ISTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

MACHINE LEARNING APPROACH FOR PREDICTING SEVERITY OF OBSTRUCTIVE SLEEP APNEA SYNDROME

M.Sc. THESIS

Onurhan HAMZAOĞLU

Department of Computational Science and Engineering

Computational Science and Engineering Programme

DECEMBER 2019

ISTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

MACHINE LEARNING APPROACH FOR PREDICTING SEVERITY OF OBSTRUCTIVE SLEEP APNEA SYNDROME

M.Sc. THESIS

Onurhan HAMZAOĞLU
(702161009)

Department of Computational Science and Engineering

Computational Science and Engineering Programme

Thesis Advisor: Prof. Dr. Fethiye Aylin SUNGUR

DECEMBER 2019

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ BİLİŞİM ENSTİTÜSÜ

OBSTRÜKTİF UYKU APNESİNİN ŞİDDETİNİN TAHMİNLENMESİNDE MAKİNE ÖĞRENMESİ YAKLAŞIMI

YÜKSEK LİSANS TEZİ

Onurhan HAMZAOĞLU
(702161009)

Hesaplamalı Bilim ve Mühendislik Anabilim Dalı

Hesaplamalı Bilim ve Mühendislik Programı

Tez Danışmanı: Prof. Dr. Fethiye Aylin SUNGUR

ARALIK 2019

Onurhan HAMZAOĞLU, a **M.Sc**. student of İTU Informatics Institute student ID 702161009, successfully defended the thesis entitled **"MACHINE LEARNING APPROACH FOR PREDICTING SEVERITY OF OBSTRUCTIVE SLEEP APNEA SYNDROME"**, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Fethiye Aylin SUNGUR** İstanbul Technical University	.....
Jury Members :	**Assoc. Prof. Adem TEKİN** Istanbul Technical University	.....
	Assoc. Prof. H. Hakan GÜREL Kocaeli University	.....

Date of Submission : 15 November 2019
Date of Defense : 13 December 2019

To my family and friends,

FOREWORD

First of all I would like to thank my advisor Prof. Dr. F. Aylin Sungur for her open-mindness and encouragements throughout this journey. Whichever excuse I came, she gave me courage and guidance with her optimism and advice.

Secondly, I am really thankful for my aunt Prof. Dr. Sebahat Genç for sharing this idea with me in the first place, and be there for me whenever I need. Also, I would like to thank Associate Prof. Mehmet Ali Habeşoğlu for sharing his research, and giving really precious insights.

Also, I am very happy to have such a wonderful friend Burcu Ulutaş who supports me at every moment in this research. Finally, and most importantly, I cannot be grateful enough to my family for being there for me all the time, from beginning to the end. Without them, none of these would be possible. I am one of the luckiest man alive to have such an amazing mother Elif Hamzaoğlu, and father Hızır Hamzaoğlu.

December 2019

Onurhan HAMZAOĞLU
(Data Scientist)

TABLE OF CONTENTS

Page

FOREWORD ..... **ix**
TABLE OF CONTENTS ..... **xi**
ABBREVIATIONS ..... **xiii**
SYMBOLS ..... **xv**
LIST OF TABLES ..... **xvii**
LIST OF FIGURES ..... **xix**
SUMMARY ..... **xxi**
ÖZET ..... **xxv**
1. INTRODUCTION **1**
1.1 Literature Review 2
1.3 Aim of the Study 3
2. OBSTRUCTIVE SLEEP APNEA SYNDROME **5**
2.1 Overview 5
2.2 Symptoms 5
2.3 Causes & Risk Factors 6
2.4 Diagnosis 6
2.5 Treatment 7
3. CLASSFICATION ALGORITHMS IN MACHINE LEARNING **9**
3.1 Logistic Regression 9
3.2 Nãive Bayes 10
3.3 K-Nearest Neighbors 10
3.4 Decision Trees 11
3.5 Ensemble Methods 12
3.5.1 Random forests 12
3.5.2 AdaBoost 12
3.5.3 Voting 13
3.6 Multi-layer Perceptron 13
3.7 Cross Validation 14
3.8 Evaluation Metrics 14
3.9 Sampling Methods 16
4. RESULTS AND DISCUSSION **21**
4.1 Development Environment 21
4.2 Gathering Data 22
4.3 Exploratory Data Analysis 23
4.4 Feature Selection 27
4.5 Model Creation 31
4.5.1 Multiclass prediction results 32
4.5.2 Binary prediction results 36
5. CONCLUSION **39**
REFERENCES **41**
CURRICULUM VITAE **45**

ABBREVIATIONS

AUC : Area Under Curve
AHI : Apnea-Hypopnea Index
ANN : Artificial Neural Network
DT : Decision Tree
ESS : Epworth Sleepiness Scale
KNN : K-Nearest Neighbors
LR : Logistic Regression
ML : Machine Learning
MLP : Multi-layer Perceptron
NB : Näive Bayes
OSA : Obstructive Sleep Apnea
PSG : Polysomnography
RF : Random Forest
ROC : Reciever-Operator Characteristic
RUS : Random Under Sampling
SVM : Support Vector Machine
SMOTE : Synthetic Minority Over-Sampling Technique

SYMBOLS

θ^T	**:** Weight vector
x	**:** Input vector
d	**:** Distance Function
i	**:** *i*th sample
P(C|*x*)	**:** Conditional probability of C given input vector x
P(C)	**:** Prior probability of class C

LIST OF TABLES

Page

Table 3.1 : Table with single row and centered columns. **15**
Table 4.1 : Feature names, descriptions and data types. **22**
Table 4.2 : Numerical and categorical variables. **23**
Table 4.3 : Train and test sets. **24**
Table 4.4 : Descriptive statistics of numerical features. **24**
Table 4.5 : ANOVA F-test results. **29**
Table 4.6 : Contingency table daytime_sleepiness vs asphyxiation. **29**
Table 4.7 : Significant dependent categorical features. **30**
Table 4.8 : Selected features. **31**
Table 4.9 : Base model performances. **32**
Table 4.10 : Number of instances per class after sampling. **34**
Table 4.11 : Best achieved performances after sampling 3-class. **36**
Table 4.12 : Best achieved performances after sampling 2-class. **37**

LIST OF FIGURES

Page

Figure 3.1 : Logistic function **9**
Figure 3.2 : Root, nodes, and leaves in a decision tree **11**
Figure 3.3 : Random forest created using three different decision trees.. ..... **12**
Figure 3.4 : Neural network structure. **13**
Figure 3.5 : 5-Fold Cross-Validation schema. **14**
Figure 3.6 : Undersampling and oversampling. **17**
Figure 3.7 : An example of TomekLink samples. **18**
Figure 3.8 : Oversampling using SMOTE.. ..... **18**
Figure 4.1 : Experiment workflow **21**
Figure 4.2 : Distribution of target class **23**
Figure 4.3 : Box plot of variable age.... ..... **25**
Figure 4.4 : Box plot of variable bmi. **25**
Figure 4.5 : Box plot of variable min_spo2.. ..... **25**
Figure 4.6 : Box plot of variable ess... ..... **26**
Figure 4.7 : Pairwise relationships of numerical variables... ..... **28**
Figure 4.8 : Correlation between features via heat map... ..... **30**
Figure 4.9 : Chi-squared test results using heat map... ..... **31**
Figure 4.10 : RF(left) and DT(right) feature importance. **33**
Figure 4.11 : Confusion matrices of base models **34**
Figure 4.12 : Confusion matrices of sampled data-multiclass... ..... **36**
Figure 4.13 : Confusion matrices of sampled data-binary... ..... **38**

MACHINE LEARNING APPROACH FOR PREDICTING SEVERITY OF OBSTRUCTIVE SLEEP APNEA SYNDROME

SUMMARY

Sleeping is a very important activity in human life. Considering that an average person spends one third of his or her life sleeping, we can see that it has a significant importance in our lives. The recommended sleep time for a healthy life 9-10 hours for people in puberty age, and 6-8 hours for adults. Not only the getting necessary time of sleep is important, but also the quailty of the sleep is very effective in order to feel rested as well as keeping physical and mental health balanced. In this context, awareness of sleep diseases come to the fore. Obstructive sleep apnea (OSA)is a common sleep disorder that is inability to breathe partially or completely during sleep as a result of blockage of the upper airway repeatedly. Its consequences such as hypertension, cardiovascular disesase, and heart failure may be fatal. Treatment methods of OSA may vary in severity. For *mild* patients lifestyle changes like losing weight, quit smoking, less alcohol intake are suggested. For *moderate-severe* cases, CPAP machine is the gold standart method for treatment unless surgery is necessary. The severity of OSA is decided using AHI. It is called *mild*if 5≤AHI≤15, *moderate*for 15≤AHI≤30, and *severe* if AHI≥30. Tests for diagnosing OSA take a lot of time, and expensive. It is necessary to find easy, effective solutions in detection process. Machine learning is to give a program the ability to manage a task without explicity programming it. With the increase of digital data, and computing power ML's popularity increased in recent years.The common applied ML solution fields are healthcare, e-commerce etc. In this study, ML solutions are applied to approximately 700 OSA patients with 19 features. Two approaches are done (a)Multi-class prediction (b)Binary-class prediction by counting *moderate* and *severe* classes as one class and *mild* class another. After applying data analysis and feature selection methods, and counting field experts' suggestions, 9 variables are fed into ML models which are minimun oxygen saturation(*p≤0.05)*, body mass index(*p≤0.05), ESS* (*p≤0.05)*, age as numerical features, and gender(*p≤0.05), daytime sleepiness*(*p≤0.05), diabetes*(*p≤0.05)*, and *dyslipidemia*(*p≤0.05), smoking* as categorical features.Working with imbalanced is a roadblock to get good results as well as a hard task. Some algorithms more robust to imbalanced data such as DT and RFs. After applying base models to training data, LR, RF, MLP and Voting classifiers are seem to overfit to training data since they look like able to predict different classes in nearly perfect in training data but none of them is able to predict minority class correctly on test data. Oversampling methods are come in handy to overcome this kind of situation. Random under sampling(RUS), Tomek Links, Synthetic Minority Oversampling Technique(SMOTE) and its variants i.e. SVM-SMOTE, Borderline-SMOTE are applied to data combined with boosting method. As a result, almost all models created by using sampled data showed improvement in prediction of minority class inspite of base models. RUS achieved best discrimination with respect to minorty predcition in binary-class prediction. As in mutli-class prediction, SVM-SMOTE and Borderline-SMOTE showed best improvement. Best metrics achieved for multi-class prediction

with sampling are precision 74%, recall 73%, and 58% overall accuracy. As in binary-class prediction weighted precision calculated as 97.8%, recall as 98.5% and overall accuracy as 88% and weighted accuracy as 70% with the most number of correct predictions in minority class. This study showed an improvement in OSA prediction on hard-to-learn imbalanced data using machine learning methods, and demonstrated that the discrimination power can be increased by using a limited number of variables with oversampling approaches when necessary. Using more balanced and diversed data, and increasing sample size is most likely result in improved performance.

OBSTRÜKTİF UYKU APNESİ ŞİDDETİNİN TAHMİNLENMESİNDE MAKİNE ÖĞRENMESİ YAKLAŞIMI

ÖZET

Uyku insan yaşamında çok önemli bir faaliyettir. Ortalama bir insanın hayatının üçte birini uyuyarak geçirdiğini göz önünde bulundurursak, yaşamımızda ne kadar önemli bir yeri olduğunu görebiliriz. Sağlıklı bir yaşam için önerilen uyku süresi ergenlik çağındaki insanlar için 9-10 saat arasında değişirken yetişkinler için 6-8 saat önerilir. Fiziksel ve zihinsel sağlığı dengede tutmak için, sadece gerekli miktarda uyumak değil, aynı zamanda en verimli şekilde uyumak da dinlenmiş hissetmek için çok etkilidir. Bu bağlamda uyku hastalıkları konusunda bilinçlenme ön plana çıkmaktadır.

Bu çalışma, OSA'nın şiddetinin belirlenmesinde şu faydaları sağlayacaktır. Her şeyden önce, zahmetli ve zaman alıcı testler gerek duyulmadığı taktirde elimine edilecek ve tanı maliyetleri düşürülecektir. İkincisi, tercih edilebilir olması sağlık uzmanlarına karar verme aşamasında kolaylık sağlayacaktır.

Obstrüktif(tıkayıcı) uyku apnesi OSA, üst solunum yolunun tekrar tekrar tıkanması sonucu uyku sırasında kısmen veya tamamen nefes yolunun kapanması sonucu nefes alamam durumuna bağlı yaygın bir uyku bozukluğudur. OSA'nın yaygınlığı artış göstermekte ve beraberinde bazı tehlikeler getirmektedir. Erkeklerde daha yaygın görülen OSA, ilerleyen yaşlarda görülme olasıkığı da artmaktadır. Trafik kazalarına neden olmakta, yaşam kalitesini olumsuz etkileyen OSA, aynı zamanda bazı tehlikeli hastalıkları da tetiklemektedir. Hipertansiyon, kardiyovasküler hastalıklar ve kalp yetmezliği gibi sonuçlar yaratma tehlikesi olduğundan ölümcül olabilir. OSA'nın tedavi yöntemleri hastalığın şiddetine göre değişmektedir. Hafif OSA hastaları için kilo verme, sigarayı bırakma, daha az alkol alımı gibi yaşam tarzı değişiklikleri önerilmektedir. Uyku sırasında kullanılabilecek olan ağız aletleri ile hava yolunun açık kalmasının sağlanması veya pozisyonel uykuyu korumak için araçların kullanılması tavsiye edilebilir. Orta-şiddetli vakalarda sürekli pozitif hava basıncı-CPAP makinesi cerrahi gerekmedikçe tedavi için altın standart yöntemdir. Bunun da çözüm olmadığı durumlarda boyun bölgesini genişletmek için fazlalık dokuyu alma amacıyla ameliyat yöntemleri uyguanabilir.

OSA'nın şiddeti apne-hipopne endeksi AHI kullanılarak şu şekilde sınıflandırılır:

- $5 \leq AHI \leq 15$ değeri için hafif,
- $15 \leq AHI \leq 30$ için orta,
- $AHI \geq 30$ ise şiddetli

Kişi en az 10 saniye nefes alamıyorsa bu duruma apne denir. Öte yandan, hipopne, üst solunum yolunun kısmen kapanmasına karşılık gelir ve her ikisi durum da kanda çözünen oksijen doygunluğunda düşüşe neden olur.

OSA teşhisi aşamasında uygulanantanı yöntemleri uygulanması zahmeti, çok zaman alan ve pahalı yöntemlerdir. Bu nedenle, bu zahmetli süreci azaltacak veya ortadan kaldıracak daha hızlı, kolay ve etkili çözümler bulma arayışı doğmuştur.

Makine öğrenmesi, bir programa açıkça programlamadan bir görevi yönetme yeteneği vermektir. Dijital verilerin artması ve bilgisayar teknolojilerinin gelişmesiyle birlikte ML'nin popülaritesi son yıllarda artmıştır. ML yaygın olarak sağlık, e-ticaret, robotik

, görüntü işleme vb. alanlarda kullanılmaktadır. Bu çalışmada, yaklaşık 700 OSA hastasının 19 farklı özelliğini içeren veri kümesi ile ML modelleri oluşturularak şiddet tahminlemesi yapılmıştır. Süreçte iki farklı yaklaşım izlenmiştir. (a) Çok sınıflı tahmin (b) Orta ve şiddetli sınıfları bir sınıf ve hafif sınıf diğer olmak kaydı ile yapılan ikili sınıf tahmin. Veri analizi ve özellik seçimi aşamalarında hedef değişken ile bağımsız değişkenler arasındaki ilişkiyi ölçmek amacıyla ANOVA F-testi ve Chi-kare testi gibi istatistiksel yöntemlerin uygulanması sonucu ve önceki araştırmalardan yapılan çıkarımlar ve uzmanların da fikirleri doğrultusunda 9 değişken seçilmiştir. Bunlar:

- Nümerik özellikler olarak:
    - minimum oksijen satürasyonu ($p \leq 0.05$),
    - vücut kitle indeksi ($p \leq 0.05$),
    - ESS ($p \leq 0.05$)
    - yaş.
- Kategorik özellikler olarak:
    - cinsiyet ($p \leq 0.05$),
    - gündüz uykululuk hali($p \leq 0.05$),
    - diyabet ($p \leq 0.05$)
    - dislipidemi ($p \leq 0.05$)
    - sigara kullanma durumu

Dengesiz veri setleri ile çalışmak çoğu ML yöntemi için engel yaratabilir ve sonuçları yanıltıcı olabilir. DT ve RF gibi bazı algoritmalar dengesiz veriye daha dayanıklıdır. Öğrenme verilerine temel ML algoritmaları uygulandıktan sonra, LR, RF, MLP ve Oylama sınıflandırıcıları eğitim verilerinde neredeyse mükemmel şekilde sınıfları öğrenip tahmin edebildiği gözlenmiştir. Ancak hiçbir model test verisinde aynı performansı gösteremeyip azınlık sınıfını doğru olarak tahmin edememektedir. Bu tür durumların üstesinden gelmek için aşırı örnekleme yöntemleri kullanışlıdır. Çoğunluk sınıfının sayısının azaltılarak azınlık sınıfı boyutuna getirilmesi ile dengeyi sağlayan rastgele aşağı örnekleme (RUS), Tomek Bağlantıları, ve sentetik örneklemler üreterek veri boyutunu arttıran ve azınlık sınıfını çoğunluk sınıfı seviyesene çeken yukarı örnekleme yöntemi olarak da Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) ve SVM-SMOTE, Sınırdoğrusu-SMOTE gibi SMOTE varyantları,ile veri dengesizliğini giderme ve verinin ayırıcı özelliğini arttırma amacıyla uygulandıktan sonra yeni eğitim kümeleri oluşturuldu.

Bu işlemlerin sonunda seçilen değişkenler kullanılarak, bahsedilen çoklama yöntemleri uygulandıktan sonra oluşturulan modellerin performansları ölçülmüş ve sonuçlar benzer çalışmaların sonuçları ile karşılaştırılmıştır. Sonuçlar yüzde cinsinden verilmiştir.

Sonuç olarak, örnek veriler kullanılarak oluşturulan hemen hemen tüm modeller, temel modellere rağmen azınlık sınıfının öngörülmesinde iyileşme göstermiştir. Örnekleme ile çok sınıflı tahmini için elde edilen en iyi ölçümler, Sonuç olarak, çoklu sınıf yaklaşımda, %71 genel doğruluk, %58 dengeli doğruluk, %73 ve %74 ağırlıklı hatırlama ve hassasiyet elde edildi. İkili sınıf yaklaşımda %88 genel doğruluk, %70 dengeli doğruluk %92 ve %98 ağırlıklı hatırlama ve hassasiyet elde edildi.

Bu çalışma, makine öğrenme yöntemleri kullanılarak öğrenmesi zor dengesiz veriler üzerindeki OSA tahmininde bir gelişme olduğunu göstermiştir. Gerektiğinde örnekleme yöntemleri sınıflar arasındaki dengesiz dağılımın önüne geçilmesi ile sınırlı sayıda değişken kullanılarak ayrımcılık gücünün arttırılabileceğini göstermiştir. Gelecekteki çalışmalarda daha dengeli, çeşitli ve aydınlatıcı veriler kullanmak ve örneklem büyüklüğünü arttırmak büyük olasılıkla performansın artmasının önünü açacaktır.

1. INTRODUCTION

Sleeping is a very important activity in human life. Considering that an average person spends one third of his or her life sleeping, we can see that it has a significant importance in our lives. The recommended sleep time for a healthy life 9-10 hours for people in puberty age, and 6-8 hours for adults. Not only the getting necessary time of sleep is important, but also the quality of the sleep is very effective in order to feel rested as well as keeping physical and mental healthbalanced.In this context, awareness of sleep diseases come to the fore.

Sleep diseases reduce the sleep quality of people and negatively affect their daily life and health. Most people who have or are predisposed to have a sleep disorder are not aware of it.In a study made in Canada, 2009, it is found that approximately 1 million Canadian adults shows signs of sleep apnea [1]. In a research conducted under the name of National Healthy Sleep Awareness Project, Dr. R. Bart Sangal stated that 25 million American adults suffers from sleep apnea [2]. In Turkey, it is found that the prevalence of sleep apnea was measured as 13.7% by applied Q&A tests [3].

Sleep apnea is divided into three categories.

Obstructive Sleep Apnea(OSA): OSA is the most common type of all three. It occurs as a result of blockage at the upper airway. Excessive relaxation of muscles in throat, large tonsils etc may cause OSA.

Central Sleep Apnea(CSA): CSA occurs when the brain cannot send proper signals to the muscles which are responsible for respiratory.

Mixed Sleep Apnea: Mixed sleep apnea is the situtation where OSA and CSA is seen together.

OSA can be seen in patients with different severities. The severity of OSA is divided into 3 categories with respect to apnea-hypopnea index (AHI). The diagnosis and determination of severity is done by a method called polysomnography test. Although this test is one of the most effective methods for achieving successful results, it is uncomfortable, time-consuming and expensive method in terms of application.

1.1 Literature Review

In a study by Gallo [5], the authors worked on predicting the severity of OSA both with regression methods to estimate AHI and directly class itself using classification methods. They used demographic values, test results, and symptoms as features collected from nearly 300 patients. After doing feature selection, they run their models on 19 features. One of the difficulties they encountered was working with imbalanced data in terms of label distribution. They tried to overcome this problem using over-sampling techniques. Another roadblock is the lack of number of patients since more data usually improve performance with high dimensional space. At the end, they achieved 44.7% classification accuracy on test data whereas regression results give 22.17 RMSE score as best.

Research represented by Leite [6], authors used Bayesian Network in order to diagnose OSA which is a sensible choice since they worked on a small dataset formed by 86 patients for training and 33 patients for validation. Bayesian approaches are powerful in order to find relationship between features using conditional probability provided that features are not correlated with each other. In conclusion, they achieved 80%, 81.3%, and 81.4% AUC score from multiple logistic regression, two Bayesian Network Classifiers, and Tree Augmented Naive Bayes respectively.

Sahin [7], authors of this research focused on to come up with a formula in order to predict AHI and to prevent unneeded PSG. Using 390 patients' data, they narrowed down the selected features to 5 which are BMI, waist circumference, neck circumference, oxygen saturation (SpO2), and tonsil size. They applied multiple linear regression to selected data and concluded with optimized weight parameters with each feature and 68% explanatory power.

According to research presented by Yücelbaş [8], using machine learning algorithms as well as deep learning techniques such as artificial neural networks (ANN) on electrocardiogram (ECG) signals for predicting OSA severity have promising results. Total of 10 experiments applied on two different datasets, they applied ANNs and reported average success rate of 97.2% and 90.18% respectively.

Karamanlı [9], used multi-layer perceptron (MLP) to diagnose OSA in the study on approximately 200 patients. In the study using five clinical variables, accuracyvalue reported as 86.6%.

1.2 Aim of the Study

There are two accepted methods for the diagnosis of OSA one of which is PSG test performed in laboratory whereas home testsare applied at home using portable monitors [10]. However, the evaluation of the results of the PSG test is done manually and takes 1-3 hours even when performed by a health professional [11]. On the other hand, problems that may arise during home tests such as breakage of the electrodes connected to the patient will decrease the reliability of test and cause the test to be repeated.

The difficulty in the implementation of polysomnography and home tests and expensive cost of the tests have made it necessary to find solutions that are easily applicable, accessible, effective, updateable and fast. The main purpose of this study is to create a solution to estimate the severity of OSA easily and effectively with a machine learning model to be trained using data that can be easily obtained from patients.

This study will provide the following benefits in determining the severity of OSA. First of all, the troublesome and time-consuming tests will be eliminated and diagnosis costs will be reduced. Secondly, the fact that it is preferable will facilitate health professionals' decisions. Finally, the artificial intelligence environment, which has become more popular in recent years, will be integrated into the health sector in the context of sleep apnea.

As a secondary objective of this study, it is aimed that the resulting solution will serve as a guide to artificial intelligence software that can be developed in the future.

2. OBSTRUCTIVE SLEEP APNEA SYNDROME

2.1 Overview

OSA is inability to breathe partially or completely during sleep as a result of blockage of the upper airway repeatedly. This may occur 5 to 100 times in an hour of sleep time. An OSA patient has symptoms of overweight or obesity, diabetes and hypertension together with daytime sleepiness, insomnia, snoring loudly [12,13]. These sympotms may have very serious consequences such as hypertension, cardiovascular disesase, and heart failure [12,14].

The level of seriousness of OSA varies from patient to patient. The severity of OSA is decided using apnea-hypopnea index (AHI). It is called apnea if person is not able to breathe at least 10 seconds. On the other hand, hypopnea corresponds partial closure of upper airway and both causes drop in oxygen saturation (SpO_2) in blood [15]. Using AHI value, OSA severity will be decided as: *mild* if 5≤AHI≤15, *moderate* if 15≤AHI≤30, *severe* if AHI>30 [14].

2.2 Symptoms

One of the first symptoms of OSA is loud snoring, which is followed by long and sustained breathlessness. Sleep quality decreases with sudden awakening at night due to choking sensation. With the decrease in sleep quality, the patient experiences lack of focus and sleepiness during the day. Thesemay lead to work accidents and increase in number of traffic accidents due to falling asleep behind the wheel.In a study evaluating approximately 950 truck drivers, it was concluded that traffic accidents is more likely to occur in drivers with OSA, and excessive daytime sleepiness determinants [15]. Other effects of OSA include the urge to go to the toilet at night (nocturia), headaches after waking up, feeling of depression, and decreased sexual desire. It is advised that the individual showing these symptoms should see an sleep expert as soon as possible.

2.3 Causes and Risk Factors

The relaxation of the muscles associated with the tongue and soft palate causes partial or complete obstruction of the upper respiratory tract, named sleep apnea if lasts at least 10 seconds, during sleep. The emergence of OSA can be caused by the use of substances such as alcohol and medicine, as well as physical and genetic characteristics of the person.

Obesity is the most important risk factor. Being overweight may increase the thickness of the throat circumference and increase the likelihood of sleep apnea and may cause apnea in different sleep positions. Having narrow upper airway and large tonsils are effective in the formation of OSA. Gender and age are other known factors of OSA. As stated by Garvey, the incidence of OSA is higher in men than in women and the risk of OSA increases with age [16]. In addition, the use of substances such as sedatives and antidepressants are among the factors that trigger OSA. As Punjabi has mentioned, snoring is more common in smokers and alcohol consumption facilitates the closure of the upper airway which causes the emergence of OSA [17].

2.4 Diagnosis

When diagnosing OSA, expert analyzes the symptoms within the information provided by the patient. Physical measurements such as weight and neck circumference are collected and patient's mouth, nose and chin structure are also examined. There are 2 basic tests performed during the definitive diagnosis stage which are polysomnography(PSG) and home test. The final decision is made after the patient's AHI is measured.

Polysomnogprahy: In PSG test, some features of the patient are observed with the help of a monitoring device. The patient's body movements during sleep, breathing pattern, and ambient sound are monitored. The PSG test, which is shown as the gold standard in the diagnosis of OSA, records blood oxygen levels, heart rhythm using ECG, brain signals using EEG, and eye movements using EOG [18]. It may have to be performed one or more times and requires overnight stay in a hospital environment. However, it may cause discomfort for some patients.

Home Tests: In at home sleep test, unlike PSG test, only breathing parameters are being monitored. It is aimed to record the difficulties and interruptions in breathing

during sleep and blood oxygen level. It is easier to perform than PSG, but it may require a repeat test if disconnection occurs during the test. It is 3-5 times less expensive than PSG, and the patient is sleeping in a familiar environment, reducing stress and contributing to the accuracy of the collected data [19].

2.5 Treatment

Treatment of respiratory disorders during sleep is aimed to correct apnea and hypopneas and to prevent future complications, to eliminate symptoms and to improve sleep quality. For the patients not in danger as in if they have *mild* OSA, simple lifestyle changes and behavioral treatments are recommended. We have mentioned that obesity is an OSA risk factor and it causes predisposition to OSA by causing narrowing of the upper respiratory tract and fat around the neck. In a study of 25 elderly people diagnosed with obesity, they showed a decrease in OSA severity after applied diet and weight loss studies for about three months [20]. It is shown that a 10% reduction in weight results in a 26% reduction in AHI [21]. Patients should be warned about alcohol intake, use of sedative drugs, and smoking should be discontinued due to its adverse effects on both OSA and general health. OSA patients should be instructed not to drink alcohol within 4 hours of sleep, and to use sedatives with cautionsincethey cause excessive relaxation of the muscles in the upper respiratory tract [22]. Another important consideration is the sleep position. In one study, apnea and hypopneas were found to vary by half according to the sleeping position, and position-related apnea was found in approximately 27% of the patients [23].

Positive airway pressure (PAP) is one of the most used treatment of OSA. In this method, air pressure is delivered using a machine through a mask that is placed patient's face that covers mouth and nose. As a result, the formation of apnea and snoring is prevented. The most used type is called *Continuous positive airway pressure* (CPAP) machine which continously push a air through mouth and nose with adjustable pressure value. The mask type varies from patient to patient and the pressure value can be adjustable according to the severity of OSA and patient's comfort.Although CPAP is the most effective method, it is frequently observed that patients discontinue treatment because of nasal dryness, mask mismatch, and imbalance of pressure value [24]. To prevent these problems, machines with pressure regulators and humidifiers

have been developed. If these conditions occur, the sleep specialist should be consulted and treatment should not be discontinued immediately.

Mouth devices are used to prevent airway closure during sleep. Some of them open the throat by adjusting the tongue position and others by pulling the lower jaw forward and facilitate breathing. In patients with high OSA severity and these treatments do not work, surgical weight loss, tissue removal from the throat and correction of anatomical disorders of the jaw can be applied.

3. CLASSIFICATION ALGORITHMS IN MACHINE LEARNING

3.1 Logistic Regression

Although its name has the word regression, logistic regression is a classification method that uses probabilities to classify inputs. It uses S-shaped (Figure 3.1) logistic function (3.1) that takes any input and maps it into (0, 1) range as probability values. Given an input vector *x* and weight vector $\boldsymbol{\theta}$, the hypothesis function $\boldsymbol{h_\theta(x)}$(3.1) and its applied logistic function $\boldsymbol{g}$ is defined in equation (3.2).

$$\boldsymbol{h_\theta(x) = \theta^T x} \tag{3.1}$$

$$\boldsymbol{g\left(h_\theta(x)\right) = \frac{1}{1 + e^{-h_\theta(x)}}} \tag{3.2}$$

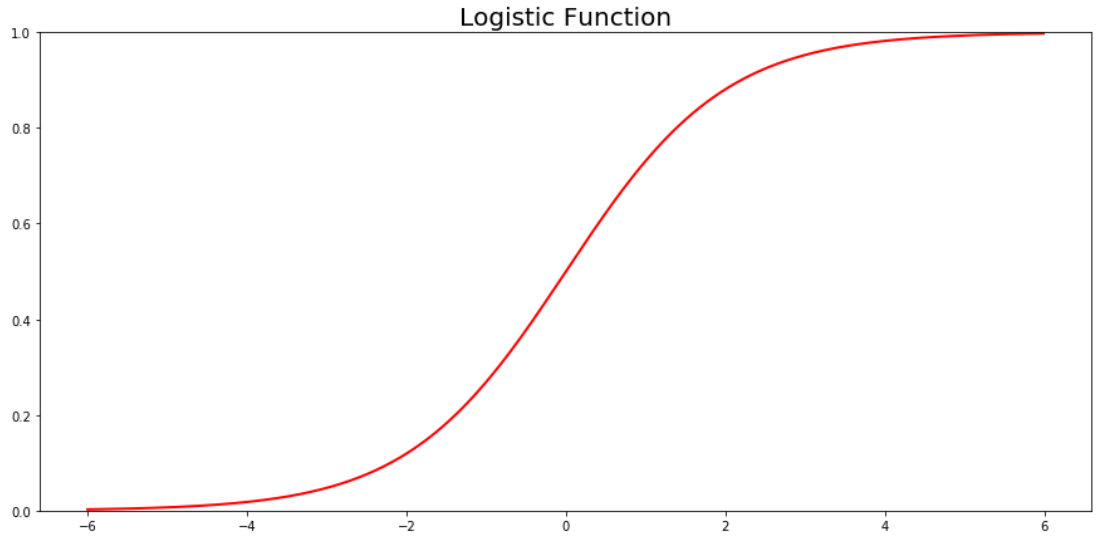


Figure 3.1 : Logistic function.

It shows the probability of an instance that belongs to class *i* based on the probability value is greater than or equal to 0.5 or not. In order to get best fitted line, the curve with the maximum likelihood is selected. Logistic regression can work with both discrete and continuous data. It is also useful to identify which variables making a contribution to the classification.

3.2 Näive Bayes

Naïve Bayes classification method is based on Bayesian Theorem (3.3) in probability theory. It has two assumptions. The first one; features given are independent from each other. However, it is very unlikely in real-world scenerios. The other one is it assumes that each predictor has same importance that means they make equal contributions to the model's outcome.

$$P(C_i|x) = \frac{P(x|C_i) \times P(C_i)}{P(x)} \tag{3.3}$$

where $P(C_i|x)$ is *posterior* probability that input vector *x* belongs to class *i*.$P(C_i)$ is the *prior* probability. $P(x)$ is called evidence. It has no effect on the result, used as normalization factor. Decision function assigns the input vector *x* to the maximum posterior probability calculated by the formula (3.4):

$$argmax_i P(C_i) \times \prod_{k=1}^{n} P(x_k|C_i) \tag{3.4}$$

3.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a fast, easily trainable, and powerful method for classification tasks. It has two important parameters to be decided: number of neighbors to be taken into account as *k*, and the distance function *d*. Using *d* the given input is classified as the closest point's class. For *k*>1 majority voting makes the final decision. Mantattan distance (3.5) and Euclidean distance (3.4) are widely used as distance function, but they can be replaced with respect to the problem. For given two *n* dimensional data points x_1, x_2:

$$d_{(manhattan)}(x_1, x_2) = |x_{11} - x_{22}| + |x_{12} - x_{22}| + \cdots + |x_{1n} - x_{2n}|$$

$$d_{(euclidean)}(x_1, x_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \ldots + (x_{1n} - x_{2n})^2)}$$

3.4 Decision Trees

In general, the way that Decision Trees (DT) works is DT asks a question to the predictor, and it splits instances based on their answers to that question. While doing that, it considers the nodes' purity. The one way to calculate how pure the internal node is using *Gini* index (3.7).

$$Gini = 1 - \sum_j p_j^2 \tag{3.7}$$

where p_j is the probability of the class j in that node. Gini index calculates how much impure the subsets/nodes after splitting. Notice that lower value is better. The attribute gives the lowest *impurity* value will be the decider of that node. If there is no improvement in *impurity* after splitting, then the node becomes a leaf. An example is shown in Figure 3.2. The first node is called the *root node* which is the feature with lowest impurity. Root is divided to *internal nodes* and *leaves* based on the decision thresholds.

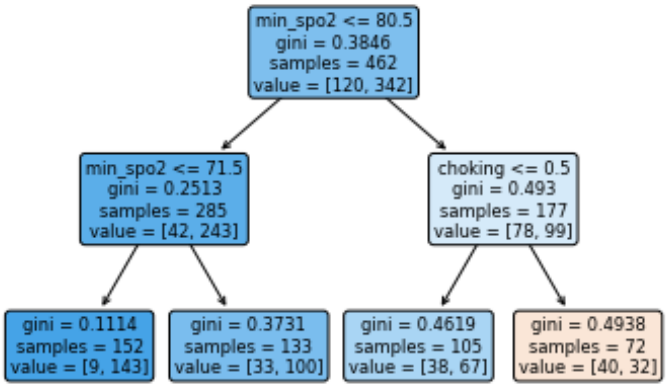


Figure 3.2: Root, nodes, and leaves in a decision tree.

3.5 Ensemble Methods

3.5.1 Random forests

Random Forests (RF) is a problem solver for DT classifiers since DTs are tend to overfit to training data. RFs are made of many DTs. Creating *n* decision trees with random samples and random features reduces the chance of overfitting. Each grown tree in forest have equal saying to the final decision.

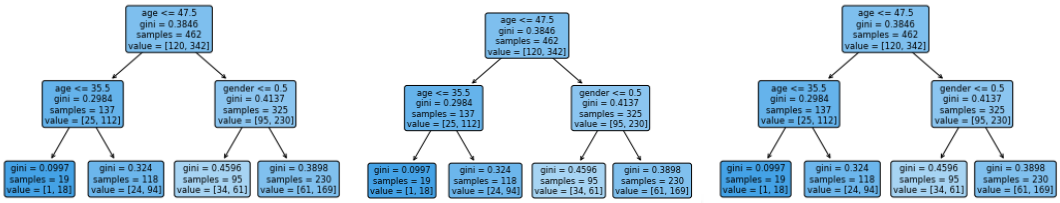


Figure 3.3: Random forest created using three different decision trees.

In Figure 3.3, three DTs constructed from randomly selected features and instances form a random forest. The most frequent predicted class by DTs in forest will be the predicted value.

3.5.2 AdaBoost

Boosting is a way to combine weak classifiers to create a better, strong classifier by focusing on the misclassified examples and giving more importance them. One of the boosting methods we will apply to our dataset is called Adaptive Boosting or AdaBoost. AdaBoost combines a lot of **weak learners** to make classification. The weak learners are almost always stumps i.e. DTs created using only one feature. AdaBoost is forest of stumpts whereas RFs are forest of DTs. Moreover, in the RFs , trees are fully grown and their depth may differ from one another. Unlike RFs, in AdaBoost some stumps get more saying in classification than others. It assigns sample weights to each specific sample. Then weight of the misclassified sample is updated (increased) for next stump. In contrast to RFs, each stump is made by taking the previous stumps mistakes into account, in other words influenced by them. We will combine AdaBoost with different sampling strategies, and weak learners, and examine the results.

3.5.3 Voting

In Voting Classifier the main idea is to combine different classifier. Each classifier makes it own prediction. The final decision is made by voting. There are two types of voting: *hard vote* and *soft vote*. In hard vote, majority of predictions decides the predicted class. On the other hand, in soft vote, the average of probabilites are considered.

3.6 Multi-layer Perceptron

Multi-layer Perceptron (MLP) is a system that simulates the structure of the human brain. It consists of three main structures. These are the input layer, the hidden layer, and the output layer. Each layer consists of neurons (Figure 3.4). The structure of a MLP system a.k.a. Artificial Neural Network (ANN) can be shown in figure below:

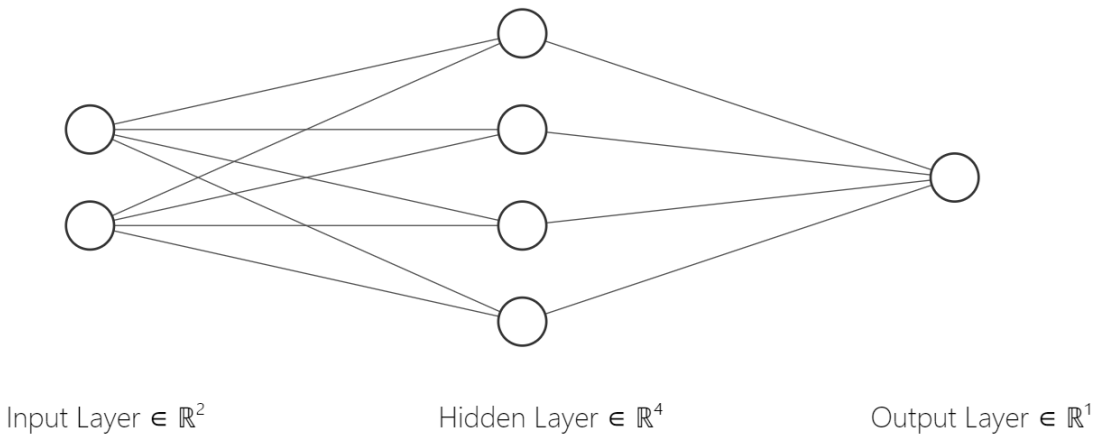


Figure 3.4: Neural newtork structure.

Each neuron is connected via weights and each layer has different weights. A neuron takes the input using linear combination with weights, adds bias to result. After that it passes the result through the *activation function*. This process is called forward proapagation. Final results is compared to expected result, then each weight in each layer is updated propagating back to network. This is called back propagation. At the end, optimized weights are the final state of the network.

3.7 Cross Validation

In general, the cross validation is used in fitting process to measure the performance of the model. It is mainly used while tuning the model's *hyperparameters*. Aim is to find best subset of hyperparameter combination, and measure the performance of the model using different parts of the data to get a general sense. First, training data is devided into *k* number of subsets. Then, the model is trained using *k-1* subsets together, and test its performance with that last part, and repeat the whole process with different subset in total *k* times. Expected performance will be the average of all those *k* scores. It is illustrated at the Figure 3.5 below:

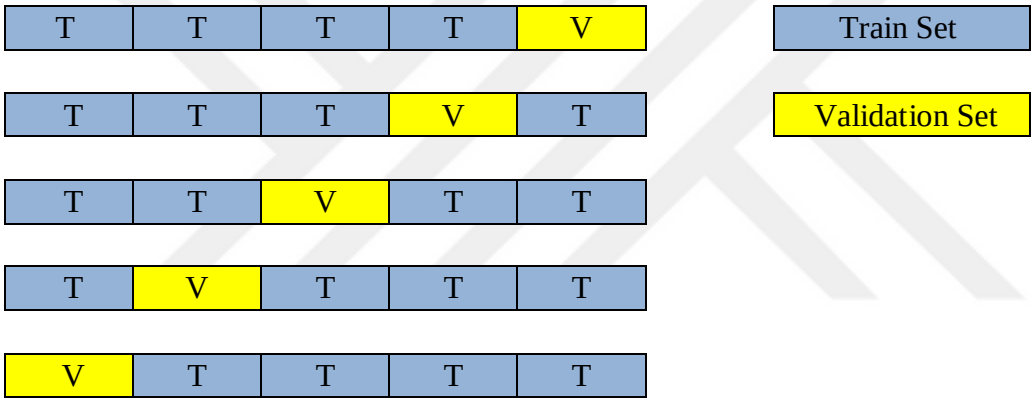


Figure 3.5: 5-Fold Cross-Validation schema.

*Stratified K-Fold CV*was used to ensure that each target class was in the training and validation sets. Thus, the class distribution ratio in each part remained the same, and the different samples provided generalization of the model. In this project, we will use 10 -Fold stratified cross- validation to measure model performances.

3.8 Evaluation Metrics

In general, the Metrics used to measure model performance in classification methods may vary according to the objective of the problem. One of the simple and effective way to see how model's doing is using *confusion matrix* (Table 3.1). In a confusion

matrix, the diagonal containts where the model predictions are correct with respect to classes, and rest is the model predictions are wrong.

Table 3.1: Confusion matrix.

Confusion Matrix	Actual Classes	
Predicted Classes	True Positive(TP)	False Positive(FP)
	False Negative(FN)	True Negative(TN)

True Positive: TP is where the acutal class is positive class and model predicted correctly.

False Positive: FP is where the actual class is negative and model predicted it as positive. It is also known as *false alarm.*

False Negative: FN is where the actual class is positive and model predicted it as positive, also known as *miss.*

True Negative: TN is where actual class is negative and model predicted it as negative.

Confusion matrix is generally used to calculate metrics: *accuracy*, *precision*, and *recall*.

Accuracy: measures the ratio of correct predictions out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.8}$$

Precision:shows how accurate the positive predictions are.

$$Precision = \frac{TP}{TP + FP} \tag{3.9}$$

Recall: shows the ratio of positive classes that are correctly predicted.

$$Recall = \frac{TP}{TP + FN} \tag{3.10}$$

F1-score: is the harmonic mean of precision and recall.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.11}$$

ROC-AUC: The Receiver Operating Characteristic (ROC) is a measure of a classifier's predictive quality that compares and visualizes the tradeoff between the model's sensitivity (recall) and specificity. ROC plot or ROC curve shows the true positive rate (TPR) -recall-sensitivty versus the false positive rate (FPR).

$$FPR = \frac{FP}{TN + FP} \tag{3.12}$$

Area under the curve (AUC) is a metric that shows the relationship between true positives and false positives, in other words measure of seperability. The higher AUC value generally means better model.

3.9 Sampling Methods

When dealing with imbalanced datasets, sampling methods are come in handy to be a solution that address this problem. If a dataset is imbalanced, many ML algorithms are tend to bias towards majority class. Because, the model's parameters will be updated more frequently for that class. One way to balance things is to use *class weights*. For example, given a dataset with binary target values say *1* and *0*, let's say wehad10 instances for class *1* and 90 instances for class *0*. In this case, we may think to give bigger class weight to class *1* and smaller weight for class *0* (e.g. 0.9 for class-1, 0.1 for class-0). Therefore, model will make bigger updates to its parameters for misclassified exapmles in class-1 than class-0. In other words, model will be punished more for wrong predictions made to minority class.

Figure 3.6: Undersampling and oversampling.

Other way to deal with imbalanced datasets is called *sampling* (Figure 3.6). To reduce the effect of majority class on model, *undersampling* may be applied to shrink its size to minority class by sampling instances from majority class. Undersampling can be applied different ways. For example; randomly sample majority class samples to match minority class size, or more advanced approaches like *TomekLinks* may be used. However, undersampling may cause loss of information and results in poor model performance. Secondly; using *oversampling*, number of instances of minority class can be increased using random sampling. But, this creates duplicated instances and possibly lead to overfitting to train data. *Synthetic Minority Over-sampling TEchnique (SMOTE)* is an algorithm to oversample minority class by generating synthetic data. Let us review some of undersampling and oversampling methods.

Random Under Sampling(RUS): RUS randomly discards instances from the majority class in order to create balance in imbalanced dataset. However, there is a risk that it may remove informative samples, and increase of the classifier variance.

Tomek Links: TL is defined as a connection or distance between two opposite class instances. There is said to be a link between two points a and b if:

$$d(a,b) < d(a,c) \;\; and \;\; d(a,b) < d(b,c) \tag{3.12}$$

where c is any point that belongs to any class and d is the distance function. TL aims to reduce noise while removing majority class instances. Notice that this method is not completely undersamples the majority but acts as a cleaning method.

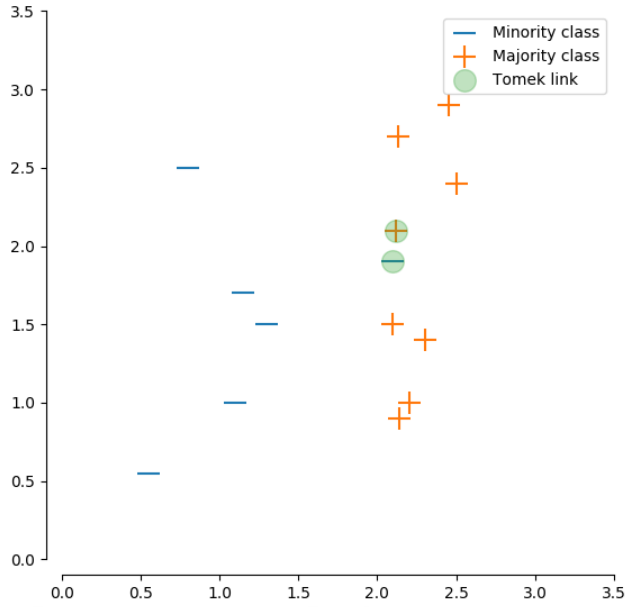


Figure 3.7: An example of TomekLink samples.

In Figure 3.7, it is seen that negative class is very close to positive class creates a link (green circle) with majority class sample. So, this two sample are tagged, and decision can me made as, remove which belongs to majority class or both samples.

SMOTE: SMOTE is a technique to address overfitting problem caused by randomoversampling with generating synthetic data points for minority class [25].

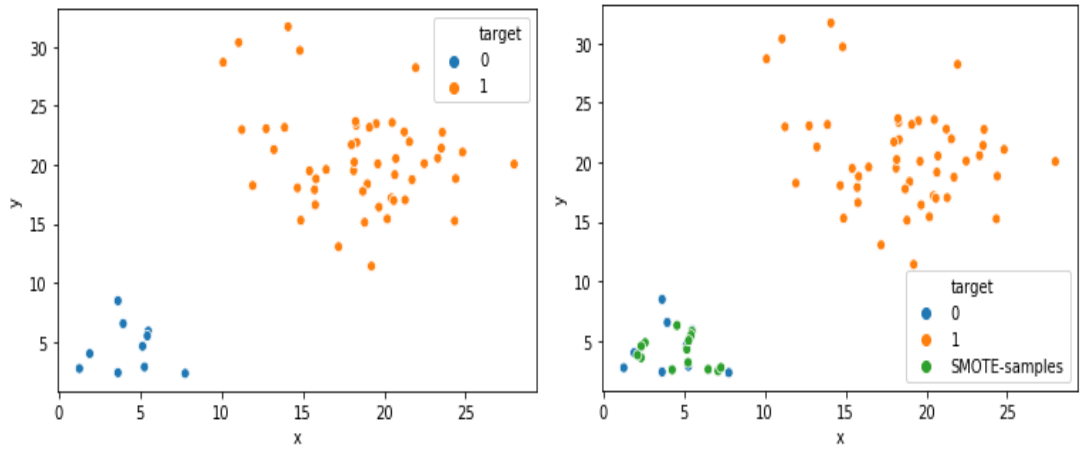


Figure3.8: Oversampling using SMOTE.

When oversampling, a random sample taken from minority class and focus on its *k* neighbors, in the case shown in Figure 3.8, *k* is equal to 3. After that, considering the vector between these *k* neighbors and the sample is multiplied by a random number in

(0, 1) interval to create new, synthetic data instance. There are many variants of SMOTE using different approaches. *SVM-SMOTE* is one of them which focuses instances of minorty class that are close to decision boundary created by SVM and oversamples them only. *Borderline-SMOTE* aims to increase the number of minority instances that border majority ones in order to allow predictor clearly distinguish different class instances [26]. *ADASYN(Adaptive Synthetic Sampling)* is oversamples minority examples which are difficult to classify and harder to learn [27]. It is encouraged to apply oversampling followed by undersampling i.e. SMOTE-Tomek (SMOTE oversampling followed by TomekLink).

4. RESULTS AND DISCUSSION

In this section, the steps shown in Figure 4.1 will be followed.

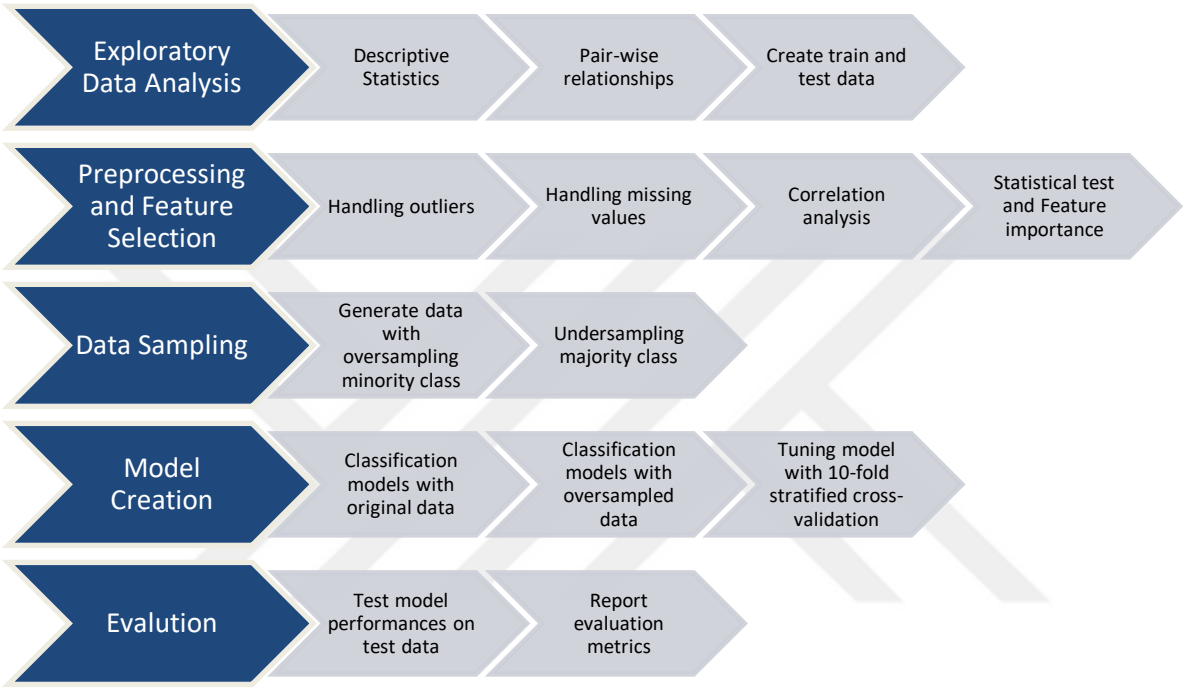


Figure 4.1: Experiment workflow.

4.1 Development Environment

Python is a open-source, high-level, dynamically typed, object oriented programming language used for multiple purposes such as scientific problem solving, web and internet development, desktop GUI applications, network programming, software and game development and so on. In recent years, as the popularity of data science has increased, Python has quickly become known for the convenience provided by its scientific and data science libraries such as *numpy* for linear algebra and matrix operations, *pandas* for structural and tabular data processing, *scikit-learn* for machine learning and data science and many more. In this research, for data manupilation, analysis and model creation processes, Python will be used as main tool.

Anaconda is an open-source platform to easily set and develop machine learning environments. It is supported by big operating systems such as Windows, Linux and

Mac OS X. It comes with pre-installed scientific packages mainly used by data scientist with programming languages like Python and R. It makes it easier to manage libraries and environments. It also has parallel computing libraries such as *Dask*, and effective visualization libraries like *Matplotlib, Bokeh* etc. It uses a notebook environment called *Jupyter (Jupyter Notebook)*which is an easily interactable, cell by cell execution based coding interface that also supports markdown languages i.e. HTML and also LaTeX. All these makes it possible to create, analyze, develop, and present a project lifecycle in one environment.

4.2 Gathering Data

Our dataset is contains informations collected from OSA diagnosed patients. It is obtained from a research on OSAperformed at Başkent University. It contains 763 instance of patients and 20 variables including target variable. The table 4.1 provides explanations of the variables.

Table 4.1: Feature names, descriptions and data types.

Feature Name	Description	Data Type
age	Age of patient	int64
gender	Gender of patient	int64
bmi	Body mass index	float64
snoring	Snoring status	int64
daytime_sleepiness	Feeling sleepy in daytime	int64
asphyxiation	Choking status	int64
fatigue	Feeling tired	int64
headache	Headache status	int64
amnesia	Forgetfulness	int64
anger	Anger status	int64
hypertension	Having hypertension	int64
cardiac	Heart disease	int64
diabetes	Having diabetes	int64
reflux	Stomachache	int64
lung_dis	Lung disease	int64
dyslipidemia	Disruption in the amount of lipids in the blood	int64
smoking	Smoking status	int64
min_spo2	Minimum Oxygen Saturation	float64
ess	Epworth Sleepiness Scale	int64
osas_severity	OSA severity	int64

In this variables, *osas_severity* is our target variable (1:mild, 2:moderate, 3:severe). Among others, we divide them as numerical and categorical variables, and apply different feature selection methods to get a sense about which features to include or exclude from our model including expert opinions. Numercial and categorical features in data are shown Table 4.2.

Table 4.2: Numerical and categorical variables.

Feature Type	Feature Name
Numerical	age, bmi, ess, min_spo2
Categorical	gender, snoring, daytime_sleepiness, asphyxiation, fatigue, headache, amnesia, anger, hypertension, cardiac, diabetes, reflux, lung_dis, dyslipidemia, smoking

4.3 Exploratory Data Analysis

In this section, features behaviour was investigated. Firstly, the distribution of target variable was examined.

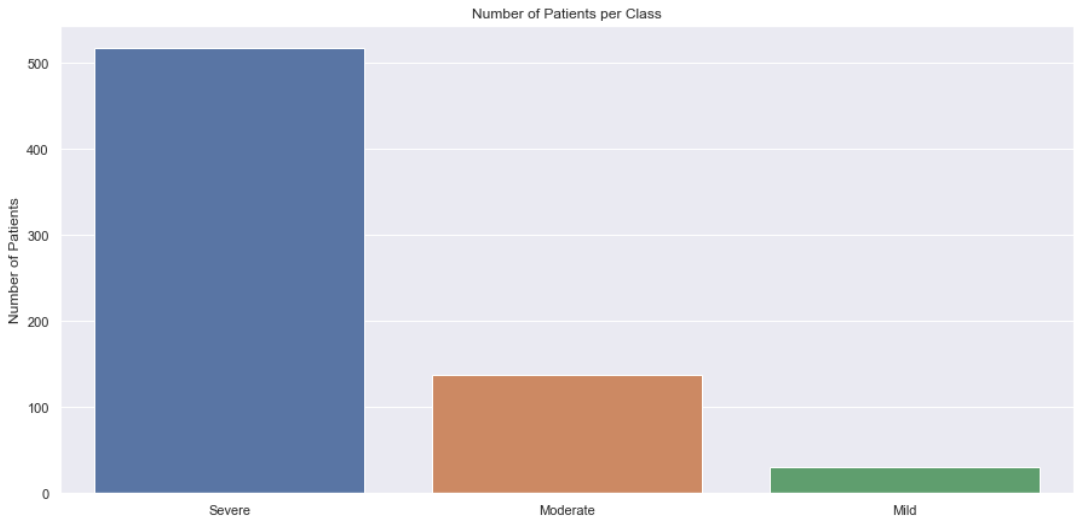


Figure 4.2: Distribution of target class.

It can be deduced from Figure 4.1 that the majority of the labels are from *severe* class and the data is imbalanced. It may cause problems for some classification algorithms. The over-sampling and under-sampling techniqueswill be applied as mentioned in section 3.9. The dataset will be splitted into train and testsets with 80% for training

20% for testing. It is advised to do exploration on train data, decide preprocessing, and feature selection steps and apply them to test (unseen) data to avoid data leakage. While doing splitting into train and test, stratify method will be applied. In that way, the ratio of labels in both sets will be keptequal. After splitting and stratifying both sets are created as shown in Table 4.3.

Table 4.3: Train and test sets.

Severity	Train Set(n=610)	Test Set(n=153)	Ratio(~)
Severe	460	116	75%
Moderate	122	30	20%
Mild	28	7	5%

It should be examined if there are some outliers in numerical features. Numerical features were*age, bmi, ess,* and *min_spo2*. The summary of statistics of these features are given in Table 4.4.

Table 4.4: Descriptive statistics of numerical features.

	age	**bmi**	**min_spo2**	**ess**
count	609	546	574	555
mean	54.47	34.34	74.08	7.86
std	11.01	5.83	12.46	5.89
min	23	22.31	0.5	0
25%	46	30.39	67	3
50%	55	33.7	78	7
75%	62	37.65	83	12
max	87	58.61	92.2	21

First, notice that even though train data has 610 samples, in Table 4.3, *age, bmi, min_spo2, ess* have 609, 546, 574, 555 instances respectively. From that, we can conclude that there are missing values in that features. Another thing to point out is minimum value of *min_spo2* is 0.5. This is very unlikely and it probably is an incorrect value. Boxplots are very handful to give insights about outliers or detecting incorrectly stored values. The boxplots for each numeric feature are drawn in Figure 4.2, 4.3, 4.4, 4.5 are shown below:

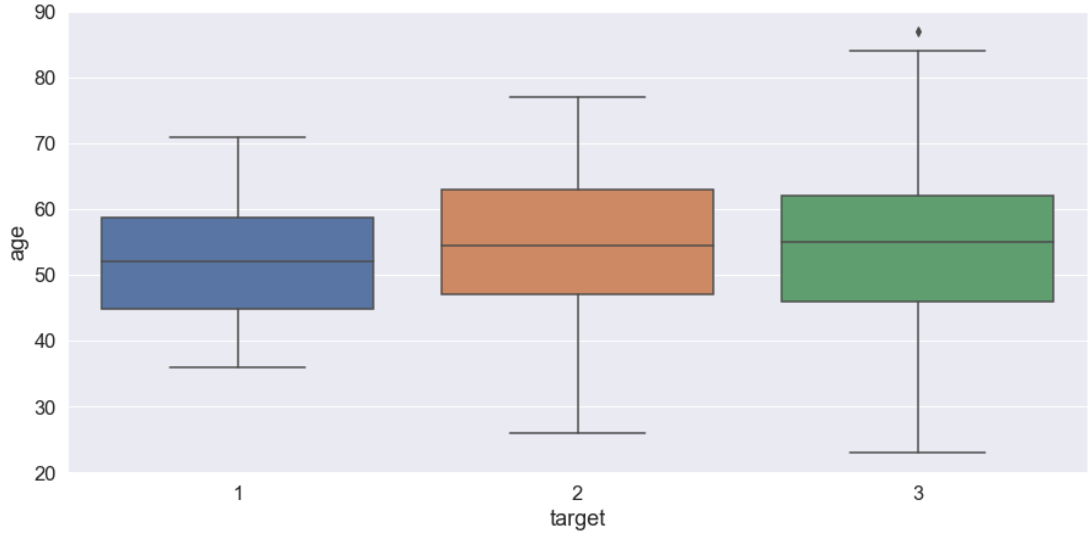


Figure 4.3 : Box plot of variable *age*.

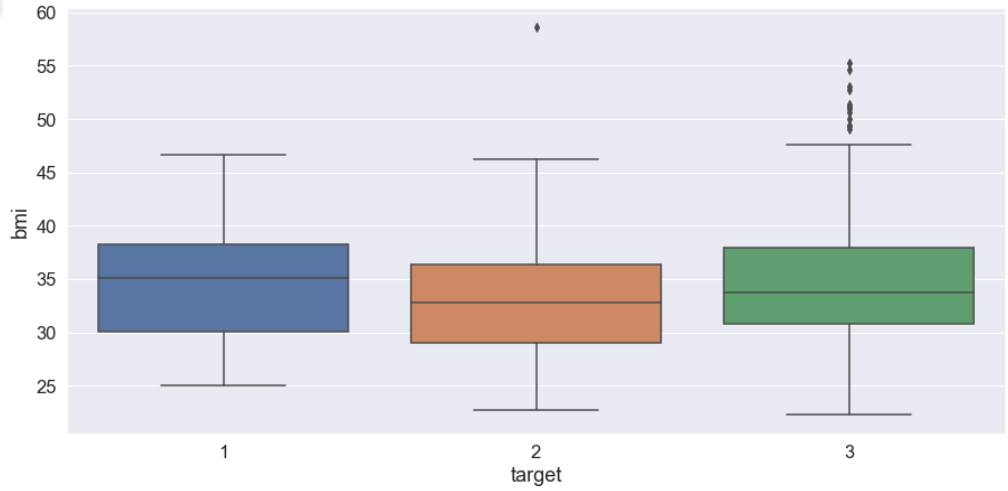


Figure 4.4 : Box plot of variable *bmi*.

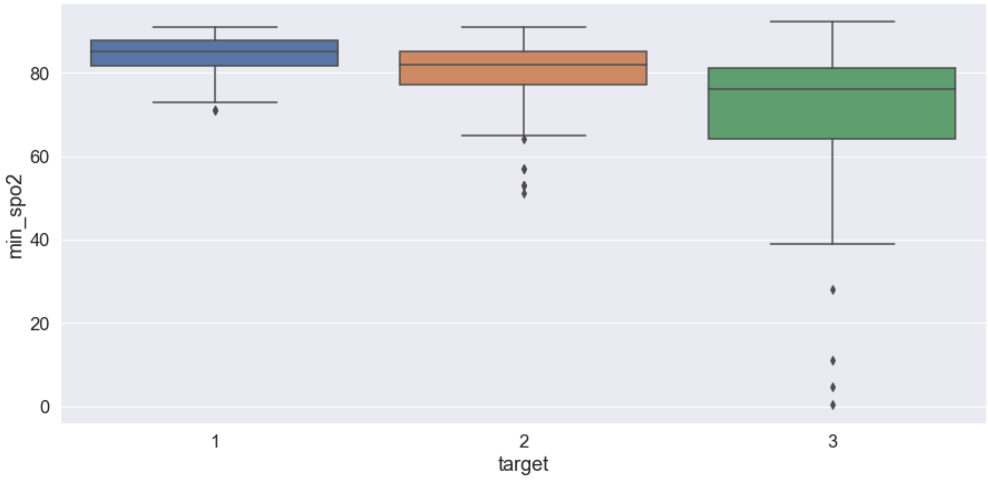


Figure 4.5 : Box plot of variable *min_spo2*.

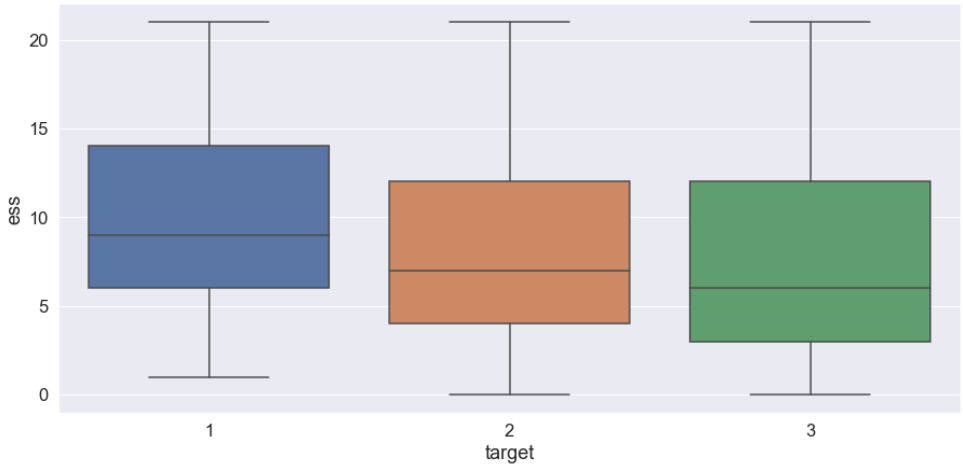


Figure 4.6 : Box plot of variable *ess.*

Target 1, 2, 3 in x-axis represents severities mild, moderate, severe respectively. In Figure 4.3 some instances are behave like outliers where severity is severe (class=3). They do not seem incorrect. Hence, we may consider replacing them with the mean value of *bmi* or use binning to suppress them. We will treat this kind of decisions as hyperparameters, too. In Figure 4.4, there are some outliers for target 2 and 3. But in target 3, minimum oxygen saturation values below 40 are likely incorrect values, especially 0.5. Hence, we will replace those with average min_spo2 value of target 3. Then we apply all those transformations to test set, too. Lastly, pair-wise relationships between numerical variables will beexaminedand their distributionsare illustrated in Figure 4.6.

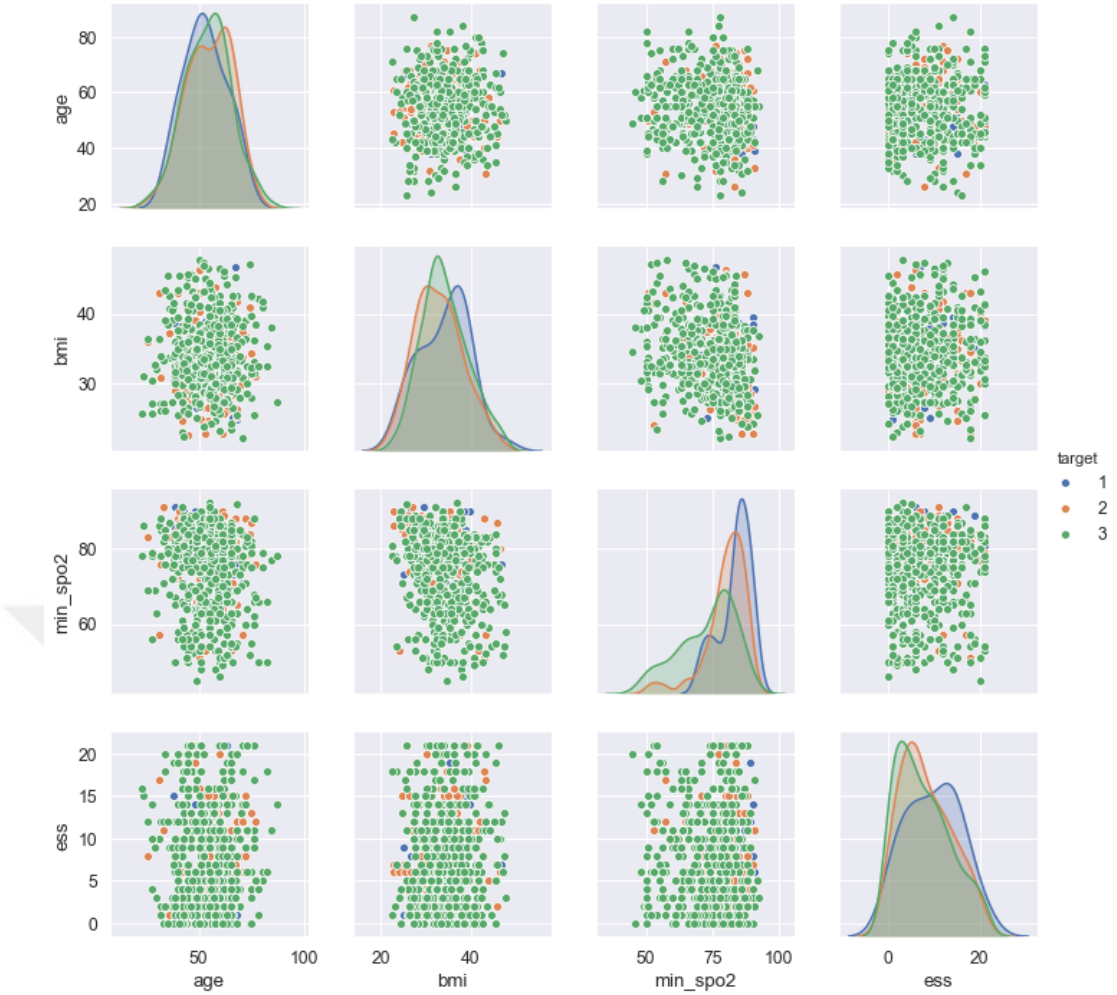


Figure 4.7 : Pairwise relationships of numerical variables.

In Figure 4.6, it looks numerical variables are not correlated to each other. However, kernel density estimation (KDE) plots shows that for each class, the variables are show minor discriminative power. It can be noticed that there is a slight right skewness for feature *bmi* and left skewness for feature *min_spo2*. We may choose to apply log transformation to bmi, and power transformation to min_spo2 to make them closeto a normal distribution since some ML may benefit from that.

4.4 Feature Selection

Feature importance or feature selection is a very important step in ML pipeline. Although it does not have a huge impact on small data sets, removing features that are found to be unnecessary may improve performance and training time. In this step, we study the correlations between features, check how important for the prediction they are, and discuss some feature ranking methods. The methods applied in feature

selection may vary with respect to type of the feature and compared feature's type i.e. numerical versus categorical or vise versa. Selecting important features and removing unnecessary ones usually improves model performance and reduces training time. First, for numerical features *Pearson's correlation* coefficients will be calculated. Pearson's correlation shows linear correlation between features and its value lies between -1 and 1 (1 for perfect positive correlation, -1 for perfect negative correlation). The correlation matrix for numerical features are plotted to see whether there are features highly correlated with each other. It is illustrated in Figure 4.7 below:

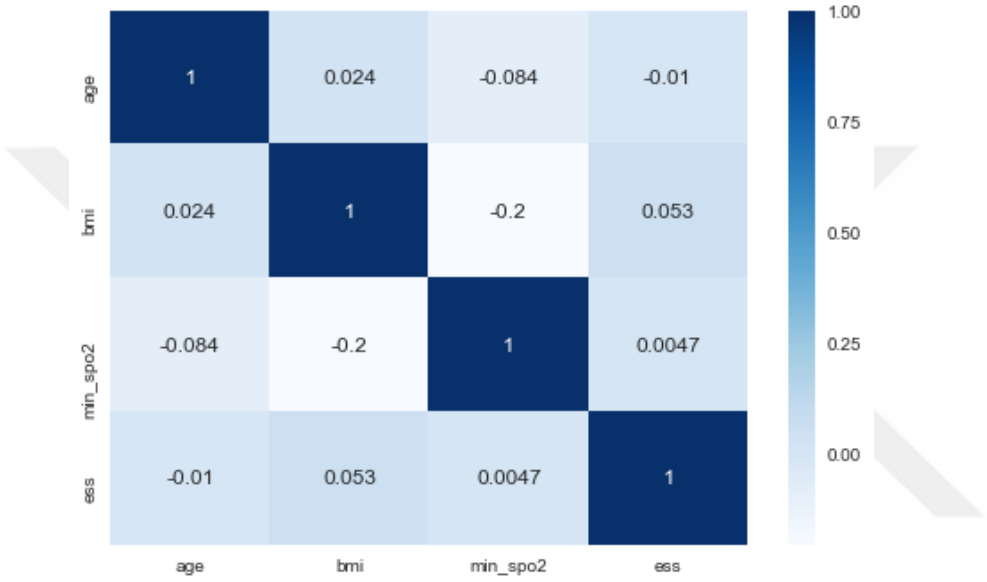


Figure 4.8 : Correlation between features via heatmap.

Notice that there are no correlations between numerical features. Next, we will check the relationship between numerical values and target variable. A statistical test called **ANOVA F-test** will be applied since there is a comparison between quantitative values and a categorical value. ANOVA F-test scores represents how the means of each group are significantly different if numerical features grouped by targets. In Table 4.5, it is seen that there are two features provide significance condition ($p \leq 0.05$) which are **min_spo2**, **bmi**. Even though **ess** is not less than *0.05*, it canstill be included since it is marginally significant ($p \leq 0.1$).

Table 4.5: ANOVA F-test results.

Feature	ANOVA	p_values
min_spo2	329.833	0.0000
bmi	32.927	0.0379
ess	28.817	0.0569
age	0.3969	0.6726

Hence, there are 3 features will be included in our model as in numerical. Next, we will examine the relationship between the target variable and categorical features. Since it is a categorical to categorical comparison, we will use another statistical test called **Chi-squared** (χ^2) **test** (4.1).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4.1}$$

In equation 4.1, O_i is number of observations in class *i*, E_i is the number of expected observations in class *i*. The motivation is to select m-features with highest χ^2-score considering statistical significance.Before applying χ^2 test, *contingency table* for each feature with respect to other feature will be created. Contingency tables are used for summarization of two categorical variables based on their occurence as seen in Table 4.6.

Table 4.6: Contingency table daytime_sleepiness vs. asphyxiation.

daytime_sleepiness \ asphyxiation	0	1
0	95	34
1	231	211

After applying χ^2 test, the dependent features will be included in model. As shown in Figure 4.8, heatmap plot, 0 means dependent while 1 means independent.

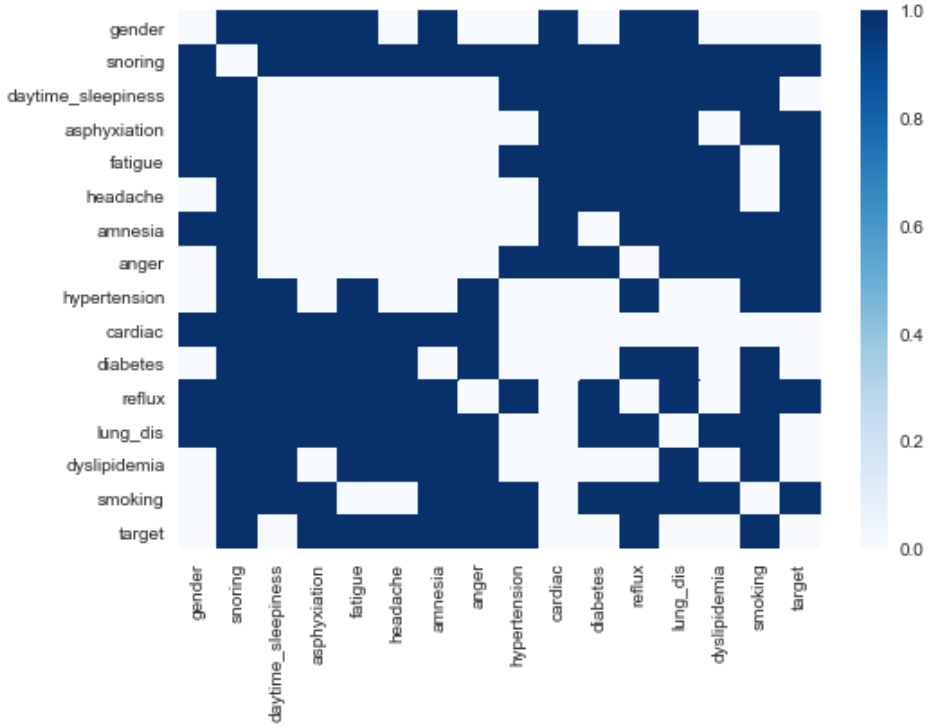


Figure 4.9: Chi-squared test results using heatmap.

As a result, four features in Table 4.7 will be considered to be included in model. Notice that between feature dependence are ignored.

Table 4.7: Significant dependent categorical features.

Feature	p-value
gender	0.0033
daytime_sleepiness	0.0162
diabetes	0.0203
dyslipidemia	0.0399

RF and DTs methods have feature importance attribute. In order to get a sense of how features are treated, we fit a RF and DT model to our training data. Relative importances of features using trees are shown in Figure 4.9 below.

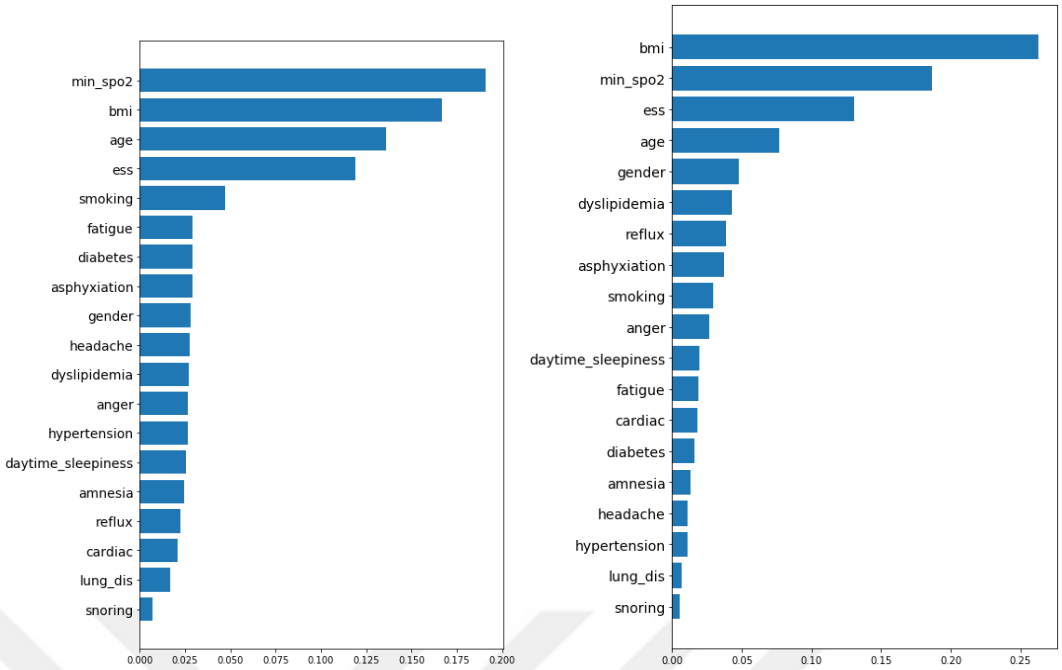


Figure 4.10: RF(left) and DT(right) feature importances.

It can be seen that the four numerical attributes are most import ones in both models. In RF, categorical features shows nearly same importance except *smoking*. Considering our Chi-squared test results, the selected four categorical attributes have nearly similar importance in RF whereas in DTs, *gender* is most important categorical value.

As a result, in addition to the three numerical variables and four categorical variables selected in Table 4.8 via *ANOVA F-test* and *Chi-squared* tests, *smoking* and age can be added to the model to test their effects since their importance are high in tree results.

Table 4.8: Selected features.

Numerical	Categorical
min_spo2	gender
ess	daytime_sleepiness
bmi	diabetes
(age)	dyslipidemia
	(smoking)

4.5 Model Creation

In this section, we will make two different approaches to the prediction. (a) Multiclass classification: where we create models to predict each three class seperately and measure performance. (b) Binary classification: where we will treat *moderate-severe*

(2-3) instances as one class and *mild(1)* another since the minority class is very hard to predict and *moderate* and *severe* patients will more likely to have similar treat. We will create models for each approach. Then, all models will be trained for best hyperparameters using cross validation with and without sampling. After that, we will calculate their performance on test data to see whether performance is improved or not. As mentioned before, we will focus on metrics precision, recall, f1-score, and ROC-AUC value. Finally, we will examine confusion matrices to analyze TPR and FPR values and trade-off between them.

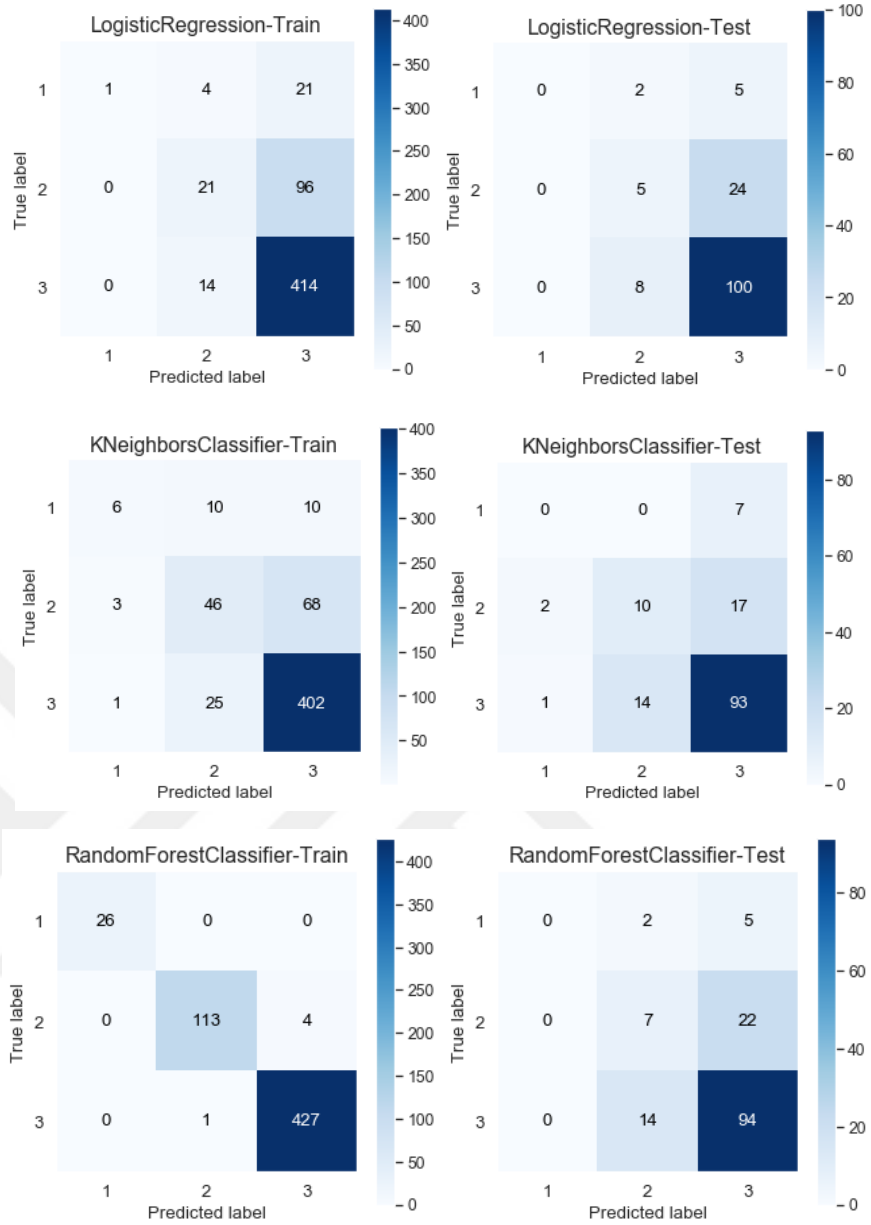
4.5.1 Multiclass prediction results

Models are created for each ML classification algorithm we mentioned in Section 3.5. In Table 4.9, evalution results are shown. The PR, RC, and F1 weighted average were evaluated to take into account the class imbalance. Accuracy is not preferred metric since it may be misleading in imbalance situations. However, we will share it for first outcome.

Table 4.9: Base model performances.

Mean CV Scores		weighted		
	accuracy	precision	recall	f1
LogisticRegression	**75.10%**	66.80%	**75.10%**	**68.60%**
Naïve Bayes	38.30%	**77.20%**	38.30%	47%
KNN	**70%**	64.60%	**70%**	66.50%
DT	61.60%	65.10%	64.10%	63.40%
RF	**70.40%**	65.70%	**73.10%**	67.10%
AdaBoost	69.20%	66.50%	69.20%	**67.60%**
MLP	**76%**	65.10%	**74.80%**	**67.30%**
Voting	**75.70%**	67%	**76%**	**69%**

LR, KNN, RF, MLP, and Voting classifiers have better results comparing to other methods. Let us review their predictions on train and test data by analyzing their confusion matrices in Figure 4.10.


LogisticRegression-Train
True label
Predicted label
1 1 4 21
2 0 21 96
3 0 14 414
1 2 3
400 350 300 250 200 150 100 50 0
LogisticRegression-Test
True label
Predicted label
1 0 2 5
2 0 5 24
3 0 8 100
1 2 3
100 80 60 40 20 0
KNeighborsClassifier-Train
True label
Predicted label
1 6 10 10
2 3 46 68
3 1 25 402
1 2 3
400 350 300 250 200 150 100 50
KNeighborsClassifier-Test
True label
Predicted label
1 0 0 7
2 2 10 17
3 1 14 93
1 2 3
80 60 40 20 0
RandomForestClassifier-Train
True label
Predicted label
1 26 0 0
2 0 113 4
3 0 1 427
1 2 3
400 350 300 250 200 150 100 50 0
RandomForestClassifier-Test
True label
Predicted label
1 0 2 5
2 0 7 22
3 0 14 94
1 2 3
80 60 40 20 0

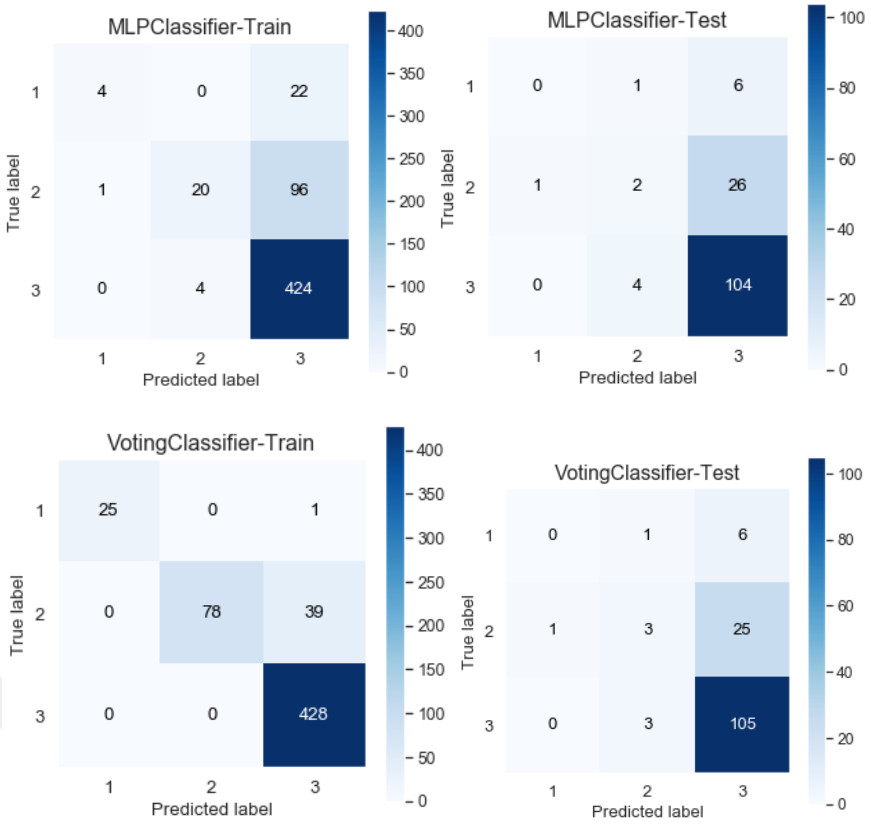


Figure 4.11: Confusion matrices of base models.

RF, MLP, and Voting classifiers are fitting good to training set. Although their metrics look promising, all three models are having trouble to predict class-1 (mild) in test set which means patients with mild OSA are going to be suggested to take an advanced treatment method like PSG while it is not the case. It was expected since there is class imbalance. Next, the sampling methods will be applied to see whether it may lead to any improvements on this matter or not.

Table 4.10: Number of instances per class after sampling.

Class	Original	SMOTE	SVM-SMOTE	SMOTE-Tomek	RUS
1	26	428	247	386	26
2	117	428	272	387	26
3	428	428	428	428	26

After sampling, class distributions are even or close in different methods (Table 4.10). Using generated data with AdaBoost models fitted, resultingmetrics are shown in Table 4.11 and confusion matrices are shown in Figure 4.13.

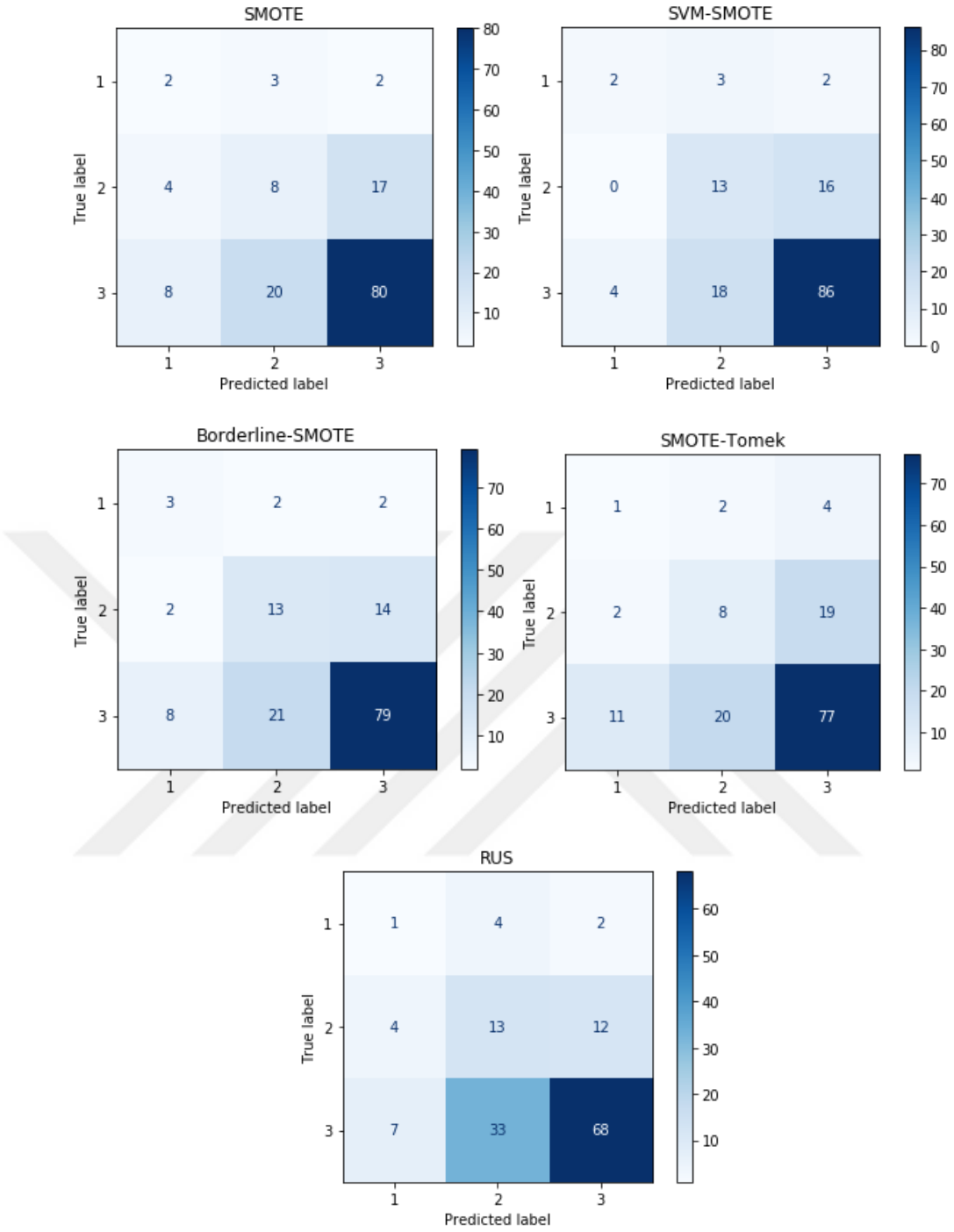


Figure 4.12: Confusion matrices of sampled data-multiclass.

As shown in Figure 4.11, after sampling, the model was able to separate some of the *mild* class. In particular, *SVM-SMOTE* and *Borderline-SMOTE* show most progress with the least amount of compromise from the majority class. The model identified the rare class better.In this case, we can deduce that the model overfitting is slightly reduced. However, in doing so it misclassified some of the instances *moderate* and *severe*.

Table 4.11: Best achieved performances after sampling 3-class.

	precision	recall	f1-score	AUC
SMOTE	68.22%	62.5%	64.27%	63.15%
SVM-SMOTE	74.02%	70.14%	70.66%	65.88%
Border-SMOTE	73.81%	70.83%	71.4%	66.55%
SMOTE-Tomek	66.7%	62.5%	64.34%	60.52%
RUS	68%	57%	61%	59%

4.5.2 Binary prediction results

In this part, we will combine *moderate* and *severe* classes and count them as a one class, and *mild* class as another to see whether we can predict more instances from the minority class in test set. Confusion matrices are shown in Figure 4.14.

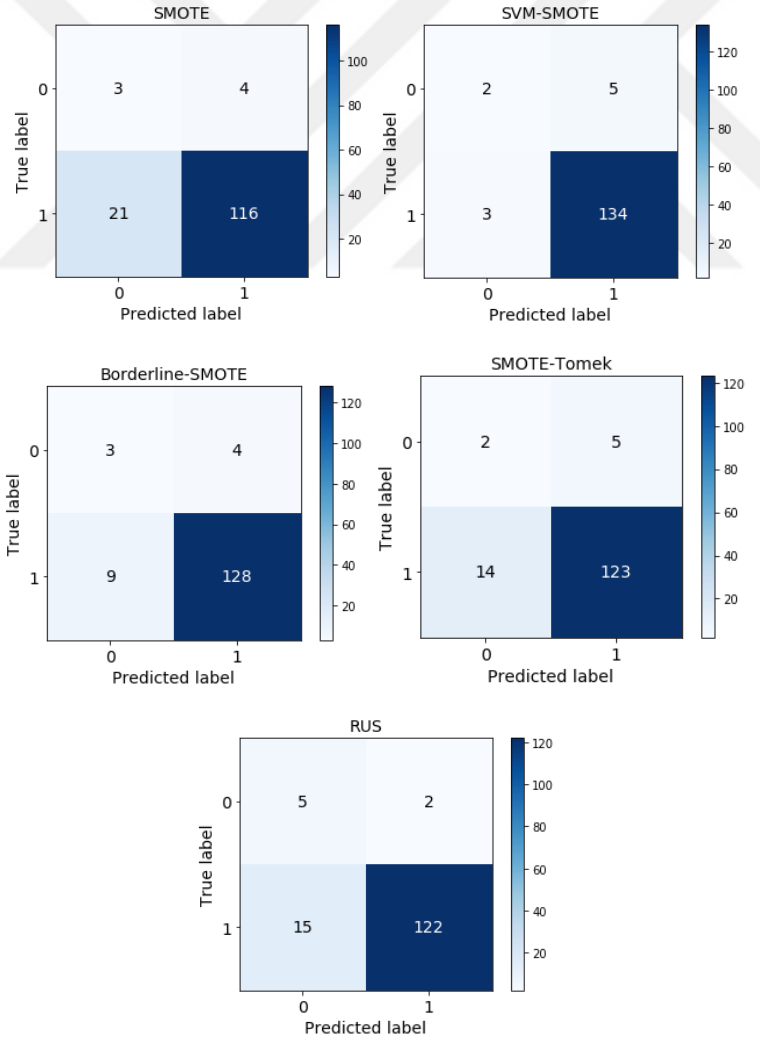


Figure 4.13: Confusion matrices of sampled data-binary.

When compared to multiclass prediction results, almost every sampling result using binary prediction shows improvement on predicting the minority class. The best improvement is achieved by RUS even though it misclassifies some of the majority class. As a result, both multiclass and binary predictions approaches after applying oversampling ended up with slightly overcoming that overfitting problem that base models suffer. In the future, with gathering more and discriminative data, results can be expected to be better.

Table 4.12: Best achieved performances after sampling 2-class.

	pre	**rec**	**f1**	**ba**
SVM-SMOTE	96.4%	98.5%	97.1%	52.9%
Borderline-SMOTE	96.8%	97.8%	96.8%	68.1%
RUS	97.9%	92.7%	94%	70%

In a similar study, OSA severity estimation results were reported as 44.7% accuracy using a total of 19 variables, including demographic data and test results from 313 patients. Compared to this study, it can be seen that the accuacy and recall results obtained in this study are better in both approaches based on Table 4.11.

In another study, linear regression model using data reduced to 5 variables, BMI, waist circumference, neck circumference, spo2, tonsil size, using 390 patient data, yielded 68% explanation power. In a study that used variables that were also observed in other studies that were effective in the determination of OSA severity, similar results were obtained in this study.

In a different study, 26 variables including age, sex, blood pressure, diabetes, hypertension, BMI, ESS as well as PSG and ECG data were used for the prediction of OSA severity. In this study using nearly 5000 patients' data, 3 different classification algorithms were used: Naive Bayes, KNN, and SVM. For the purpose of two-class prediction (AHI≤5, AHI>5), clinical features were obtained using 59% sensitivity and 96.5% specificity for SVM, 57.5% sensitivity and 73.7% specificity for NB. In addition, spectrophotometric properties were used to obtain 43.4% sensitivity and 83.5% specificity for SVM, 39% sensitivity and 82.7% specificity for NB [28]. Unlike the data used in this study, a much larger data was used, and the results obtained by including PSG and ECG signal data in addition to similar variables in this study were promising.

In a study involving approximately 450 children, it concluded that some clinical parameters had more relevance to severe OSA. However, it was reported that there were poor results in predicting OSA severity using multiple linear regression [29].

Data	N	Problem	acc	pre	rec	sen	spe
19 variables, demographic & test	313	multiclass	44.7%	X	X	X	X
BMI, waist circumference, neck circumference, spo2, tonsil size	390	multiclass	68%	X	X	X	X
26 variables including age, sex, blood pressure, diabetes, hypertension, BMI, ESS as well as PSG and ECG data	~5000	two-class	X	X	X	59%, 57.5%, 43.4%, 39%	96.5%, 73.7%, 83.5%, 82.7%
Clinical parameters	450(children)		X	X	X	X	X
Voices data during PSG	120	two-class, multiclass	92.5%, 88.3%	X	X	X	X
9 demographic & test	**763**	**two-class, multiclass**	**70%, 58.2%**	**97.8%, 74.4%**	**98.5%, 73%**	**X**	**71.5%, 78.5%**

Figure 4.14: Comparison with different study results.

In a study using voice data collected by recording the voices of the patients during the PSG test, OSA severity estimation results for the 4-group and 2-group were 88.3% and 92.5%, respectively. The aim was to distinguish severity with a device that can record audio, such as a mobile phone, without the need for PSG data. The results were obtained using artifical neural networks formed using 120 patients' data [30]. Unlike this study, the results obtained using sound data are promising and can be useful in the future with easy implementation.

In Figure 4.14, some studies have better results since they include advanced features and has more data in number. In studies with similar features, models created in this research show superiority and/or close performance.

5. CONCLUSION

Sleep diseases have been present for many years but are not recognized by people unless they are very affected. Obstructive sleep apnea is the most common of these. Patients suffering from OSA may cause occupational accidents and car accidents as a result of carelessness in daily life due to decreased sleep quality. In addition, it reduces the quality of life of patients, adversely affects work and learning performance and may also cause other diseases. As a result of the studies, the mean incidence of OSA varies between 3-5%. While this figure seems to be scarce, its impact on individual life can be huge or even fatal.

Various studies have been carried out in the detection and prevention of this disorder that occurs in most people based on similar characteristics. Questionnaire tests were developed in the patients based on symptoms that may include physical measurements and / or expected symptoms. During the detection of OSA, signal data were collected from patients with electrical devices from different organs during sleep for examination.

In this study, 19 different characteristics of 763 different patients were examined. As a result of the analysis, machine learning models were trained by using minimum oxygen saturation in blood, body mass index, Epworth sleepiness test score, gender, daytime sleepiness, diabetes, lipidity data as model input. Age and smoking status of the patient were also included in the model. In the decision stage, the class with the highest probability value was determined as the model output. The biggest challenge encountered during the ML model training was the imbalanced data. As seen in other studies, the number of patients with severe OSA is generally higher. In such cases, the basic ML models were found to be biased towards the majority class in the decision stage and failed to predict the minority class. With the application of the methods proposed in the literature for the solution of the imbalance problem, the estimation rate in the minority class increased. As a result, in 3-class approach, 71% overall accuracy, 58% balanced accuracy, 73% and 74% weighted recall and precision are achieved. In

2-class approach 88% overall accuracy, 70% balanced accuracy 92% and 98% weighted recall and precision are achieved respectively.

As a result, the determination of OSA severity in a model that will be formed by using machine learning can be obtained faster preliminary results regarding the disease status, the treatment process can be shaped and unnecessary applications as well as unnecessary expenses and wrong treatments can be prevented. With the feedback given to the experts as a result of this study, it will be suggested to collect the data in more detail for similar studies that can be done in the future and by using more data, it will be possible to achieve better results independent of the unbalanced data. Moreover, an automated model environment that can be improved for necessary needs will be suggested in the future which can be kept up-to-date to continuously to the new data and monitor results.

REFERENCES

[1] **Public Health Agency of Canada.** (2010). Fast Facts from the 2009 Canadian CommunityHealth Survey—Sleep Apnea Rapid Response (Report No: HP35-19/1-2010E-PDF). Canada: Public Health Agency.

[2] **Url-1** *<https:/aasm.org/rising-prevalence-of-sleep-apnea-in-u-s-threatens-public-health>*, date retrieved 04.11.2019

[3] **Demir, A.U., Ardic, S., Firat, H. et al**. (2015). Sleep Biol. Rhythms 13: 298. https://doi.org/10.1111/sbr.12118

[4] **Kapur, V. K., Auckley, D. H., Chowdhuri, S., Kuhlmann, D. C., Mehra, R., Ramar, K., & Harrod, C. G.** (2017). Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. Journal of clinical sleep medicine: JCSM : official publication of the American Academy of Sleep Medicine, 13(3), 479–504. doi:10.5664/jcsm.6506

[5] **Mencar, C., Gallo, C., Mantero, M., Tarsia, P., Carpagnano, G. E., Foschino Barbaro, M. P., & Lacedonia, D**. (2019). Application of machine learning to predict obstructive sleep apnea syndrome severity. Health Informatics Journal. https://doi.org/10.1177/1460458218824725

[6] **L. Leite, C. Costa-Santos and P. P. Rodrigues,** "Can We Avoid Unnecessary Polysomnographies in the Diagnosis of Obstructive Sleep Apnea? A Bayesian Network Decision Support Tool," 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, New York, NY, 2014, pp. 28-33.doi: 10.1109/CBMS.2014.30

[7] **Sahin, M., Bilgen, C., Tasbakan, M. S., Midilli, R., & Basoglu, O. K.** (2014). A clinical prediction formula for apnea-hypopnea index. International journal of otolaryngology, 2014, 438376. doi:10.1155/2014/438376

[8] **Yücelbaş, Ş., Yücelbaş, C., Tezel, G., Özşen, S., Küççüktürk, S., Yosunkaya, Ş.**, Pre-determination of OSA degree using morphological features of the ECG signal,Expert Systems with Applications,Volume 81,2017,Pages 79-87,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2017.03.049

[9] **Karamanli, H., Yalcinoz, T., Yalcinoz, M.A**. et al. (2016). Sleep Breath 20: 509. https://doi.org/10.1007/s11325-015-1218-7

[10] **Epstein, L. J., Kristo, D., Strollo, P. J., Jr, Friedman, N., Malhotra, A., Patil, S. P., …** (2009). Adult Obstructive Sleep Apnea Task Force of the American Academy of Sleep Medicine. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine, 5(3), 263–276.

[11] **Aşık, M., Bostancı, A., & Turhan, M.** (2014). Comparison of Manual and Automated Scoring Techniques in Polysomnography. Turk Arch Otolaryngol, (52), 17–21. Retrieved from https://www.turkarchotolaryngol.net/sayilar/57/buyuk/17-21.pdf

[12] **Crummy, F., Piper A.J.**, Naughton MT Obesity and the lung: 2 · Obesity andsleep-disordered breathing*Thorax* 2008; 63:738-746.

[13] **Bonsignore, M. R., Baiamonte, P., Mazzuca, E., Castrogiovanni, A., & Marrone, O**. (2019). Obstructive sleep apnea and comorbidities: a dangerous liaison. *Multidisciplinary respiratory medicine, 14*, 8. doi:10.1186/s40248-019-0172-9

[14] **Bhadriraju S., Kemp C. R. Jr., Cheruvu M. & Bhadriraju S**. (2008). Sleep apnea syndrome: implications on cardiovascular diseases. Crit Pathw Cardiol. 7, 248–253

[15] **Garbarino, S., Durando, P., Guglielmi, O., Dini, G., Bersi, F., Fornarino, S., … Magnavita, N.** (2016). Sleep Apnea, Sleep Debt and Daytime Sleepiness Are Independently Associated with Road Accidents. A Cross-Sectional Study on Truck Drivers. PloS one, 11(11), e0166262. doi:10.1371/journal.pone.0166262

[16] **Garvey, J. F., Pengo, M. F., Drakatos, P., & Kent, B. D.** (2015). Epidemiological aspects of obstructive sleep apnea. Journal of thoracic disease, 7(5), 920–929. doi:10.3978/j.issn.2072-1439.2015.04.52

[17] **Punjabi, N.M.** The epidemiology of adult obstructive sleep apnea, Proc. Am. Thorac. Soc., 5 (2008), p. 140

[18] **Mazzotti, D. R., Lim, D. C., Sutherland, K., Bittencourt, L., Mindel, J. W., Magalang, U., … Penzel, T.** (2018). Opportunities for utilizing polysomnography signals to characterize obstructive sleep apnea subtypes and severity. Physiological measurement, 39(9), 09TR01. doi:10.1088/1361-6579/aad5fe

[19] **Url-2** <*hopkinsmedicine.org/health/wellness-and-prevention/what-to-know-about-an-at-home-sleep-test*>, data retrieved 06.11.2019

[20] **Dobrosielski, D. A., Patil, S., Schwartz, A. R., Bandeen-Roche, K., & Stewart, K. J.** (2015). Effects of exercise and weight loss in older adults with obstructive sleep apnea. Medicine and science in sports and exercise, 47(1), 20–26. doi:10.1249/MSS.0000000000000387

[21] **Peppard, P.E., Young, T., Palta, M., Dempsey, J., Skatrud, J.** Longitudinal Study of Moderate Weight Change and Sleep-Disordered Breathing. JAMA. 2000;284(23):3015–3021

[22] **Mulgrew, A.T., Sigurdson, K., Ayas, N.T.**, Adjunctive and alternative therapies. In: obstructive sleep apnea diagnosis and treatment (Kushida C.A., ed), Informa Healthcare, New York, 2007:233-46

[23] **Mador M.J., Kufel T.J., Magalang U.J., et al**. Prevalence of positional sleep apnea in patients undergoing polysomnography. Chest 2005;128:2130-7

[24] **Zozula, R., Rosen R.** Compliance with continuous positive airway pressure therapy: assessing and improving treatment outcomes. Curr Opin Pulm Med 2001;7:391-8

[25] **Chawla, N. V., Bowyer K.W., O.HallL., Kegelmeyer, W. P.**, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, 321–357, 2002.

[26] **Han H., Wen-Yuan, W., Bing-Huan, M.**, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," Advances in intelligent computing, 878–887, 2005.

[27] **He, H., Yang, B., Garcia, E.A., Shutao, L.** (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the International Joint Conference on Neural Networks. 1322 - 1328. 10.1109/IJCNN.2008.4633969.

[28] **Eiseman, N. A., Westover, M. B., Mietus, J. E., Thomas, R. J., & Bianchi, M. T.** (2011). Classification algorithms for predicting sleepiness and sleep apnea severity. Journal of Sleep Research, 21(1), 101–112. https://doi.org/10.1111/j.1365-2869.2011.00935.x

[29] **Mitchell, R. B., Garetz, S., Moore, R. H., Rosen, C. L., Marcus, C. L., Katz, E. S., … Redline, S.** (2015). The Use of Clinical Parameters to Predict Obstructive Sleep Apnea Syndrome Severity in Children. JAMA Otolaryngology–Head & Neck Surgery, 141(2), 130. https://doi.org/10.1001/jamaoto.2014.3049

[30] **Kim, T., Kim, J.-W., & Lee, K.** (2018). Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques. BioMedical Engineering OnLine, 17(1). https://doi.org/10.1186/s12938-018-0448-x

CURRICULUM VITAE

Name Surname **:** Onurhan HAMZAOĞLU

Place and Date of Birth **:** Istanbul/Turkey, 1993

E-Mail **:** onurhan.hamzaoglu@gmail.com

EDUCATION **:**

- **B.Sc.** **:** 2016, Istanbul Technical University, Faculty of Science and Letters, Mathematics Engineering
- **M.Sc.** **:** 2019, Istanbul Technical University, Faculty of Informatics, Computational Science and Engineering