T.R.
VAN YUZUNCU YIL UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

COMBINATION OF PCA WITH SMOTE OVERSAMPLING FOR CLASSIFICATION OF HIGH-DIMENSIONAL IMBALANCED DATA

M.Sc. THESIS

PREPARED BY: Guhdar Abdul-Aziz Ahmed MULLA
SUPERVISOR : Asst. Prof. Dr. Yıldırım DEMİR
CO-SUPERVISOR: Dr. Masoud Muhammed HASSAN

VAN-2021

T.R.
VAN YUZUNCU YIL UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

COMBINATION OF PCA WITH SMOTE OVERSAMPLING FOR CLASSIFICATION OF HIGH-DIMENSIONAL IMBALANCED DATA

M.Sc. THESIS

PREPARED BY: Guhdar Abdul-Aziz Ahmed MULLA

VAN-2021

ACCEPTANCE AND APPROVAL PAGE

This thesis entitled "**Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data**" presented by Guhdar Abdul-Aziz Ahmed MULLA under supervision of Asst. Prof. Dr. Yıldırım DEMİR in the department of statistics has been accepted as a M. Sc. thesis according to Legislations of Graduate Higher Education on 18/01/2020 with unanimity of votes members of jury.

Chair: Prof. Dr. H. Eray ÇELİK	Signature:
Member: Assoc. Prof. Dr. Hamit MİRTAGİOĞLU	Signature:
Member: Asst. Prof. Dr. Yıldırım DEMİR	Signature:

This thesis has been approved by the committee of The Institute of Natural and Applied Science on ..……./...../…..….. with decision number .....

Signature
……..……..
Prof. Dr. Suat ŞENSOY

THESIS STATEMENT

All information presented in the thesis obtained in the frame of ethical behavior and academic rules. In addition, all kinds of information that does not belong to me have been cited appropriately in the thesis prepared by the thesis writing rules.

Signature

Guhdar Abdul-Aziz Ahmed MULLA

ABSTRACT

COMBINATION OF PCA WITH SMOTE OVERSAMPLING FOR CLASSIFICATION OF HIGH-DIMENSIONAL IMBALANCED DATA

MULLA, Guhdar Abdul-Aziz Ahmed
M. Sc. Thesis, Statistics Department
Supervisor: Asst. Prof. Dr. Yıldırım DEMİR
Co-supervisor: Dr. Masoud Muhammed Hassan
January 2021, 71 pages

Imbalanced data classification is a common issue in data mining where the classifiers are skewed towards the larger data class. Classification of high-dimensional skewed (imbalanced) data is of great interest to decision makers as it is more difficult to. Dimension reduction method, which is a process in which variables are reduced, allows high dimensional datasets to be interpreted more easily with a certain loss. Furthermore, classification of high dimensional imbalanced data has become a recurring problem. In this study, a new method combining SMOTE oversampling with Principal Component Analysis (PCA) is proposed to solve the imbalance problem in high dimensional data. Six classification algorithms consisting of Decision Tree (DT), Support Vector Machines (SVM), K-Nearest Neighbor Method (K-NN), Naive Bayes (NB), Logistic Regression (LR) and Artificial Neural Networks (ANN) and six different datasets were used to check the efficiency of the proposed method and determine the performance of the classifiers. Respectively, raw datasets, converted datasets by PCA, SMOTE and SMOTE+PCA (SMOTE and PCA) methods, were analyzed with the given algorithms. Analyzes were made using WEKA.

Analysis results suggest that almost all classification algorithms improve their classification performance by using PCA, SOMTE and SMOTE+PCA methods. However, the SMOTE method gave more efficient results than PCA and PCA+SMOTE methods for data rebalancing. Experimental results also suggest that SVM and K-NN classifiers provided higher classification performance compared to other algorithms.

Keywords: Classification, Dimensionality reduction, Imbalanced classes, Machine learning, PCA, SMOTE oversampling.

ÖZET

YÜKSEK BOYUTLU DENGESİZ VERİLERİN SINIFLANDIRILMASI İÇİN SMOTE AŞIRI ÖRNEKLEME İLE PCA'NIN KOMBİNASYONU

MULLA, Guhdar Abdul-Aziz Ahmed
Yüksek Lisans Tezi, İstatistik Bölümü
Tez Danışmanı: Dr. Öğr. Üyesi Yıldırım DEMİR
İkinci Danışman: Dr. Masoud Muhammed Hassan
Ocak 2021, 71 sayfa

Dengesiz verilerin sınıflandırması, sınıflandırıcıların daha büyük veri sınıfına doğru çarpıtıldığı veri madenciliğinde yaygın bir konudur. Yüksek boyutlu çarpık (dengesiz) verilerin sınıflandırılması, daha zor olduğundan karar vericiler için büyük ilgi görmektedir. Değişkenlerin azaltıldığı bir süreç olan boyut indirgeme yöntemi, yüksek boyutlu veri setlerinin belirli bir kayıpla daha kolay yorumlanmasını sağlamaktadır. Ayrıca, yüksek boyutlu dengesiz verilerin sınıflandırılması tekrarlanan bir sorun haline gelmiştir. Bu çalışmada, yüksek boyutlu verilerde dengesizlik problemini çözmek için SMOTE aşırı örnekleme ile Temel Bileşen Analizini (PCA) birleştiren yeni bir yöntem önerilmiştir. Önerilen yöntemin etkinliğini kontrol etmek ve sınıflandırıcıların performansını belirlemek için Karar Ağacı (DT), Destek Vektör Makineleri (SVM), En Yakın Komşu (K-NN), Naive Bayes (NB), Lojistik Regresyon (LR) ve Yapay Sinir Ağlarından (ANN) oluşan altı sınıflandırma algoritması ve altı farklı veri kümesi kullanılmıştır. Sırasıyla, ham veri setleri, PCA, SMOTE ve SMOTE+PCA yöntemleriyle dönüştürülen veri setleri verilen algoritmalarla analiz edilmiştir. Analizler WEKA programlama dillerinden yararlanılarak yapılmıştır.

Analiz sonuçları, neredeyse tüm sınıflandırma algoritmalarının PCA, SOMTE ve SMOTE+PCA yöntemlerini kullanarak sınıflandırma performanslarını iyileştirdiğini göstermektedir. Bununla birlikte, SMOTE yöntemi, verilerin yeniden dengelenmesi için PCA ve PCA+SMOTE yöntemlerinden daha etkili sonuçlar vermiştir. Ayrıca deneysel sonuçlar, SVM ve K-NN sınıflandırıcılarının diğer algoritmalara kıyasla daha yüksek sınıflandırma performansı sağladığını göstermektedir.

Anahtar kelimeler: Boyut indirgeme, Dengesiz sınıflar, Makine öğrenimi, PCA, SMOTE Aşırı örnekleme, Sınıflandırma.

ACKNOWLEDGMENT

First of all, I would like to express my greatest and hearty thanks to God for giving me strength and credibility to complete this research. I would like to express my sincere gratitude to my supervisor Dr. Öğr. Üyesi Yıldırım DEMİR and co-supervisor Dr. Masoud Muhammed Hassan for their continuous support for my MSc study and research, for their patience, motivation, enthusiasm and immense knowledge. Their guidance helped me throughout the time of research and the writing of this thesis. Special thanks to my family. Truly, words cannot describe how thankful and grateful I am to my parents. I would also like to express my deepest gratitude to Mr. Amar Yahya and Mr. Sana Salih for her support, encouragement, and patience. Finally, my thanks go to my friends and every person who helped me during this research journey.

2021

Guhdar Abdul-Aziz Ahmed MULLA

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

SYMBOLS AND ABBREVIATIONS

Some symbols and abbreviations used in this study are presented below, along with descriptions.

Symbols	Description
M	Mean of Variables
E	Expectation Value
$\sum$	Covariance Matrix
$\boldsymbol{\sigma^2}$	Variance
σ	Standard Deviation
λ	Eigenvalue
a	Eigenvector
S	Sample Covariance Matrix

Abbreviations	Description
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
PCA	Principal Component Analysis
ICA	Independent Component Analysis
LLE	Locally Linear Embedding
NB	Naïve Bayes
LR	Logistic Regression
ANN	Artificial Neural Network
SVM	Support Vector Machine
K-NN	K-Nearest Neighbor
DT	Decision Tree

Abbreviations	Description
MI	Mutual Information
FS	Fisher Score
EC	Eigenvector Centrality
CFS	Score for Correlation
RF	Random Forest

1. INTRODUCTION

One of the sciences developed in order to examine the phenomena in nature and to solve existing or potential problems is data analysis science. The aim of data analysis science is to explain the subject with a certain probability by using methods and theories suitable for data structure with a limited number of observations and to shed light on future researches. There is a lot of data in every science field in nature and access to these data is getting easier day by day. However, the question of how much of the obtained data can be used or how important it always comes to the fore. In addition to obtaining information, mankind beings consume data even in their daily work and produce more than they consume.

Due to the intensive use of data, it is necessary to classify the data in order to make it useful and generate new information. Classification is the grouping of a product or data according to determined distinguishing features through algorithms. It is often impossible to manually classify data. Because millions of data types are formed related with a field even in just one day. However, the very functional operation of the algorithms has made it easier to classify the data. At the same time, subjecting a data to many algorithms to classify it can produce different results. The effective classification depends on which algorithm will be applied to the data set. Thus, it is likely that not every algorithm will give the same accuracy to all data sets and that the algorithms used to model the data are too low over the data set. In this direction, machine learning and classification algorithms come to the fore (Baran, 2020).

Data mining is the process of extracting useful information and patterns from data (Gundecha and Liu, 2012). In addition, data mining is defined as a process of discovering useful or actionable knowledge from large scale data (Osisanwo et al., 2017). Variety of data mining techniques and algorithms has been proposed to deal with a number of problems occurring in different fields of science. Classification is one of these algorithms suggested in data mining and is widely used in different areas of research. With the classification, it is aimed to build models that can predict the class of new cases. These data mining models are mathematical expressions that are used to understand and studied

more easily of data (Santos et al., 2006). Data mining and machine learning techniques are used to extract and discover interesting, inexplicable, hidden features from large amounts of data. For example, in the medical field, data mining methods have been commonly used for pattern identification, grouping, clustering, and prediction, and these are the principal constitute of the application of these methods to other science fields (Swapna et al., 2018).

Classification uses a set of pre-classified examples to create a model that can describe the population used in large datasets. Diagnosing diseases, identification of spam emails, fraud detection and credit risk applications are as the examples of such analysis. Training and test data are needed for data classification method. The training data is analyzed by the classification algorithm to learn about the unknown parameters of the model. On the other hand, testing data are used to check the performance of the classification model on unseen data, and hence the accuracy of the classification rules is calculated. If the accuracy is appropriate, the rules for new data tuples can be applied to make decision. In order to determine the set of parameters required for proper discrimination, the classifier-training algorithm uses these pre-classified examples. Then the algorithm encodes the parameters into a model called a classifier (Baitharu et al., 2015). There are several different algorithms for classification that exist in the literature. Decision Tree (DT), Support Vector Machines (SVM), K-Nearest Neighbor Method (K-NN), Naive Bayes (NB), Logistic Regression (LR) and Artificial Neural Networks (ANN) are the most important and the most commonly machine learning algorithms for classification process. Many of these advanced classification algorithms typically give a high classification accuracy with empirical training data and new test data. In view of such a high accuracy trend, it can be difficult for model developers to devise a better classification method. In addition, the provision of such high predictive precision may mean that machine learning techniques can correctly solve almost any classification problem. However, in solving every problem, such high prediction accuracy cannot be seen (Lorena and Lehmann, 2019). Therefore, investigating the issues exist in the classification algorithms, such as imbalance class problem, is necessary to improve the classification performance.

One of the most difficult issues in the classification algorithms is the classification of imbalanced data. Because a problem occurs in binary classification when the sample sizes is not equal in both classes. In other words, one class has many samples called majority,

while the other class has relatively few samples called minority. However, this problem may not very important a problem if there is very little difference between samples from the positive and negative groups. In addition, when data are imbalanced, the majority class usually takes into account the key features of interest to learn from while ignoring the impact of the minority in the dataset. Nonetheless, most conventional issue cannot be solved by classification algorithms, as while they are designed to achieve high overall precision, they are most likely to misclassify positive class samples, and this is a disadvantageous of classification under imbalanced data. To find a good answer with good precision for both the positive and negative groups has become a significant research area (Tahir et al., 2019). In order to classify imbalanced data, different oversampling or under sampling should be used first. Synthetic Minority Oversampling Techniques (SMOTE) is a method used to rebalance the dataset. This method offers an optimal solution for oversampling based imbalanced data distribution problem. The basic SMOTE assumption is based on how to find parallels between the features of the minority and majority classes (Mustafa et al., 2017).

Another issue with classification is the high dimensionality of data, where there are lots of redundant features in the dataset. The goal of using dimension reduction methods is to reduce the irrelevant features from the data before applying any classification algorithms. The objective is to obtain a compact, precise, data representation that decreases or removes statistically redundant information ingredients. The reduction of measurements is central to an array of data processing objectives. Input selection in classification and regression problems is a mission-specific dimension reduction form. High-dimensional data visualization involves mapping to a lower dimension, usually three or less. Principal Component Analysis (PCA) is a very-well known classic method for linear dimension reduction. One performs an orthogonal transformation for the correlation of vectors and projects spanning the subspace corresponding to the highest eigenvalues by such eigenvectors in PCA. This conversion makes the components of the signal unrelated, and the projection along the high-variance directions maximizes variance while minimizes the mean squared residual between the initial signal and its dimension reduction approach (Kambhatla and Leen, 1997).

In this research, it is aimed to solve the problem of classification for high-dimensional imbalanced data. In the study, the classification problem is investigated (using the PCA and SMOTE methods) by reducing the number of unnecessary features with

rebalancing simultaneously. In accordance with this purpose, six well-known classification algorithms (DT, SVM, K-NN, NB, LR and ANN) were used for different imbalanced datasets where the number of sample in one class (majority) is substantially higher than the number of sample in the other class (minority). In the extremely imbalanced data sets, the amount of data in the minority classes according to the other majority classes is insufficient to obtain adequate information. Weka program, which is a tried/tested open-source data mining software that can be accessed via a graphical user interface, and it is standard terminal applications, was used in data analysis. Thanks to this program, the data have classified by rebalancing and reducing its dimension. The WEKA project aims to provide an extensive set of algorithms for machine learning and data preprocessing instruments for researchers. This supports applications to quickly test and compare various machine learning process methods on new datasets (Hall et al., 2009).

In addition to chapter one, the rest of the thesis is organized as follows. Chapter two is literature review. Chapter three gives a detailed description and theoretical background to the preprocessing, dimensionality reduction, rebalancing data, machines learning, classification algorithms and explains the proposed oversampling method in detail. Chapter four presents the findings of the experimental results carried out in the project. Chapter five presents the discussion and conclusion of the study and discusses further research that could be carried out in the future.

2. LITERATURE REVIEW

Classification is a significant pattern recognition activity. A variety of classification learning algorithms have been well developed and successfully applied to many application domains, including decision tree, backpropagation neural network, Bayesian network, nearest neighbor, support vector machines, and the newly documented associative classification. However, for most classifier learning algorithms, the imbalanced class distribution of a dataset has encountered a severe difficulty that implies a reasonably balanced distribution. The imbalanced knowledge is defined as having far more cases than those in certain groups. In rare cases, classification rules that predict small classes appear to be rare, undiscovered or ignored; thus, test samples belonging to small classes are more frequently misclassified than those belonging to the dominant classes. The proper classification of the samples in small groups also has a higher importance in such applications than in the opposite case (Sun et al., 2009). Below is a survey about some related works for dealing with the classification of imbalanced data in the past few years.

Luque et al. (2019), their approach goes beyond simply studying case studies and develops a systematic analysis of this impact by simulating the results obtained using binary classifiers. A set of functions and numerical indicators were attained which enables the comparison of the behavior of several performance metrics based on the binary confusion matrix when they faced with imbalanced datasets. In their study, they proposed a new way to measure the imbalance which surpasses the Imbalance Ratio used in previous studies. From the simulation results, several clusters of performance metrics had been identified that involve the use of Geometric Mean or Bookmaker Unforcedness as the best null-biased metrics if their focus on classification successes (dismissing the errors) presents no limitation for the specific application where they were used. From the simulation results, they quantitatively justified guide to select performance metrics in the presence of imbalance classes, and the accuracy of the model had been developed.

(Naseriparsa and Kashani, 2014), in this research they proposed a method for combination of unsupervised dimensionality reduction methods with resampling and the results were tested on Lung-Cancer dataset. In the first step, PCA was applied on Lung-

Cancer dataset to compact the dataset and eliminate irrelevant features and in second step SMOTE resampling is carried out to balance the class distribution and increase the variety of sample domain. Finally, Naïve Bayes classifier was applied on the resulting dataset and the results were compared and evaluation metrics were calculated. The experiments showed the effectiveness of the proposed method across four evaluation metrics: Overall accuracy, False Positive Rate, Precision, and Recall.

Maldonado et al. (2019), investigated the effects of high-dimensionality on SMOTE oversampling into a new distance metric dependent formalization only on the related variables of the problem. The qualifications via feature ranking methods, such as feature ranking methods, important variables are achieved. Mutual Information (MI) as the Fisher Score (FS), Eigenvector Centrality (EC) and the Score for Correlation (CFS) were used. They explored the effect of distance metrics that are more fitting for High-dimensional data, such as the Euclidean norm, and the distances from Manhattan and Chebyshev. In this study, a new extension for the oversampling of SMOTE is presented in binary classification to deal with the class-imbalance problem. Synthetic examples are provided by SMOTE by first defining the nearest neighbors from the minority community, k nearest neighbors within this class, the reference sample is then interpolated with a randomly chosen item from its neighborhood.

Buda et al. (2018), in this study they used three benchmark datasets of increasing complexity, MNIST, CIFAR-10 and ImageNet, to investigate the effects of imbalance on classification and perform an extensive comparison of several methods to address the issues: oversampling, under sampling, two-phase training, and thresholding that compensates for prior class probabilities. Their main evaluation metric was area under the receiver operating characteristic curve (ROC AUC) adjusted to multi-class tasks since overall accuracy metric is associated with notable difficulties in the context of imbalanced data. Based on results from their experiments, they concluded that (i) the effect of class imbalance on classification performance was detrimental; (ii) the method of addressing class imbalance that emerged as dominant in almost all analyzed scenarios was oversampling; (iii) oversampling should be applied to the level that completely eliminates the imbalance, whereas the optimal under sampling ratio depends on the extent of imbalance; (iv) as opposed to some classical machine learning models, oversampling did not cause overfitting; (v) thresholding should

be applied to compensate for prior class probabilities when overall number of properly classified cases was of interest.

Lorena and Lehmann (2019), in this paper, they reviewed resampling methods and analyzed measures which could be extracted from the training datasets in order to characterize the complexity of the respective classification problems. Their used methods in recent literature were also reviewed and discussed, allowing to prospect opportunities for future work in the area. Finally, descriptions were given on an R package named Extended Complexity Library (ECoL) that implements a set of complexity measures and was made publicly available.

Basgall et al. (2018), this study presents SMOTE-BD, fully scalable preprocessing approach for imbalanced classification in Big Data. It was based on one of the most widespread preprocessing solutions for imbalanced classification, namely the SMOTE algorithm, which creates new synthetic samples according to the neighborhood of each example of the minority class. Their novel development was made to be independent of the number of partitions or processes created to achieve a higher degree of efficiency. Experiments conducted on different standard and Big Data datasets showed the quality of the proposed design and implementation.

Abdoh et al. (2018), used cervical cancer risk factors to build classification model using Random Forest (RF) classification technique with the synthetic minority oversampling technique (SMOTE) and two feature reduction techniques: Recursive Feature Elimination and PCA. Most medical data sets used in their study were imbalanced because the number of patients was much less than the number of non-patients. Because of the imbalance of the used data set, SMOTE method was used to solve this problem. The data set consists of 32 risk factors and four target variables: Hinselmann, Schiller, Cytology, and Biopsy. After comparing the results, they found that the combination of the Random Forest classification technique with SMOTE improved the classification performance.

Mustafa et al. (2017), proposed a new and efficient combined algorithm based on FD_SMOTE (Farther Distance Based on Synthetic Minority Oversampling Techniques) and Principal Component Analysis (PCA), which successfully reduces the high dimensionality and balances the minority class. The proposed algorithm had been investigated on biomedical data and it given the desired results in terms of dimensionality and data balancing. In this study, the quality of dimensionality reduction and balanced data had been

evaluated using assessment metrics like covariance, Accuracy (ACC) and Area Under the Curve (AUC). It had been observed from the numerical results that the performance of the algorithm achieved the best accuracy with metrics of ACC and AUC.

Liu et al. (2020), focused on tackling class imbalance problems and finding variable correlations. Two Fraud Detection datasets on Kaggle used to build classifiers and analyze the impact of different data processing techniques. Through their process, they realized recent findings of fraud detection, they got to know more about different data processing methods, and they implemented distinct types of classifiers. They confirmed the significance of class imbalance tackling and variable correlations analyzing. Finally, they explored into correlations between each variable and improved performance a little. Their experimental results well proved that tackling the class imbalance approach had a positive effect on imbalanced data sets, and digging into the relationship among variables would also be useful.

Mohammed et al. (2020), proposed a method to analyze, diagnose, and classify imbalanced diabetes patients using six machine learning algorithms for a new real diabetes dataset. The newly created dataset, called ZADA, was obtained from medical records of about 7000 patients in Zakho city of Iraq. Their proposed classification analysis was based on the three normalization methods along with the resampling (SMOTE) method to tackle the class imbalance problem. Various experiments were conducted to find the best algorithm with the best performance according to the distribution of minority classes. Results showed that the resampling method and the normalization techniques had a positive effect on classification model performance.

Zheng (2020), investigated the use of the Synthetic Minority Oversampling Technique (SMOTE) with different classifiers. The SMOTE method was implemented for comparison: regular SMOTE, Borderline-SMOTE, SVM-SMOTE, K-Means SMOTE, which were combined with four classification algorithms under the classical and Neyman Pearson (NP) paradigms to build predictive models on the heart disease data, and the performance of these models was compared. Their results showed that the SVM-SMOTE and the Borderline-SMOTE outperform other SMOTE variants, and the NP classification was superior in controlling the type I error effectively.

Scott et al. (2019), proposed a new method called GAN-SMOTE, a novel approach to synthetic minority class oversampling used a generative adversarial network that can be applied to boost the performance of classifiers learning from small and imbalanced one-hot

encoded datasets. This study also introduced techniques that were key for ensuring the stability and variance of the generative adversarial network in this setting. The proposed method demonstrates meaningful improvement on a well-studied graphical dataset with significant class imbalance.

3. MATERIALS AND METHODS

In this chapter reviewed some existing well-known data mining and machine learning methods to deal with the class imbalance problem for high dimensional data. We briefly explain the machine learning, type of learnings, and classification algorithms. This chapter also reviews the background of the six well-known classification algorithms: Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT) and Artificial Neural Network (ANN). Furthermore, we briefly explain the problem of high-dimensional, dimensionality reduction, feature extraction, Principal Component Analysis (PCA) method, class imbalance problem, and the SMOTE resampling method for rebalancing data.

3.1. Materials

To investigate the efficiency and the reliability of the proposed methods, this study applied our methods on six different datasets. The number of datasets (six datasets) were selected based on the imbalanced percentage values between the negative class and positive class; the low values were 2.6%, 10% for the first and second datasets respectively. The medium values were 25.4%, 59.4%, 56.8% for the third, fourth and fifth datasets respectively. And the highest values 92% for the sixth dataset. Here, we briefly describe the main characters and information of the six datasets as follows.

3.1.1. Dataset 1: heart disease attack

This data set is from 1986 and it contains 76 variables, including the predicted variable, but all published experiments refer to using a subset of 14 of them and 1025 patient. The ‘‘target’’ field refers to the presence of heart disease in the patient. It is integer valued 0 = No disease and 1 = Yes disease. Table 3.1 describes the main characteristics of dataset 1. The dataset was taken from address https://www.kaggle.com/johnsmith88/heart-disease-dataset on 12.06.2020.

Table 3.1. General characteristics of dataset 1 (Heart Disease Attack)

Variable	Describe Variable
Age	Age of the patient in years
Sex	1 = male, 0 = female
Cp	Chest pain type
Trest bps	Resting blood pressure (in mm Hg on admission to the hospital)
Chol	Serum cholesterol in mg/di
Fbs	Fasting blood sugar and gt: 120 mg/di (1 = true, 0 = false)
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina (1 = yes, 0 = no)
Old peak	ST depression induced by exercise relative to rest

3.1.2. Dataset 2: heart disease

This is another version of Dataset 1. In this data set has 14 variable and 303 patients. Table 3.2 describes the main characteristics of dataset 2. The dataset was taken from address https://www.kaggle.com/ronitf/heart-disease-uci on 12.06.2020.

Table 3.2. General characteristics of dataset 2 (Heart Disease UCI)

Variable	Describe Variable
Age	The person's age in years
Sex	The person's sex (1 = male, 0 = female)
Cp	The chest pain experienced (value 1: typical angina, value 2: typical angina, value 3: non-angina pain, value 4: asymptomatic)
Trest bps	The person's resting pressure (mm Hg on admission to the hospital)
Chol	The person's cholesterol measurement in mg/di
Fbs	The person's fasting blood sugar (>120 mg/di, 1 = true, 0 = false)
Restecg	Resting electrocardiographic measurement (0 = normal, 1 = having (ST-T wave abnormality, 2 = showing probable of definite left ventricular hypertrophy by Estes criteria)
Thalach	The person's maximum heart rate achieved
Exang	Exercise induced angina (1 = yes, 0 = no)
Old peak	ST depression induced by exercise relative tortes
Slope	The slope of the peak exercise ST segment (value 1: up sloping, value 2: flat, value 3: down sloping)
Ca	The number of major vessels (0 – 3)
Thal	A blood disorder called thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect)
Target	Heart disease (0 = no, 1 = yes)

3.1.3. Dataset 3: breast cancer wisconsin (diagnostic)

Features in this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The dataset was taken from address https://www.kaggle.com/uciml/breast-cancer-wisconsin-data on 12.06.2020.

The main characters and information of this dataset is as follows.

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For sample, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Class distribution: 357 benign, 212 malignant.

3.1.4. Dataset 4: hepatitis domain

In this dataset, the hepatitis database and BILIRUBIN problem, we have the following features: BILIRUBIN is continuous variable (the number of its “values" in the ASDOHEPA.DAT file is negative); "values" are quoted because when speaking about the continuous variable there is no such thing as all possible values. However, it

represents which so called "boundary" values; according to these "boundary" values the variable can be discretized. At the same time, because of the continuous variable, one can perform some other test since the continuous information is preserved. Table 3.3 describes the main characteristics of dataset 3. The dataset obtained from address https://sci2s.ugr.es/keel/imbalanced.php#sub20 on 12.06.2020.

Table3.3. General characteristics of dataset 4 (Hepatitis Domain)

Variable	Describe Variable
Class	Die, Live
Age	10, 20, 30, 40, 50, 60, 70, 80
Sex	male, female
Steroid	no, yes
Antivirals	no, yes
Fatigue	no, yes
Malaise	no, yes
Anorexia	no, yes
Liver Big	no, yes
Liver Firm	no, yes
Spleen Palpable	no, yes
Spiders	no, yes
Ascites	no, yes
Varices	no, yes
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alk Phosphate	33, 80, 120, 160, 200, 250
Sgot	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	no, yes

3.1.5. Dataset 5: cardiology categorical

In this data set has 14 variables and 232 patients (records), in which 50 of them were sick and 182 of them were healthy. Table 3.4 describes the main characteristics of dataset 5. This dataset obtained from address https://sci2s.ugr.es/keel/imbalanced.php#sub20 on 12.06.2020.

Table 3.4. General characteristics of dataset 5 (Cardiology Categorical)

Variable	Describe Variable
Age	29 - 76
Sex	male, female
Chest pain type	86, 45, 19, 82
Blood pressure	94 - 180
Cholesterol	126 - 564
Fasting blood sugar < 120	false - true
Resting ecg	hybe, normal, abnormal
Maximim heart rate	88 - 202
Angina	false - true
Peak	0 - 6.2
Slope	flat, up, down
Colored vessels	0 - 3
Thal	rev, normal, fix
Class	sick, healthy

3.1.6. Dataset 6: Lymphography

Table 3.5 describes the main characteristics of the dataset 6 (Lymphography-normal-fibrosis) and its variables. This dataset obtained from address https://sci2s.ugr.es/keel/imbalanced.php#sub20 on 12.06.2020.

Table 3.5. General characteristics of dataset 6 (Lymphography)

Variable	Describe Variable
Lymphatics	normal, arched, deformed, displaced
Block_of_affere	no, yes
Bl_of_lymph_c	no, yes
Bl_of_lymph_s	no, yes
By_pass	no, yes
Extravasates	no, yes
Regeneration_of	no, yes
Early_uptake_in	no, yes
Lym_nodes_dimin	[0.0, 3.0]
Lym_nodes_enlar	[1.0, 4.0]
Changes_in_lym	bean, oval, round
Defect_in_node	no, lacunar, lac_margin, lac_central
Changes_in_node	no, lacunar, lac_margin, lac_central
Changes_in_stru	no, grainy, drop_like, coarse, diluted, reticular, stripped, faint
Special_forms	no, chalices, vesicles
Dislocation_of	no, yes
Exclusion_of_no	no, yes
No_of_nodes_in	[1.0, 8.0]
Class	positive, negative

3.2. Machine Learning

Machine learning is about discovering and designing mathematical models and data-learning algorithms. There are two main types of learning in machine learning algorithms, supervised and unsupervised. Supervised learning paradigm focuses on classification goals and consists of modeling optimum mapping between the data domain and the information set, and designing learning algorithms. The classification requires a set of (labeled) training data, a set of validation data, and a set of test data. It will address the meanings and functions of certain data sets. However, to explain briefly, the training data set helps to find the optimal parameters of a model, the validation data set helps prevent model overfitting and the test data set helps detecting and predicting (Suthaharan, 2016).

Machine learning is the science of having computers to learn and behave like humans do, by feeding them data and knowledge in the form of observations and real-world experiences, and developing their learning over time in an autonomous fashion. “The practice of using algorithms to parse data, learn from it, and then make a decision or prediction about something in the world is machine learning at its simplest”. Depending on what task you are trying to accomplish and the type and amount of data that you have available, there are many algorithms for getting machines to learn, from using simple decision trees to clustering to artificial neural network layers (Chetana et al., 2020). The two well-known types of learning are briefly explained as follows:

1. Supervised Machine Learning.
2. Unsupervised Machine Learning.

3.2. 1. Supervised machine learning

In supervised machine learning, it is important to mark the inputs of the systems to extract information from those inputs. In the training cycle the relationships between the inputs and the labels given to them are investigated. Classification is one of the most widely used supervised data mining techniques, where the label is the class to which this record belongs for each record. The classifiers extract the relationships between the value of the variables that define that record, and the class that is classified as a member of that record. This information is then extended to new documents not listed for the purpose of

determining a class for them. This prediction can help to estimate the future actions of that new record, depending on the overall characteristics of the records of that class (Suthaharan, 2014).

Classifier needs labeled data set for the extraction of knowledge, so that this dataset is used to train the classifier. The classified dataset is split in two sections to provide further steps. The first is used for the training phase, and the second is used for classifier testing.

3.2.2. Unsupervised machine learning

This type of machine learning did not require the records in the dataset to be labeled because these algorithms tend to find the relation between the records themselves, depending on the values that characterize each record. Clustering is one of the most common unsupervised data mining techniques where records are distributed, based on their variable values, into groups in the dataset. Each record in these groups is more similar to the other records within that category than the other records in the other groups. Clustering thus creates homogeneous register group (Liu and D'Aquin, 2017).

The term modeling refers to data processing both in mathematics and statistics. Modeling is aimed at creating a parametrized mapping between the data domain and response collection. This mapping may be a parameterized function or a parameterized process that learns device characteristics from the (labeled) input data. In the context of machine learning the term algorithm is a confusing term. The algorithm's word, for a computer scientist, means systematic step-by-step instructions for a machine to solve a problem. The modeling may have many algorithms in machine learning to drive a program, but the term algorithm here refers to a learning algorithm. The learning algorithm is used to train, validate, and test the model using a given data set to find an optimal value for the parameters, validate it, and evaluate its performance (Suthaharan, 2016). Figure 3.1 shows the general diagram of machine learning modeling.

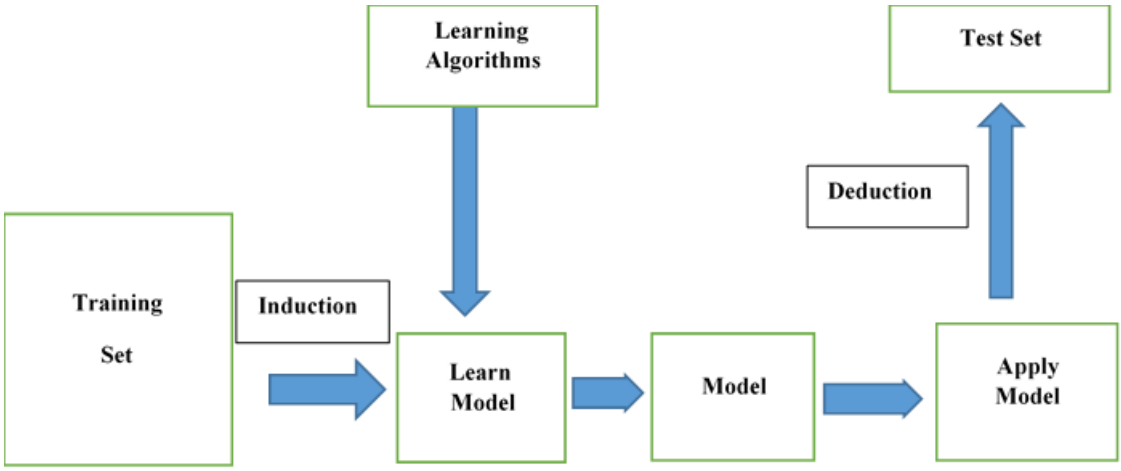

Figure 3.1. General diagram of machine learning modeling.

3.3. Classification

Classification is a process of determining classes of certain objects on the basis of their characteristics, where semantic classes are previously known (Mythili et al., 2014). It is of assigning new observed samples by examining the features of the sample to an existing defined class. It can then make decisions for the unseen cases by constructing a model from previous data (Iyer and Sumbaly, 2015). Therefore, it is the finding a model (or function) that describes and distinguishes data classes or concepts. It uses data mining techniques to investigate the relationship between the values of the variables for each row and the label given to that row. There are several different classification algorithms, such as NB, LR, SVM, K-NN, DT, and ANN. These algorithms use different representations of these relationships, so that when new rows are fed to the classifier, the extracted knowledge can be applied to predict a label for that row, depending on the variable values that characterize the row. These relationships differ from one dataset to another, so it is critical that the classifiers are trained using a labeled training dataset (Agrawal and Agrawal, 2015). Classification consists of predicting, based on a given data, a certain result. With a view to predicting the consequence is that the algorithm processes a training set containing a set of variables and their respective characteristics. Outcome, usually referred to as an variable of target or prediction. The algorithm attempts to uncover relationships between the variables that would allow the result to be predicted. Next is the algorithm: Provided a data set not previously seen, called a prediction system, containing the same set of variables, Except for

the variable of prediction-not yet understood. The algorithm analyzes and generates the input and a forecast. "The precision of the prediction determines how" good" the algorithm is. In, for example, a Medical Database The training set must have previously reported relevant patient information; Whether or not the patient has a heart attack is the prediction trait. The general concept behind the Data Mining classification is to predict the study of the target class dataset for training (Haghanikhameneh et al., 2012). We briefly explain these six classification algorithms as follows.

3.3.1. Naïve bayes

A Naive Bayes (NB) classifier is a simple probabilistic classifier based on the application of the Bayes theorem with strict (naive) independence assumptions (from Bayesian statistics). The benefit of the naive Bayes classifier is that only a small amount of training data is needed to estimate the parameters required for classification (Kaur and Oberai, 2014).

We will briefly explain in this section how the NB technique works mathematically and logically. Due to probabilistic but still strong precision, the Naïve Bayes classifier has found its way into many applications today (Vidhya and Aghila, 2010; Maertens et al., 2017). Conditional probabilities are calculated based on a mathematical theory is a Bayesian classifier. For example, in text mining classification by Naïve Bayes classifier, the presence or absence of a word determines the effect of the prediction in the textual document. In other words, each word that is processed is given a probability of belonging to a certain group. The likelihood is determined from the occurrences in the training manual of the word where the categories are already defined. When all these probabilities are determined, it is possible to define a new document by the number of probabilities for any category of each word that exists within the document. They are referred to as "naive" since the algorithm assumes that all words happen independently of each other. Let assume a set of i document in vectors $D = \{d_1, \dots, d_i\}$, classified by a set of C classes of k, $C = \{C_1, \dots, C_k\}$, the Bayesian classifier estimates the probabilities of each C_k class provided to the d_j document as:

$$\mathrm{P}\left(\frac{C_k}{d_j}\right) = \frac{P(C_k)\, P\left(d_j \mid C_k\right)}{P\left(d_j\right)} \tag{1}$$

Where $P(d_j)$ is the probability that a randomly selected document has its representation of vector d_j, and the probability that a randomly selected document belongs to class C_k, the calculation of $P(d_j|C_k)$ is problematic because the sum of possible document d_j is very high. Naïve Bayes assumes, in order to simplify the calculation of $P(d_j|C_k)$, that the likelihood of the word of the phrase is independent of that of the other terms occurring in the same text (Lewis, n.d.1998). Although this may seem an over simplification, Naïve Bayes actually presents outcomes that are very competitive with those obtained through more elaborate techniques. In addition, since only words and not combinations of words are used as predictors, this Naïve simplification makes it much more effective to calculate the data model associated with this strategy than other non-naïve Bayesian methods. It is possible to calculate $P(d_j|\ C_k)$ as the product of the probabilities of each word that appears in the document using this simplification.

3.3.2. Logistic regression

Logistic Regression (LR) operates very much like linear regression, but with variable binomial response. The main advantage of this algorithm is that you can use continuous explanatory variables and it is simpler to use them. Simultaneously treat more than two explanatory variables. To explore the implied relation between a reaction the LRA is variable and has one or more explanatory variables suitable for research. Taking one explanatory case, the X variable with one Y binary consequence variable, the logistic variable model predicts the X Logit Y representing a natural chances logarithm of *Y*. The main formula of logistic regression model can be defined as follows (Peng et al., 2002; Yildiz et al., 2012)

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \tag{2}$$

The left-hand side is called the log-odds or logit. The LR model has a logit that is linear in X. Hence:

$$\pi(X) = E\left(\frac{Y}{X}\right) = \frac{e^{\alpha+BX}}{1+\ e^{\alpha+BX}} \tag{3}$$

Where π is the probability of the outcome of interest given as $X = x_1, x_2, \dots, x_k$, α is a parameter representing the *Y*-intercept, and β is a parameter of the slope, X can be

qualitative (categorical) or quantitative variables, and Y is always qualitative or categorical. The equation (4) can be expressed and extended from simple to multiple linear regression as follows:

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (4)$$

Therefore,

$$\pi(x) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \quad (5)$$

3.3.3. Support vector machine

SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural Risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is a hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be (Vapnik, 2000).

We consider data points of the form $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$.

Where $y_n = 1/-1$, a constant denoting the class to which that point x_n belongs. *n*= number of samples. Each x_n is p-dimensional real vector. The scaling is important to guard against variable with larger variance. We can view this Training data, by means of the dividing (or separating) hyperplane, which takes.

$$w.x + b = 0 \quad (6)$$

Where b is scalar and w is p-dimensional vector. The vector w points perpendicular to the separating hyperplane. Adding the offset parameter b allows us to increase the margin. Absent of b, the hyperplane is forced to pass through the origin, restricting the solution. As we are interesting in the maximum margin, we are

interested SVM and the where b is scalar and w is p-dimensional vector. The vector w points perpendicular to the separating hyperplane. Adding the offset parameter b allows us to increase the margin. Absent of b, the hyperplane is forced to pass through the origin, restricting the solution. As we are interested in the maximum margin, we are interested SVM and the parallel hyperplanes. Parallel hyperplanes can be described by equation

$$w.x + b = 1$$

$$w.x + b = -1$$

If the training data are linearly separable, we can select these hyperplanes so that there are no points between them and then try to maximize their distance. By geometry, we find the distance between the hyperplane is $2/|w|$. So we want to minimize $|w|$.

$$w.x_i - b \geq 1 \quad \text{or} \quad w.x_i - b \leq -1$$

This can be written as

$$y_i(w.x_i - b) \geq 1, \qquad 1 \leq i \leq n \tag{7}$$

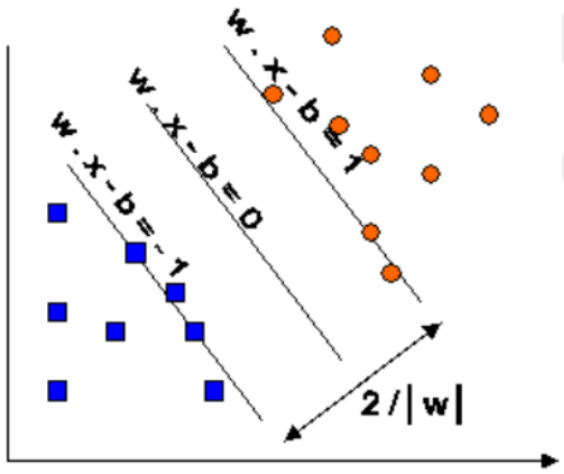

Figure 3.2. Maximum margin hyperplanes for a SVM trained with samples from two classes.

Samples along the hyperplanes are called Support Vectors (SVM). A separating hyperplane with the largest margin defined by $M = 2/|w|$ that is specifies support vectors means training data points closets to it. Which satisfy?

$$y_j[w\,T.x_j + b] = 1, \qquad j = 1,2,\dots,n \tag{8}$$

Optimal Canonical Hyperplane (OCH) is a canonical Hyperplane having a maximum margin. For all the data, OCH should satisfy the following constraints

$$y_i[w\,T.x_i + b] \geq 1, \qquad i = 1,2,\dots,l \tag{9}$$

Where *l* is Number of Training data point. In order to find the optimal separating hyperplane having a maximum margin, A learning machine should minimize $\|w\|2$ subject to the inequality constraints

$$y_i[w\,T.x_i + b] \geq 1, \qquad i = 1, 2, \dots, l$$

This optimization problem solved by the saddle points of the Lagrange's Function l0

$$LP = L(w, b, \alpha) = 1/2\|w\|2 - \sum\alpha_i\,(y_i(w\,Tx_i + b\,) - 1)$$

$$= \frac{1}{2}wT\,w - \sum \alpha_i\,(y_i(wT\,x_i + b) - 1) \qquad i = 1, 2, \dots, l \qquad (10)$$

Where α_i is a Lagranges multiplier. The search for an optimal saddle points (w_0, b_0, α_0) is necessary because Lagranges must be minimized with respect to w and b and has to be maximized with respect to non-negative α_i ($\alpha_i \geq 0$). This problem can be solved either in primal form (which is the form of *w* and *b*) or in a dual form (which is the form of α_i). Equation number (9) and (10) are convex and KKT conditions, which are necessary and sufficient conditions for a maximum of equation (9). Partially differentiate equation (10) with respect to saddle points (w_0, b_0, α_0).

$$\partial L/\partial w_0 = 0$$

$$i.e \qquad w_0 = \sum\alpha_i\,y_i\,x_i \qquad i = 1, 2, \dots, l \qquad (11)$$

and

$$\partial L/\partial b_0 = 0$$

$$i.e \qquad \sum\alpha_i\,y_i = 0 \qquad i = 1, 2, \dots, l \qquad (12)$$

Substituting equation (11) and (12) in equation (10). We change the primal form into dual form.

$$Ld(\alpha) = \sum \alpha_i - \frac{1}{2}\sum \alpha_i\alpha_j\,y_i\,y_j\;\;x_i\,T\,x_j \qquad j = 1, 2, \dots, n \qquad (13)$$

In order to find the optimal hyperplane, a dual lagrangian (Ld) has to be maximized with respect to nonnegative α_i (i.e. α_i must be in the nonnegative quadrant) and with respect to the equality constraints as follow

$$\alpha_i \geq 0, \qquad i = 1, 2, \dots, l$$

$$\sum \alpha_i\;\;y_i = 0$$

3.3.4. K-Nearest neighbor

K-Nearest Neighbor (K-NN) classification is one of the most critical and simple classification strategies and should be one of the key decisions to analyze a classification when there is essentially no earlier knowledge on the appropriation of the knowledge (Samanthula et al., 2015). K-NN classification was established while an analysis of the likelihood of densities for accurate parametric approximations that is unknown or difficult to determine was needed.

The K-NN classifier is generally based on the Euclidean distance between the specified training samples and a test sample (Yoon and Friel, 2015). Let x_i be an input sample with m features $(x_{i1}, x_{i2}, \dots, x_{in})$, n is the total number of input samples $(i = 1, 2, \dots, n)$ and m the total number of features (Fix & Hodges, 1951). The Euclidean distance between sample x_i and $x_j (j = 1, 2, \dots, n)$ is defined;

$$\mathrm{D}(x_i, x_j) = \sqrt{(x_1 - x_1)^2 + (x_2\ x_2)^2 + \cdots + (x_n - x_n)^2} \quad (14)$$

Where x_i and x_j are subjects to be compared with n characteristics. There are also other methods to calculate distance such as Manhattan distance

The value of k in the K-NN algorithms can be preferred for neighbors. The appropriate choice of k has an important influence on the K-NN algorithm for diagnostic results. A big value of k reduces the effect of random error-induced variance, however there is a possibility that small but meaningful trends can be overlooked. The secret to choosing the correct k value is to achieve a balance from overfitting to under fitting (Zhang, 2014).

3.3.5. Artificial neural networks

Artificial Neural Networks (ANN) is another machine learning algorithm that has become the most critical part of science and innovation nowadays. It shows the computer's ability to comprehend the structure of using a mathematical or statistical model of data. The artificial neural network, however, is not just a simple neural network, it involves more processes in a typical process and deals with complex patterns in a large amount of data. This includes a thorough understanding of the structure of methods and systematic algorithms (Farizawani et al., 2020).

The number of layers in ANN typically depends on the complexity of a problem that needs to be solved at the beginning. Four key steps in the neural network are involved: (i) initialization, (ii) activation function, (iii) weight training, and (iv) number of iterations used in classification task. The activation functions, however, change based on that problem to be solved. The elements of the neural network basic information processing begin with linked neurons by connections. Each connection has its own weight, and there is more than one neuron in each layer, the weight and the changed weight, respectively. The ties will pass from the input neuron towards the other nodes. This is referred to as the neural network for feed-forward (FF). A relation has a numerical and related weight that connect it through the output. Both related connections are referred to as perceptions. After the input links are given, weighted separately, they will be summed up as an activation mechanism to shape together. The diagram below (Figure 3.3) shows a feedforward and feedforward elementary of the NN method Algorithms to back propagate.

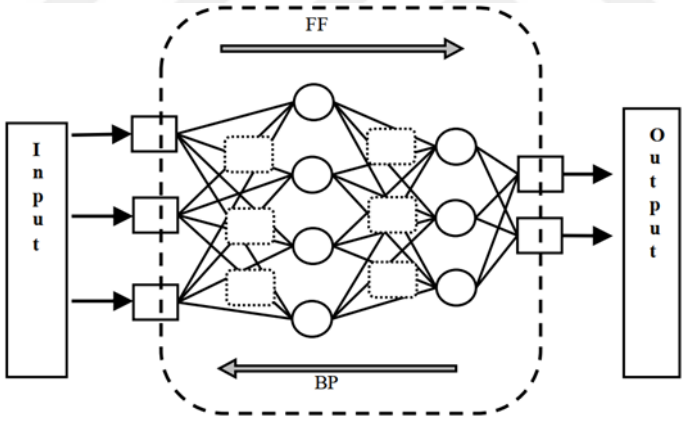

Figure 3.3. Artificial neural network process.

In the meantime, as shown in the following graph, the ANN method was defined in detail by inserting formulation for each input and proceed to the next layer of the network using the same formula. Then, the error corrections would be omitted since the aim of the process is to find the optimum performance value, based on the value of the given input. In order to get better, the application of formulations plays a main role. Using any other similar NN formulations to compare result (Farizawani et al., 2020).

3.3.6. Decision trees

Decision Trees (DT) are trees which classify samples by sorting them according to the values of the function. Each node in a decision tree represents a function in an samples to be categorized, and each branch represents a value that can be inferred by the node; samples are classified beginning at the root node and sorted based on their values. Learning from the decision tree, uses a decision tree as a predictive model that maps assumptions about an object to conclude on the target value of the item. Usually, decision tree classifiers employ post-pruning techniques that measure the efficiency of decision trees, since they are pruned using a validation set. Every node can be removed and allocated to the most common class of training samples (Osisanwo et al., 2017).

Decision tree learning is a technique for approximating discrete-valued target functions, in which a decision tree represents the learned function. Trees learnt the rules to enhance human readability may also be re-represented as collections of if-then rules. These methods of learning are among the most popular inductive inference algorithms and have been successfully applied to a wide variety of learning tasks. Diagnosis of medical training cases to determine the credit risk of loan applicants (Mitchell, 1999).

Using Decision Tree; we begin by defining a measure widely used in information theory, called entropy, in order to define information, gain precisely. A set of arbitrary sample. Provided the *S* list, which includes positive the entropy of (S) relative to this, and negative examples of a certain target term, Boolean scoring is defined as follows

$$Entropy\ (S) = -\sum_{i=1}^{n} P_i\ log_2(P_i) \qquad i: 1, 2, \dots, n \qquad (15)$$

where

S: Let *S* be a resource

P: A probability distribution

n: Simple Volume

For example, if s is a collection of 14 examples of some Boolean concept. Including 9 positive and 5 negative examples (we adopt the notation $[9+, 5-]$ to summarize such a sample of data). Then the entropy of S relative to this Boolean classification is

$$Entropy\ (9+, 5-) = -(9/14)\ log_2(9/14) - (5/14)\ log_2(5/14) = 0.940$$

Notice that the entropy is 0 if all members of S belong to the same class.

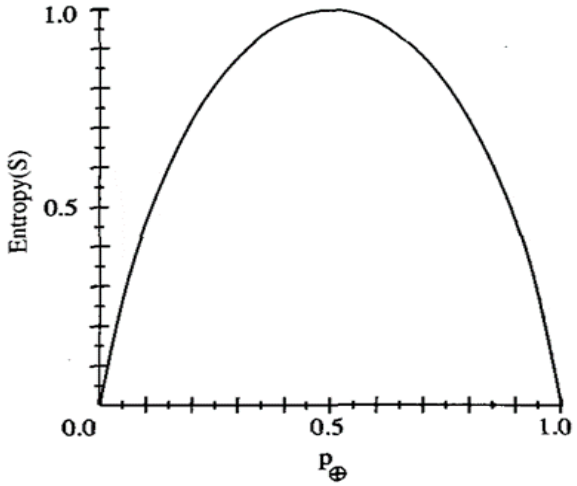

Figure 3.4. The entropy function relative to a Boolean classification as the proportion, pe, of positive examples varies between 0 and 1.

Figure 3.4 shows the form of the entropy function relative to a Boolean classification, as p varies between 0 and 1.

One interpretation of entropy from information theory is that it specifies the minimum number of bits of information needed to encode the classification of an arbitrary member of S (i.e., a member of S drawn at random with uniform probability). For example, if P, is 1, the receiver knows the drawn example will be positive, so no message need be sent, and the entropy is zero. On the other hand, if $P_{\oplus}$ is 0.5, one bit is required to indicate whether the drawn example is positive or negative. If $P_{\oplus}$ is 0.8, then a collection of messages can be encoded using on average less than 1 bit per message by assigning shorter codes to collections of positive examples and longer codes to less likely negative examples.

Thus far we have discussed entropy in the special case where the target classification is Boolean. More generally, if the target variable can take on c different values, then the entropy of S relative to this c-wise classification is defined as

$$Entropy\ (s) = \sum_{i=1}^{c} -p_i\ log_2\ p_i \qquad (16)$$

where p_i is the proportion of S belonging to class i. Note the logarithm is still base 2 because entropy is a measure of the expected encoding length measured in bits. Note also

that if the target variable can take on C possible values, the entropy can be as large as $logC$.

3.4. Dimensionality Reduction

The use of high dimensional dataset with hundreds of thousands of features is not uncommon in the various applications of machine learning and data mining. Modern datasets are, in other words, most much in high-dimensional space. Extracting information from immense data calls for new strategies. The more complex the datasets, the higher the computation time and the more difficult it is to interpret and analyze. Classification on high-dimensional data has therefore become a recurrent problem; as it occurs in various data mining applications for which a decision is required (Lin et al, 2010).

The effect of the high dimensionality is poorly understood on classification. Most datasets like microarray, DNA, etc. have thousands or more features while usually the sample size is tens or less than hundred. When the dimensionality is small, simplest classificatory is break down. Miller reported that there is a well-known phenomenon that a prediction model based on thousands of variables (m) can be quite unstable but has a relatively small sample size N (Silipo et al., 2014).

The above problem shows the importance of reducing dimensionality on classifying data of high dimensions. Reduction of dimensionality is a mechanism in which the random variable under consideration is decreased. The dimensionality reduction gives some advantages:

- For high-dimensional data, most machine learning and data mining techniques may not be effective.
- As the dimension increases, the accuracy and efficiency of the application degrades rapidly.
- High cost of computation.
- Helps avoid overfitting (training on highly related features instead of contingent features).
- Increasing data set time for study.

There are many dimensionality reduction techniques, but we will use using just one technique in this thesis.

3.4.1. Feature extraction

In feature extraction, all available variables (features) are used in the extraction of the function, and the data is converted to a reduced dimension space using a linear transformation. Its principal purpose is to replace the initial variable with a smaller set of underlying variables. In this thesis, we used PCA in the extraction of the function. PCA is one of the most commonly used techniques for data processing and compression extraction of functionality. PCA can be used to decrease the dimensionality of a dataset by identifying new variables smaller than the original but still maintaining much of the original data set. PCA extracts new variables which are linear combinations of the original variables by discovering new linear orthogonal combinations of the original variables with the greater variance. The main objective of the analysis of the components is to find directions that maximize the variance of the original data (Beniwal and Arora, 2012).

3.4.2. Principal component analysis

PCA is a statistical method that uses an orthogonal transformation to translate a set of observations of variables that may be associated into a set of linearly uncorrelated variables called the principal component. The principal component number is less than or equal to the original variable number. This transformation is described in such a way that the first main component has the greatest possible variance (i.e. accounts for as much of the data variability as possible), and each following component in turn has the greatest possible variance under the restriction that the previous component is orthogonal. The resulting vectors are an orthogonal basis set that is uncorrelated. The main components are orthogonal, as they are the own vectors of the symmetric covariance matrix. The relative scaling of the original variables is subject to the PCA (Mohammed et al., 2017).

3.4.3. Using principal component analysis

When applying the PCA, the variance of the random variables is shown by the eigenvalue of the covariance matrix that are located at the main diagonal of matrix $\sum$. In matrix $\sum$ the eigenvalues are sorted according to their magnitudes, from bigger to smaller. That means, the PCA with bigger variance comes first. For example, it is given the variances of the seven components in Table 3.6.

Table 3.6. Variances of the seven principal components

k	1	2	3	4	5	6	7
Var (k)	0.5162	0.7604	0.8446	0.9187	0.9678	0.9962	1

When applying the principal component and the variance retained by the first *k* principal components are explain in the Table 3.6. This means that if we select only one of the first principal component, then we retain only 51.6% of the variance. If we select the first two principal components, then 76% of the variance will be retained, and so on. By knowing this, we can select how many principal components to choose. For example, if we need to retain 90% at least of the variance, then we shall choose the first four principal components, if we need 99% of the variance, then we shall select the first six principal components. Table 3.7 shows the general shape of high dimensional data with *N* samples and m variables.

Table 3.7. The shape of a dataset consisting of *N* samples, each has m variable

		Variables			
		a_1	a_2	...	a_m
Samples	x_1	a_{11}	a_{12}	...	a_{1m}
	x_2	a_{21}	a_{22}	...	a_{2m}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	x_N	a_{N1}	a_{N2}	...	a_{Nm}

The main purpose of using the PCA is to reduce the high-dimensional data with a dimension m into a lower dimensional data of dimension k, where $k \leq m$.

Let $x_1, x_2, \dots, x_m$ be m continuous predictors. Principal component analysis can be described as follows:

1. Input C_{mxm}, the covariance matrix of $x_1, x_2, \dots, x_m$

2. Calculate the eigenvectors and eigenvectors of the covariance matrix. Sort the eigenvalues (and corresponding) in descending order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$.
3. Derive new predictors. Suppose the elements of the first component v_1 are $v_{11}, v_{22}, \ldots, v_{1m}$, then the new derived predictors are $\frac{v_{11}}{\sqrt{\lambda_1}} x_1 + \frac{v_{12}}{\sqrt{\lambda_1}} x_2 + \cdots + \frac{v_{1m}}{\sqrt{\lambda_1}} x_m$

3.5. The Class Imbalance Problem

The learning algorithms are widely used during the data mining process's pattern extraction phase. As this deal with real word data, several issues have emerged regarding the application of existing and well-established learning algorithms to real data. Among these issues is learning in the presence of class imbalances which is an important practical issue. Many learning algorithms were designed to assume well-balanced class distributions, i.e. no significant differences in the probabilities of class prior. However, in real world data this is not always the case where one class could be represented by a large number of examples, while the other is represented by only a few.

Generally, when one class represents a circumscribed concept, the problem of imbalanced data sets occurs, while the other represents the counterpart of that concept, so that examples from the counterpart class outnumber heavily examples from positive concept class. In this case, the inductive bias of learning algorithms that are not specifically designed to address class imbalances tends to be focused in the class represented by the greatest number of examples.

Several research papers reported class imbalances of 1 percent in the minority class and 99 percent in and up in the majority. Learning algorithms tend to induce classifiers with very low overall error levels in these situations, by simply classifying each new example as belonging to the majority class. These classifiers are obviously useless, as the minority class with unusual cases is the one, generally interested in predicting well.

Various proposals address the problem of class imbalance from an algorithmic point of view, primarily by adapting existing algorithms and techniques to the special

characteristics of imbalanced data. Such ideas include cost-sensitive instruction, one-class classificatory, and classifier ensembles, among others. The objective of cost-sensitive learning, is to minimize misclassification costs. Class imbalance can be dealt with in a similar way by allocating higher classification costs to the classes represented by just a few examples. Although, by adjusting the expected class ratio, misclassification costs are transformed into a class distribution, a complicating factor is that the costs associated with each misclassification are generally not known in advance (Prati et al., 2009).

There are different methods to use for solving the imbalance problem, and the most common one is the SMOTE method.

3.5.1. SMOTE method

Synthetic Minority Over-Sampling Technique (SMOTE) is an over-sampling process involving the generation of synthetic data. Its main idea is to create new examples of minority classes by interpolating between several examples of the minority class lying together. More precisely, SMOTE randomly selects an example of a group, E_i, and its neighbors. An example E_i is selected from the nearest neighbor set, and a new example is created using the equation below:

$$Enew = E_i + (E_i - E_i)\delta \quad (17)$$

Where δ in the interval {0,1} is a randomly selected constant. It is important to note that since the selected nearest neighbor may be from a class other than the E_i class, SMOTE may increase the space of the minority class, allowing for the development of synthetic examples which further spread into the space of the majority class(Chawla et al., 2002).

A basic example of SMOTE method is shown in Figure 3.5. The sample x_i from minority class is chosen as the basis for generating new synthetic data points. Focused on the distance metric is selected, with several nearest neighbors of the same class (points x_{i1} to x_{i4}). Eventually, a randomized interpolation is performed to obtain new samples r_1 to r_4.

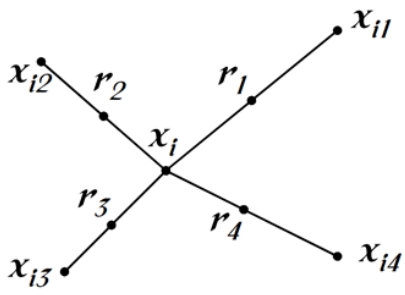

Figure 3.5. An illustration of how to create the synthetic data points in the SMOTE algorithm.

3.2. Design of the Proposed Classification Model

Classification of high-dimensional data with class imbalance problem is with great interest to researchers in different fields of science. Dealing with these two problems (high-dimensionality and class imbalance) for classification algorithms is challenging, yet very beneficial. In this thesis, we proposed a new method based on combining two well-known techniques: (PCA for dimensionality reduction) and (SMOTE for rebalancing) to tackle these two problems together. Figure 3.6 shows the diagram of the proposed method to choose the best model for classification and to comparing classification results before and after using dimensionality reduction by PCA as dimensionality reduction, and comparing the classification results and after rebalancing dataset by using the SMOTE oversampling method. The method starts with the use of the original dataset and then we will check if it is imbalanced or not. If the dataset is imbalanced, then we will balance it by using the SMOTE method. After this process, we will use the PCA to reduce the dimensionality of the data. But if the dataset is already balanced then we will directly use the PCA. After the data have been rebalanced and their dimensionality is reduced, we will split the dataset into training and testing to find out the training percentage which is 66%, along with the testing percentage which is 34%. By the end of these process, we will classify the dataset using the six classification algorithms which are (NB, LR, ANN, SVM, K-NN, DT). Thus, we will evaluate all the six classification models by using their evaluating measurements which are: Accuracy, F-measure, ROC area (the area under the ROC curve). We used ten-fold cross validation for evaluation: the knowledge was automatically split into ten pieces of equal size and the training and testing the procedure was carried out ten times, each component being

the test data once and the remaining pieces for training. After the preparation of the data for classification and the assessment, our method gives each test record to the most likely class. The next and last step is to compare them separately by comparing classification performance before and after using the PCA, SMOT and combined SMOT with PCA according to the evaluating of the classification models, to find out if the accuracies of the classification models have been improved or not.

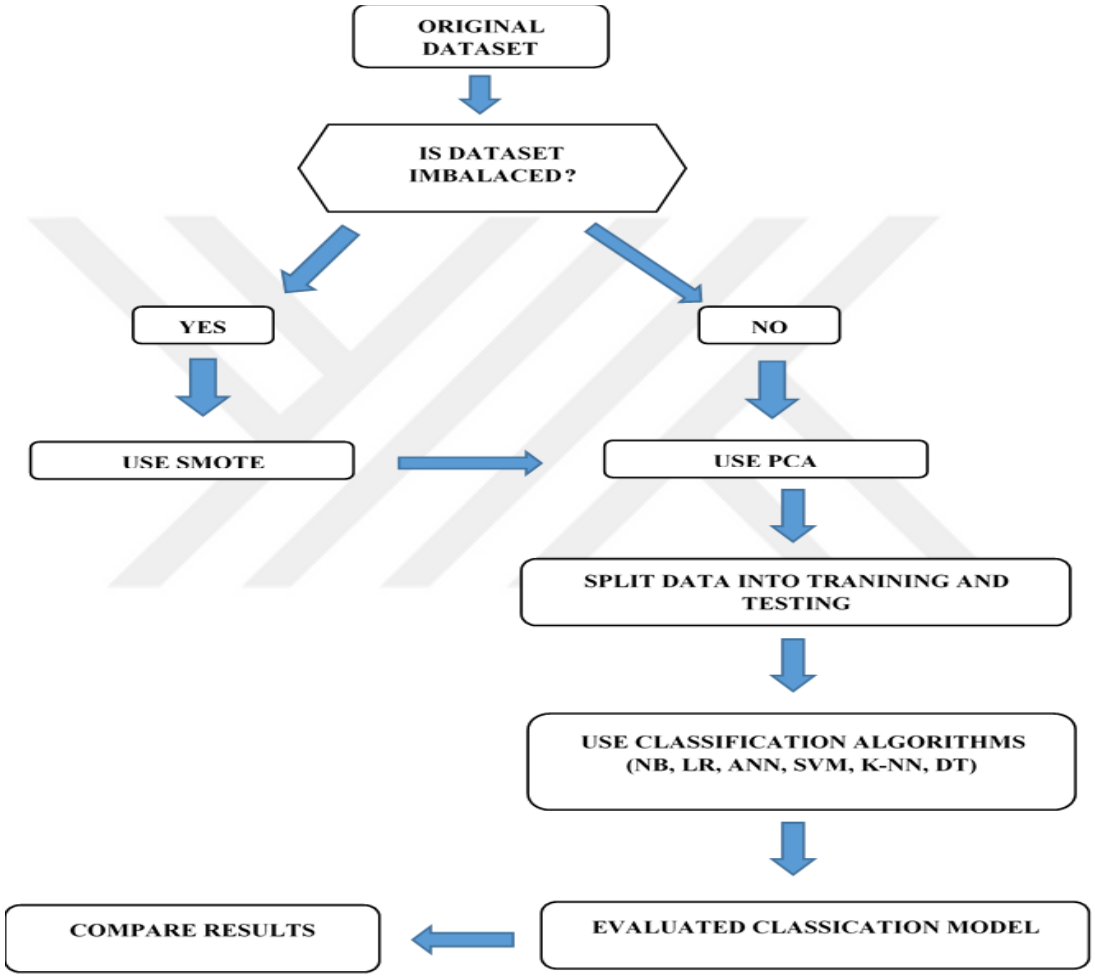

Figure 3.6. Design of the proposed classification model.

To check the efficiency and performance of each classification model investigated, we use the confusion matrix (defined below) to calculate different evaluation metrics as follows.

3.6.1. Confusion matrix

It's just a representation of the above parameters in a matrix format. Better visualization is always good.

Table 3.8. Confusion matrix

		Predicted
Actual	True Positive False Positive	False Negative True Negative

3.6.2. Accuracy

In this measure, classification data mining algorithms has proposed as a fitness function to evaluate each subset that is produced.

The formulation of measuring the accuracy of each subset is described as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

Where false negative (*FN*) denote the positive sample, which are incorrectly classified into the negative class, true positive (*TP*) represent the correct classifications of the positive example, false positive (*FP*) represent incorrect classification of the negative example into the positive class; while true negative (*TN*) is the correct classifications of negative example.

3.6.3. Precision

Great! Now let us look at Precision first.

$$precision = \frac{TP}{TP + FP}$$

What do you notice for the denominator? The denominator is actually the Total Predicted Positive! So, the formula becomes.

$$Precision = \frac{true\ positive}{total\ predicted\ positive}$$

Immediately, you can see that Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive. Precision is a good measure to determine, when the costs of False Positive is high. For sample, email spam detection. In email spam detection, a false positive means that an email that is non-spam

(actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

3.6.4. Recall

So let us apply the same logic for Recall. Recall how Recall is calculated.

$$Recall = \frac{TP}{TP + FN}$$

$$Recall = \frac{true\ positive}{total\ actual\ positive}$$

There you go! So Recall actually calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative. For sample, in fraud detection or sick patient detection. If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank. Similarly, in sick patient detection. If a sick patient (Actual Positive) goes through the test and predicted as not sick (Predicted Negative). The cost associated with False Negative will be extremely high if the sickness is contagious.

3.6.5. F-Measure

Now if you read a lot of other literature on Precision and Recall, you cannot avoid the other measure, F- Measure which is a function of Precision and Recall. the formula is as follows:

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

F-measure is needed when you want to seek a balance between Precision and Recall. so what is the difference between F-measure and Accuracy then? We have previously seen that accuracy can be largely contributed by a large number of True Negatives which in most business circumstances, we do not focus on much whereas False Negative and False Positive usually has business costs (tangible & intangible) thus F-measure might be a better measure

to use if we need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives).

3.6.6. ROC area

An ROC area **(**receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (*TPR*) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (*FPR*) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots *TPR* vs. *FPR* at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives

3.7. WEKA

WEKA is a series of algorithms for data mining tasks and machine learning. It includes tools for preprocessing data, classification, regression, clustering, laws of interaction, and visualization (Markov and Russell, 2006). In this thesis, we used this software for all data analyses and classification models. The data had not been preprocessed. There is four different ways of working at WEKA.

- Simple CLI: this provides a simple command-line interface allowing direct access carrying out WEKA orders.
- Explorer: this is a WEKA data-exploration environment.
- Experimenter: this is a setting for the conduct of experiments and statistical tests between learning schemes.

- Knowledge flow: this presents a WEKA-inspired guy called "data-flow. “The user can choose WEKA components from a tool bar, place them on a toolbar design canvas and link them to form a "awareness" information collection and review flow.

WEKA is a state-of-the-art production facility for Techniques of machine learning (ML) and their application to problems of real-world data mining. That is the for-data mining activities, a series of machine learning algorithms. You apply the algorithms to a dataset directly. WEKA implements data preprocessing algorithms, classification algorithms, Regression, clustering and correlation rules; a visualization tool is also included. For the latest computer with this kit, learning schemes can also be care.

4. RESULTS

In this chapter, a demonstration of experimental results that were carried out using different classification algorithms with imbalanced high-dimensional data will be present in details. The use of high-dimensional datasets with high number of features is not unusual in the various machine learning and data mining applications. As this deal with real world data, there have been many issues concerning the application of current and well-developed learning algorithms to deal with real data. Six classification algorithms (NB, LR, ANN, SVM, K-NN and DT) were conducted in our investigations using WEKA program (Bin Othman et al., 2007) to tackle the imbalance problem with high-dimensional data. The high-dimensionality of data was first reduced using the PCA method and the imbalanced data were rebalanced using the SMOTE method. A comparison was made between classification algorithms before and after using dimensionality reduction. Furthermore, imbalanced datasets were rebalanced and a comparison between classification performance before and after balancing data was also conducted. Six different datasets were used in our investigation. Table 4.1 shows the general information about these six data sets.

Table 4.1. Information about six different datasets

Data Set	Number of Variable	Sample Size	Negative class	Positive class
Dataset 1	14	1025	499	526
Dataset 2	14	303	138	165
Dataset 3	31	569	357	212
Dataset 4	20	153	122	31
Dataset 5	14	232	182	50
Dataset 6	19	148	142	6

4.1. Experimental Setup

During this study, there were 4 different methods have been applied to the data under study. The four methods include: (1) classification of raw data, (2) classification with the use of PCA, (3) classification with the use of the SMOTE, and (4) classification with the use of the combination of both, SMOTE and PCA. Each one of these methods were applied by

using all six algorithms, which are NB, LR, ANN, SVM, K-NN and DT. For NB algorithm, the dataset was divided into 10 folds for cross-validations. The identified ratio of splitting data for training was 66%, and for testing 34% for all classifier. When we used LR algorithms, we set the ridge value in the log-likelihood was 1.0E-8 and the dataset was divided into 10 folds for cross-validations. For ANN algorithms, we used 20 validations threshold and the learning rate of weight update 0.3 where dataset was divided 10 validations. When we used SVM, we had the complexity parameter C equal 1. The epsilon round off error 1.0 E-12. For the K-NN algorithms we used k=1, and the dataset divided to 10 cross validations. When we used DT algorithms, the confidence factor used for pruning was 0.25 and determines the amount of data used for reduced-error purring 3. When we used PCA method, we reduced the first and second datasets to 7 ranker variables, while we used 9 ranker variables for the third dataset. However, when we used fourth and sixth datasets, we had 10 ranker variables. The other we used 8 rankers variables. The variance covered that we used in PCA method was 0.95. On the other hand, when we applied the SMOTE method, we used k=5 and the seed used for random sampling was equal to 1, the percentage of SMOTE sample to create was equal to 100%.

4.2. Experimental Result

Table 4.2 shows the performance of the six classification algorithms for Dataset 1, before and after using PCA, SMOTE, and PCA+SMOTE methods to deal with the imbalance problem with high-dimensional data, according to Accuracy (Acc.), F-measure (F) and ROC metrics.

Table 4.2. Comparing the performance of the six classification algorithms using (PCA, SMOTE, and PCA+SMOTE) for Dataset 1

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
NB	81.1	81.2	89.3	81.2	81.2	88.4	92.5	92.6	97.1	82	82	87.7
LR	79.7	79.7	89.7	79.7	79.7	89.7	99.4	99.4	99.2	83.2	83.1	91.2
ANN	83.6	83.6	93	83.9	83.9	92.2	93.5	93.5	99.1	85.8	85.8	91.4
SVM	79.6	79.6	79.6	80.8	80.8	80.8	99.6	99.6	99.4	83.7	83.6	81.1
K-NN	**100**	**100**	**100**	**99.7**	**99.7**	**99.3**	**99.8**	**99.8**	**100**	**99.5**	**99.5**	**99.6**
DT	83.9	83.8	88.9	85.1	85	90.1	96.1	96.2	97.8	97.9	97.9	98.9

From the results in Table 4.2, we can see that the accuracy, F-measure and ROC area measures were calculated for checking the performance of the six classification algorithms with and without using dimensionality reduction and rebalancing data methods. For the raw data, the highest Accuracy, F-measure and ROC area rate of 100%, 100% and 100% were obtained in K-NN respectively; According to the three measurements calculated, the lowest rates as 79.6% were obtained in SVM. However, when the PCA dimensionality reduction method was used, we can see that the highest Accuracy and F-measure ROC area rate of 99.7%,99.7%, and 99.3% were obtained in K-NN respectively. When the SMOTE oversampling method was applied, the highest Accuracy, F-measure and ROC rate of 99.8% 99.8%, 100% were obtained in K-NN, respectively. On the other hand, when the two methods (PCA + SMOTE) were simultaneously applied, we can see that the classification performances were further improved for almost all the classifications algorithms used, and the highest Accuracy, F-measure and ROC rates of 99.5%, 99.5%, and 99.6% were obtained in K-NN algorithm, respectively. Looking at results in Table 4.2 in more details, we can see that after reducing the dimensionality and rebalancing data, the performances of the classification algorithms have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data under existence of imbalance problem.

Figure 4.1 shows the performance of each classification algorithms for Dataset 1, before and after using PCA, SMOTE, and PCA+SMOTE using Accuracy (Acc.), F-measure (F) and ROC metrics.

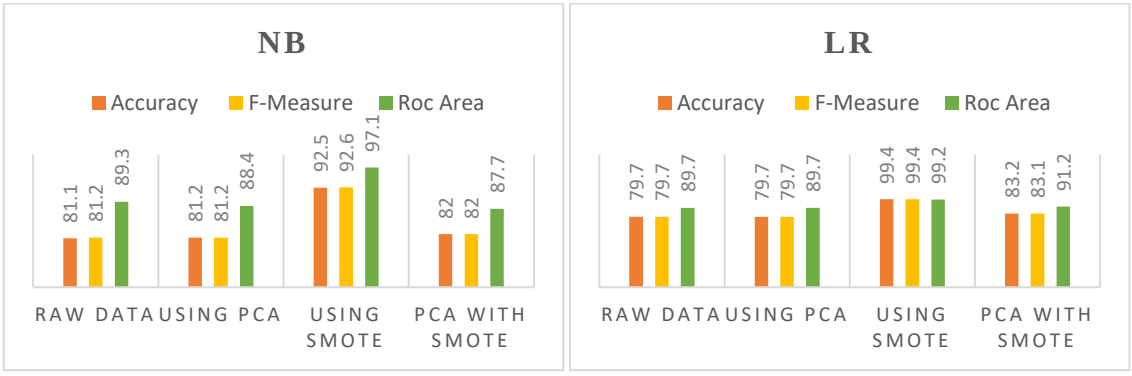

Figure 4.1. Performance of each classification algorithms for Dataset 1, before and after using (PCA, SMOTE, and PCA+SMOTE).

Figure 4.1. Performance of each classification algorithms for Dataset 1, before and after using (PCA, SMOTE, and PCA+SMOTE) (continue).

Figure (4.1) shows that when using either PCA, SMOTE, or the combination of them (PCA+SMOTE), the performances of the classification algorithms are almost always providing better classification results. This indicates that the two methods applied in this thesis are very helpful. As its clear that from the figure of NB algorithm, the Accuracy, ROC area and the F-measure methods are higher when using SMOTE method. For the LR algorithm, we can see that all three methods show higher results when applying using SMOTE. However, when we used ANN algorithm, we can see that after using PCA method, the accuracy and F-measure rate were improved but the ROC area rate was decreased; yet when using SMOTE method, all measures was developed to provide better results. For the combination PCA with SMOTE method, the accuracy and F-measure rate was increased but the ROC area was giving the worst result. SVM algorithms had been improved under the three measures used with and without using dimensionality reduction and rebalancing data methods. Only results of K-NN algorithm figure showed that after using all the three methods, the results were worst, and this is because this algorithm has already provided optimal results which does not require any

further improving. When applying the DT algorithms with the three methods that we used, the classification performance almost all the classifiers were better with either one of the three methods.

Table 4.3 shows the performance of the six classification algorithms for Dataset 2, before and after using PCA, SMOTE, and PCA+SMOTE methods to deal with the imbalance problem with high-dimensional data.

Table 4.3. Comparing the performance of the six classification algorithms using (PCA, SMOTE, and PCA+SMOTE) for Dataset 2

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
NB	80.8	80.8	**89.4**	80.1	80.1	86.1	87	87.1	**94.3**	81.1	81	87.3
LR	67.3	67.2	74	81.1	81.2	**88.9**	74.8	74.6	74.2	82.3	82.3	**90.9**
ANN	**81.1**	**81.2**	88.9	81.1	81.1	87.4	78	78.2	79.1	83.4	83.3	88.6
SVM	78.2	78.2	78.2	**82.1**	**82.1**	81.9	84.1	84	82.4	84.1	84	82.4
K-NN	80.1	80.2	84.7	76.5	76.5	77	**88.6**	**88.3**	92.7	**85.2**	**84.9**	82.2
DT	75.9	75.9	81.1	80.8	80.9	78.2	78.2	77.7	80.3	83.2	82.9	82.4

From the results in Table 4.3, we can see that the accuracy, F-measure and ROC area measures were calculated for checking the performance of the six classification algorithms with and without using dimensionality reduction and rebalancing data methods. For the raw data, the highest Accuracy and F-measure rate of 81.1% and 81.2% were obtained in ANN respectively, but the highest ROC area performance rate of 89.4% was obtained in NB. When using the PCA method, an improvement was observed in the classification performance of LR and SVM according to all measurements and DT according to Accuracy and F measurements, but a decrease was observed in the performances of the other three classifiers. When the SMOTE method was used, there was a decrease rather than an improvement solely in the performance of the ANN classifier and the ROC area value of DT according to the three measurement values. However, when combination of the two methods (PCA+SMOTE) was applied, all the classification algorithms have improved in terms of evaluation metrics applied. Looking at results in Table 4.3 in more details, we can see that after reducing the dimensionality and rebalancing data, the performances of the classification algorithms have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data under existence of imbalance problem.

Figure 4.2 shows the performance of each classification algorithms for Dataset 2, before and after using PCA, SMOTE, and PCA+SMOTE.

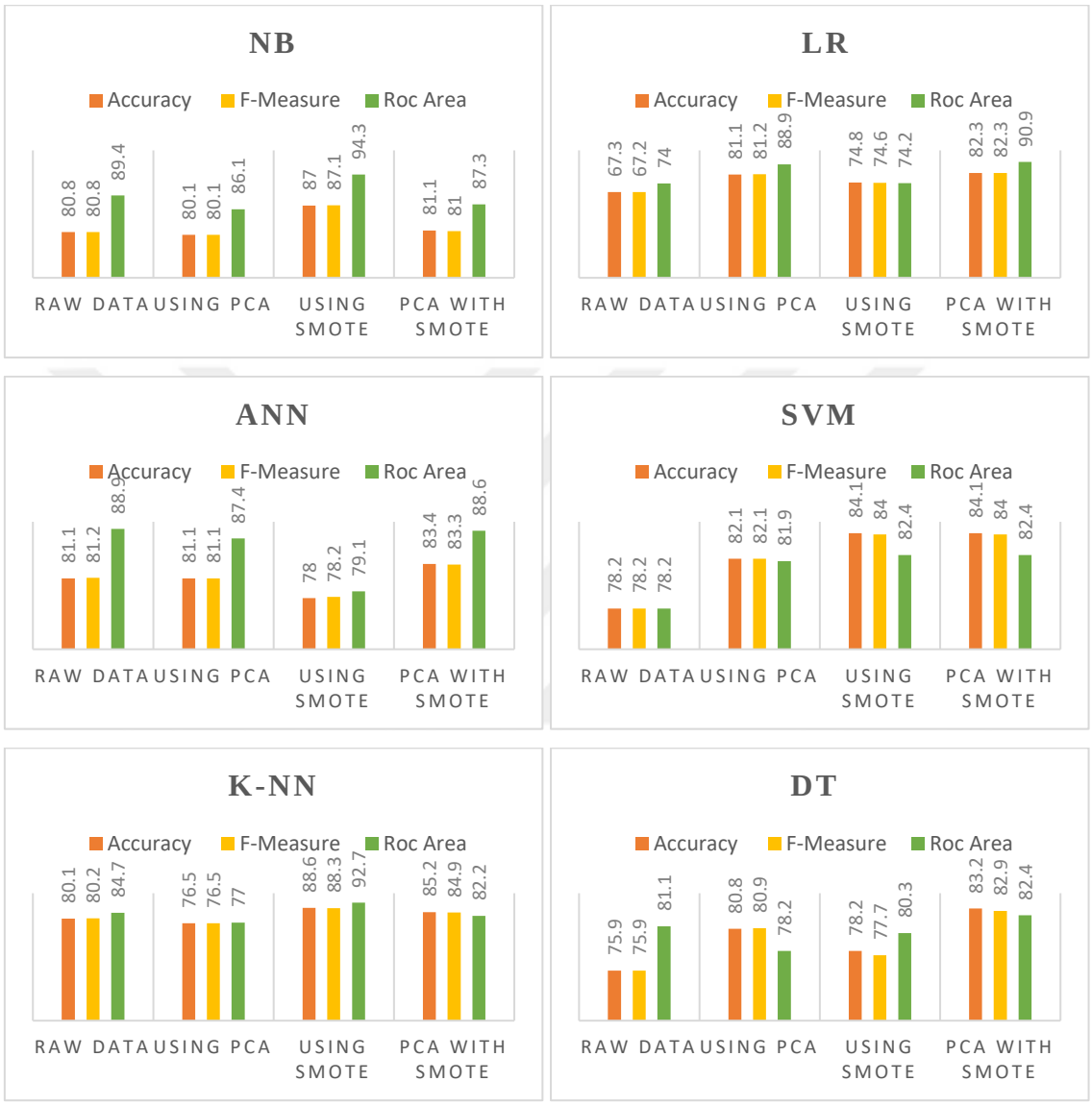

Figure 4.2. performance of each classification algorithms for Dataset 2, before and after using (PCA, SMOTE, and PCA+SMOTE).

Figure (4.2) represent that when using either PCA, SMOTE, or the combination of them (PCA+SMOTE), the performances of the classification algorithms are almost always providing better classification results. The NB figure showed that accuracy and F-measure decreased after using PCA method, but when we used the SMOTE method for the same algorithm, the classification results have improved. For the LR algorithm, we

can see that the accuracy, F-measure and ROC are increased to 81.1%, 81.2% and 88.9% when using PCA method. However, when the combination of PCA with SMOTE method was used, the LR algorithm has further improved, and provided very much better results. PCA method, has not affected on the ANN method, but after using SMOTE method, ANN gave worst results. However, when we using the combination PCA with SMOTE, the results have significantly improved. The SVM algorithm gave better result when using PCA, SMOTE and combination PCA with SMOTE. When using PCA method, K-NN algorithm has provided worst results, but the DT algorithm has provided better results. When using SMOTE method, the results of K-NN and DT algorithms have increased, but the results was decreased in DT algorithm according to ROC area.

Table 4.4 shows the performance of the six classification algorithms for Dataset 3, before and after using PCA, SMOTE, and PCA+SMOTE methods to deal with the imbalance problem with high-dimensional data.

Table 4.4. Comparing the performance of the six classification algorithms using (PCA, SMOTE, and PCA+SMOTE) for Dataset 3

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
NB	92.9	93	98	91.5	91.5	97	93.3	93.4	98.5	93.3	93.4	98.3
LR	93.8	93.9	97	97	97	98.9	96.2	96.3	97.7	97	97.1	**99.6**
ANN	96.6	96.7	99.3	**98.2**	**98.2**	**99.6**	97.4	97.4	**99.6**	97.3	97.3	**99.6**
SVM	**97.3**	**97.3**	**99.6**	95	95	93.6	**97.6**	**97.7**	97.8	**97.5**	**97.6**	97.6
K-NN	95.9	96	95.4	95.6	95.6	94.8	96.4	96.4	96.5	97.3	97.3	97.3
DT	93.3	93.3	93.1	94.5	94.5	93.2	95.3	95.4	94.6	94.1	94.1	94.4

From the result Table 4.4, we can see that the Accuracy, F-measure and ROC area measures were calculated for checking the performance of the six classification algorithms with and without using dimensionality reduction and rebalancing data methods. From the raw data, we can see that the results of all classification algorithms have showed very well performances with the use of PCA, SMOTE and combine (PCA+SMOTE) methods. The highest classification performance according to Accuracy, F-measure and ROC area rate of 97.3%, 97.3% and 99.6% were obtained in SVM, respectively. According to the three calculated criteria; In the PCA method, only LR, ANN and DT (NB, SVM and K-NN models showed a decline) classifier models were improved, while in the SMOTE method, an improvement was achieved in all classifier

models except SVM Roc measurement. On the other hand, when the combination of the two methods (PCA+SMOTE) were simultaneously applied, we can see that the highest Accuracy and F-measure rate of 97.5% and 97.7% were obtained in SVM respectively, but the highest ROC area rate of 99.6% was obtained in LR and ANN. Looking at result in Table 4.4 in more details, we can see that after reducing the dimensionality and rebalancing data, the performances of the classification algorithms have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data existence of imbalance problem.

Figure 4.3 shows the performance of each classification algorithms for Dataset 3, before and after using PCA, SMOTE, and PCA+SMOTE.

Figure 4.3. Performance of each classification algorithms for Dataset 3, before and after using (PCA, SMOTE, and PCA+SMOTE).

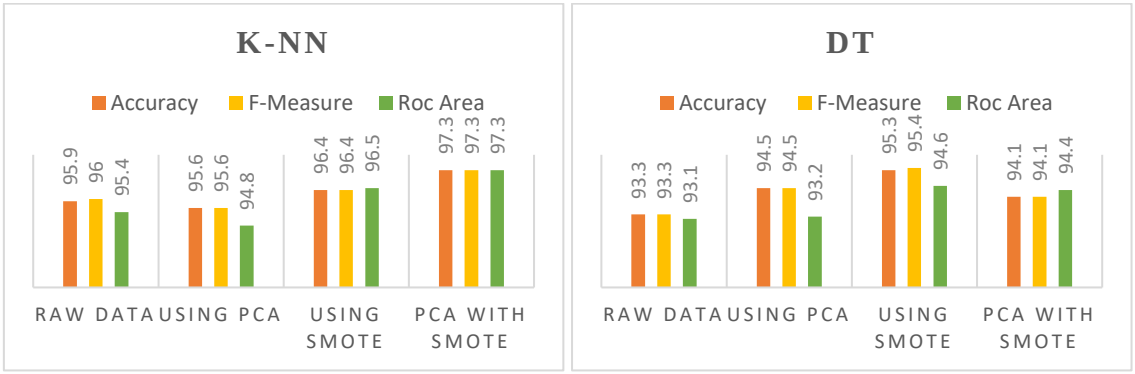

Figure 4.3. Performance of each classification algorithms for Dataset 3, before and after using (PCA, SMOTE, and PCA+SMOTE) (continue).

Figure (4.3) represent that when using either PCA, SMOTE, or the combination of them (PCA+SMOTE), the performances of the classification algorithms are almost always providing better classification results. NB figure showed that when we used PCA method, the NB algorithm gave worst result, while when we used SMOTE method the result got better. However, when we combined PCA with SMOTE method, NB results have improved. From the LR figure represents that the algorithms were developed when we used the three methods on it. ANN figure showed that when used PCA, SMOTE and PCA with SMOTE, almost all classifiers have given better performances, while the highest rate of 99.3%, 99.6% were obtained in ROC area, respectively. For SVM when we used PCA method, the SVM algorithm has not improved, while when used SMOTE and PCA and SMOTE method, the result gave better performances. However, when we used PCA method, K-NN algorithm gave worst result but the results of the DT algorithm have improved. As it is clear that from the figure of K-NN algorithm, the Accuracy, ROC area and the F-measure methods are higher when using PCA with SMOTE method. It is also clear that from the figure of the DT algorithm, the Accuracy, ROC area and the F-Measure methods are higher when using SMOTE method.

Table 4.5 shows the performance of the six classification algorithms for Dataset 4, before and after using PCA, SMOTE, and PCA+SMOTE methods to deal with the imbalance problem with high-dimensional data.

Table 4.5. Comparing the performance of the six classification algorithms using (PCA, SMOTE, and PCA+SMOTE) for Dataset 4

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
NB	83.5	84.1	**86.6**	82.2	82.1	**85.9**	85.2	85.5	90.8	85.7	85.6	**90.7**
LR	84.2	84	85.4	**88.1**	**88**	**85.9**	84.6	84.8	**91.6**	84.1	84.2	90.2
ANN	80.2	80.3	83.8	84.2	84.2	85.4	81.4	81.4	89.6	84.6	84.8	90.1
SVM	**86.1**	**85.7**	75.7	85.5	84.7	72.9	**85.7**	**85.8**	84.5	84.1	83.9	80.9
K-NN	80.9	80.6	66.7	80.9	81	70	84.6	84.7	81.3	85.2	85.5	84.2
DT	76.9	76.5	62.3	87.5	87.1	81.9	82.5	82.5	81.3	**87.4**	**87.4**	81.8

From the result Table 4.5, we can see that the Accuracy, F-measure and ROC area measures were calculated for checking the performance of the six classification algorithms with and without using dimensionality reduction and rebalancing data methods. As for the raw data, the top accuracy and F measure percentage of 86.1% and 85.7% was obtained in SVM respectively, whereas the highest percentage of ROC area of 86.6% was gained in NB. On the other hand, once PCA method is used, the outcomes are developed in the sense that the top accuracy percentage of 88.1% was gained in LR hereafter, whereas the top percentage of F-measure of 88% was gained in LR, also the top ROC level of 85.9% was obtained in NB and LR. With the use of SMOTE, the highest Accuracy and F-measures of 85.7%, 85.8% were obtained in SVM in return, also the highest ROC area percentage of 91.6% was obtained in LR. However, the use of both PCA and SMOTE led to the improvement of classification performance of most of the used algorithms and the highest Accuracy and F-measures of 87.4%, 87.4% were obtained in the DT algorithm, whereas the highest percentage of ROC area of 90.7% was obtained in NB. Looking at result in Table 4.5 in more details, we can see that after reducing the dimensionality and rebalancing data, the performances of the classification algorithms have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data existence of imbalance problem.

Figure 4.4 shows the performance of each classification algorithms for Dataset 4, before and after using PCA, SMOTE, and PCA+SMOTE.

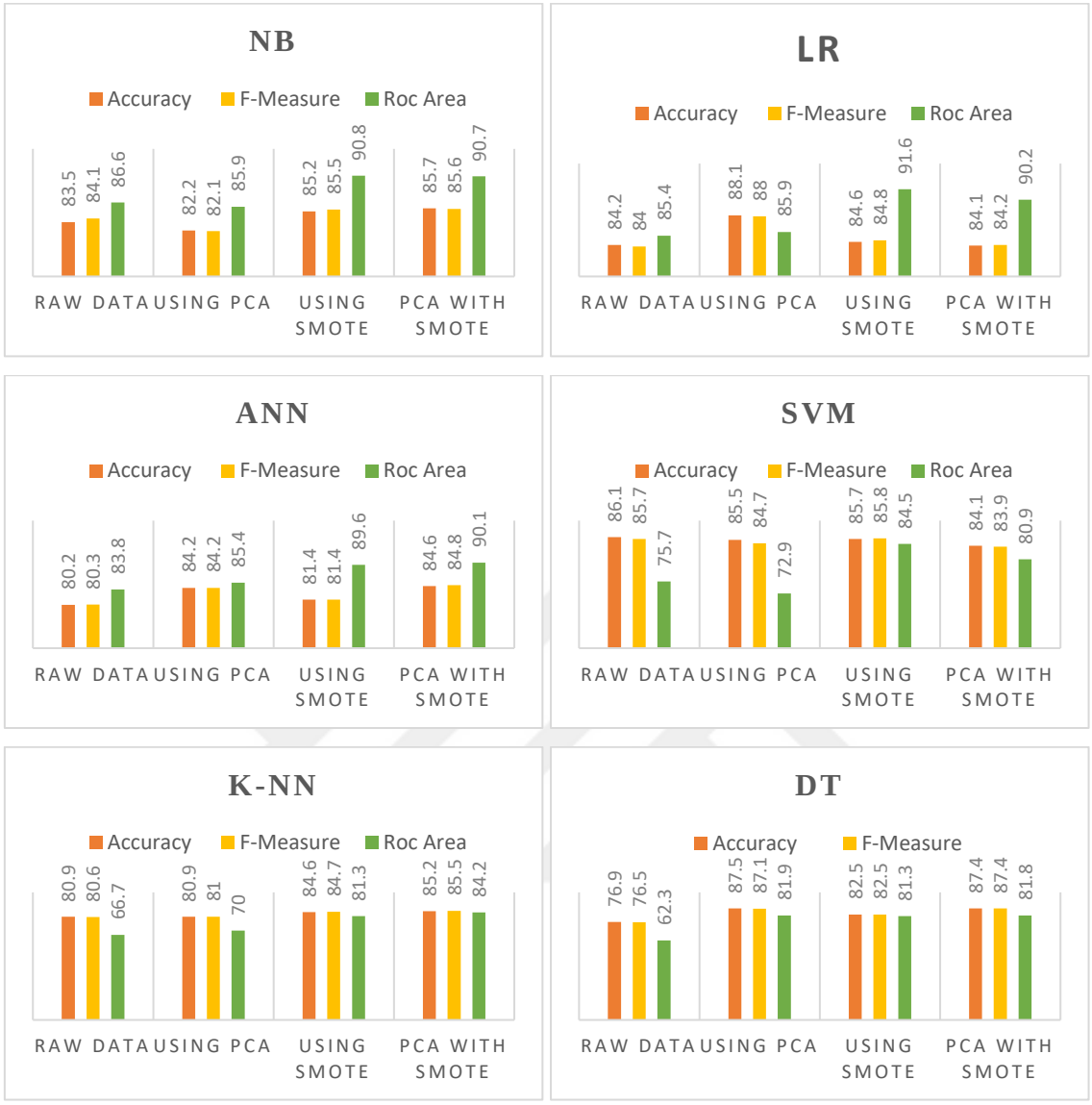

Figure 4.4. Performance of each classification algorithms for Dataset 4, before and after using (PCA, SMOTE, and PCA+SMOTE).

Figure (4.4) shows that when using either PCA, SMOTE, or the combination of them (PCA+SMOTE), the performances of the classification algorithms are almost always providing better classification results. From the figure of the NB algorithm, we can see that the results have improved in all methods that we used in this study except for the PCA method. The highest Accuracy and F-measure rate for the LR algorithm was obtained in PCA method, while the highest ROC area was obtained in SMOTE method. From the ANN figure, it showed that the results were significantly improved when using each of the three methods used in this study. The highest Accuracy, F-measure and ROC

area for the ANN algorithm were got obtained in combination of PCA with SMOTE method. Accuracy measure got worst result when used in three methods that we used but the F-measure was improved when we used SMOTE method. While the ROC area measure was giving better result in SMOTE and the combination of PCA with SMOTE methods. However, when we used PCA, SMOTE and PCA with SMOTE methods, K-NN and DT algorithms were provided better classification performances.

Table 4.6 shows the performance of the six classification algorithms for Dataset 5, before and after using PCA, SMOTE, and PCA+SMOTE methods to deal with the imbalance problem with high-dimensional data.

Table 4.6. Comparing the performance of the six classification algorithms using (PCA, SMOTE, and PCA+SMOTE) for Dataset 5

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
NB	87	87.2	**91.9**	86.6	86.4	**90.5**	89.3	89.4	**95.1**	87.2	87.2	92.5
LR	88.3	88.3	88.9	88.7	88.5	90.2	**90.7**	**90.8**	93.5	87.2	87.2	**92.6**
ANN	85.3	85.1	87.1	87.9	87.7	86	85.1	85.2	91	86.5	86.5	91.2
SVM	**90.5**	**90.3**	83.8	**90**	**89.6**	81.4	90	90	88.9	89	88.9	87.4
K-NN	84.4	84.5	76.6	87.5	87.5	82.6	86.8	86.9	85.5	**89.3**	**89.4**	88.9
DT	84.4	83.8	75.4	88.3	88	86.4	88.2	88.2	87.8	85.8	85.6	86.8

From the result Table 4.6, we can see that the Accuracy, F-measure and ROC area measures were calculated for checking the performance of the six classification algorithms with and without using dimensionality reduction and rebalancing data methods. As for the raw data, we can see that the results of all classification algorithms have showed very well model performances with the use PCA, SMOTE and the combination (PCA+SMOTE) methods. The highest performance according to Accuracy, F-measure rate of 90.5%, 90.3% were obtained in SVM respectively, also the highest performance according to ROC area rate of 91.9% was got in NB. When using PCA method, NB and SVM classifiers have not improved, while the other four classifiers gave better results. However, when we used SMOTE method, all the classifiers algorithms have improved except the SVM which gave worst results. On the other hand, when the two methods (PCA+SMOTE) were simultaneously applied, we can see that the highest Accuracy and F-measure rate of 89.3% and 89.4% were obtained in K-NN respectively, but the highest ROC area rate of 92.6% was obtained in LR. Looking at result in Table

4.6 in more details, we can see that after reducing the dimensionality and rebalancing data, the performances of the classification algorithms have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data existence of imbalance problem.

Figure 4.5 shows the performance of each classification algorithms for Dataset 5, before and after using PCA, SMOTE, and PCA+SMOTE.

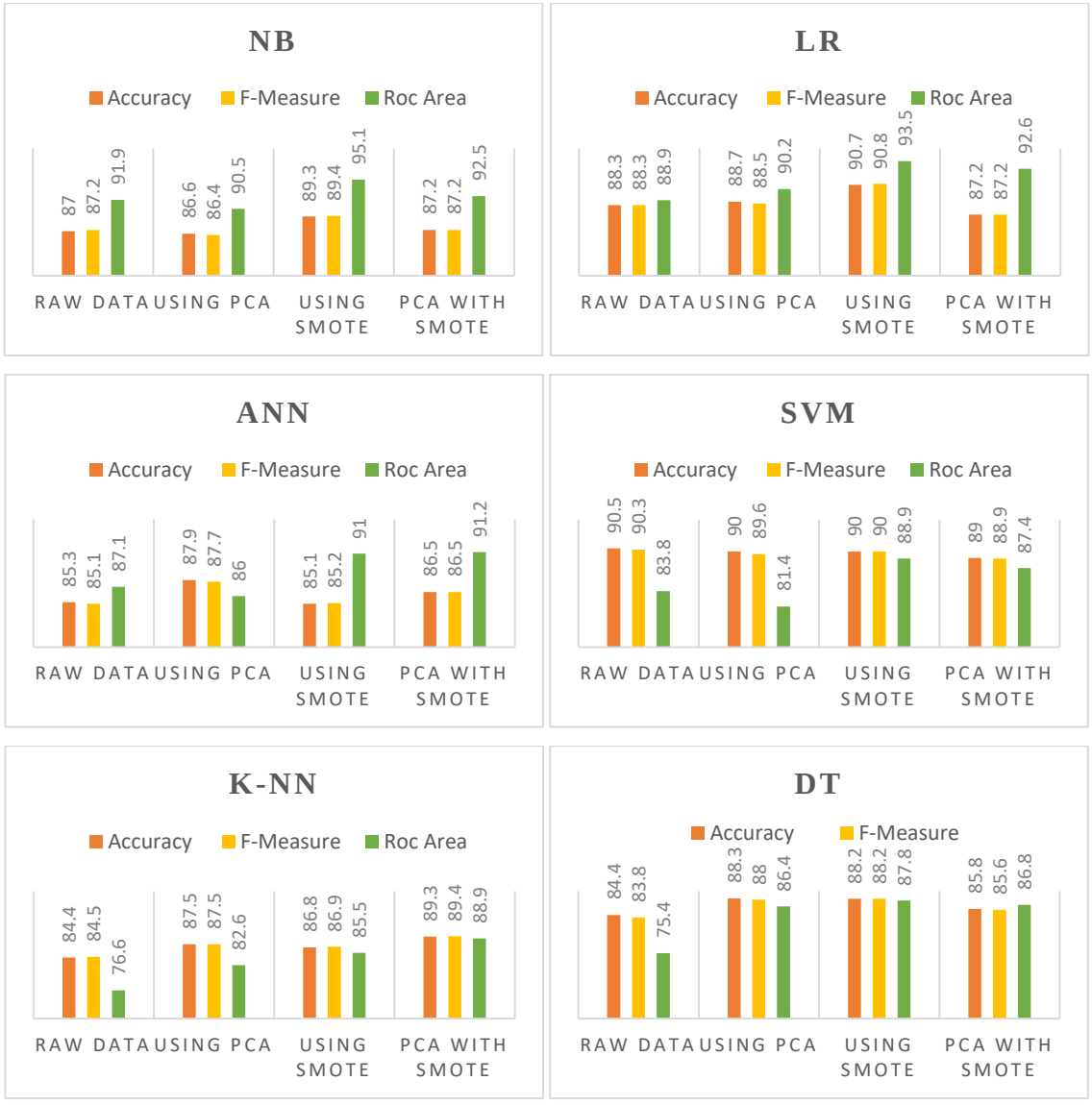

Figure 4.5. Performance of each classification algorithms for Dataset 5, before and after using (PCA, SMOTE, and PCA+SMOTE).

Figure (4.5) shows that when using either PCA, SMOTE, or the combination of them (PCA+SMOTE), the performances of the classification algorithms are almost always providing better classification results. This indicates that the two methods applied in this thesis are very helpful. As its clear that from the figure of NB algorithm, the Accuracy, ROC area and the F-measure methods are higher when using SMOTE method. For the LR algorithm, we can see that all three methods show higher results when applying SMOTE. However, from ANN algorithm figure, it showed that after using PCA method, the Accuracy and F-measure rate were improved but the ROC area rate was decreased, while when the SMOTE method is applied, all measures were developed to provide better results except for the Accuracy metric. For the combination of PCA with SMOTE method, the Accuracy and F-measure rate were increased. The SVM algorithms has not been improved with the three measures used with and without using dimensionality reduction and rebalancing data methods. The K-NN figure showed that after using all method, the classifiers have obtained better results. When the DT algorithms crossing the three methods, the results were better than before.

Table 4.7 shows the performance of the six classification algorithms for Dataset 6, before and after using PCA, SMOTE, and PCA+SMOTE methods to deal with the imbalance problem with high-dimensional data.

Table 4.7. Comparing the performance of the six classification algorithms using (PCA, SMOTE, and PCA+SMOTE) for Dataset 6

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
NB	97.9	98	**99.8**	**99.3**	**99.3**	**99.9**	97.4	97.5	99.1	**100**	**100**	**100**
LR	97.2	97.5	97.8	98.6	98.6	99.4	98	98.1	96.6	98.7	98.8	99.9
ANN	97.9	97.9	99.4	97.2	97.3	99.4	**98.7**	**98.6**	**99.5**	99.3	99.4	99.9
SVM	**98.6**	**98.5**	83.3	98.6	98.5	83.3	**98.7**	**98.6**	91.7	99.3	99.4	99.6
K-NN	**98.6**	**98.5**	89	**99.3**	**99.3**	93.3	**98.7**	**98.6**	97.2	99.3	99.3	98.3
DT	97.9	97.6	54.7	97.2	97.5	91.2	96.1	95.9	83	98.7	98.8	99.2

From the results in Table 4.7, we can see that the accuracy, F-measure and ROC area measures were calculated for checking the performance of the six classification algorithms with and without using dimensionality reduction and rebalancing data methods. For the raw data, the highest Accuracy and F-measure rate of 98.6%, 98.6 and 98.5%, 98.5 were obtained in SVM and K-NN respectively, but the highest ROC area

performance rate of 99.8% was obtained in NB. When using PCA method, NB, K-NN and LR have showed significant improvement in classification performance, while the PCA method has not provided effective on the other classifiers, and the ANN algorithm has the worst results. With the use of SMOTE method, LR, ANN, SVM and K-NN classifiers provided better result compared to the other classifiers and the other gave worst results. However, when the combination of (PCA+SMOTE) is used, all the classification algorithms have improved in terms of the evaluation metrics used. Looking at results in Table 4.7 in more details, we can see that after reducing the dimensionality and rebalancing data, the performances of the classification algorithms have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data under existence of imbalance problem.

Figure 4.6 shows the performance of each classification algorithms for Dataset 6, before and after using PCA, SMOTE, and PCA+SMOTE.

Figure 4.6. Performance of each classification algorithms for Dataset 6, before and after using (PCA, SMOTE, and PCA+SMOTE).

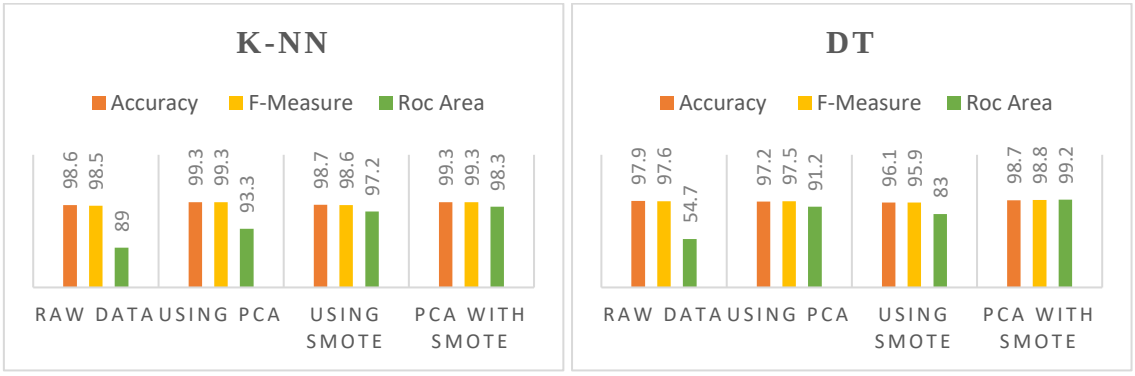

Figure 4.6. Performance of each classification algorithms for Dataset 6, before and after using (PCA, SMOTE, and PCA+SMOTE) (continue).

Figure (4.2) represent that when using either PCA, SMOTE, or the combination of them (PCA+SMOTE), the performances of the classification algorithms are almost always providing better classification results. The NB figure showed that the Accuracy and F-measure have increased after using PCA method, but when we used the SMOTE method for the same algorithm, the results have not improved, while the combination of the two methods have provided very well improvement. For the LR algorithm, we can see that the Accuracy, F-measure and ROC are increased to 98.6%, 98.6% and 99.4%, respectively, in PCA method. When we used SMOTE method and the combination of PCA with SMOTE method, the LR algorithm has been improved. With the use of PCA method, it has not affected on the ANN method, but after using SMOTE method, the ANN algorithm gave better results. However, when we used the combination of PCA with SMOTE method, the ANN algorithm gave much better results. With the PCA method, the results have not been improved in SVM, but when we used the SMOTE and the combination of the PCA with SMOTE, the performance of the classifiers have improved well. When using PCA method, the K-NN algorithm has provided better results, but the DT algorithm has worst results. While when using the SMOTE method, the K-NN algorithm increases the results of the classification, but the results was decreased in the DT algorithm. The results in K-NN and DT have improved when the combined method of the PCA with SMOTE is applied.

5. DISCUSSION AND CONCLUSION

The issue was argued by decreasing the number of unnecessary features (using the PCA method of reducing dimensionality) and thus re-balancing data (using the SMOTE method). The health sector hence whole society will get benefit through developing a strategy for rapid and more accurate model identification with a view to identifying and addressing, along with efficient implementation (Naseriparsa and Kashani, 2014).

Experimental results on the six different classification algorithms for six imbalanced high-dimensional data showed that all classification algorithms have enhanced the classification performance of datasets using either PCA, SOMTE, or PCA+SMOTE methods. However, the preferred classification algorithm with the highest performance compared to others, varied from dataset to another. Experimental results from the first dataset demonstrated that the best classification model was K-NN when normal classification, PCA, SMOTE and PCA+SMOTE methods were used. However, in case of PCA, SMOTE and PCA+SMOTE methods, the results obtained from K-NN were not very good compared to other algorithms. On the other hand, when the PCA method was used, the accuracy of NB, ANN, SVM and DT algorithms were improved, but K-NN and LR models did not show that improvement. Additionally, when using the SMOTE and the PCA+SMOTE methods, the performances of all classification algorithms were improved except for the K-NN algorithm.

For the second dataset, the ANN algorithm has provided the best results for raw data before applying any rebalancing or dimensionality reduction methods, but when the PCA method was used, the best results were provided by the SVM algorithm. However, when the SMOTE and the PCA+SMOTE methods were used, the K-NN classifier provided better results than the other algorithms. The results from NB algorithm were poor at PCA method, but when the SMOTE and combined PCA with SMOTE methods were used, the results of the NB algorithm have significantly improved. In the same dataset, the LR algorithm gave better results in combined PCA with SMOTE method. We noticed that the PCA method has not affected the ANN algorithm results, while the SMOTE method declined the ANN results, only the combined PCA with SMOTE method has significantly improved the ANN results. Additionally, we noticed that the SVM and DT algorithms results were improved

using all the three methods which we used in this study. Results obtained from the K-NN algorithm showed improvement using the SMOTE and the combined PCA with SMOTE, while no improvement was obtained using the PCA method.

Regarding the third dataset, all models provided significant results. In this dataset the best classification results were obtained from the SVM model from row data. However, when the PCA method was used the best model was ANN. For the SMOTE and the combined PCA with SMOTE method, the best model was SVM. When the PCA was used for the third dataset, three algorithms (NB, SVM and K-NN) provided poor results but the other three algorithms (DT, LR and ANN) were better. Using the SMOTE and the combined (PCA with SMOTE) methods have significantly improved the classification performances for all algorithms.

The best algorithm for the fourth dataset when raw data was used obtained from the SVM algorithm. For the PCA method, the best algorithm was the LR. However, the best results were obtained from the SVM algorithm under the use of the SMOTE method, but for the combined (PCA with SMOTE) methods, the DT algorithm was the best model. In this dataset, results from NB and SVM algorithms were declined under the use of the PCA method. However, when the SMOTE and the combined (PCA with SMOTE) methods were applied, all algorithms were significantly improved in terms of the classification performances.

Regarding the fifth dataset, the SVM algorithm in raw data and in the PCA was the best model compared to other algorithms. When SMOTE method was used, results obtained from the LR algorithm were better than the other algorithms. Under the use of the combined PCA with SMOTE method, the K-NN algorithm was superior. However, the results of the NB algorithms were declined at PCA method, while using SMOTE and the combined PCA with SMOTE method, the NB algorithm has significantly improved to better provide classification results. In the same dataset, the LR algorithm gave better result in PCA and SMOTE methods. All the three methods have improved the ANN algorithm results. In contrast, the SVM algorithm results were declined using the three methods. The results of the K-NN algorithm has improved under the use of the three proposed methods; however, the improvement was higher when using SMOTE and the combined PCA with SMOTE methods.

All algorithms showed improvement in results using SMOTE, PCA, and combined (PCA with SMOTE) methods in dataset six. In this dataset, the best results of raw data were obtained from SVM and K-NN algorithms. For the PCA method, the best model was NB and K-NN. The ANN, SVM and K-NN algorithms have provided better results under the use of the SMOTE method. The best algorithm was the NB for the combined PCA with SMOTE method. When the PCA was used, the ANN algorithms provided poor results but the other models gave significant results. When the SMOTE and the combined PCA with SMOTE methods were used, all algorithms showed significant improvement.

The experimental results from analyzed data indicated that when using PCA, SOMTE, or PCA+SMOTE methods, the classification performance of datasets are improved. However, when using these three methods, the SVM and K-NN models showed higher performance compared to other algorithms (Mustafa et al, 2017).

We hope that this study will contribute to the studies to be done with high dimensional unbalanced data in different fields. Future works may thus concentrate on applying this method to other real-world problems. In the future we will use other method of dimensionality reduction with SMOTE method, such as Nonlinear Principal Component Analysis (NLPCA), Independent Component Analysis (ICA) and Locally Linear Embedding (LLE). In this study we had used one method of oversampling used. As another feature work, one can use other methods such as ADASYN Adaptive Synthetic Sampling method for imbalanced data and Gaussian Distribution oversampling method instated of SMOTE method. In addition, it can be used in combination with the SMOTE method in methods such as NLPCA, ICA and LLE in the classification of larger data sets with more complex higher dimensional and class imbalances. In this experimental study, we only used six different datasets, but in the future, we will use more complex and big datasets to further check the efficiency of our method. Thus, besides an effective application in the field of health, it will be beneficial by developing an effective strategy in diagnosing and defining more accurate models in diagnosis and diagnosis.

REFRENCES

Abdoh, S. F., Rizka, M. A., Maghraby, F. A., 2018. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. ***IEEE Access, 6***: 59475-59485.

Agrawal, S., Agrawal, J., 2015. Survey on anomaly detection using data mining techniques. ***Procedia Computer Science, 60***: 708-713.

Baitharu, T. R., Pani, S. K., 2015. A comparative study of data mining classification techniques using lung cancer data. ***IJCTT, 22***(2): 91-95.

Baran, M., 2020. ***Makine Öğrenmesi Yöntemleriyle Çoklu Etiketli Verilerin Sınıflandırılması*** (Master thesis). Sivas Cumhuriyet Üniversitesi, Sivas.

Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A., Herrera, F., 2018. SMOTE-BD: An exact and scalable oversampling method for imbalanced classification in big data. ***Journal of Computer Science & Technology, 18***(3): 203-209.

Beniwal, S., Arora, J., 2012. Classification and feature selection techniques in data mining. ***IJERT, 1***(6): 1-6.

Bin Othman, M. F., Yau, T. M. S., 2007. Comparison of different classification techniques using WEKA for breast cancer. ***Biomed 06, IFMBE Proceedings 15*** (Editors: F. Ibrahim, N.A. Abu Osman, J. Usman, N.A. Kadri). Springer-Verlag Berlin Heidelberg, pp.520-523.

Buda, M., Maki, A., Mazurowski, M. A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. ***Neural Networks, 106***: 249-259.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique. ***Journal of Artificial Intelligence Research, 16***: 321-357.

Chetana, V. L., Kolisetty, S. S., Amogh, K., 2020. A Short Survey of Dimensionality Reduction Techniques. ***RACSPA-2019.*** 21-22 October 2019, Andhra Pradesh, India. 3-14.

Farizawani, A. G., Puteh, M., Marina, Y., Rivaie, A., 2020. A review of artificial neural network learning rule based on multiple variant of conjugate gradient approaches. ***2nd JICETS 2019***. 25-27 November 2019, Bandung, Indonesia.1-13.

Fix, E., Hodges, J. L., 1951. ***Discriminatory Analysis: Nonparametric Discrimination, consistency properties.*** Prepared at the University of California. Contract No, AF41, Texas. 43.

Gundecha, P., Liu, H., 2012. Mining social media: a brief introduction. ***https://pubsonline.informs.org/doi/pdf/10.1287/educ.1120.0105. INFORMS TutORials in Operations Research.*** Date accessed: 25.08.2020.

Haghanikhameneh, F., Panahy, P. H. S., Khanahmadliravi, N., Mousavi, S. A., 2012. A comparison study between data mining algorithms over classification techniques in squid dataset. ***IJAI, 9***(12): 59-66.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The WEKA data mining software: an update. ***SIGKDD Explorations, 11***(1): 10-18.

Iyer, A., Jeyalatha, S., Sumbaly, R., 2015. Diagnosis of diabetes using classification mining techniques. ***IJDKP, 5***(1): 1–14.

Kambhatla, N., Leen, T. K., 1997. Dimension reduction by local principal component analysis. ***Neural Computation, 9***(7): 1493-1516.

Kaur, G., Oberai, E. N., 2014. Naive Bayes classifier with modified smoothing techniques for better spam classification. ***IJCSMC, 3***(10): 869-878.

Lewis, D. D., 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. ***Machine Learning: ECML-98, 10th European Conference on Machine Learning Chemnitz.*** 21-23 April 1998, Germany. 4-15.

Lin, Y. T., Chen, Y. H., Yang, Y. H., Jao, H. C., Abiko, Y., Yokoyama, K., Hsu, C., 2010. Heme oxygenase-1 suppresses the infiltration of neutrophils in rat liver during sepsis through inactivation of p38 MAPK. ***Shock, 34***(6): 615-621.

Liu, D., Sun, R., Ren, H., 2020. Efficient Fraud Detection Classification: Class Imbalance and Attribute Correlations. ***The Frontiers of Society, Science and Technology, 2***(11): 96-103.

Liu, S., d'Aquin, M., 2017. Unsupervised learning for understanding student achievement in a distance learning setting. ***In 2017 IEEE Global Engineering Education Conference (EDUCON***). 25-28 Apr 2017, Athens, Greece.1373-1377.

Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P., Ho, T. K., 2019. How Complex is your classification problem? A survey on measuring classification complexity. ***ACM Comput. Surv., 52***(5): 1-34.

Luque, A., Carrasco, A., Martín, A., de las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. ***Pattern Recognition, 91***: 216-231.

Maertens, R. M., Long, A. S., White, P. A., 2017. Performance of the in vitro transgene mutation assay in MutaMouse FE1 cells: Evaluation of nine misleading ("False") positive chemicals. ***Environ. Mol. Mutagen., 58***(8): 582-591.

Maldonado, S., López, J., Vairetti, C., 2019. An alternative SMOTE oversampling strategy for high-dimensional datasets. ***Applied Soft Computing Journal, 76***: 380-389.

Markov, Z., Russell, I., 2006. An introduction to the WEKA data mining system. ***ACM SIGCSE Bulletin, 38***(3): 367-368.

Mitchell, T. M., 1999. Machine learning and data mining. ***Communications of the ACM, 42***(11): 30-36.

Mohammed, A. J., Hassan, M. M., Kadir, D. H., 2020. Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method. ***International Journal of Advanced Trends in Computer Science and Engineering, 9***(3): 3161-3172.

Mohammed, M., Khan, M. B., Bashier, E. B. M., 2017. ***Machine Learning: Algorithms and Applications***. *Crc Press,* Boca Raton, 206.

Mustafa, N., Memon, E. R. A., Li, J. P., Omer, M. Z., 2017. A classification model for imbalanced medical data based on PCA and farther distance based synthetic minority oversampling technique. ***IJACSA***, ***8***(1): 61-67.

Mythili, M. S., Shanavas, A. M., 2014. An Analysis of students' performance using classification algorithms. ***IOSR-JCE, 16***(1): 63-69.

Naseriparsa, M., Kashani, M. M. R., 2014. Combination of PCA with SMOTE resampling to boost the prediction rate in lung cancer dataset. ***International Journal of Computer Applications, 77***(3): 33-38.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. ***IJCTT, 48***(3): 128-138.

Peng, C. Y. J., Lee, K. L., Ingersoll, G. M., 2002. An introduction to logistic regression analysis and reporting. ***The journal of Educational Research, 96***(1): 3-14.

Prati, R. C., Batista, G. E., Monard, M. C., 2009. Data mining with imbalanced class distributions: concepts and methods. ***4th Indian International Conference on Artificial Intelligence (IICAI-09)***. 16-18 December 2009, Tumkur, India. 359-376.

Scott, M., Plested, J., 2019. GAN-SMOTE: A Generative Adversarial Network approach to Synthetic Minority Oversampling for One-Hot Encoded Data. ***Australian Journal of Intelligent Information Processing Systems, 15***(2): 29-35.

Samanthula, B. K., Elmehdwi, Y., Jiang, W., 2015. K-nearest neighbor classification over semantically secure encrypted relational data. ***IEEE Transactions on Knowledge and Data Engineering, 27***(5): 1261-1273.

Santos, M. F., Cortez, P., Pereira, J., Quintela, H., 2006. Corporate bankruptcy prediction using data mining techniques, Section 9. ***Data Mining VII. Data, Text and Web Mining and Their Business Applications*** (Editor: A. Zanasi, C. A. Brebbia, N. F. F. Ebecken). WIT Press, Southampton. 460.

Silipo, R., Adae, I., Hart, A., Berthold, M., 2014. Seven Techniques for Dimensionality Reduction Missing Values, Low Variance Filter, High Correlation Filter, PCA, Random Forests, Backward Feature Elimination, and Forward Feature Construction. ***https://studylib.net/doc/18241563/seven-techniques-for-dimensionality-reduction.*** KNIME. Date accessed: 12.07.2020.

Sun, Y., Wong, A. K. C., Kamel, M. S., 2009. Classification of imbalanced data: A review. ***International Journal of Pattern Recognition and Artificial Intelligence, 23***(4): 687-719.

Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ***Performance Evaluation Review, 41***(4): 70-73.

Suthaharan, S., 2016. Machine learning models and algorithms for big data classification. ***Integr. Ser. Inf. Syst, 36***: 1-12.

Swapna, G., Vinayakumar, R., Soman, K. P., 2018. Diabetes detection using deep learning algorithms. ***ICT Express, 4***(4), 243-246.

Tahir, M. A. U. H., Asghar, S., Manzoor, A., Noor, M. A., 2019. A classification model for class imbalance dataset using genetic programming. ***IEEE Access, 7***: 71013-71037.

Vapnik, V. N., 2000. ***The Nature of Statistical Learning Theory*** Second Edition. Springer, New York. 314.

Vidhya, K. A., Aghila, G., 2010. A Survey of Naïve Bayes Machine Learning approach in Text Document Classification. ***IJCSIS, 7***(2): 206-211.

Wang, H., Shi, Y., Zhou, X., Zhou, Q., Shao, S., Bouguettaya, A., 2010. Web service classification using support vector machine. ***2010 22nd IEEE International Conference on Tools with Artificial Intelligence.*** 27-29 Oct. 2010, Arras, France. 3-6.

Yildiz, M., Bozdemir, M. N., Kiliçaslan, I., Ateşçelik, M., Gurbuz, Ş., Mutlu, B., Onur, M. R., Gürger, M., 2012. Elderly trauma: the two years experience of a university-affiliated emergency department. ***European Review for Medical and Pharmacological Sciences, 16***(Suppl 1): 62-67.

Yoon, J. W., Friel, N., 2015. Efficient model selection for probabilistic K nearest neighbor classification. ***Neurocomputing, 149***(PB): 1098-1108.

Zhang, Z., 2014. Too much covariates in a multivariable model may cause the problem of overfitting. ***Journal of Thoracic Disease, 6***(9): 196-197.

Zheng, X., 2020. ***SMOTE Variants for Imbalanced Binary Classification: Heart Disease Prediction*** (Master thesis). University of California, Los Angeles.

(Genişletilmiş Türkçe Özet)

EXTENDED TURKISH SUMMARY

YÜKSEK BOYUTLU DENGESİZ VERİLERİN SINIFLANDIRILMASI İÇİN SMOTE AŞIRI ÖRNEKLEME İLE PCA'NIN KOMBİNASYONU

MULLA, Guhdar Abdul-Aziz Ahmed
Yüksek Lisans Tezi, İstatistik Bölümü
Tez Danışmanı: Dr. Öğr. Üyesi Yıldırım DEMİR
İkinci Danışman: Dr. Masoud Muhammed Hassan
Ocak 2021, 68 sayfa

ÖZ

Dengesiz verilerin sınıflandırması, sınıflandırıcıların daha büyük veri sınıfına doğru çarpıtıldığı veri madenciliğinde yaygın bir konudur. Yüksek boyutlu çarpık (dengesiz) verilerin sınıflandırılması, daha zor olduğundan karar vericiler için büyük ilgi görmektedir. Değişkenlerin azaltıldığı bir süreç olan boyut indirgeme yöntemi, yüksek boyutlu veri setlerinin belirli bir kayıpla daha kolay yorumlanmasını sağlamaktadır. Ayrıca, yüksek boyutlu dengesiz verilerin sınıflandırılması tekrarlanan bir sorun haline gelmiştir. Bu çalışmada, yüksek boyutlu verilerde dengesizlik problemini çözmek için SMOTE aşırı örnekleme ile Temel Bileşen Analizini (PCA) birleştiren yeni bir yöntem önerilmiştir. Önerilen yöntemin etkinliğini kontrol etmek ve sınıflandırıcıların performansını belirlemek için Karar Ağacı (DT), Destek Vektör Makineleri (SVM), En Yakın Komşu (K-NN), Naive Bayes (NB), Lojistik Regresyon (LR) ve Yapay Sinir Ağlarından (ANN) oluşan altı sınıflandırma algoritması ve altı farklı veri kümesi kullanılmıştır. Sırasıyla, ham veri setleri, PCA, SMOTE ve SMOTE+PCA yöntemleriyle dönüştürülen veri setleri verilen algoritmalarla analiz edilmiştir. Analizler WEKA programlama dillerinden yararlanılarak yapılmıştır.

Analiz sonuçları, neredeyse tüm sınıflandırma algoritmalarının PCA, SOMTE ve SMOTE+PCA yöntemlerini kullanarak sınıflandırma performanslarını iyileştirdiğini göstermektedir. Bununla birlikte, SMOTE yöntemi, verilerin yeniden dengelenmesi için PCA ve PCA+SMOTE yöntemlerinden daha etkili sonuçlar vermiştir. Ayrıca deneysel sonuçlar, SVM ve K-NN sınıflandırıcılarının diğer algoritmalara kıyasla daha yüksek sınıflandırma performansı sağladığını göstermektedir.

Anahtar kelimeler: Boyut indirgeme, Dengesiz sınıflar, Makine öğrenimi, PCA, SMOTE Aşırı örnekleme, Sınıflandırma.

1. GİRİŞ

Veri madenciliği, verilerden yararlı bilgi ve modeller çıkarma sürecidir (Gundecha ve Liu, 2012). Ayrıca, veri madenciliği büyük ölçekli verilerden yararlı veya eyleme geçirilebilir bilgiyi keşfetme süreci olarak da tanımlanmaktadır (Osisanwo ve ark., 2017). Farklı bilim alanlarında ortaya çıkan bir dizi problemi çözmek için çeşitli veri madenciliği teknikleri ve algoritmaları önerilmiştir. Sınıflandırma, veri madenciliğinde önerilen bu algoritmalardan birisi olup farklı araştırma alanlarında yaygın olarak kullanılmaktadır. Sınıflandırmanın amacı, yeni vakaların sınıfını tahmin edebilen modeller oluşturmaktır. Sınıflandırma ile yeni vakaların sınıfını öngörebilecek modellerin oluşturulması amaçlanmaktadır. Bu veri madenciliği modelleri, verilerin daha kolay anlaşılması ve incelenmesi için kullanılan matematiksel ifadelerdir (Santos ve ark., 2006). Veri madenciliği ve makine öğrenimi teknikleri, büyük miktarda verilerden ilginç, açıklanamayan, gizli özellikleri çıkarmak ve keşfetmek için kullanılmaktadır. Örneğin, tıp alanında veri madenciliği yöntemleri, örüntü tanımlama, gruplama, kümeleme ve tahmin için yaygın olarak kullanılmaktadır ve bunlar, bu yöntemlerin diğer bilim dallarına uygulanmasının temelini oluşturmaktadır (Swapna ve ark., 2018).

Literatürde birçok farklı sınıflandırma algoritması bulunmaktadır. Karar Ağacı (DT), Destek Vektör Makineleri (SVM), K-En Yakın Komşu Yöntemi (KNN), Naive Bayes (NB), Lojistik Regresyon (LR) ve Yapay Sinir Ağları (ANN), sınıflandırma işlemi için en önemli ve en yaygın kullanılan makine öğrenme algoritmalarıdır. Bu gelişmiş sınıflandırma algoritmalarının çoğu, tipik olarak deneysel eğitim verileri ve yeni test verileriyle yüksek bir sınıflandırma doğruluğu sağlamaktadır. Ancak sınıflandırma algoritmalarındaki farklılık ve dengesiz verilerin sınıflandırılması gibi konuların araştırılması, sınıflandırma performansının iyileştirilmesi için önemlidir (Lorena ve Lehmann, 2019).

Sınıflandırma algoritmalarındaki en zor problemlerden birisi dengesiz verilerin sınıflandırılmasıdır. Zira örneklem büyüklükleri her iki sınıfta eşit olmadığında ikili sınıflandırmada bir sorun ortaya çıkmaktadır. Bununla birlikte, pozitif ve negatif gruplardan alınan numuneler arasında çok az fark varsa, bu çok önemli bir sorun olmayabilir (Tahir ve

ark., 2019). Dengesiz verileri sınıflandırmak için önce aşırı yüksek veya düşük derecede örnekleme kullanılmalıdır. Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE; Synthetic Minority Oversampling Techniques), veri setini yeniden dengelemek için kullanılan bir yöntemdir. Bu yöntem, aşırı örneklemeye dayalı dengesiz veri dağılım sorunu için optimal bir çözümü sunmaktadır. Temel SMOTE varsayımı, azınlık ve çoğunluk sınıflarının özellikleri arasındaki paralelliklerin nasıl bulunacağına dayanmaktadır (Mustafa ve ark., 2017).

Sınıflandırma ile ilgili bir diğer sorun ise veri kümesinde çok sayıda düşük etkili değişkenin bulunması yani yüksek boyutluluktur. Bu sorunun çözümü için boyut indirgeme yöntemlerinden biri olan Temel Bileşen Analizi (PCA) kullanılabilir. Böylece istatistiki olarak anlamsız olan bazı değişkenler veri setinden çıkarılarak veri seti boyutu azaltılmaktadır (Kambhatla ve Leen, 1997).

Bu araştırmada yüksek boyutlu dengesiz veriler için sınıflandırma probleminin çözülmesi amaçlanmıştır. Çalışmada, eş zamanlı olarak yeniden dengeleme ile gereksiz özelliklerin sayısı azaltılarak (PCA ve SMOTE yöntemleri kullanılarak) sınıflandırma problemi araştırılmıştır. Bu amaç doğrultusunda, çoğunluk ve azınlık örnek sayılarından oluşan farklı dengesiz veri kümeleri için altı iyi bilinen sınıflandırma algoritması (DT, SVM, K-NN, NB, LR ve YSA) kullanılmıştır. Veri analizlerinde, grafik kullanıcı arabirimi üzerinden erişilebilen denenmiş/test edilmiş bir açık kaynak veri madenciliği yazılımı olan WEKA programı kullanılmıştır.

2. MATERYAL VE YÖNTEM

Araştırma materyali olarak, negatif sınıf ve pozitif sınıf arasındaki dengesiz yüzde değerlerine göre (veri setlerine ait dengesizlik durumu %2.6 ile %92 arasında değişmektedir) seçilen altı farklı veri seti kullanılmıştır. Bu veri setleri https://www.kaggle.com/datasets adresinden alınmıştır. Yüksek boyut ve sınıf dengesizliğine sahip bu veri setleri kullanılarak önerilen yöntemlerin güvenilirliği ve etkinliği araştırılmıştır.

Yüksek boyutlu ve sınıf dengesizliği problemine sahip verilerin sınıflandırılması, bu iki problemden dolayı genellikle kolay değildir. Bu nedenle boyut indirgeme için PCA ve yeniden dengeleme için ise SMOTE yöntemleri kullanılmıştır. Ayrıca bu çalışmada, yüksek

boyutluluk ve verilerde sınıf dengesizliği problemlerini aynı anda ele almak için bu iki yöntem birleştirilerek elde edilen yeni yöntem ile en iyi model belirlenmeye çalışılmıştır.

Önerilen yöntemde (PCA+SMOTE) ilk önce orijinal veri seti kullanılmakta ve ardından veri setinin dengesiz olup olmadığını kontrol edilmektedir. Veri seti dengesizse, SMOTE yöntemini ile veri seti dengelendikten sonra PCA ile veri setinin boyutu indirgenmektedir. Veriler dengelendikten ve boyutları düşürüldükten sonra, eğitim ve test (%34'ü test ve %66'sı eğitim) veri kümesi olmak üzere ikiye ayrılmıştır. En son olarak, veri seti altı sınıflandırma algoritması (NB, LR, ANN, SVM, K-NN, DT) kullanarak sınıflandırılmaktadır.

Uygulamaların tümünde ilk önce veri setlerine herhangi bir yöntem uygulanmadan ham veriler ile sınıflandırma algoritmalarının performansları Doğruluk, F ve ROC alan (ROC eğrisi altında kalan alan) değerlerine göre kontrol edilmiştir. Daha sonra sırasıyla, boyut indirgeme yöntemi (PCA), veri dengeleme yöntemi (SMOTE) ve boyut indirgeme ile veri dengeleme yöntemleri (PCA+SMOTE) beraber kullanılarak dönüştürülen veri setleri ile sınıflandırma algoritmalarının performansları üç ölçüm değerine göre kontrol edilmiş ve elde edilen sonuçlara göre sınıflandırıcıların performansları karşılaştırılmıştır. Böylece, her bir veri seti için bir uygulama yapılmış ve her bir uygulama da dört alt uygulamadan oluşmuştur.

3. BULGULAR

Birinci uygulama için kalp krizi (veri seti 1) veri seti kullanılmış ve analiz sonuçları tablo 4.2'de verilmiştir. Tablo 4.2'de ham veriler için en yüksek Doğruluk, F-ölçüm ve ROC alanına ait oranlar %100 K-NN algoritması ile elde edilmiştir; hesaplanan üç ölçüme göre ise en düşük oranlar %79.6 ile SVM'den elde edilmiştir. Ayrıca, PCA, SMOTE ve PCA+SMOTE yöntemleri kullanıldığında da en iyi performansı K-NN göstermiştir. En yüksek Doğruluk, F ölçüm ve ROC alan oranları PCA için sırasıyla %99.7, %99.7 ve %99.3 olarak, SMOTE için sırasıyla %99.8, %99.8, %100 olarak ve PCA+SMOTE için sırasıyla % 99.5, %99.5 ve %99.6 olarak hesaplanmıştır. Böylece birinci uygulamanın dört alt uygulaması için de en iyi sınıflandırma performansını K-NN gösterirken, SMOTE yönteminde K-NN diğer iki (PCA ve PCA+SMOTE) yöntemden daha iyi performans göstermiştir.

İkinci uygulamada kalp hastalarına ait (veri set 2) veri seti kullanılmış ve analiz sonuçları tablo 4.3'de verilmiştir. Tablo 4.3'de ham veriler için en yüksek Doğruluk ve F-ölçüm değerleri sırasıyla %81.1 ve %81.2 ANN ile ve en yüksek ROC alan değeri ise %89.4 NB ile elde edilmiştir. PCA yöntemi kullanıldığında, LR ve SVM'nin tüm ölçümlerine göre, DT'nin ise Doğruluk ve F ölçümüne göre sınıflandırma performanslarında bir iyileşme izlenirken, ancak diğer üç sınıflandırıcıya ait performanslarda bir gerileme izlenmiştir. SMOTE yöntemi kullanıldığında, üç ölçüm değerine göre de sadece ANN sınıflandırıcısının performansında ve DT'nin ROC alan değerinde bir iyileşmeden ziyade bir gerileme yaşanmıştır. Ayrıca ANN, uygulamanın bu bölümünde (SMOTE) diğer iki bölümünden (PCA ve PCA+SMOTE) daha kötü bir performans göstermiştir. Son olarak, iki yöntemin kombinasyonu (PCA+SMOTE) kullanıldığında, hesaplanan ölçüt değerlerine göre tüm sınıflandırıcı performanslarında bir iyileşme gözlemlenmiştir.

Üçüncü uygulamada meme kanseri (veri set 3) veri seti kullanılmış ve analiz sonuçları tablo 4.4'de verilmiştir. Tablo 4.4'de sınıflandırma algoritmalarının tümünün, ham veriler ve diğer üç yöntem ile çok iyi bir performans gösterdiği görülmektedir. Ham veriler kullanıldığında, Doğruluk, F-ölçüm ve ROC alanı için en yüksek oranların sırasıyla %97.3, %97.3 ve %99.6 olarak SVM ile bulunduğu ve buna göre üçüncü uygulamada ham veriler için en iyi modelin SVM olduğu söylenebilir. Hesaplanan üç ölçüte göre de; PCA yönteminde sadece LR, ANN ve DT (NB, SVM ve K-NN modellerinde gerileme görülmüştür) sınıflandırıcı modellerinde bir iyileşme sağlanmışken, SMOTE yönteminde ise SVM Roc ölçümü hariç tüm sınıflandırıcı modellerinde bir iyileşme sağlanmıştır. Öte yandan, iki yöntemin (PCA + SMOTE) kombinasyonu birlikte kullanıldığında, Doğruluk ve F-ölçüm değerlerinin en yüksek oranları sırasıyla %97.5 ve %97.7 ile SVM'de elde edildiğini, ancak ROC alanı için en yüksek değer %99.6 ile LR ve ANN'den elde edildiği görülmektedir.

Dördüncü uygulamada hepatit ile ilgili (veri set 4) veri seti kullanılmış ve analiz sonuçları tablo 4.5'de verilmiştir. Tablo 4.5'de ham veri için Doğruluk ve F ölçümün en yüksek yüzdesi sırasıyla%86.1 ve%85.7 olarak SVM'den elde edilirken, en yüksek ROC alanı yüzdesi ise %86.6 ile NB'den elde edilmiştir. Öte yandan, PCA yöntemi kullanıldığında, Doğrulu ve F ölçüm değerleri için en yüksek oranlar sırasıyla %88.1 ve %88 ile LR'den, ROC alanı için en yüksek oran ise %85.9 ile NB ve LR'den elde edilmiştir. SMOTE kullanıldığında, Doğruluk ve F ölçümlerine ait en yüksek oranlar %85.7, %85.8

olarak SVM'den elde edilirken, en yüksek ROC alan yüzdesi ise %91.6 olarak LR'den elde edilmiştir. Bununla birlikte, PCA ve SMOTE ortak kullanıldığında, kullanılan algoritmaların çoğunun sınıflandırma performansında iyileşme sağlamıştır ve Doğruluk ve F ölçümlerinin en yüksek değerleri %87.4, %87.4 ile DT algoritması ile elde edilirken, ROC alanı için en yüksek değer %90.7 ile NB'den elde edilmiştir.

Beşinci uygulama için kardiyoloji kategorik (veri seti 5) veri seti kullanılmış ve analiz sonuçları tablo 4.6'de verilmiştir. Tablo 4.6'de ham verilerle yapılan analizlerde tüm sınıflandırma algoritmalarının çok iyi sonuçlar verdiği, dolaysıyla PCA, SMOTE ve PCA+SMOTE yöntemleri de güvenilir bir şekilde bu veri setine uygulanabilir. Ham veriler için en yüksek Doğruluk ve F-ölçüm değerleri sırasıyla %90.5, %90.3 olarak SVM'den elde edilirken, en yüksek ROC alan performansı ise %91.9% ile NB'den elde edilmiştir. PCA yöntemini kullanıldığında, Doğruluk ve F ölçümü için en iyi performan SVM algoritmasına ait olmasına rağmen SVM ve NB sınıflandırıcıları ilk duruma göre iyileşme göstermemiştir. Ayrıca, SMOTE yöntemini kullanıldığında, SVM hariç diğer tüm sınıflandırıcı algoritmalarında bir iyileşme görülmektedir. Öte yandan, iki yöntem aynı anda kullanıldığında, en yüksek Doğruluk ve F-ölçüm değerini K-NN verirken, en yüksek ROC alan değerini ise LR vermektedir.

Altıncı uygulamada için Lenfografi-normal-fibroz (veri seti 6) veri seti kullanılmış ve analiz sonuçları tablo 4.7'de verilmiştir. Tablo 4.7'de ham veriler için DT, ROC alan performansı hariç genel olarak tüm sınıflandırıcı algoritmaları çok iyi bir performans göstermiştir. PCA yöntemi kullanıldığında, NB, K-NN ve LR, sınıflandırma performansında önemli gelişme görülürken, özellikle ANN algoritmasında bir gerileme görülmektedir. Ayrıca DT, ROC ölçüm değerinde de yüzde yüze yakın bir ilerleme gözlenmiştir. SMOTE yöntemi kullanıldığında, üç ölçüme göre de LR, ANN, SVM ve K-NN sınıflandırıcılarının performanslarında bir ilerleme görünmüştür. PCA+SMOTE kombinasyonu kullanıldığında, tüm sınıflandırma algoritmaları kullanılan değerlendirme ölçütlerine göre çok iyi bir performans göstermiştir. Bu veri seti için tüm alt uygulamalardan çok iyi sonuçlar ve özellikle de PCA+SMOTE uygulamasında mükemmel sonuçlar elde dilmiştir.

4. TARTIŞMA ve SONUÇ

Altı dengesiz yüksek boyutlu veri üzerindeki altı farklı sınıflandırma algoritmasına ait deneysel sonuçlara göre, PCA, SMOTE ve PCA+SMOTE yöntemlerini kullanan tüm

sınıflandırma algoritmalarının genel olarak veri setlerinin sınıflandırma performansını artırdığı söylenebilir. Ayrıca, sınıflandırma algoritmalarının performansı veri setine göre değişiklik göstermekle birlikte bu üç yöntem kullanılarak yapılan analizlerde SVM ve K-NN modellerinin diğer dört modele göre daha yüksek bir performans gösterdiği belirlenmiştir. Sonuç olarak, boyut indirgendikten ve veriler yeniden dengelendikten sonra, sınıflandırma algoritmalarına ait performansların önemli derecede iyileştiği ve böylece yüksek boyutlu verilerde dengesizlik sorunuyla baş etmede önerilen yöntemin önemini olduğu görülmektedir.

Bu çalışmanın, farklı alanlarda yüksek boyutlu dengesiz verilerle yapılacak çalışmalara katkı sağlayacağını ümit etmekteyiz. Ayrıca, daha karmaşık yüksek boyutlu ve sınıf dengesizliğine sahip büyük veri kümelerinin sınıflandırılmasında, doğrusal olmayan temel bileşen analizi (NLPCA) ve bağımsız bileşen analizi (ICA) gibi yöntemlerde SMOTE yöntemi ile kombine edilerek kullanılabilir. Böylece, özellikle sağlık alanında etkili bir uygulamanın yanı sıra tanı ve teşhis konusunda hızlı ve daha doğru model tanımlamada etkili bir strateji geliştirerek fayda sağlayacaktır.

CURRICULUM VITAE

Guhdar Abdul-Aziz Ahmed Mulla was born in 1992, in Duhok province, Iraq. He completed primary, secondary and high school in Duhok district. He graduated from Statistic Department in Duhok University - College of Administration and Economic in 2016. He started his Postgraduate study at the Department of Statistic, Institute of Natural Applied Sciences at Van Yuzuncu Yil University in Turkey on February 2019.