electronic nose. *Sensors and Actuators B: Chemical, 166- 167*, 721–725. doi:10.1016/j.snb.2012.03.047

Haciefendioğlu, Ş. (2012). *Makine öğrenmesi yöntemleri ile glokom hastalığının teşhisi.* Master's thesis, Selçuk Üniversitesi, Konya.

Hershman J. (2020). *Overview of the Thyroid Gland.* Retrieved May 25, 2021, from https://www.msdmanuals.com/home/hormonal-and-metabolic-%20disorders/thyroid-gland-%20disorders/overview-of-the-thyroid-gland

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions On Neural Networks*, *13*(2), 415-425.

Ionita, Irina & Ioniță, Liviu. (2016). Applying Data Mining Techniques in Healthcare. *Studies in Informatics and Control. 25*. 385-394.

Jonklaas, J., Bianco, A. C., Bauer, A. J., Burman, K. D., Cappola, A. R., Celi, F. S. & Sawka, A. M. (2014). Guidelines For The Treatment Of Hypothyroidism: Prepared By The American Thyroid Association Task Force On Thyroid Hormone Replacement. *thyroid*, *24*(12),1670-1751.

Kiranmai, S. & Ahuja, Laxmi. (2018). Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. Protection and Control of Modern Power Systems. *Protection and Control of Modern Power Systems, 3,29.*

Kousarrizi, M. N., Seiti, F., & Teshnehlab, M. (2012). An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification. *International Journal of Electrical & Computer Sciences IJECS-IJENS*, *12*(01), 13-20.

Korhan, N. (2016). *Diagnosis Of Thyroid Disease Via Support Vector Machines.* Doctoral dissertation, Istanbul Technical University, Istanbul.

Matthew Hoffman, (2021). *Human Anatomy*. Retrieved August 11, 2021, from https://www.webmd.com/women/picture-of-the-thyroid.

Mircia, E., Imre, S., & Balaci, T. (2010*). Broad Research in Artificial Intelligence and Neuroscience,* (Vol. 1), Bacau: EduSoft.

Murat Gültekin, Güledal Boztaş (2014). T.C Sağlık Bakanlığı Türkiye Halk Sağlığı KurumuTürkiye Kanser İstatistikleri.

Patil, N., Rehman, A., & Jialal, I. (2021). *Hypothyroidism*. Retrieved August31, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK519536/.

Pearce, E. N., Farwell, A. P., & Braverman, L. E. (2003). Thyroiditis in *New England Journal of Medicine*, *348*(26*),* 2646-2655.

Sultana, J., & Jilani, A. K. (2018). *Predicting Breast Cancer Using Logistic Regression And Multi-Class Classifiers In International Journal Of Engineering & Technology*, 7(4.20), 22- 26.

Titapiccolo JI, Ferrario M, Cerutti S, Barbieri C, Mari F, Gatti E, Signorini MG. (2013), Artificial Intelligence Models ToStratify Cardiovascular Risk In Incident Hemodialysis Patients. *Expert Systems with Applications, 40*(11): 4679-4686.

Quinlan R. (1987). *UCI Machine Learning Repository*. Retrieved June, 2020, from https://archive.ics.uci.edu/ml/datasets/thyroid+disease.

Wei, G., Zhao, J., Feng, Y., He, A., & Yu, J. (2020). A novel hybrid feature selection method based on dynamic feature importance. *Applied Soft Computing, 93*, 106337.

Yıldırımkaya, M., Özata, M., Yılmaz, K., Kılınç, C., Gündoğan, M. A., & Kutluay, T. (1996). Lipoprotein (a) Concentration in Subclinical Hypothyroidism Before and After Levo- Thyroxine Therapy. *Endocrinejournal, 43(6),* 731-736.