T.C.
ISTANBUL KULTUR UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

MODELING EDUCATIONAL DATA WITH MACHINE LEARNING METHODS

MA Thesis by

Ayşe İlknur DİLEK

Department : Mathematics and Computer Sciences M.S.

Programme :Mathematics and Computer Science

Supervisor: Asst.Dr. Mehmet Fatih UÇAR

JUNE 2022

T.C.
ISTANBUL KULTUR UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

MODELING EDUCATIONAL DATA WITH MACHINE LEARNING METHODS

MA Thesis by

Ayşe İlknur DİLEK

(2000006380)

Department : Mathematics and Computer Sciences M.S.

Programme :Mathematics and Computer Science

Supervisor and Chairperson : Asst.Dr.Mehmet Fatih UÇAR

Members of Examining Committee : Assoc.Dr.Ozan KOCADAĞLI,
Asst.Dr.Levent CUHACI

JUNE 2022

FOREWORD AND ACKNOWLEDGEMENT

In this study, the factors affecting academic success were evaluated with the help of machine learning algorithms by using the data set obtained with the help of a questionnaire, one of the data collection tools, and the training data were modeled with the algorithms used.

I would like to thank my esteemed advisor Dr. Mehmet Fatih UÇAR, who supported me throughout the process with his knowledge, experience, devoted work, ability to eliminate problem situations, tolerance and always positive energy, for his efforts throughout my graduate education.

Prof. Dr. Remzi Tunç MISIRLIOĞLU, who shared his valuable information with us during my graduate education, who was my teacher during my undergraduate education, whose ideas I benefited from in the process, and who guided me with his exemplary stance, my course teachers Assoc. Dr. Suleyman Hikmet CAGLAR, Dr. Canan AKKOYUNLU, Dr. Uğur GÖNÜLLÜ, Dr. Mehmet Fatih UÇAR, and all my teachers who have contributed to me both academically and psychologically in my education process so far.

To my advisor, Dr. Mehmet Fatih UÇAR, who will devote their precious time during my thesis defense process, to my esteemed faculty members Dr. Levent CUHACI, Assoc. I would like to thank Ozan KOCADAĞLI.

I would like to thank all of my students, whom we crossed paths with throughout my teaching career, who helped me to look at the concept of "student point of view" throughout the process, and I hope my research result will be beneficial for them.

I would like to express my endless thanks to my very valuable and esteemed manager Hüseyin SARI, who set an example for us with his work discipline, all his work being equipped, planned, organized, instructive personality, moral values and humanistic approach, and from whom I learned the importance of the role of correct and effective communication in problem solving. .

I would like to thank my dear friends, who supported me, gave me strength and spiritually, when I was discouraged during this whole process.

With my suggestion, I would like to express my endless thanks to my colleague and my only sister Burcu DİLEK, who is also a mathematician and our student years are at the same time, for her psychological and academic support during this difficult process.

I can not pay for my efforts to come to these days, I feel their support at every stage of my life, who is always by my side in all the decisions I make, I have never experienced negative situations that may occur throughout my academic life thanks to their sacrifices and devotions, and I use time management against them the most in this process, who are with me in every difficulty. To my dear mother Gülten DİLEK, who always increased my motivation with her strength, I wanted to be a mother like her in every moment of my life, and I believe it is a great chance to be her daughter, and to my dear mother Gülten DİLEK, who raised me with great sacrifices, who taught me the value of working with her life story, struggle and determination, I would like to express my love, respect and gratitude to my dear father İbrahim DİLEK.

Contents

LIST OF ABBREVIATIONS

EKK : Least Squares Method
$\hat{\beta}$: least squares estimator
$\hat{\beta}k$ ::Ridge estimator
LASSO Least absolute shrinkage and selection operatör
PCA: Principal Component Analysis
LDA: Linear Discriminant Analysis
LR :Logistic Regression
RF :Random Forest
:Regression Coefficient
SVM :Support Vector Machine
ANN: Articial Neural Networks
KNN:K Nearest Neighbour Algorithm
DT:Decision Tree
SFFS :Sequential Forward Floating Selection
SBFS :Sequential Backward Floating Selection
VIF :Variance Inflation Factor
ROC:Receiver Operating Characteristics
AUC: Area Under the Receiver Operating Characteristics)
MSE :Mean Square Error
etc :Et Cetera, Other Similar Things

List of Tables

List of Figures

GENİŞLETİLMİŞ ÖZET

Eğitim Verilerinin Makine Öğrenmesi Algoritmaları Kullanılarak Modellenmesi

Ayşe İlknur DİLEK

2000006380

GENİŞLETİLMİŞ ÖZET

Çalışmanın Amacı

Ülkemizde akademik başarının önemi her geçen gün artmakla birlikte akademik başarıyı etkileyen faktörler çeşitlilik göstermektedir. Bu çeşitlilik; farklı alanlarda, farklı faktörlerle olmakla birlikte bu değişkenlerin bir arada değerlendirilmesinin ve bunun sonucunda tahmin algoritmaları kullanılarak akademik başarıyı yordayan değişkenlerin kendi içlerinde birbirlerini etkileme ve hedef değişken olan akademik başarıyı etkileme gücü problemin konusunu oluşturmuştur. Bu çalışmada amaç; Lise öğrencilerinde akademik başarıyı etkileyen demografik, sosyoekonomik, tutum, sosyal aktivite, motivasyon, sağlık ve spor, akademik başarı kategorilerinde yer alan anket soruları yardımı ile akademik başarının çalışmanın büyük çoğunluğunda hedef değişken olarak yer alması ve bu faktörlerin akademik başarı hedef değişkenini etkileme derecesinin tespit edilip hangi makine öğrenmesi modellerinin bu gücü anlamlı bir şekilde yorumlayabildiği değerlendirilmesi amaçlanarak bu çalışmanın sonucunda akademik başarıyı etkileyen faktörlerin ve etkileme derecelerinin belirlenerek eğitim sistemine, özellikle öğrenciye, katkı getirmesi amaçlanmıştır.

Araştırma Soruları

Akademik başarıyı etkilediği varsayılan faktörler olan demografik, sosyoekonomik, tutum, motivasyon, sosyal aktivite, sağlık ve spor kategorisinde yer alan soruların kendi kategorisi içerisinde her birinin akademik başarıyı etkileme gücü, etkileme derecesi nedir? Akademik başarıyı etkileyen faktörlerin birbirlerini etkileme derecelerini hesaplayınız? Denetimli öğrenme modellerinden olan Regresyon çeşitlerinden Multilineer Regresyon, Ridge ve Lasso regresyonlarının başarı oranları ve değerlendirilmesi nedir? Denetimli öğrenme modellerinden Sınıflandırma algoritması modellerinden olan Karar ağacı, Rastgele orman, En yakın Komşular, Destek vektör makinaları algoritmalarından hangileri başarılıdır, başarı oranları nelerdir, değerlendirilmesi nedir? Kolektif öğrenme modellerinin başarı oranları nelerdir,değerlendirilmesi nedir? Derin Öğrenme modellerinden olan Yapay sinir ağları modelini değerlendiriniz.

Akademik başarının artırılmasına yönelik çalışmalar her geçen gün artmakla birlikte teknolojinin gelişmesi ile birlikte bilgisayar bilimleri, akademik başarıyı etkileyen faktörlerin değerlendirilmesinde büyük katkılar sağlamaktadır. Makine öğrenmesi algoritmaları kullanılarak eğitim verilerinin modellendirilmesi ve veri madenciliği ve Yapay zekanın birleşimiyle verilerin sınıflandırma, tahmin ve kümeleme çalışmaları yapılmaktadır.

Çalışmaların ulusal ve uluslararası düzeyde sürekli gelişerek artması bu konudaki akademik araştırmaların niteli ve niceliğini geliştirerek bilgiye kolay ulaşılabilinmesine de katkıda bulunmuştur. Bu çalışma yapılırken Ulusal tez merkezi, uluslararası düzeydeki tezler, çeşitli branşlarda olmak şartıyla makaleler (özellikle sosyal bilimlerdeki makaleler çok fazlasıyla taranmıştır.).Dergide yayınlanan makaleler, dergi köşe yazıları incelenmiştir. Kütüphane ziyaretleri yapılarak kaynaklara direkt ulaşım sağlanılmakla birlikte online yayınlar ve online makalelere, çevrimdışı verilere uzaktan eğitim kapsamında erişim sağlanılmıştır. Konu ile ilgili adı geçen sözcükler detaylı bir şekilde incelenmiştir. Aşağıda incelenen tezler arasında konu kapsamı, içerik, kullanılan algoritmalar açısından bu çalışmaya benzer 3 çalışmadan bahsedilmiştir.

Türkiye'de Yalova ilinde 3 farklı ortaokulda uygulanan anket sonucunda öğrencilere demografik, sosyaekonomik, sağlık, spor, sosyal, aktivite, not başarı durumları ile ilgili sorular yöneltilmiş . Türkçe, Matematik ve dönem sonu not ortalamaları hedef değişken alınarak sınıflandırma ve regresyon kullanılarak tahmin algoritmaları sonucunda yordama gücü öznitelik seçiminin de uygulanması ile birlikte anlamlı sonuçlar elde edilmiştir. (Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi Murat GÖK1, * 1Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA)

Portekizde, 2005 2006 yılları arasında, iki devlet okulunda yapılan araştırmada öğrenci dağılımı 9 yıllık temek eğitim sonrasındaki gruptur. Matematik ve Portekizce notları ülkedeki eğitim sistemleri 3 aşamada değerlendirilmiş olup G1,G2,G3 olarak isimlendirilmiştir.G3 final notudur. Bu değişkenler hedef değişken olmakla birlikte Karar ağaçları ,Rastgele Orman ,yapay sinir ağları ve Destek Faktör makinaları olmak üzere farklı sınıflandırma algoritmaları kullanılmış tahmin yapılmıştır. Özellikler arasında kullanılan algoritmalarla anlamlı tahminler çıkarılabilmekle birlikte daha az etkileyen değişkenlerin var olduğu da gözlenmiştir. Ayrıca ANN ve SVM yöntemlerinin gürültülü girdilere, aykırı değerlere değişkenlerine karşı daha hassas yöntemler oldukları gözlenmiştir.

İncelenen üçüncü çalışma Kaggle platformundan hazır data kullanmış ve Karar ağacı, Rastgele orman ile sadece Lojistik regresyon kullanarak tahmin çalışmaları yapmıştır. Bu çalışmada 395 ve 245 öğrenci sayıları olmak üzere iki farklı veri seti kullanılmıştır. Tüm özellikler bu veri seti için aynıdır. En iyi doğruluk oranı Karar Ağacı algoritmasına aittir. Data setleri ayrı ayrı değerlendirilmekle birlikte 649 öğrenci total olarak da değerlendirilmiş.3 farklı veri seti seti kullanıldığı zaman ise en fazla sayıda öğrenci sayısıyla en yüksek doğruluk değeri yine Karar ağacına aittir.

Yöntem

Bu çalışmada öğrenmeyi etkileyen faktörler farklı kategorilerde olmak koşulu ile ayrıntılı bir şekilde açıklanmıştır. Makine öğrenme modellerinden Denetimli, Denetimsiz, öğrenme kavramları açıklanmıştır. Makine öğrenmesi Denetimli öğrenme modellerinde sınıflandırma algoritmaları olan Karar Ağacı,Rassal Orman,K-en yakın komşular,Lojistik regresyon ,Destek vektör makinaları,Regresyon algoritmaları olan Multilineer regresyon ,Ridge ve Lasso regresyonları ,Kolektif öğrenme modelleri ve Derin öğrenme modellerinden yapay sinir ağları modelleri açıklanmıştır

Kaggle'dan edinilen veri ilk önce kullanılabilir olacak şekilde hazır hale getirilmiştir. Makine öğrenme algoritmaları ile Denetimli öğrenme modellerinden olan Sınıflandırma, Regresyon,Kolektif öğrenme modelleri uygulanmış ve başarılı sonuçlar elde edilmiştir. Derin öğrenme modeli olan Yapay Sinir Ağları modelleri veri setine uygulanmıştır. Tahmin,

sınıflandırma ve kümeleme çalışmaları sonucunda model performansı sınıflandırma algoritmaları için doğruluk değerleri ve çeşitleri, Roc eğrisi, karmaşıklık matrisi kullanılarak değerlendirilmiştir. Regresyon modelleri olan Multilineer regresyon, Ridge ve Lasso regresyon modelleri eğitim ve test seti sonuçlarına göre değerlendirildiğinde sonuç değerlerinin aynı olduğu gözlemekle birlikte Ortalama kareler hata katsayısına göre değerlendirildiğinde en iyi çalışan regresyon modelinin Ridge Regresyon olduğu kararına varılmıştır. . Derin öğrenme algoritması olan Yapay sinir ağ modelinde perceptron kulanılarak başarılıbir sonuç elde edilmiştir .

Sonuç ve Değerlendirme: Regresyon modelleri kendi içerisinde, sınıflandırma modelleri kendi içerisinde değerlendirilerek en iyi performansla çalışan modeller değerlendirildiğinde; Regresyon modelleri içerisinde Multilineer Regresyon, Lasso Regresyon ,Ridge Regresyon modelllerinin eğitim ve test sonuçları (her üçünün de aynı) sırasıı ile 0.87 ve 0.77 dir. Ortalama kareler hata katsayısı değerleri incelendiğinde içerisinde Multilineer Regresyon 6.40, Ridge Regresyon 6.41, Lasso Regresyon 6.39 ortalama kareler hata katsayısına sahiptir. Regresyon modellerinde değerlendirme yapıldığında diğerlerinden açık ara fark olmamak üzere skorlar değerlerine bakılarak en iyi performansla çalışan sınıflandırma modeli Lasso Regresyon olmuştur. Sınıflandırma modelleri kendi içlerinde değerlendirildiğinde;

Karar Ağacı algoritması değerlendirildiğinde Doğruluk değeri : 0.89 Roc eğrisi altında kalan alan değeri :0.97

Rassal Orman algoritması değerlendirildiğinde Doğruluk değeri : 0.91 Roc eğrisi altında kalan alan değeri :0.97

Destek Vekör Makinası algoritması değerlendirildiğinde Doğruluk değeri : 0.92 Roc eğrisi altında kalan alan değeri :0.97

XgBoost algoritması değerlendirildiğinde Doğruluk değeri : 0.90 Roc eğrisi altında kalan alan değeri :0.97

AdaBoost algoritması değerlendirildiğinde Doğruluk değeri : 0.86 Roc eğrisi altında kalan alan değeri :0.95

Bagging algoritması değerlendirildiğinde Doğruluk değeri : 0.92 Roc eğrisi altında kalan alan değeri :0.97

Lojistik regresyon algoritması değerlendirildiğinde Doğruluk değeri : 0.94 Roc eğrisi altında kalan alan değeri : 0.97

K- En yakın komşular algoritması değerlendirildiğinde Doğruluk değeri : 0.80 Roc eğrisi altında kalan alan değeri :0.84 sonuçlarına ulaşılmıştır.

Yapay Sinir Ağları algoritması değerlendirildiğinde Doğruluk değeri : 0.94 Roc eğrisi altında kalan alan değeri :0.89

Bu çalışmanın sonunda ; Türkiye'de farklı okul türleri, farklı sınıf düzeyleri, farklı bölgelerden oluşan geniş bir örneklemle öğrenmeyi etkileyen faktörler farklı kategorilerde ve geniş bir şekilde yer almak şartı ile ,öğrenmeyi etkileyen faktörlerin başarılı algoritmalar ve modeller ile birlikte toplanan veri setine uygulanması ve bu çalışmada anlamlı sonuçlar veren geliştirdiğimiz model ve algoritmaları uygulayarak ülkemizde eğitime katkı sağlamaktır.

Anahtar Kelimeler: Makine öğrenmesi, Derin Öğrenme, Yapay zeka, Yapay sinir ağları, Çoklu Lineer regresyon, Polinomsal regresyon, Lojistik regresyon, Lasso and Ridge regresyonları, Karar ağacı, Rastgele Orman,Destek Vektör Makinaları, En yakın K komşuları,

Yapay sinir ağları, K ortalama algoritmaları,Topluluk öğrenmesi,

Bilim Dalı Sayısal Kodu : 20515

University : Istanbul Kültür University
Institute : Institute of Graduate Education
Department :Mathematics and Computer Science
Programme : Mathematics and Computer Science
Supervisor : Asst.Dr.Mehmet Fatih UÇAR
Degree Awarded and Date : MA – June 2022

ABSTRACT

MODELING EDUCATIONAL DATAS WITH MACHINE LEARNING METHODS

Ayşe İlknur DİLEK

In our country, the effect of the academic success of the student, especially in the secondary education period, on the stage of choosing the profession he will have in the future and on the academic career goal is an undeniable reality. Academic success is affected not only by the data belonging to the academy, but also by many different categories. It is affected by many factors, especially methodological, and this diversity increases with individual differences. Regression and Classification from supervised learning models and Clustering algorithms from unsupervised learning models were applied to the data set. Multiple linear regression, polynomial regression, Lasso and Ridge regressions,Decision Tree, Random Forest, Support Vector Regression as regression methods, Decision Tree, Random Forest, Support Vector Machine, Logistic regression, K Nearest Neighbors methods were used as classification methods. As Clustering methods we are used K means algorithms, hierarchical method as unsupervised learning methods. In addition Artifical Neural Network, a deep learning algorithm, were applied to the data set. In the study, these factors and sub-factors were evaluated categorically and machine learning was used. Various determinations were made with estimation algorithms by establishing relations that predict the academic achievement target variable . By evaluating the data results, it is aimed to determine which factors affecting success are significant according to the sample group studied, which variables affect success individually and categorically, and the degree of influence, and as a result, it is aimed to contribute to education.

Keywords: Machine Learning, Deep Learning, Artificial intelligence, Artificial Neural Networks, Multiple linear regression , Polynomial regression, Logistic regression, Lasso and Ridge regressions, Decision tree, Random Forest, Support Vector Machine, Artifical Neural Network,Bagging,XgBoost,AdaBoost

Science Code : 20515

1 INTRODUCTION

Academic success is the student's performance that can be measured with different techniques at certain periods on a course basis. The grades that the student has on the basis of the course are the concrete data formed as a result of the student's performance.

In the questionnaire, in which the factors affecting academic success were used, artificial intelligence and machine learning algorithms were used to calculate the degree of influence of the variables on each other and the degree of influence of the variables on the academic achievement variable, which is the target variable. By the classification of the variables, it is determined which technique would be appropriate to apply to the relevant data set. In our research, the factors affecting the academic success and the degree of influence were determined by different techniques and approaches. Using the information here, the factors affecting the academic success of the students are determined, so that the negative reinforcers that affect the success are also eliminated.

In our thesis , although the factors that affect academic success are accepted as the target variable, the effects of each of the survey questions on each other as a separate independent variable are also mentioned using graphics and tables .

1.1 Problem Statement

While the importance of academic success in our country is increasing day by day, the factors affecting academic success are increasing in different categories. Although there are different factors in different fields, The power of evaluating these variables together and, as a result, the effect of the variables that predict academic success by using estimation algorithms, and the power of affecting academic achievement, which is the target variable, constituted the subject of the problem.

1.2 Aim Of The Thesis

The purpose of this study is to determine the factors affecting academic achievement in high school students, to determine the degree of influence of these factors on the academic achievement target variable and to determine which techniques are appropriate. For this purpose, answers to the following questions were sought: What are the factors affecting academic achievement during high school education? What are the degrees of influence? According to the success rates, which techniques from artificial intelligence applications are suitable for interpreting data? Sub-problem sentence: Calculate the degree to which the factors affecting academic success affect each other? What are the success rates of Regression, Clustering, Classification algorithms? Which technique is more suitable?

1.3 Research Significance

In the research, demographic, social, academic, psychological and personal questions that affect academic success were determined as a result of the answers obtained by asking the

students with the help of a questionnaire, the effect of the variables on academic success and the effect power of the variables were determined as a result of the Machine learning algorithms. Determining the factors that affect the learning and academic success in this way, which hinders the academic success of the students. It is important to eliminate the factors, to bring success-enhancing methods to the agenda, to increase motivation.

1.4 Assumptions

It was assumed that the high school students participating in the research gave their answers sincerely and accurately. It is assumed that the sample group is representative of the population.

1.5 Limitations

This research is limited to the responses of High School students in two schools in Portugal.

1.6 Definitions

Training data: The set of observations presented for the algorithm to learn.

Academic Success: Academic success is the achievement of one's goals in the field of education.

Machine Learning: It is the modeling of systems with computers that make predictions by making inferences on data with mathematical and statistical operations.

Modeling: It is the description of the system using machine learning algorithms.

1.7 Universe-Example

Sample: Students studying in two specific high schools in Portugal.

Universe: High school students

Method: the research model, universe-sample (Study Group), data collection tools, data collection techniques and data analysis are explained.

2 BACKGROUND

2.1 FACTORS AFFECTING ACADEMIC SUCCESS

Academic success is the student's performance that can be measured with different techniques at certain periods on a lesson basis. The grades that the student has on a lesson basis are the concrete data formed as a result of the student's performance. Academic data is affected by many factors, especially cognitive, affective, kinetic, individual, environmental, and methodological, and this diversity increases with individual differences. In this study, it is aimed to make various determinations with estimation algorithms by categorically evaluating these factors and sub-factors, using machine learning algorithms, and establishing relations that predict the academic achievement target variable. By evaluating the results of this study, it is aimed to determine which factors affecting success are significant according to the sample group studied, which variables affect success individually and categorically, and the degree of influence, and as a result, it is aimed to contribute to education.In this part of the study, the factors affecting academic achievement and included in the survey questions were examined.

2.1.1 Attitude

Today, among the important determining factors of national language and foreign language learning, the features related to the learner are effective among the factors that affect success, just like the effect on learning Mathematics. If students have a positive attitude towards that language, especially in learning a foreign language, they are successful in learning the language easily, being inquisitive and curious, adding new information to the knowledge they have learned and improving themselves. It will help the system in determining the degree to which the student reaches that goal. In this study, it is predicted that besides the cognitive characteristics of the student, the affective characteristics of the lesson are at least as important as the cognitive effects. The most important affective feature is the attitude towards the lesson.

Studies have shown that the attitudes of the individual towards his mother tongue are positive. The reasons for this are cultural heritage, the awareness of being a nation, the fact that we are born with a language, the ability to communicate with the people around him thanks to this language, the effect of the individual on the continuation of his life and the sense of belonging. Although correlation values will be examined as a result of our research and prediction algorithms will be established by applying different methods, at this point, if factors such as native language, foreign language learning, the basic structure of the language, the family of the language, the similarity of the native language and the relevant foreign language, and grammar rules are taken into account, a supporting power It is undeniable. The more the student masters his mother tongue, the easier it will be to learn a foreign language. The student will be able to learn a new student language at the level of proficiency in native language. Attitude will be directly proportional at this point. .According to this,

Academic achievement, which is our target variable in this study, is affected by many cognitive factors as well as effective factors. One of these variables is the attitude towards the lesson, which arises from the learner-related features from the effective features. Although there is a generally correct ratio between the attitude towards the lesson and the academic success, this judgment can change from lesson to lesson.

Although there are many factors in different fields that affect success in lesson.Among these concepts, Attitude variable is the most easily expressed by students compared to others, it covers other effective concepts the most and it is the item on which a lot of work has been done. When the studies examining the correlation between the mathematics achievement of the students and the attitude towards the Mathematics lesson are examined and the feedback received as a result of the communication with the students, it is observed that the attitude towards Mathematics outweighs the cognitive factors. In this study, it was aimed to investigate to what extent students' attitudes towards lessons affect success by examining many factors affecting success. According to the resulting data, it is aimed to include affective elements in order to increase lesson achievement. Lessons and attitudes towards lessons research results show a lot of variability. Therefore, it is possible to come up with results that will contribute to this general perspective by using too many techniques while evaluating the elements and variables in our research. When the results of the research done so far are examined, when the studies in which there are significant relationships between the variables of achievement and attitude towards lessons are examined, it is observed that as the positive attitude towards Mathematics increases, the success in Mathematics increases. However, there are also research results in which there are no significant or weak relationships between lesson achievement and attitude. Estimating the relationship between the different techniques and algorithms used and these variables is one of the primary objectives of the study.

Today, among the important determining factors of national language and foreign language learning, the features related to the learner are effective among the factors affecting success, just like the effect on learning Mathematics. The importance of the student's attitude towards the relevant lesson is revealed. If students have a positive attitude towards that language, especially in learning a foreign language, they are successful in learning the language easily, being inquisitive and curious, adding new information to the knowledge they have learned and improving themselves. Otherwise, success decreases. Therefore, determining the student's attitudes towards the relevant courses will help the system in terms of determining the goal of the course and the degree to which the student reaches that goal. In this study, it is predicted that besides the cognitive characteristics of the student, the effective characteristics towards the lesson are at least as important as the cognitive effects. In our study, effective characteristics, while liking the lesson, being interested in the lesson, and academic self-confidence towards the lesson, perhaps the most important effective feature is the attitude towards the lesson.

Studies have shown that the individual's attitudes towards own native language are positive. The reasons for this can be shown as cultural heritage, the consciousness of being a nation, the fact that it is a language that we are born with, the ability to communicate

with the people around it thanks to this language, the effect on the continuation of the individual's life and the sense of belonging. Although correlation values will be examined as a result of our research and prediction algorithms will be established by applying different methods, at this point, if factors such as native language , foreign language learning, the basic structure of the language, the family in which the language is located, the similarity of the native language and the relevant foreign language, and grammar rules are taken into account, a supporting power can be found. it is undeniable. The more the student masters own native language , the easier it will be to learn a foreign language. For this reason, the academic success of the students was also wanted to be examined. By using these data, the success correlation of the native language and foreign language will be examined. To what extent the student will be able to learn a foreign language at the level of proficiency in their native language and how successful they will be is among the questions that are wondered, and it will be answered as a result of the study within the sample we will use as a result of the prediction algorithms. Although the attitude factor is in the category of effective data, we often hear the following statement from the students. "The more I succeed, the more I like the course". In general, although it is not expected that the positive attitude towards the mother tongue will be at the same level for a general sample in the foreign language, it is noteworthy that some students have a high level of passion for learning a foreign language and their capacity for self-development. In this research, using estimation algorithms, interpretations will be made by taking into account the strong relationships of the variables with each other.

2.1.2 Demographic

The presence of the student's demographic information in the study is extremely important in terms of defining and categorizing the student profile. In addition, it has been supported by various studies that demographic characteristics are among the factors affecting success. Although cultural differences are effective, for example, if we evaluate in terms of mathematics achievement in our country and give an example with a specific unit content, it has been observed that male students are more inclined to solve questions and approach the question with faster and more accurate results in solving speed problems. From this point of view, information such as the student's age and gender is valuable in terms of identifying the student and being the factors affecting his or her success.

2.1.3 Socioeconomic

The social and economic welfare level of a country has an extremely important place on the effectiveness of the education quality, efficiency and reflection on the students as well as the comfortable life of the citizens of the country. In this study, in a part of the questionnaire created in order to evaluate social and economic aspects, questions that have the degree of social and economic impact were also included. In this study, it is among the aims of the study to estimate to what extent demographic variables such as economic level, gender, parental education, occupation, family economic income, number of siblings in the family, which are thought to have an impact on the academic success of the individual, affect the academic success of the individual by using various methods using machine learning

algorithms. Depending on the social and economic variable, among its general objectives, it is aimed to raise awareness about these effects when students are exposed to the negative effects of this variable and to minimize these effects in terms of the student's educational welfare. Equality of opportunity in education, which is one of the basic principles of national education in our country, is an inclusive principle that includes these variables. It is one of the implicit goals of the research that this variable and the degree of its effect on success also provide benefits in terms of increasing the projects carried out to support socioeconomically disadvantaged students.

2.1.4 Social Support -Social Activity

The fact that students engage in social activities throughout their education process also generally has a positive effect on their academic achievement. It is likely that the academic success of the student, who is involved in social activities, will increase if he/she makes a time plan. For example, we observe that the life of the student who is interested in sports is disciplined and he carries this discipline to many fields. Social activity questions are also included in our study and this situation is tested with our existing sample. In our country, considering this factor, various incentives are given to students in terms of social activities. Human is a social entity. As a result of being social, they may have various sensory needs, and when these needs are not met, they have to struggle with feelings such as anxiety, anxiety and being alone. At this point, social support is a phenomenon that exists in order to prevent these negative feelings from occurring. Human beings need this support very much at almost every age, especially in today's conditions. Especially in the competitive environment of young people, the continuation of their education processes, the status of their social relations, economic conditions and many other factors are effective, and it has become necessary for the individual to have social support for a healthy life. Social support includes family members, friends, teachers, specialists, and people in our social network who can help us. At the point of motivation towards the lesson, which is one of the factors affecting the academic success of the students, especially the social support received from the teacher positively affects the academic success of the students. Considering the power of this variable to affect success, related questions were also included in our research. At this point, when evaluating, there is an evaluation prediction with more than one variable. For example, a female student's need for social support is higher than a male student's, indicating that social support will be more effective at that point. These variables were to be evaluated together when interpreting the data results.

2.1.5 Learning Types, Motivation

Permanent learning that is not based on rote, learning at the cognition level and above, and academic achievements as a result; It takes place as a result of effective and planned studies. The target question asked to be successful in almost every subject, the purpose for which the relevant course will be studied and the knowledge in which area and how it will benefit you as a result of the process are one of the prerequisites of motivation to study that

course. Motivation is an indispensable variable for academic success and student-teacher and parents should cooperate in order to provide the necessary motivation. In addition, the teacher's lesson plan includes the part of informing the students about the target. Attention to the lesson and motivation is the basis of effective lecture listening. The student's knowledge of efficient study techniques is another important factor affecting academic success. While each lesson has its own unique study techniques, each student's lesson understanding techniques are also different from each other. Students who can match these two dynamic elements correctly will be more successful academically. After providing the necessary conditions with the factors described so far, the student should have the ability to study in a planned and efficient way to ensure academic success for the relevant course. For this reason, the most helpful factor is undoubtedly time management. Unconscious internet use, unnecessary time spent in traffic between home and school, can be among the situations that can prevent students from studying among the time losses, which are one of the biggest problems of today. In order to prevent these losses, the student should plan a time and act according to this plan. The fact that the working environment is a suitable area for work is also one of the factors affecting success. Personal time management Another factor that is as important as arranging the time and affecting the efficiency is the organization of the working environment. Success is inevitable in studies carried out by using a suitable working environment, sufficient motivation, correct time management and correct technique.

2.1.6 Health And Sports

There is a significant relationship between the physical and mental development of students and their regular nutrition. In our research, it was aimed to test this factor with questions such as breakfast habits and daily protein intake and similar questions targeting this item. While regular nutrition and sports are so important for human life, they are of vital importance for young people at the age of development. In this context, students are encouraged to engage in healthy nutrition and sports through health information lessons, public service announcements and activities.

2.1.7 Emotional Intelligence, Life Satisfaction and Academic Success

Although emotional intelligence is an effective concept throughout human life, the correct management of emotional intelligence is a variable that affects academic success. The concept of life satisfaction, which is another affective element, is also an element of emotional intelligence, and it is variables that mutually affect each other. It is about the student's general view of life, how well he defines life, how positive he can look, the energy to live life, the degree of satisfaction with his life, the meaning he gives to life. As a result, life satisfaction, emotional intelligence and academic achievement are among the factors that mutually affect each other. For example, when a student achieves low academic success, his life satisfaction decreases and he is badly affected emotionally. Contrary to this situation, life satisfaction decreases during an emotionally challenging period, and as a result, a decrease in academic achievement can be observed. These variables can directly affect each other dynamically. Our research includes questions measuring these variables.

2.2 BIG DATA ARTICICAL INTELLIGENCE

With the development of today's technology, the possibility of making inferences from the data and making predictive predictions using the data has also increased. Thanks to this opportunity, the data available; Thanks to the analysis studies, the methods and algorithms used, they can not stay on their own and take the knowledge to the next level.

Boyd and Crawford notice that "Big Data is less about data that is big than it is about a capacity to search, aggregate, and crossreference large data sets". [1].

There are many different definitions of Artificial Intelligence when the literature is searched for Artificial Intelligence. Although there is information about how the data in the brain is received and stored, it has not been fully explained yet, especially how the data is transferred. While the algorithms of the human thought system are such a curiosity, the definition of artificial intelligence had to constantly update itself and different definitions were formed. Among the reasons for this are the definition of intelligence, how exactly the process of thinking using human intelligence is realized, the transfer of this thinking system to a machine, and perhaps most importantly, the definition is constantly updating itself while keeping up with technological developments. If the artificial intelligence concepts in the literature are examined, artificial intelligence is:

Artificial intelligence encompasses machine learning. In machine learning, there is more specifically the use of learned information.

The transformation of data into competitive advantage is what makes "Big Data" such an impactful revolution in today's business world. [2]

When machine learning algorithms are examined in detail, it is seen that the basis of all algorithms is based on mathematical models. In machine learning; Predictive predictions are made by applying mathematical and statistical operations on the data. Machine learning uses algorithms to identify relationships, similarities and differences in data. If we compare it to human learning, the more people communicate with people and the more life experiences they have, the higher their learning will be. In the same way, machine learning will make more accurate predictions the more data it has and the more trials are done. In machine learning, the process is dynamic, so its integration into data is useful in terms of adaptability in case of data change.

Machine learning can reach a result thanks to the algorithms it uses. Algorithm is one of the important steps applied in solving an existing mathematical problem and it is a solution method that contains functions for the solution. In machine learning, it can be defined as the procedure applied to machine inputs to obtain the outputs given by the machines. As it is known, a problem has more than one solution. In our study, using different estimation algorithms, the estimation levels of these algorithms will be evaluated. The suitability of the estimation algorithms used; It is related to factors such as the degree of estimation of the target variable, the operating speed of the algorithm, and the algorithm's sample-inclusive

results.

As in the problem solving steps in machine learning, first of all, analyzes are made about the data we have, and a suitable model is created according to the results obtained from the analysis. The data set is introduced to the model to be used, it is aimed to train and learn the model used on the data set. As a result of the process, the model comes to the level of understanding and interpreting the data set data in the appropriate format and gives the appropriate output. In machine learning, this process is dynamic, data can be removed from the data set or data can be added, the system will continue to work. In our study, the most suitable model will be decided according to the success rates by using various models and algorithms.

2.3 MACHINE LEARNING

Data mining: Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. [3]

Machine Learning: It is the process of evaluating and solving the problem that occurs as a result of the computer learning about a similar event that it is about to encounter and transferring the knowledge and experience it has learned to the new event. Machine learning investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data.[4]

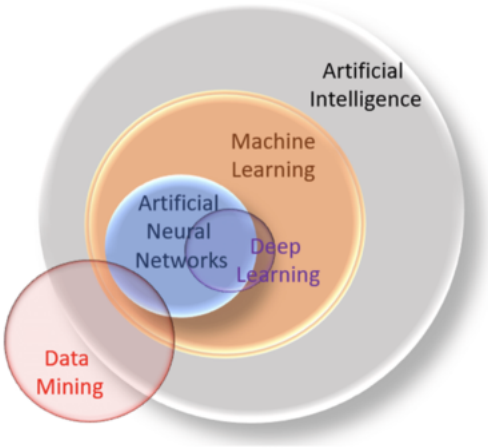


Figure 1: Correlation of Data Mining ,ANN,Deep Learning and Machine Learning

Deep Learning: Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to "learn" from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.[5]

Artifcial Intelligence: Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind.[6]

Artifical Neural Network : Artifical Neural Network is a computational model that consists of several processing elements that receive inputs and deliver outputs based on their predefined activation functions " [7] Machine learning is a discipline that dramatically improved. It has a lots of types which base on learning. Provided that the same procedure is applied in our study, machine learning studies progress within the framework of a certain discipline. The first step in this study is the existence of appropriate data and documentation

processes (collecting, categorizing and usable data). The next step is to determine which learning method will be used and to process the data with appropriate modeling, methods and algorithms. The final stage is the evaluation stage. The performance of the machine learning used here is tested. Factors such as the suitability of the model, its ability to categorize or predict, its ability to predict results, reliability and validity are evaluated. Learning; It is divided into 4 according to the factor of being human support:

2.3.1 Supervised Learning

In general, the data set consists of dependent and independent variables. In this learning, the machine that learns the relationship, rule and plot in the data set given to the machine creates a rule as a result of its evaluation between the independent variables and the dependent variables. The function operator and the mechanism used in mathematics are very similar. The match between the image set and the domain set in the function is formed by the function rule. The learning action of machine learning is based on finding the function rule. If the domain and value set are considered as independent variables and dependent variables, respectively, these sets are observed and the rule that emerges when the relationship between them is determined is machine learning. In the next process, it is determined which target variable result will be reached by any new independent variable. This value can be either a numerical value or a categorical data. Supervised learning is generally expressed in two different ways:

Regression: It is the degree to which the unit effect in the independent variables affects the target variable quantitatively. Multiple linear regression, polynomial regression, logistic regression, Lasso and Ridge regressions are used in our study.

Classification: It is the process of estimating a categorical variables Decision tree, Random Forest, Naive Byes, Support Vector Machine, K nearest Neighbors, Artifical Neural Network, Logitic Regression are the classification methods used in our study.

The supervision in the learning comes from the labeled examples in the training data set. We send machine data set and conclusion of data set then programma give us between the data set and conclusuion data set correlation named regression and clustring.The main aim is we hope that the machine find out the function about between data set and conclusıin of data set.Tecnically supervised learning study at two process which are training and test.The process of Training work out like that: the machine get data and find out by using supervised learning algorıtms then the machine predict .

Index	Features			Target
	input_1	input_2	input_n	
1	x	x	x	y
2	x	x	x	y
3	x	x	x	y
4	x	x	x	y
5	x	x	x	y
6	x	x	x	y
7	x	x	x	y
8	x	x	x	y
9	x	x	x	y
10	x	x	x	y

Table 1: Independent and Dependent variables on data set with ındex and feature names

2.3.1.1 Train Set

The sample taken is the most intense part of the data set. The sample taken from the data set varies between 60% and 90%.

2.3.1.2 Test Set

This stage is the evaluation stage. The success rate of the model and the algorithms used in accordance with it will be tested here.

The implementation is as follows: The target variable to be tested in the application is removed at this stage. Variables other than the target variable in the test set are added to the established model. Here, the aim is for the model to accurately predict the target variables. and the performance level of the model is determined.

2.3.1.3 Validation Data set

The data in the validation data set is taken from the train data set. Because it will be difficult to deal with large data in cases where the amount of data may be large, this sample will make our work easier. It works on about 20% of data. The model that was studied on the train dataset before and learned by the machine is applied.

While working on the train data set, choosing the right application model and determining the appropriate algorithms for this model, this model selected in the validation data set is developed and made more suitable.

Table 2 is a horizontal representation of table 1, showing how it is divided into train, validation, and test.

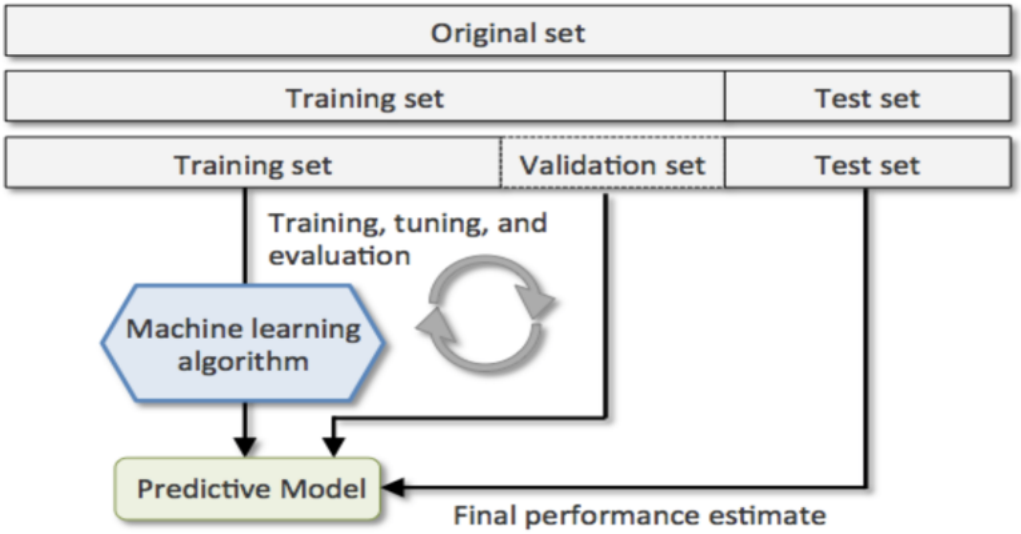


Figure 2: The model of train ,test and validation set

2.3.2 Unsupervised Learning

In unsupervised learning, the system learns from the data itself without being supervised. The data are not labeled. There are no target variables. Unsupervised learning allows more complex operations than supervised learning. The performance evaluation of the model is more specific and relevant to the relevant field. K means, Hierarchical Clustering Analysis, Principal Component Analysis are examples of clustering.

2.3.3 Semi-Supervised Learning

It is a type of machine learning that is between unsupervised learning and supervised learning. It is also a type of supervised learning when working with unlabeled data as in unsupervised learning.

2.3.4 Reinforcement Learning

It is a type of learning that imitates the learning that takes place biologically in the human brain. Reinforced learning is based on the reward-punishment system. The main purpose in reinforced learning is to receive the highest reward in the environment. The model is dynamic. Will it try new ways to get a high reward? The choices between these two options will determine the performance of the model. The best performance is to try new ways by gradually moving from the previous award point to reach the highest reward.

2.4 MULTICOLLINEARITY AND FEATURE SELECTION

Some times in multivariate regression equations the success rate of the algorithm may be low. This was the case in the first place in the data sets we examined. Why is the success rate of an algorithm compatible with the model under normal conditions? The most important reason for this problem is the multi-connection problem in this technique. What is the multi-connection problem? It is the situation when the relationship between the variables predicting the target variable is high, and sometimes the features of some variables include the other. In the multivariate regression method, the variables are expected to have elements that are as independent as possible and especially not affect each other. At this point, if there is a correlation between the variables for which regression will be used, and especially if this correlation coefficient is high this compromises the accuracy of the regression. As a result, the target variable gives rise to the illusion that the intended variables are not well represented, or it leads to the illusion that the appropriate technique is not used in the model. However, the cases where the variables are orthogonal are unfortunately few in number. In general, multicollinearity problem is seen in multiple linear regression.

In order to solve the multicollinearity problem, it is first necessary to determine which variables cause this situation. There are various techniques to find out which features of the problem. The most commonly used technique is to calculate the coolinearity coefficients of the variables included in the regression analysis. method is used.

The variance increase factor is the most widely used method for finding problems.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Since the regression coefficient will be used in VIF, the explanation and formulas of the regression coefficient and the corrected regression coefficient are given below.

2.4.1 Regression Coefficient

While constructing the regression curve, we should draw such a line that this line can represent all the data in our data set as much as possible. In short, this line must pass through the closest location to the points. For this, the distances of all the points in the data set to the line are calculated. The line where this distance is minimum is the regression line. The technique is the least squares method. The right points should represent our points, but of course there are too many points that are not on the line. These points are called waste. .We call every point that can't hit the right point as waste. After the mean value of the system is found, the sum of the distances of the points in the data set from the mean is another factor that affects the success of the regression curve. The regression coefficient is the sum of the distances of the wastes from the line, subtracted from the ratio of the distances from the data to the mean. the lower the R^2 value, the higher the result. This result is an indication that the regression line is drawn so successfully.

A high value of R^2 indicates that the regression model is successful. The value of R^2 is called the regression goodness fit index. The analysis used until now was simple regression. the denominator of the fraction belonging to the formula will tend to increase and the value of the fraction to decrease, and therefore the value of the regression goodness index will tend to increase.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coeffient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

It will ensure that the model fits perfectly. However, this is not a correct indicator. The corrected R^2 coefficient is used in models that use multiple variables. The rationale for the correction coefficient is to go over the unnecessary variables that have no effect on the existing target variable in the system. This corrected value is in data sets with multiple variables. Which will give more reliable and valid results about the fit of the regression model, while under normal conditions, more than one variable can affect any event we encounter in our daily life.

$$R^2_{adjusted} = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-m-1}\right)$$

$R^2_{adjusted}$ = the adjusted multiple correlation coefficient
R^2 = the original multiple correlation coefficient
n = the number of cases
m = the number of variables

At the beginning of these determinations is the variance inflation factor—VIF (variance width factor). The common degree of variance between the variables that control the target property is checked. Target variable It finds the variance of the predictor variables or the non-common variance of two variables as a coefficient that goes from 1 to infinity. By looking at the value of VIF between any two variables, the correlation and degree of relationship between the variables is found. According to the formula, if VIF is $1 : 1-R^2$, for example, if the VIF value is 5 correlation coefficient. It was 80 percent. This shows that the relationship between the two variables is stronger than the middle, at this point, a multicollinearity problem has occurred between these variables. The higher the VIF value, the higher the multi-connection problem.

2.4.2 Feature Selection

In machine learning, the success rates of the methods may differ even if the same methods are used in the problems where the same data set and the same dependent variable are the

target. It is extremely important for the research result to be selected according to the criteria, to determine which features will be included in the data set in which the own data will be created, together with the correct modeling. That's why creating and choosing the right variables from the beginning is the most important part of this job. The feature decision is more important than the selected model, the algorithms used, the number of variables used and the size of the data set, in terms of affecting the success of the result. The aim is to guide us towards the result. It is the selection of quality and right features that will lead you to the right way. The most important contribution to the success of the model is the algorithm and techniques used, as well as the cleaned features set.

When choosing a feature, we can make various combinations and not take some of it at all. Based on this idea, it allows a lot of diversity for problem diversity. For example, let's have 6 features whose names are the letters of the alphabet, a, b, c, d, e, f. These features can be combined in different combinations. There are 63 different situations in which it will be used. In this case, even if the same methods are used with the same data set, it will give 63 different results for a data set that has 6 features only from the criterion of select feature. If the appropriate features are determined in the selection process and the right features are included in the research, the result will be successful. The large number of variables may decrease the working performance of the model.

Using too many variables can reduce the performance of the model. The large number of variables in the model may complicate the model. Our desire is to work with variables with a simple and clear representation power. It is also one of the conditions where the model is desired to work quickly. The complexity of the model is among the factors causing the over-learning problem, which is one of the most important problems, and the number of variables that we can control, at least at the beginning, is among the factors that cause it.Thanks to its qualified features, it enables the model to learn the variables rather than memorizing them, thus preventing the over-learning problem. Considering the above-mentioned situations, the answer to the question of why we shouldn't take all the raw features in the data set is given. Feature selection is examined under 3 headings.

2.4.2.1 Basic Cleaning Methods

The data cleaning part is explained in this section.

Disable the feature

As for what we should pay attention to when choosing a feature, and what steps we should follow, we must disable the data that we do not have at a high rate in the first place. For example, if the question of the relevant feature in our questionnaire was left blank by most of the students, its presence in our data set will not contribute to the result and even cause the method to be misinterpreted. However, in small data losses, we do not need to directly remove it. There are also techniques for completing these data.

It is appropriate to disable categorical data that have all the same outputs or all outputs

different. For example, if we examine the success rate according to the provinces, all of our results will be "Istanbul" in the first place. In this case, the system cannot infer from this answer. In this case, the answers will be equally distributed with 81. In this case, the system will not be able to find the answer to success according to the provinces. In short, the fact that all of our data is the same and our data is the opposite, all of them are different. Therefore, this data should be removed from the data system. The repetition of the same value as in the example of Istanbul, and the problem of all answers being the same, if it belongs to a numerical value in our existing data set, then the variance calculation is observed, and if the variance of the feature is low, the relevant feature is disabled.

Variable elimination and selection according to statistical methods
Target variable and related feature correlation

The relationship between the variables and the target variable is extremely important for the model to be successful. Independent variables that affect the target variable to the extent that they cause it are valuable for us. It is aimed to find the variables that can affect the target variable the most. Therefore, the variables that affect the target variable at the highest degree are selected by looking at the rate of the variables affecting the target variable one by one using statistical methods. The methods used for numerical variables and categorical variables are different.

Pearson Correlation:
In correlation, the degree of relationship between the target variable and the variables we want to query is determined. It is used for numerical variables. This coefficient is a measure of the degree of relationship between the target variable and data variable. This coefficient is a coefficient that varies between -1 and +1. While there is a strong positive relationship between the relevant variables as you go from 0 to +1, as you go from 0 to -1, there is a strong positive correlation between them. There is a strong negative correlation. It is used for numerical values.

Chi-square Test:
This test is used for categorical data. The stages of use of the test are as follows. First, the relationship between the target variable and the independent variable is graded.. After measuring the relationship between the categorical variable and the target variable with the chi-square test, the best k features or the features that fall into the best k percentile will be taken.The selection is made using functions. It is also used in the decision tree and random forest methods and in the tree pruning part. [8]

Anova test:
That is, we use it to measure the relationship between a categorical variable and a numerical variable. The effect ratio of the existing groups of the categorical variable to the numerical variable is checked. Anova test: It is used to determine whether there is a significant change between categorical variables that are thought to affect the target variable. The Anova test alone provides limited information. In our research, there is data on numerical values that represent academic success. The purpose of the anova test is to measure the

existence of significant changes in cases where these numerical values are affected by the categorical data in our data. For example, the father's profession variable will categorically separate the data from the variables that affect academic success in our research. Answer options are 1) Paid Employee/Worker 2) Officer 3) Education sector employee 4) Engineer 5) Health sector employee 6) Law sector employee 7) Trade 8) Other 9) When the options not working are examined, is there any difference in terms of affecting the result? Thanks to the Anova test, we prevent type 1 error. Type 1 error is the interpretation of categorical data as existing when there is no significant difference between the variables. Thanks to the Anova test, for example, the answers to the father's profession question ,Type 1 error will be avoided since positive error will not occur, since a positive error will not occur between the two. that is, if there is a factor affecting academic achievement, we cannot learn from the Anova test the information between which variables this significant difference occurs. By applying different tests in addition to the Anova test, it is learned from which elements of the variable this difference arises. Post hoc tests to determine between which groups the difference detected in the Anova test is in the image below, the methods used change depending on whether the input and output variables are numerical and categorical.

In the data cleaning and statistical methods described so far, individual variables have been disabled or by applying statistical operations on individual variables, the effect and degree of effect on the target variable has been found. This is not a costly method, but when a few different variables come together that will not affect the result alone. In other words, while these variables are not significant on the target variable on their own, they can create significant effects when combined.

In the techniques described in feature selection 1, when the relationship between the feature and the target variable is examined, the features are looked at individually. The effect of the spiral structures formed as a result of the combination of these features on the target variable will be examined. As a disadvantage of helical structures, examining different combinations of features is one of the superior features of the technique, as it is not examined individually for each feature. Selecting a set of meaningful properties by adding and subtracting variables one after the other (Wrapped Methods) There are multiple submodels of this technique.
i) Sequential Forward Selection (SFS)
ii) Sequential Backward Selection (SBS)
iii) Add $L - R$ Subtract (plus l – minus r)
iv) Sequential Forward Floating Selection (SFFS)
v) Sequential Backward Floating Selection (SBFS)

2.4.3 Sequential Forward Selection (SFS))

In our first technique, the variable that affects the target variable the most is found by looking at the features one by one. This variable is our most important feature that affects the target variable and the first degree. Then, the binary combinations are examined by adding the 2nd variable next to this feature that affects it at the 1st degree. Among the binary

combinations that affect the 1st degree, the best combination that affects the target variable is found. Afterwards, this search system continues sequentially. Thanks to the stopping criterion, the model with a certain number of combinations at a certain point is the model with the most representative features in which it affects the target variable. At the end of the technical implementation, these features are selected.

2.4.4 Sequential Backward Selection (SBS)

In this method, calculation is made by creating a model in which all features are used at the first moment, as opposed to forward selection. It eliminates the worst performing variable and re-builds the model with the missing variable, and then calculates a model performance by using all the features simultaneously each time. Afterwards, it tries to build the model by removing each variable in turn and eliminates that variable that causes the worst performance. Then, it again removes the variables in order, detects the feature that affects the model performance the worst and eliminates that variable. While these processes continue, the extraction process is stopped the first time the performance of the new model decreases. The features obtained as a result of this process are determined as our best features. plus l – minus r selection

2.4.5 Sequential Forward Floating Selection (SFFS)

In this method, it starts from the empty set. The algorithm works on a dynamic structure that changes as the name suggests in the sliding method model, instead of the constant determination of the L and R values in the other method.

2.4.6 Sequential Backward Floating Selection (SBFS))

This method starts by using all the features. L and R values are determined by the algorithm. It continues as the reverse of the forward sliding selection model.[9]

2.4.7 Recursive Feature Elimination:

It is similar to the backward feature selection method. All features are ranked in order of importance towards a specific target, and the features with the least importance according to a certain criterion are eliminated. The aim here is to include only the features that maximize the performance of the model in the data set. It is used in support vector and decision tree classification methods.

2.4.8 Select From Model:

It is a feature selection model that works as an “all or nothing law”. First, a threshold value is determined according to certain parameters. The features that exceed this threshold value are continued, and other features are eliminated. The threshold value can be a fixed numerical value or a function. The average of the features, the most repeated value of the features can be given as an example for this situation.

2.4.9 Dimensional Reduction:

When the size download is reviewed in the literature, it is explained in this section with the idea that it will be available in the ”Feature Selection 3” section, although it is included in the ”feature selection” in only some sources.

2.4.9.1 Principal Component Analysis (PCA):

PCA is a linear method used in categorical data where there is unsupervised learning without a target variable. It can be used in clustering algorithms to give more accurate results. The PCA method was needed as a result of the fact that the existing independent variables, which are the main cause of the connection problem, have a strong bond among themselves, sometimes the two variables are strong, and as a result, wrong inferences are made. In this way, the features that are strong among them will be combined and act as a single feature, thus reducing the number of dimensions. Being economical is a desired feature in terms of machine efficiency.

Our goal in PCA is to represent the data set with the lowest size, having the highest variance, and to have new features for accurate results with the specialized features of the raw data set that are capable of meta-predicting. Here, like other methods, features are not eliminated, features are not disabled. Certain features existing in PCA come together to obtain a new feature set. Grammar words and can think of it as the concept of “phrases”. While each of the words has different meanings, the phrases that come together from the words can have wider, different meanings. At this point, words are not eliminated, but new word groups are obtained by bringing them together. The number of dimensions is reduced, but the effect power is increased. The reason why it is called principal component analysis is that the names of newly acquired properties are called principal components.

2.4.9.2 Linear Discriminant Analysis (LDA) :

It is used in the supervised learning model where the target variable exists and the features are controlled with the target variable. It can often be used in target variables where numeric values exist. For example, in classification algorithms, which are one of the supervised learning models, the aim is to separate the classes with sharp clear lines. We should have such features that each of them should be able to represent the class they are in. As in the PCA method, the aim is to reduce the number of dimensions and to have strong singular

features in the system.

2.5 ALGORITHMS

2.5.1 REGRESSION ALGORITHMS

2.5.1.1 Linear regression

It is a statistical method that measures the extent to which the dependent variable changes as a result of the change in the unit of the independent variables by establishing a mathematical modeling between the independent variables and the dependent variables. If the number of independent variables is one, it is called Simple Regression, if more than one, it is called Multiple Regression. In the regression model, a model is created using machine learning algorithms in the data set, which is separated under the name of train and test, and it is aimed to predict the result when new data is included. Although supervised learning is in question, the estimated value is a numerical value. [10]

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Y_i = Dependent Variable
β_0 = Population Y Intercept
β_1 = Population Slope Coefficient
X_i = Independent Variable
ϵ_i = Random Error Term
$\beta_0 + \beta_1 X_i$ = Linear Component
ϵ_i = Random Error Component

2.5.1.2 Multilinear Regression

The statistical technique measuring the extent to which the dependent variable was affected by only one independent variable was simple regression. In multiple linear regression, although the degrees of the variables are mostly first-order, the number of independent variables affecting the target variable is more than one. In order to understand the working principle of the regression method, generalizations were made to our regression topic through the simple regression title. However, even the simple events we encounter in our daily lives may not be the only reason. At this point, simple regression is not enough to meet the needs. Multiple linear regression is a type of linear regression that measures the degree to which more than one variable affects the target variable, based on the existence of a linear relationship between them. The formula of the multivariate regression model is as follows. In simple regression, the points in the data indicate a linear line, while in the multivariate regression, the points in the data scan a planar area.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:
y_i = dependent variable
x_i = expanatory variables
β_0 = y-intercept(constant term)
β_p = slope coefficients for each explanatory variable
ϵ = the model's error term (also known as the residuals)

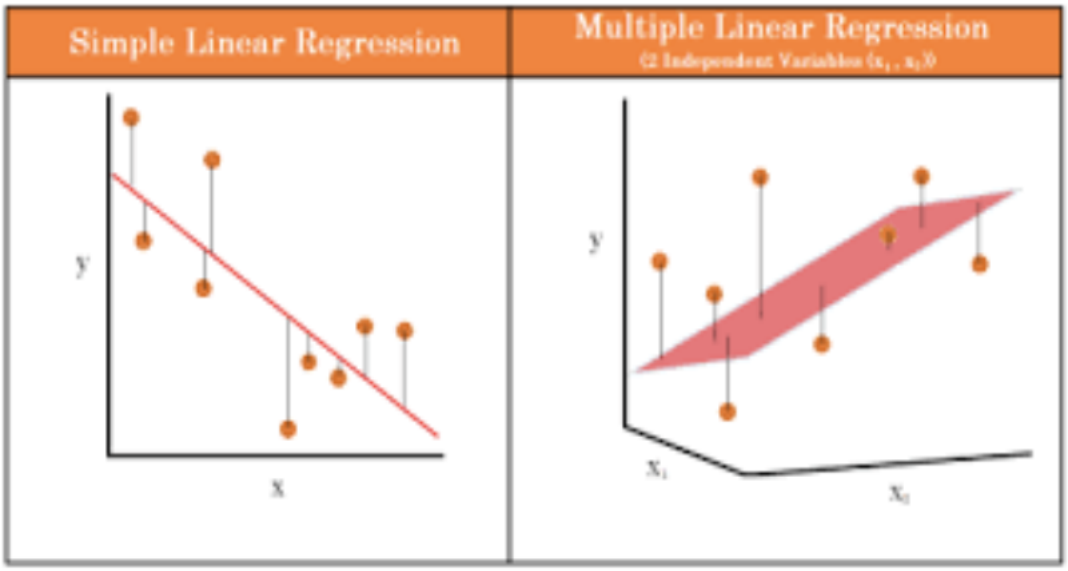


Figure 3: Compare of Simple Linear and Multiple linear graphs

2.5.1.3 Polynomial Regression

Polynomial regression is needed in cases where the independent variables affecting the target variable are not linear. In fact, the reason why most regression models have low success rates is that the model cannot adapt to linear regression and needs polynomial regression. For example, one of the variables may affect the target variable proportionally to its square, not to a one-to-one degree. In this case, linear regression cannot give us the correct predictive value.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_n X^n + \epsilon$$

The purpose of polynomial regression is to find the coefficients. The graph of the polynomial regression is not straight creates a curve

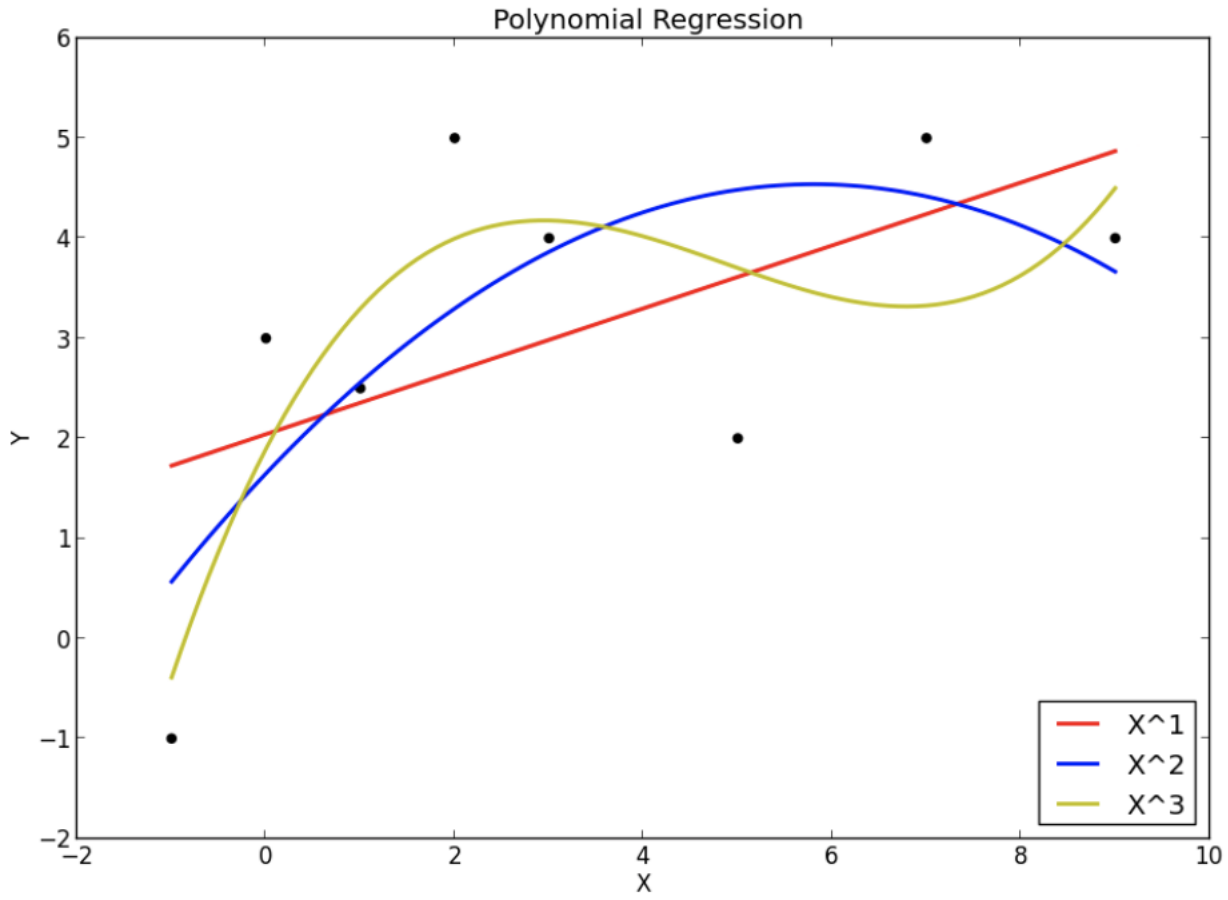


Figure 4: The graph of polynomial regression

The red graph indicates true and is linear regression. The yellow and blue graph consists of a curve and is a polynomial regression. The blue graph indicates a quadratic polynomial equation. The yellow graph represents a third-order cubic polynomial equation.

Before explaining Ridge and Lasso regressions, the factors that cause errors in algorithms and their solution methods will be explained. In addition, the reason for the emergence of Ridge and Lasso regressions will be explained with this link. The least squares method used in regression is a mathematical method that aims to minimize the error. If the mean squares sum is accepted as the error criterion, it is true that this error is accepted as the sum of two general errors, Bias and Variance. There are generally two major sources of error in machine learning. If we examine it on a sample, let's aim to develop a model with a 4% margin of error. If the existing error is 10% in the training set and 11% in the test set, is it possible to increase the number of data among the measures to be taken in this case? The answer to the question is negative. Because the system already learns wrong, it gives this error. For this reason, we should first try to increase the performance in the training set. The problem in the example described is the high bias problem. It measures the degree of inaccuracy of the model. Insufficient learning occurs in the presence of a high bias problem. Although the reasons are various, for example, it may even be possible to use a linear function when it should be represented by a 3rd degree function. For this reason, it is extremely important to analyze our data set sufficiently and to choose appropriate models and algorithms. In order to clarify the concept of Variance, if we continue with the same example, if there is an error in our model with a 4% error margin in our 1% training set and 15% in our test set, if the model is successful in the training set but unsuccessful in the test set, it means that the model memorized the training set in which it did not learn the training set. This situation is called overlearning. The target board model is examined in the image below.

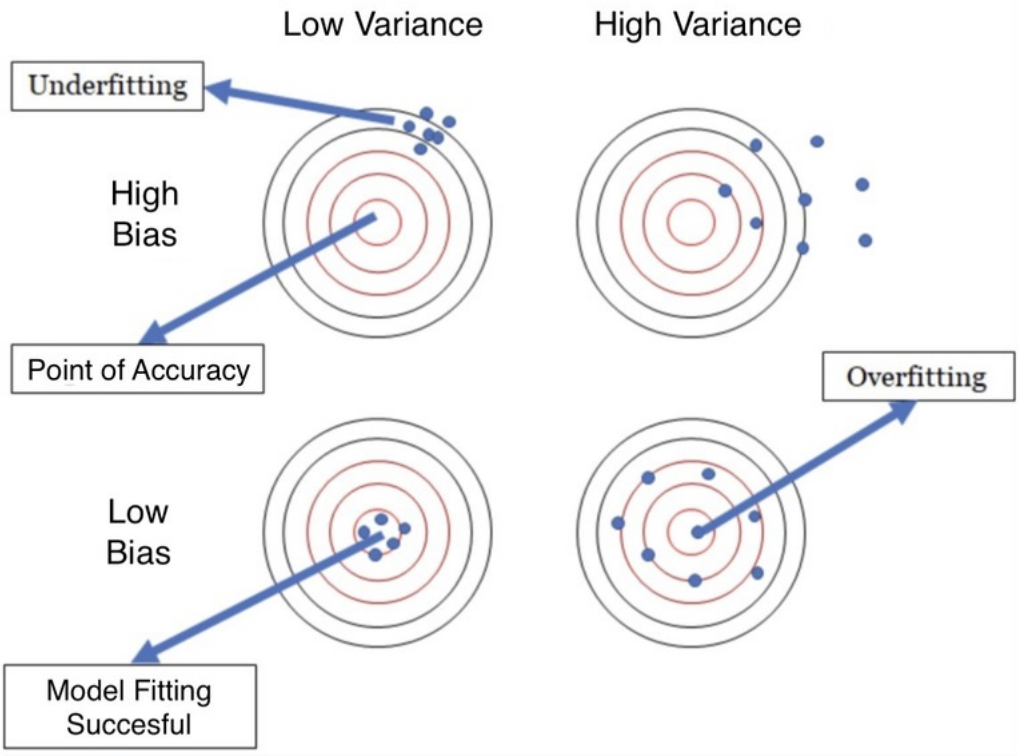


Figure 5: The dots on it are shot, the target shooting board and bias / variance relationship

In the 1st image, the shots are not in the targeted area. The deviation between the targeted area and the area where the shots are located is high. In this case, the bias value, that is, the error of error, is high. The variance value is low. The model is not well trained. There is incomplete learning. In the second image, the shots are not fully within the targeted area and the shots are not homogeneous but spread out. In short, the model has both high bias and high variance. 3. The model is successful in the image. All shots are at the target point and homogeneous. The model is well trained.

In the 4th image, the shots are in the targeted area, but the shots are not homogeneous in the area where the shots are located. In this case, the model did not learn but memorize. There is overlearning as a result of low bias and high variance. As stated in both examples, being underfitting means that the model has not even learned the training set yet. The model could not learn the training set and its variance is low, so the failure to learn is homogeneous. Among the factors that may cause underfitting are the lack of data in the data set or the inability to make an appropriate modeling or even the wrong modeling. In order to eliminate this problem, a suitable modeling should be made for the appropriate data set. In general, when the data set is sufficient, this underfitting situation arises in linear and logistic linear models. The reason is that these models are applied when the model is not linear. your model In case of underfitting, the regularizatio value should be decreased in order to increase the number of data, increase the model complexity by increasing the number

of independent variables, and simultaneously increase the weight coefficients of the variables.

The case of being overfitting is that the model learns the training set very well and fails in the test set. In this case, the bias value is low, the variance value is high. The variance value can also be called the difference between the test set and the training set learning situation. The overfitting model memorizes all the variables in the training set. Because it memorizes, it cannot transfer its learning to the test set and fails. Overfitting is proportional to the complexity of the model. The less the target variable, the more independent variables, the more complex the model. In this case, the model fails to make inferences between the variables and find the effect on the target variable. Since it cannot find any rules, it uses the option to memorize the training set. In order to reduce the complexity of the model, regalurization should be done. By reducing the weight of the variables with high weight in the model, those variables are penalized. Lasso and Ridge functions. Models with a high probability of overfitting: Decision Trees, k-Nearest Neighbors Support Vector Machines can be given as examples. Among the reasons may be that the raw data is not clean. As a solution to this situation, data cleaning and data preprocessing can be done in order to reduce the noisy data. In cases where there is a multi-connection problem, these features can be disabled. A single variable can be created from these variables. If the training set is simple and suitable for memorization, the data in the training set can be diversified by adding data.

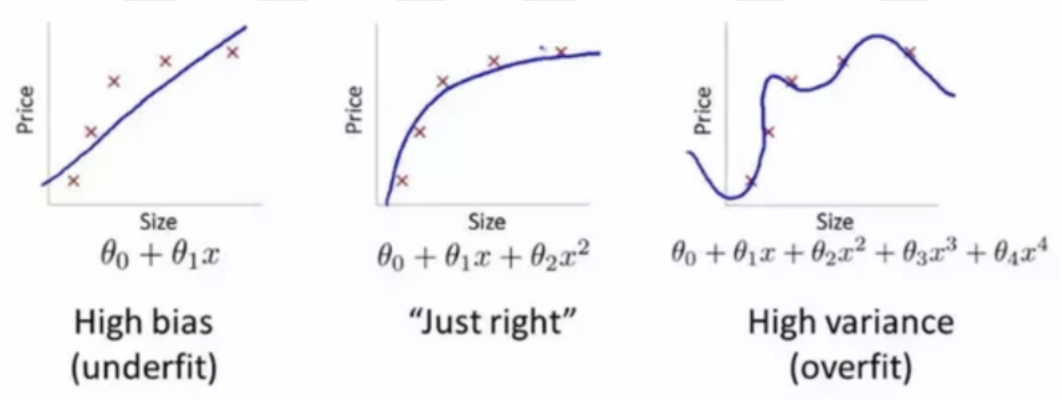


Figure 6: Complexity, Bias and variance relationship

When the graphics are examined, simple models have high stakes. The higher the complexity of the model, the higher the variance value. As the model complexity increases, the bias value will decrease, but the variance value will increase. Increasing the prediction rate on the training set, but the model fails on the test data set. This creates the bias variance dilemma. This dilemma can be clearly seen when the chart below is examined.

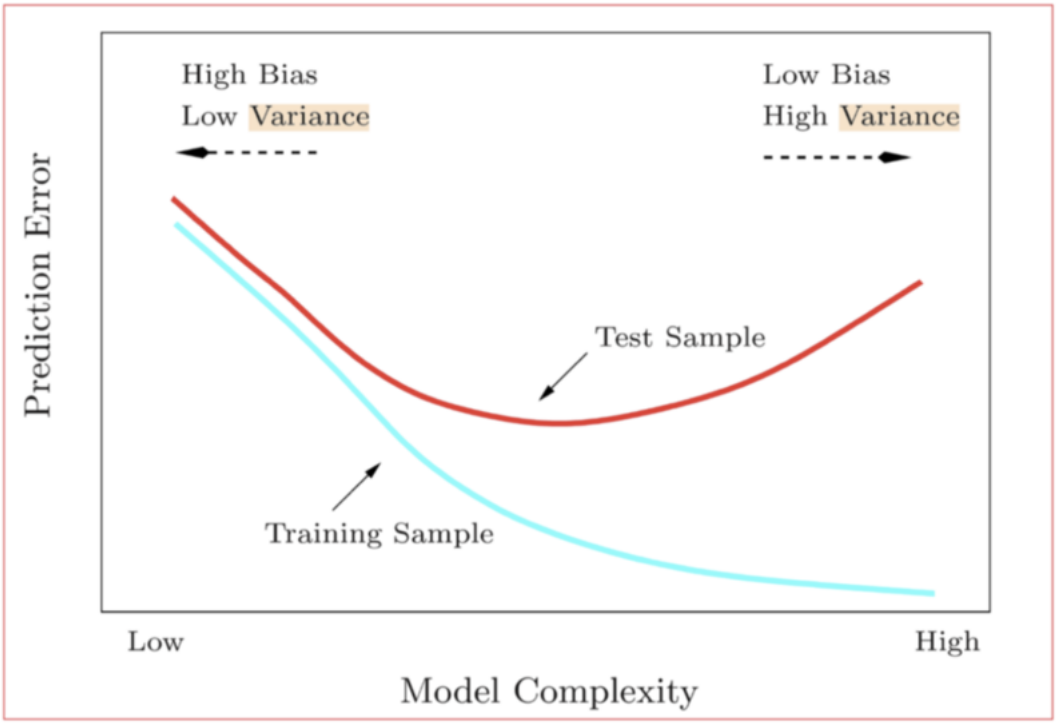


Figure 7: Complexity, Bias and variance relationship on the same graph[11]

One of the errors other than Variance and bias in machine learning is irreducible error. In this type of error, the data in the data set may be incomplete, inconsistent and noisy. In this case, the existing data should be made usable with data preprocessing and data cleaning.

MULTILINEAR CONNECTION PROBLEM

In order to solve this problem, which occurs when there is a strong correlation between the independent variables affecting the target variable, it is necessary to first determine which variables cause this situation. There are several approaches to identify which features are causing the problem. One of them is the Variance Inflection factor. In this method, independent variables are made dependent variable respectively and regression coefficients are obtained by regression with other variable. The formula is as follows.

$$VIF_i = \frac{1}{1 - R_i^2}$$

If the formula is examined, the higher the correlation between the two variables, the higher the VIF value. If the regression coefficient formula is reminded:

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coeffient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

If there is no relationship, the regression coefficient will be zero, and the VIF value will be 1. In the case where the relationship is maximum, the regression coefficient will take the value of 1. According to the limit rule in mathematics, the result of $1/0$ will be ∞, so the result will be ∞. In general, if the VIF value is greater than 10, there is a multi-connection problem. In cases where there is a multicollinearity problem, removing the variables that cause the VIF value to increase may be the solution. Variables with a linear relationship between them can be combined. The volume in the data set can be increased. Lidge and Lasso regressions, which are alternative methods and used in the study, will be explained below.

2.5.1.4 Ridge ve Lasso Regression

It is one of the methods used in the presence of a multilinear problem or when the model is overfitting. In this regression, an alternative method to the least squares method will be used.

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Least squares method

$$\hat{\beta}_{lasso} = arg_{\beta} min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j \right\}$$

The aim of the least squares method is to ensure that the residuals are minimal. Lasso regression forces the coefficients of some variables to be 0 thanks to the added error term. That is, it envisages removing those variables. In this case, feature selection is also made.

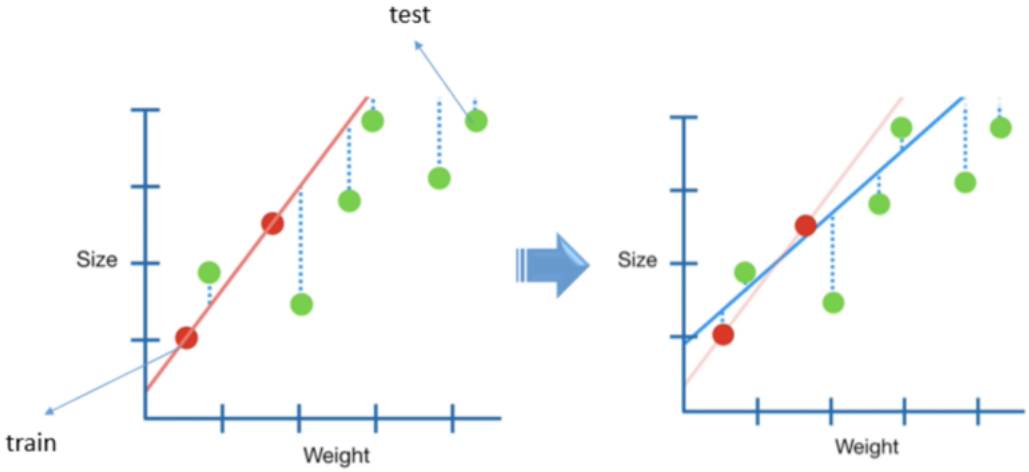


Figure 8: Fit of points on the line

When the graphs above are examined, the model in the 1st graph is fitted with EKK. When the 1st graph is examined, it is seen that the red trained points memorize the model, therefore the test data cannot adapt to the model. Our aim is to find a new truth. It should not be neglected that when looking at the graph in the EKK method, there is no problem in the bias value, there is a problem in the variance value. When finding a new truth, the existing bias value will be waived.

If it is explained with an example, for example, if the factors affecting the success in mathematics are taken into consideration, let's consider the age beta indices as motivation, sleep pattern, coffee consumption and time spent for fun. In this case, the regression formula is:

Mathematics achievement: $\beta_0 + \beta_1 * \text{Motivation} + \beta_2 * \text{Sleep patterns} + \beta_3 * \text{Coffee consumption} + \beta_4 * \text{Time spent having fun } \lambda * (|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4|)$ In Lasso Regression, where the λ value is used, as the value increases, "λ_1" and "λ_2" will move inversely as the value increases, respectively, but the variables that do not affect the target variable will be directly zero. This shows that Lasso regression disabled 2 features and fulfilled the feature selection function, as seen in this example.[12]

When examined graphically, this can be thought of as making the model fit again by reducing the weight values of the variables. The model will become higher biased but with lower variance and will be freed from being overfittting. The slope of the graph changes as the λ value changes in the lines below. The important thing here is to find the appropriate λ value. By using the cross validation technique, the most appropriate λ value is found and the model is fitted.

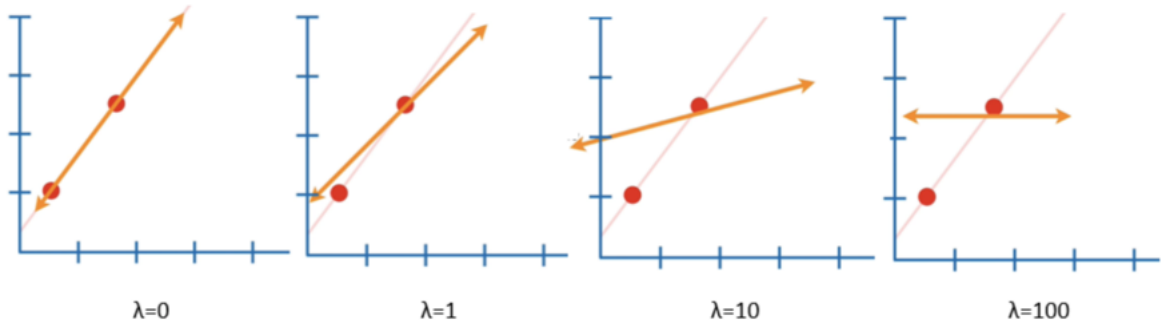


Figure 9: Graphs of the relationship between the slopes of the straight lines and the coefficient [13]

When the graph is examined, as the λ value increases, the weight values of the variables decreased and even the coefficients of some variables became 0. In this way, variable selection is realized.

Ridge Regression

Just like Lasso regression, Ridge regression offers a more biased but lower variance model compared to LCC. In Ridge Regression, the parameters with the coefficient of λ do not equal 0. In this case, no features are disabled. Ridge regression is a technique used in cases where features need to be preserved. Since it does not select features, it reveals a more complex model than Lasso regression. Since the weight coefficients of the variables will decrease inversely proportional to the λ value, the variance decreases, but some bias increases.

$$\hat{\beta}_{ridge} = arg_{\beta}min\left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j^2| \right\}$$

2.5.2 CLASSIFICATION ALGORITHMS

2.5.2.1 Logistic Regression

Logistic regression is a prediction algorithm like other regressions. In logistic regression, the effect of independent variables on the dependent variable is probabilistic as bianary. While the dependent variables in the regressions examined so far can be continuous, discrete and qualitative, in the logistic regression the dependent variable can only be qualitative. For this reason, although its name is regression, it is more suitable for the classification algorithm category. The graph below shows the logistic regression graph and its suitability for the classification model.

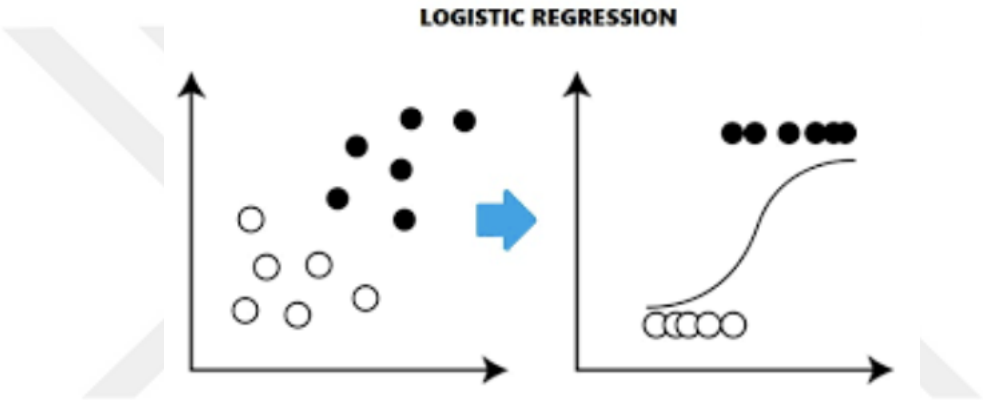


Figure 10: Representing the logarithm function of the data

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$$

,where p = probability of positive.

$$odds = \frac{p}{1-p} = \frac{ProbabilityOfPresenceOfCharacteristic}{ProbabilityOfAbsenceOfCharacteristic}$$

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

As can be understood from the formula, the logarithm ratio of the probability of an event occurring/not happening is used.

The increase in the number of variables in the logistic regression method increases the validity of the model and causes the problem of over-learning of the model. In order to prevent this situation, alternative coefficients have been developed that are compatible with the R^2 coefficient, the model and the algorithm.

The regressions aimed to minimize the sum of squares of error. The minimum sum of the squares of the waste was tried to be determined correctly. In logistic regression, it is aimed to have the highest probabilities of the observed values. Logistic regression uses the maximum probability estimation method instead of using the least squares method. This method is based on the estimation of the parameters of the existing data set.

" The basic principle behind the maximum likelihood method is the expectation: "The occurrence of a random event is because it is the event with the highest probability of occurrence." This method was invented by the British statistician Sir Ronald A. Fisher (1890-1962) in the 1920s." [14]

Among the advantages of logistic regression is that the data do not have to be normally distributed. Logistic regression is sensitive in cases where the correlation between independent variables is high, that is, in case of multicollinearity problem.
Linear and logistic functions are compared below.

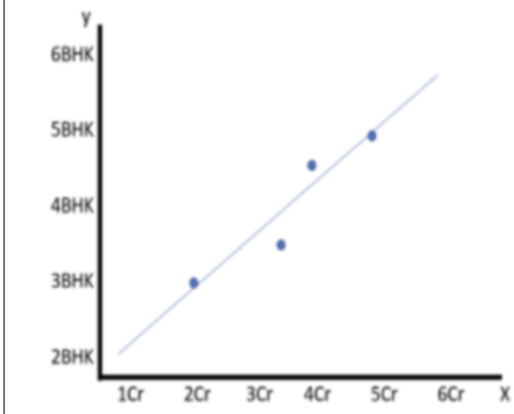
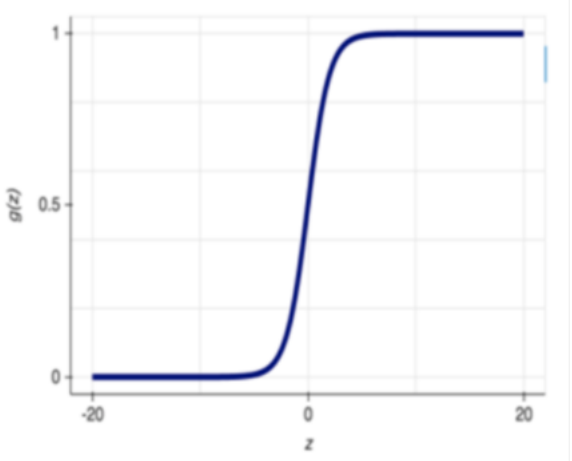
Linear Regression	**Logistic Regression**
Target is an interval variable	Target is discrete(binary or ordinal) variable
Predicted values are the mean of the target variable at the given values of the input variable	Predicted values are the probability of the particular levels of the given values of the input variable
Solve regression problems	Solve classification problems
Example: What is the Temperature?	Example : Will it rain or not?
Graph is straight line	Graph is S-curve
	

Table 2: Compare of Logistic and Linear algorithms

"Logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability."[15]

2.5.2.2 Decision Tree

We need to complete two main operations in order to create the tree structure in the bait sense. It is the process of pruning the tree, in short, to create the tree with the training set in the existing data set and then to remove the unnecessary, redundant and misleading data. Because if all the features are included in the tree structure in the decision trees, problems occur in the data set in the test phase because excessive learning occurs in the machine. Therefore, pruning of the tree should be considered as a factor that prevents excessive learning of the machine.

There are two types of pruning. The first is pre-pruning and the second is final pruning. (preprunning) Pre-pruning in decision trees is based on the principle that some branches should not exist during the creation of the tree. The tree is formed and the tree is pruned simultaneously. Or, when the tree is formed and it is decided that it is large enough, pre-pruning can be done to prevent further branching. When it reaches the final pruning, the tree is created with all its elements. Under certain conditions, the tree is pruned. For example, after the tree is formed, it is observed that some branches have no members and those branches are pruned. The important process step is the pruning process. Pruning The process can be done in two ways.

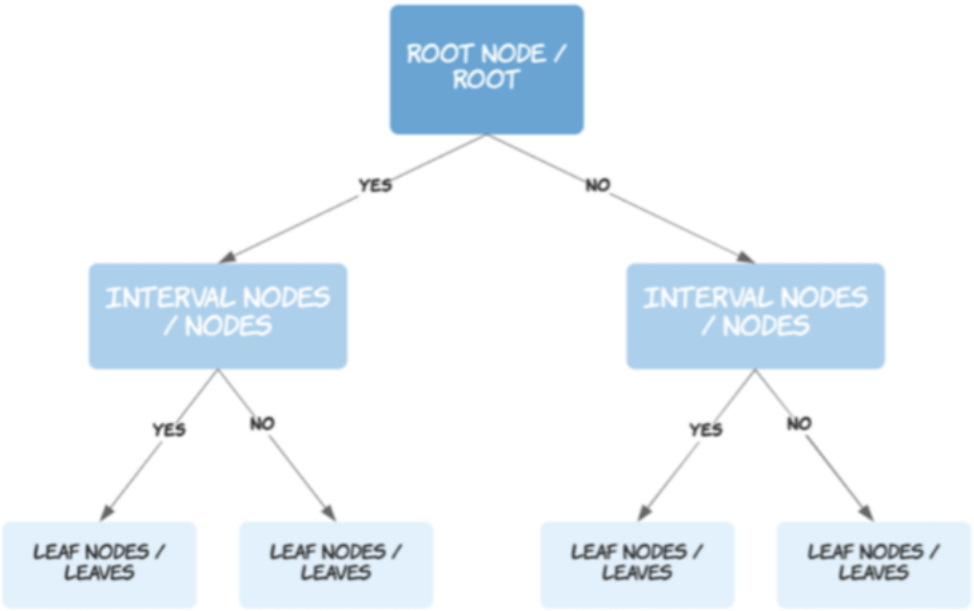


Figure 11: Roots and Nodes in Decision tree

Let us now examine the structure of decision trees. At the top of the decision trees is a structure called root. first (root or root node). Just below the root cells are nodes (interval nodes or nodes). At the bottom of the decision tree are leaves (leaf nodes or leaves). The task of the leaves is to show us the final stage of the decision. Since decision trees are based on entropy, let's explain entropy in detail. In Information Theory (Entropy in Information theory)

Entropy: It is a measure of how much uncertainty the feature in the data set. It is the concept of information theory that calculates how organized the information is in the system it is in and the amount of regularity. Irregularity means that the information does not continue steadily. In the entropy formula,

H:entrpoy

n :is the amount of information. amount

π: probability of having information

$$H = -\sum p(x) log p(x)$$

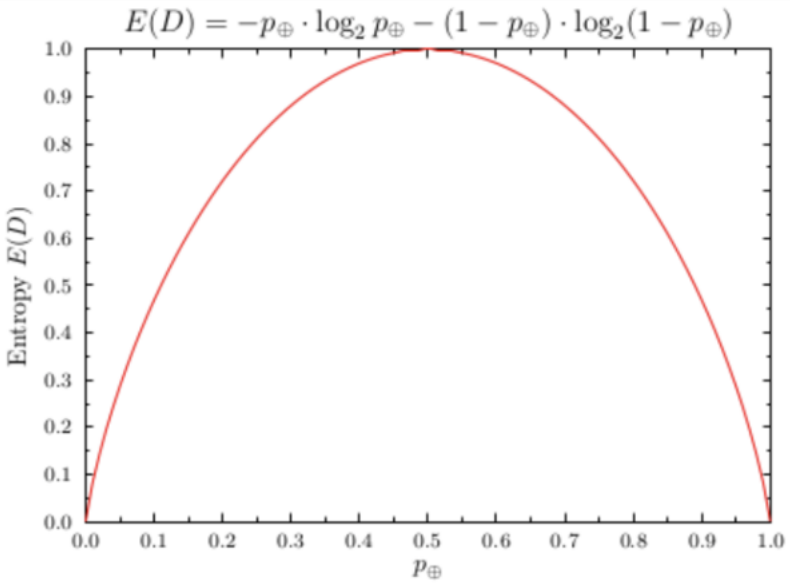


Figure 12: Entropy and Probability of Graph

Looking at the graph, it can be seen that:

It reaches its maximum value when the two probabilities are equal. The maximum disorder is at this point. Because when considered graphically, this point becomes the peak of the graph. As the probability value of the feature gets closer to 1, the entropy value decreases and the feature becomes more stable. As a result, the entropy value decreases.

Information gain is a concept that acts inversely with entropy. We can say how much value a feature in the data set will have for the class it will be in. For example, let's say

we have a feature and 5 separate classes. If it takes 5 different values, the entropy of this feature is 0, so the information gain is 1. The big gain is related to the correlation amount of the relationship between each feature and the class.

The formula for information gain is given by:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

The most important part of decision trees is deciding with which feature to choose the root of the tree. The root should be chosen in such a way that it can represent all the features in the data set. Therefore, the most important factor affecting the data set must be the feature at the root. There are many approaches that help in selecting the root. We look at which feature is important to divide in the decision trees, we find how much that probability passes on my site, and the effect of that feature in the decision tree, we divide it. It is more advantageous to divide from which point in the decision tree and according to which feature, we have 3 methods to calculate this: gain gain ratio or gini value about how balanced it divides the system after But first explain the concepts of entropy and information gain.

2.5.2.3 Random Forest

The Random Forest algorithm is a useful supervised learning algorithm that can be used in both classification and regression problems and does not cause problems in terms of compliance. The random forest algorithm first trains many decision trees and then each instance of all decision trees one by one. hundreds of decision trees, and then it trains each decision tree on a different observation sample. The random forest algorithm takes its final predictions as the average estimated value of the trees it trains individually . The accuracy of the result with the number of trees used There is a correct proportion. In order to construct a classification tree, there is the attribute that best determines the examples in the learning set. With this feature, the so-called branch and leaves of the tree are separated and a new sample set is created. A new defining attribute is found from the instances on this parsed branch and new branches are created. If all instances in each subdataset, ie on the branch, belong to the same class, there are no other attributes to parse the instances, and there are no other instances with the value in the remaining attributes, the branching process ends. Otherwise, there is a re-determining feature to parse the sub-dataset (Albayrak, 2015).

2.5.2.4 Support Vector Machine

Support Vector Machines are a machine learning method that aims to find the boundary points that classify the data of the dataset. The aim is to clearly distinguish different classes from each other with a sharp, clear line vector using boundary points with high margins.

Among the advantages of support vectors, the most important one is to be able to work with multiple independent variables and to model by distinguishing decision classes.

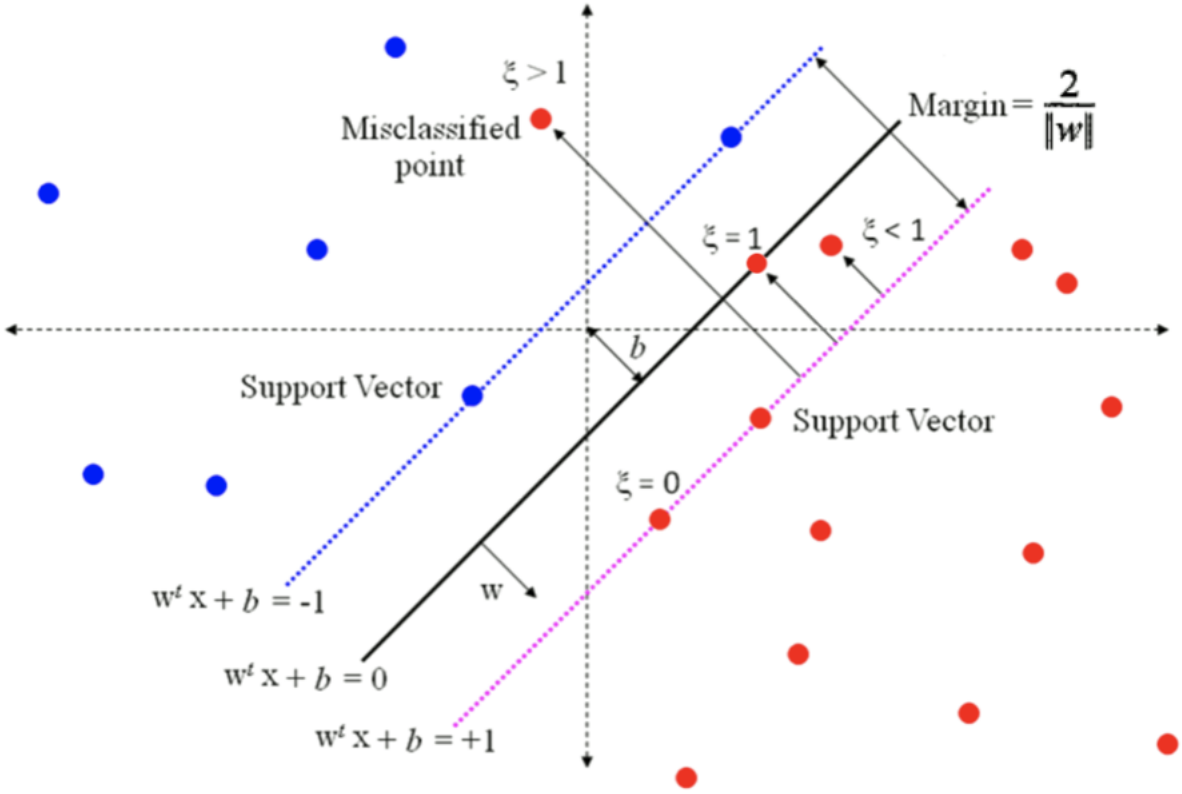


Figure 13: Margin and Support Vector

$$\hat{y} = \begin{cases} 0 & \text{if } \mathbf{w}^T.\mathbf{x} + b < 0, \\ 1 & \text{if } \mathbf{w}^T.\mathbf{x} + b \geq 0 \end{cases}$$

w : weight vector
x : input vector
b : is the deviation

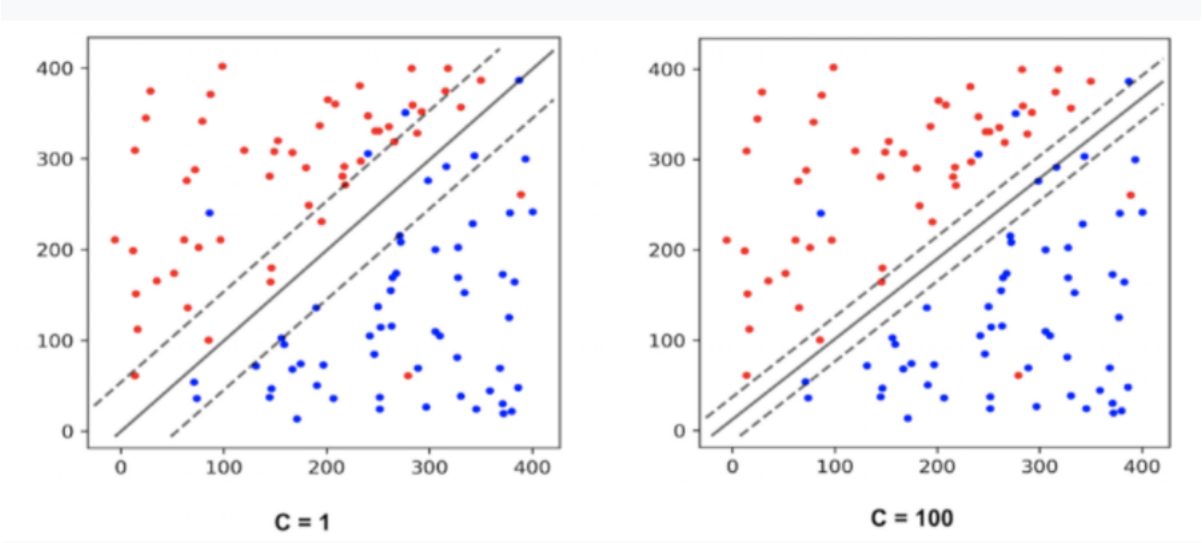


Figure 14: Correlation of C and Margin

Also, if the model is overfit, C should be reduced. if it is underfit, you need to increase its level.

In some cases, it may be insufficient to represent only our xy axis or 2 dimensional space data set, and it cannot classify our data. In such cases, we can easily separate our data from each other by using the size increase method. If we look at the example below, it is impossible to separate these two data using only the x-y axis, while the z axis is impossible with this method. When the system reaches the 3rd dimension, our data is separated from each other very easily and clearly and thus classified.

We mentioned that we can separate the data that we cannot distinguish in the x-y plane by increasing the size. This size increase trick is called Kernel Trick. The Kernel Trick method is called Polynomial Kernel to explain our data set, which we cannot explain in the x-y plane, using more dimensions.

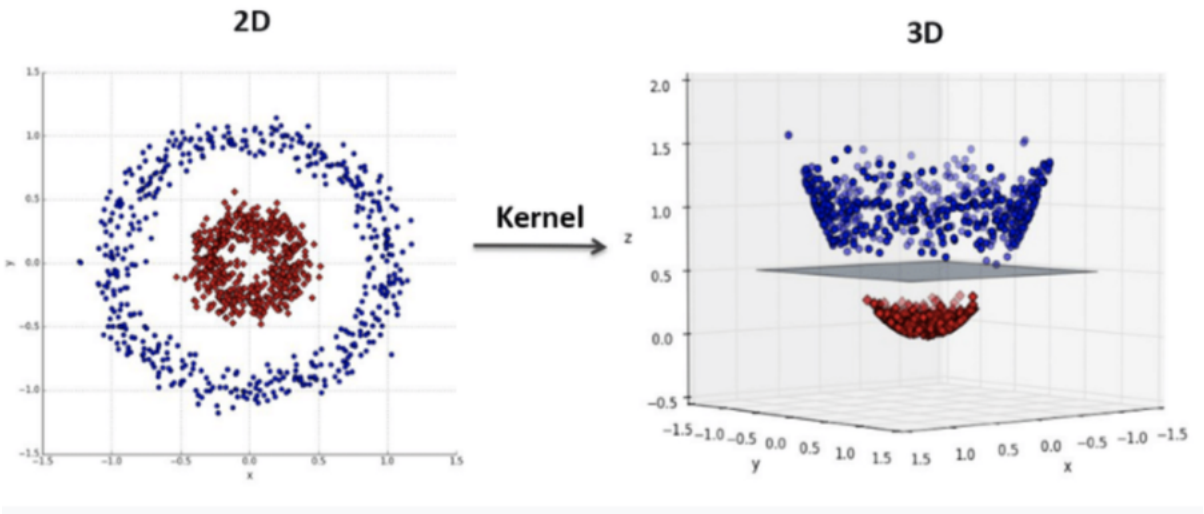


Figure 15: Kernel Trick /increas in dimension

2.5.2.5 K Nearest Neighbours

KNN is a method based on estimating the class of the new data to be added based on the information in which class the nearest neighbors of the collection formed by the independent variables in the KNN data set are dense.The method estimates on two main components.

1)K (number of neighborhoods)

2)Distance

Types plays an important role in machine learning. Distance measurements are selected depending on the type of data. Therefore, knowing which distance measurement is suitable for which data type and understanding the differences between them is extremely important for ease of application.

A properly selected metric improves the performance of the machine model, regardless of both classification and clustering.

The distance of the new data, which will be added to the sample data set, is calculated according to the existing data, and its k close neighbors are checked. Three types of distance functions are generally used for distance calculations.These are:

- Euclidean Distance function
- Manhattan Distance function
- Minkowski Distance function.

1)Euclidean Distance function: Euclidean distance is defined as the shortest distance between two representative points occupying space. Before calculating the Euclidean distance, the data set must be normalized or standardized on the data set. If this is not done, data with large values will take up space. The Euclidean distance is expressed as the sum of the squares of the differences between the components of the two vectors.

$$D_e = \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{1/2}$$

n = number of dimensions

p_i, q_i = data points

2)Manhattan Distance: Manhattan Distance is the sum of the absolute differences between points along all dimensions.

Manhattan Distance is the sum of the absolute differences between points along all dimensions.

$$d = |p_1 - q_q| + |p_2 - q_2|$$

$$D_m = \sum_{i=1}^{n} |p_i - q_i|$$

Minkowski Distance: It is a generalization of Euclidean and Manhattan distance measures and adds a parameter called "row" or "p" that allows calculation of different distance measures.

$$D = \left(\sum_{i=1}^{n} (p_i - q_i)^p \right)^{1/p}$$

KNN is suitable for noisy data where data pre-cleaning is needed a lot. The negative feature of the model, on the other hand, is not positive in terms of taking up space, as the memory to be used will be large when the available data is large, as it hides all the possibilities that may exist while doing it in the distance functions used.

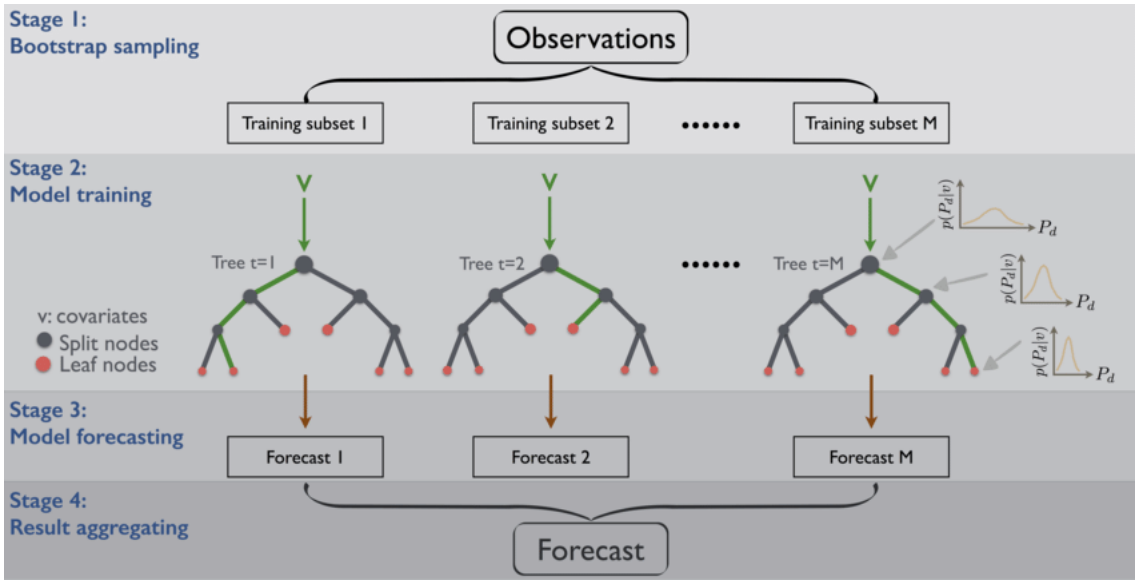
Figure 16: Bagging

2.5.3 ENSEMBLE LEARNING

In our study, single models have been examined so far. The performance of individual models has been evaluated. In the Ensemble Learning model; Instead of working with each individual model, a new, higher-performance model is created by combining or developing individual models by running them in parallel or sequentially in accordance with the technique. In this study, Bagging method and Boosting method from Ensemble Learning models, AdaBoosting and XgBoosting models from Boosting methods will be evaluated.

2.5.3.1 Bagging

In the Bagging method, random samples are taken from the training set section from the sections we have divided into training and test sets in our data set. While taking these samples; The sampled part is sent back to the training section, and the previously selected samples are replaced in the training set and new samples are created by making a selection on the added set. Although the classification method used in the bagging method remains constant, the training examples are changing. Different sample training sets working with the same classification algorithm continue simultaneously and parallel to each other. Studies do not affect each other and work independently of each other. Bagging method can be given as an example from the classification algorithms in our study. In decision trees, there is a risk of over-learning in the models we create using Entropy. In order to prevent this, the Random Forest algorithm was established with random samples taken using the Bagging method. In the bagging method, the aim is to achieve a better classification score, but by reducing the variance, the probability of overfitting is also reduced. In order to obtain the final score, it is decided by looking at the mode value in the classification algorithms.



2.5.3.2 AdaBoosting

AdaBoost(Adaptive Boosting) In the AdaBoost algorithm; Evaluation is made as a result of classification. As a result of the evaluation, the algorithm is developed by focusing on the misclassified examples. In each iteration, the weight coefficients of the correctly classified samples are decreased, while the weights of the incorrectly classified samples are increased. The second training, the wrong data of the first training; The third training classifies by targeting the erroneous data of the second training. Thus, it is aimed to deal with faulty samples and to make a more accurate classification by minimizing the amount of error. Instead of running the algorithms simultaneously as in the Bagging method and making the final decision, recursive studies that minimize the error of the previous algorithm are obtained sequentially.

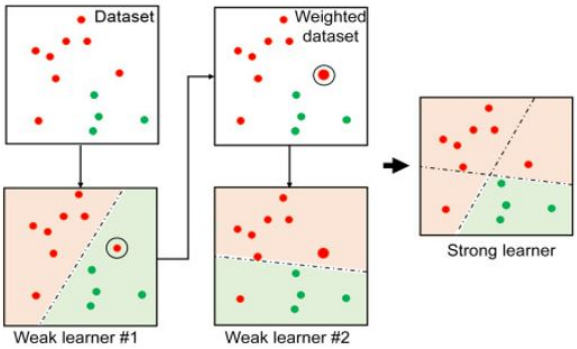


Figure 17: AdaBoosting

2.5.3.3 XgBoosting

XgBoost (Extreme Gradient Boosting)
Xgboost is an advanced model in terms of performance and speed, based on the decision tree, in which the concept of "Gradient Descent" is used to minimize the error in the system where the values found to be faulty are targeted to the system established with weak classification models at the first time."
The XGBoost algorithm is a useful algorithm for large data sets. Although there is no null value in our data set in our study, it performs successfully in the data set with null values.

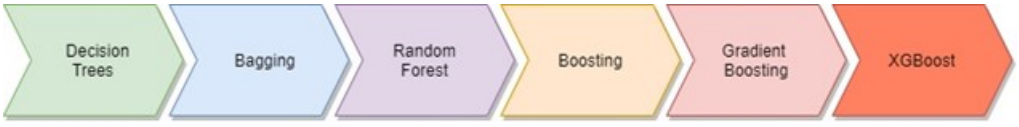


Figure 18: XgBoosting

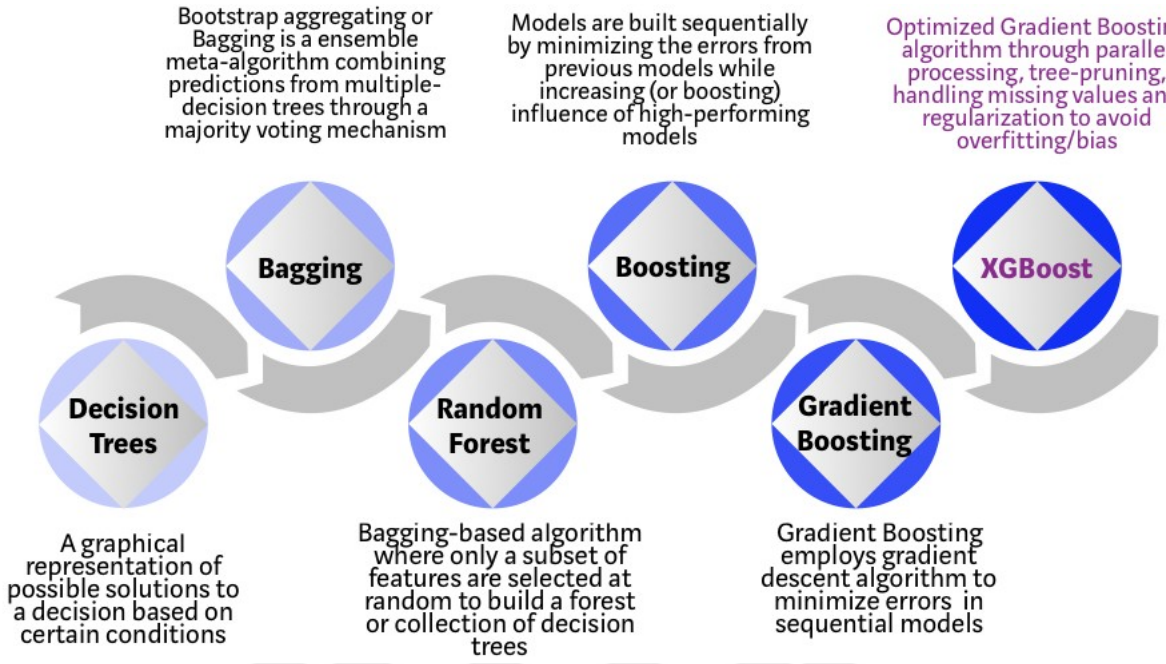


Figure 19: Evolution of Decision Tree

2.5.4 DEEP LEARNING

2.5.4.1 Artificial Neural Networks

It is a learning model that bases its algorithms, which take the formation process of artificial neural networks from the brain neural transmission system, on mathematical modeling. Artificial neural networks are biologically inspired by the human brain. In the nervous system in the brain, artificial neural networks have been created by using the transmission system model of the nerves. A fiction has been made on the axon and dendrite parts of the nerve cells. Dendrites receive messages from neighboring nerve cells. Axons are responsible for transmitting these received messages to the target organ.
The main lines of the Artificial Neural Network are as follows. Networks are formed as a result of starting nerve cells in series or parallel to each other. Generally, it has 3 parts: Input Part, Hidden Layer and Output Part. As the name suggests, the information is taken from the input layer. The layer where the algorithm/mathematical functions work is the middle layer. Purpose of this ; The weight values of the information received from the input layer are calculated and sent to the output layer. There are output values obtained as a result of algorithms applied in the output layer.
Although the emergence of the model is based on biological foundations, the working principle of the algorithm is based on mathematical modeling.
There is an input value from nerve cells adjacent to the dendrites. It corresponds to our input value in our algorithm. There are weight values throughout the networks existing in the dendrite and the weight value of each branch is different. The input value is multiplied

by the weight value in the relevant network and transmitted to the nerve cell. In all neural networks, these values are calculated and summed separately and the bias value is added. The basic operation applied in this model reveals the importance of calculating the weight parameter and bias value. The layer where these processes take place is the intermediate layer. The information processed after this layer can be transferred to the final layer as well as to another neighboring nerve cell. According to the number of layers, they are divided into two as single-layer and multi-layer. In the image below, the templates are shown according to the number of layers.

Single Layer:

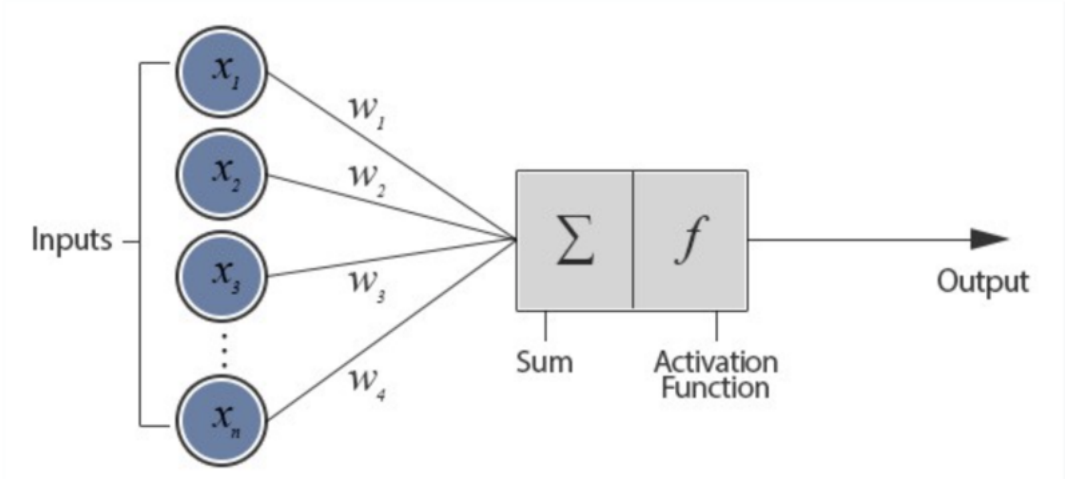


Figure 20: Single Layer

Multi Layer:

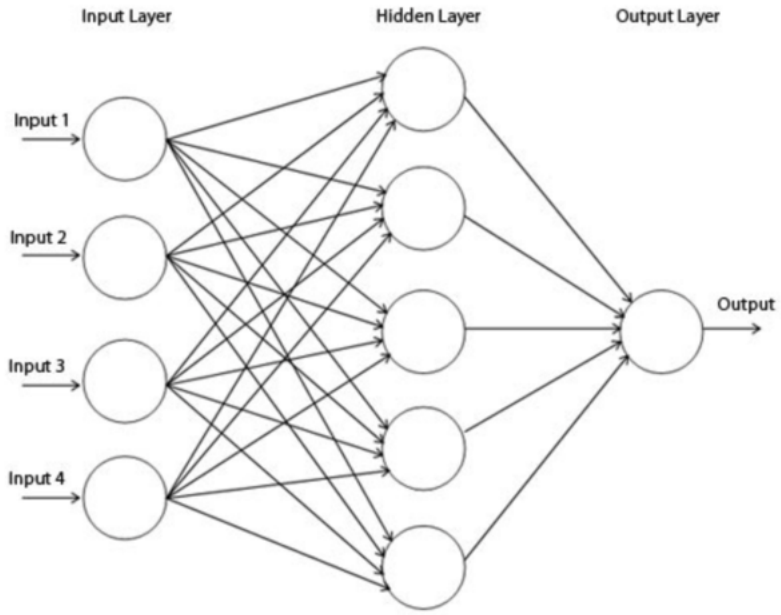


Figure 21: Multi Layer [16]

Artificial neural devices can work with high efficiency even with incomplete information. It is a model with advanced data adaptation capabilities.

3 LITERATURE REVIEW

Although studies to increase academic success are increasing day by day, computer science makes great contributions to the evaluation of the factors affecting academic success with the development of technology. Classification, estimation, and categorization studies of data are carried out by modeling training data using machine learning algorithms and a combination of data mining and artificial intelligence.

The continuous development and increase of studies at the national and international levels have also contributed to easy access to information by improving the quality and quantity of academic research on this subject. While doing this study, the national thesis center, international level theses, and articles in various branches (especially the articles in social sciences were scanned a lot). The articles published in the journal and the journal columns were examined. While direct access to resources is provided by making library visits, online publications, online articles, and offline data are accessed within the scope of distance education. The mentioned words related to the subject have been examined in detail. Among the theses examined below, 3 studies similar to this study were mentioned in terms of subject scope, content, and algorithms used.

As a result of the questionnaire applied in 3 different secondary schools in Yalova in Turkey, questions were asked about demographic, socioeconomic, health, sports, social, activity, and grade achievement status. Significant results were obtained with the application of predictive power feature selection as a result of estimation algorithms using classification and regression by taking Turkish, Mathematics, and end-of-term grade averages as target variables. [17]

In the research conducted in Portugal, between 2005 and 2006, the student distribution is the group after 9 years of primary education. The education systems in the country were evaluated in 3ages and named as G1, G2, and G3. G3 is the final grade. Although these variables are the target variables Different classification algorithms, such as decision trees, Random Forest, artificial neural networks and Support Factor machines, were used and predictions were made. While meaningful estimations can be made with the algorithms used among the features, it has been observed that there are variables that affect less. In addition, ANN and SVM methods are against noisy inputs and output variables. It has been observed that they are more sensitive methods.[18]

The third study used ready data from the Kaggle platform and made predictions using only Logistic regreson a with a Decision tree, and Random forest classification algorithms. In this study, two different data, 395 and 245 student numbers, were used. All properties are the same for these data. The best accuracy rate belongs to the Decision Tree algorithm. Although the data sets were evaluated separately, a total of 649 students were also evaluated. When 3 different data sets are used, the highest number of students and the highest accuracy value belong to the Decision tree. [19]

4 PREDICTION PROCESS

4.1 Data Set

Attribute Name	Attribute Description
Sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4a)
Mjob	mother's job (nominalb)
Fedu	father's education (numeric: from 0 to 4a)
Fjob	father's job (nominalb)
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: 3 or 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if 1 n 3, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)

Attribute Name	Attribute Description
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Table 3: Features(Data Columns)

4.2 Data Set Information

The dataset is the ready data taken from the Kaggle platform .
There are 33 columns in our data and each represents a separate feature. Features include students' demographic structure, socioeconomic status, health/sports/activity characteristics, attitude questions, and lecture notes. Of the features in our data, 16 are numerical and 17 are categorical.395 students participated in the survey study.
In cases where supervised machine learning algorithms will be used in our data, G1, G2, G3, and combined averages will be taken as target variables. In addition, independent variables will be predicted by algorithms that can make predictions among themselves.
There is no missing data in our data. The features in the data set can be extracted, compressed, or categorized by applying Scale, Feature selection, LDA, PCA methods, provided that they change with the algorithms used.

4.3 Heatmap

When the data in the data set is examined, whether there is a relationship between the data or not, the degree of the relationship is very important for interpreting the data. The heatmap used to observe the relationship between the data in the data set provides convenience to the user at this point.
Heatmap: It is the table showing the correlation of each data with all other data. The color of the table is used as it is. The red color of the table, although the main color changes, means that the correlation between the relevant data in that region is high.

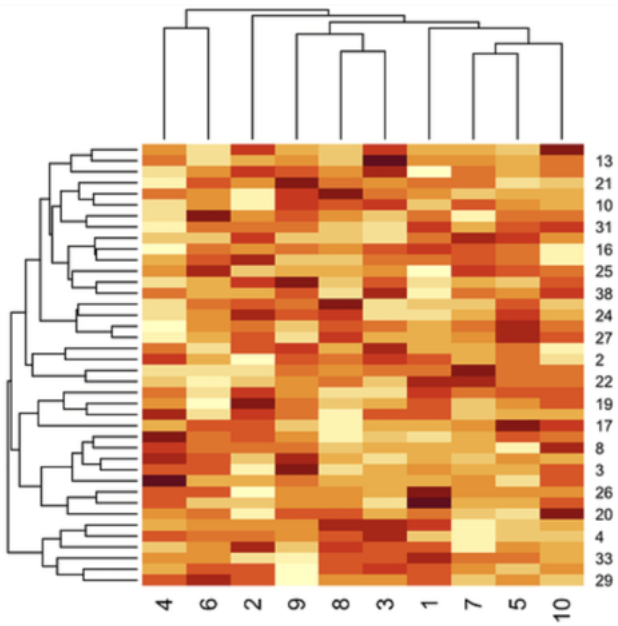


Figure 22: Heatmap

In the classification method, which is one of the supervised learning models, our target variables are categorical data. Therefore, the evaluation of the model is not numerical, like the data whose target variable is continuous, which is also one of the supervised learning models. This situation does not provide clarity in terms of the numerical output of the values and the analysis and evaluation of the predicted values in the model. In order to provide a solution to this problem, techniques have been found for the evaluation of classification models. Below, the solution options used in this study will be explained. After the classification algorithms are applied, when the data we train and the data we test are compared, probabilities emerge, depending on how many categories the data consists of. For example, there are two options for the student home address type in our data set, urban and rural. When the outputs of our model are evaluated, our model will have made right and wrong evaluations. If the right and wrong are examined, there are four possibilities. In this study, the urban life style should be considered as positive and the rural life style as negative. Our model correctly guessed the number of students living in the urban area. The name of this situation in the confusion matrix is True positive. Our model correctly guessed the number of students living in rural areas. The name for this situation in the confusion matrix is True negative. Our model misunderstood the number of students living in the urban area. The name for this situation in the confusion matrix is False positive . Our model misunderstood the number of students living in the urban area. The name for this situation in the confusion matrix is False negative. The confusion matrix for categorical data with binary result is as follows.

4.4 Confusion Matrix

Figure 23: Confusıon Matrix

TP:True Positive
TN:True Negative
FP:False Positive
FN:False Negative

When we want to evaluate the results obtained from the confusion matrix, we encounter terms that quantify the success of the model. If these terms are to be explained;
Accuracy: It is the rate at which the model generally knows the entire data set correctly. Because TP and TN values are the values that the model knows correctly. $TP + FP + TN + FN$ values are the values in the whole data set. Just looking at the accuracy value can be misleading. For example, let's assume that there are 100 students in the data set and 1 person lives in the countryside and 99 people live in the city. As a result of the model's inability to correctly predict 1 student living in the countryside, the accuracy value is 99%, but the model is truly unsuccessful. There are formulas with different metrics to eliminate the problem in this situation. Choosing the formulas given below according to the data in the model by properly interpreting gives accurate information about the performance of the model. It is especially used when the variation of the target variable is high.

Recall = $TP/(TP + FN)$
Precision = $TP/(TP + FP)$
Specificity = $TN/(TN + FP)$
Sensitivy= $TP/(TP + FN)$
Negative predictive value: $TN/(TN + FN)$

In addition to these, there is another value obtained using the harmonic mean value. The name of this value is the F1-score value.

F1-score = 2 * Precision * Recall / (Precision + Recall) [20]

Logarithmic Loss: It is an expression of the represented probability of the predicted value in the model. The smaller the Loss value, the higher the model success.

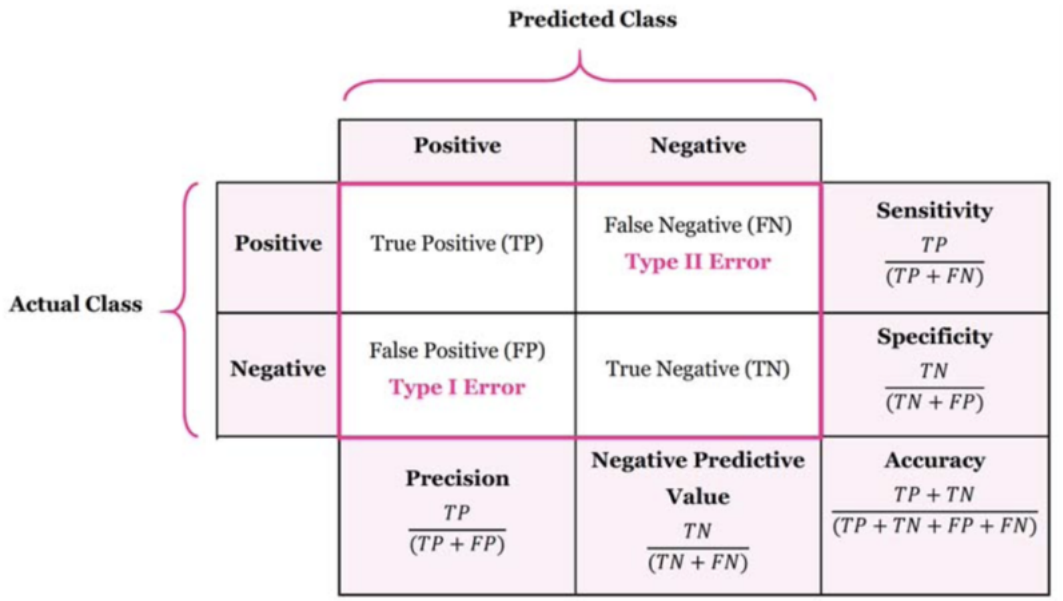


Figure 24: formula on Confusion Matrix

4.5 ROC CURVE

ROC AUC Curve (Receiver Operating Characteristics-Area Under the Receiver Operating Characteristics) is a probabilistic template that indicates how well the data can be separated in classification methods. It is a visual that summarizes all the techniques that are the evaluation criteria of the classification method described so far. AUC is the area under the ROC curve. The purpose of classification is to keep the class boundaries as far from each other as possible, so that the data can be clearly known to which class they belong. Therefore, as the amount of AUC area increases, the separation of data from each other becomes clearer. Becomes specific and the performance success of the model increases. Below is the ROC curve. It is important to interpret the Roc curve correctly and effectively to determine the classification success of the model.

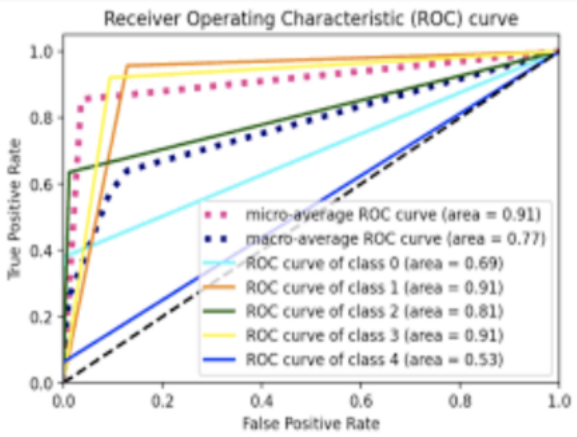


Figure 25: ROC Curve

5 STUDENT ACADEMİC PERFORMANCE PREDICTION-EVALUATE

5.1 Prediction of Decision Tree

Decision tree is used on the end of year final grade prediction dataset which has %70 train and %30 test variables. We have 99 test variable to predict with our machine and compare it with the real values.

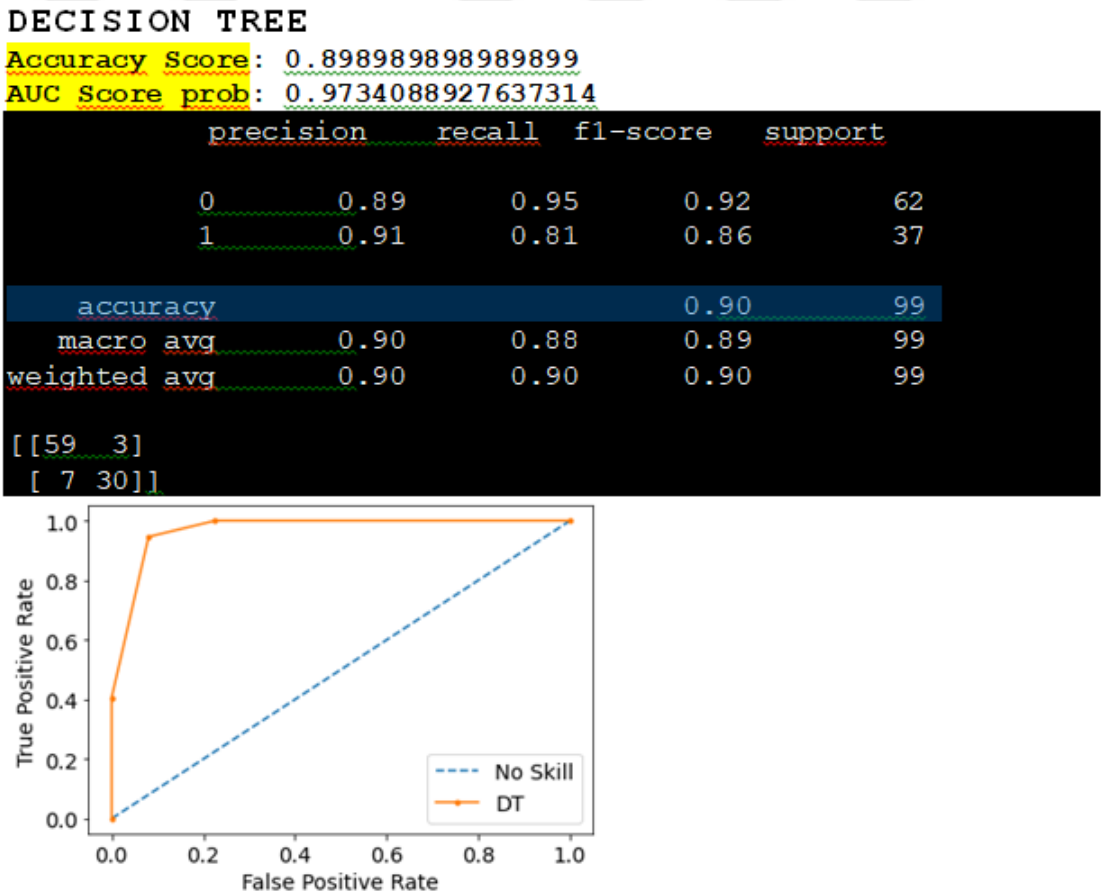


Figure 26: Score of Decision Tree

Test size is 99 so, if we look at our confusion matrix TP = 59 , FP = 3 , FN = 7, TN = 30.
precision = TP/(TP+FP) = 0.89
recall = TP/(TP+FN) = 0.95

f1-score, works with precision and recall f1 score is combined version of them.
(2*(precision*recall)/(precision+recall)) f1-score = 0.92
AUC SCORE PROB = 0.97 score for probability of classes.

5.2 Prediction of Support Vector Machine

Support vector machines is used on the end of year final grade prediction dataset which has %70 train and %30 test variables. We have 99 test variable to predict with our machine and compare it with the real values.

SUPPORT VECTOR MACHINE

Accuracy Score: 0.9292929292929293

AUC Score prob: 0.9734088927637314

	precision	recall	f1-score	support
0	0.97	0.92	0.94	62
1	0.88	0.95	0.91	37
accuracy			0.93	99
macro avg	0.92	0.93	0.93	99
weighted avg	0.93	0.93	0.93	99

```
[[57  5]
 [ 2 35]]
```

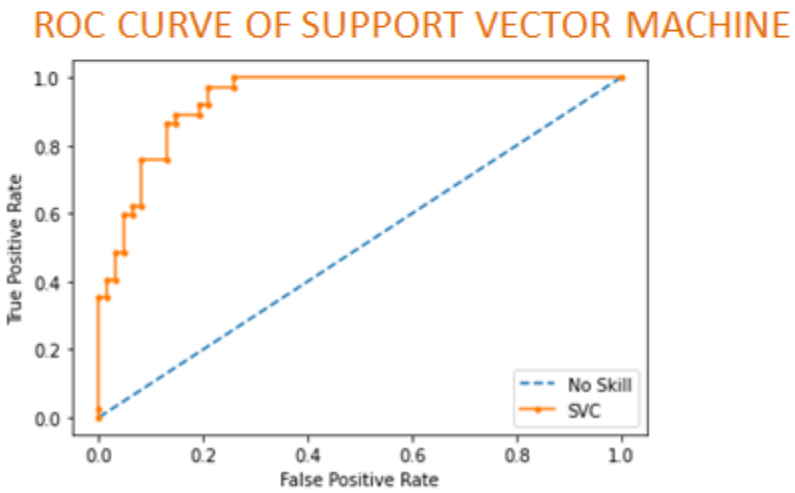


Figure 30 Score of Support Vector Machine]Score of Support Vector Machine

Test size is 99 so, if we look at our confusion matrix TP = 57 , FP =5 , FN = 2, TN = 35.
par 80 of 99 values are predicted right (TP+FN).
We can calculate accuracy score with this information 92/99 = 0.848
precision = TP/(TP+FP) = 0.97
recall = TP/(TP+FN) = 0.92
f1-score, works with precision and recall f1 score is combined version of them.
(2*(precision*recall)/(precision+recall))f1-score=0.94
AUC Score=0.97

5.3 Prediction of Regression Algorithms

5.3.1 Prediction of Multilinear Regression Algorithms

Multilineer Regression is used for predicting the final scores on our dataset.

```
In [16]: # Linear Regression
         from sklearn.linear_model import LinearRegression


         lr = LinearRegression()
         lr.fit(X_train, y_train)
         y_pred = lr.predict(X_test)
         print("Training set score: {:.2f}".format(lr.score(X_train, y_train)))
         print("Test set score: {:.2f}".format(lr.score(X_test, y_test)))
         #print("Actual value:y_test)

         Training set score: 0.87
         Test set score: 0.77

In [20]: from sklearn import metrics
         y_pred = lr.predict(X_test)
         print(metrics.mean_squared_error(y_pred,y_test));

         6.408594911748713
```

Figure 27: Prediction of Multilinear Regression

As we see, train score 0.87 and the test score is 0.77
Calculated MSE for multiple lineer regression is 6.40

5.3.2 Prediction of Ridge Regression

```
In [25]: # Ridge Regression
         from sklearn.linear_model import Ridge
         from sklearn import metrics

         ridge = Ridge().fit(X_train, y_train)
         y_pred = ridge.predict(X_test)
         print("Training set score: {:.2f}".format(ridge.score(X_train, y_train)))
         print("Test set score: {:.2f}".format(ridge.score(X_test, y_test)))
         print(metrics.mean_squared_error(y_pred,y_test))

         Training set score: 0.87
         Test set score: 0.77
         6.417604733107373
```

Figure 28: Prediction of Ridge Regression

Train score 0.87 and the Test score is 0.77
Calculated MSE for multiple lineer regression is 6.40 .

5.3.3 Prediction of Lasso Regression

```
In [29]: # Lasso Regression
         from sklearn.linear_model import Lasso
         lasso = Lasso(alpha=0.001, max_iter=1000000).fit(X_train, y_train)
         print("Training set score: {:.2f}".format(lasso.score(X_train, y_train)))
         print("Test set score: {:.2f}".format(lasso.score(X_test, y_test)))

         Training set score: 0.87
         Test set score: 0.77

In [30]: from sklearn import metrics
         y_pred = lasso.predict(X_test)
         print(metrics.mean_squared_error(y_test,y_pred))

         6.397146101483409
```

Figure 29: Score of Lasso Regression

EVALUATION OF REGRESSIONAL ALGORITHMS	Train Score	Test Score	Mean Absolute Error	Mean Squared Error
Linear Regression	0.87	0.77	1.54	6.40
Ridge Regression	0.87	0.77	1.54	6.41
Lasso Regression	0.87	0.77	1.53	6.39

Figure 30: Compare of Multilinear ,Lasso and Ridge Regression

5.4 Prediction of K Nearest Neighbors

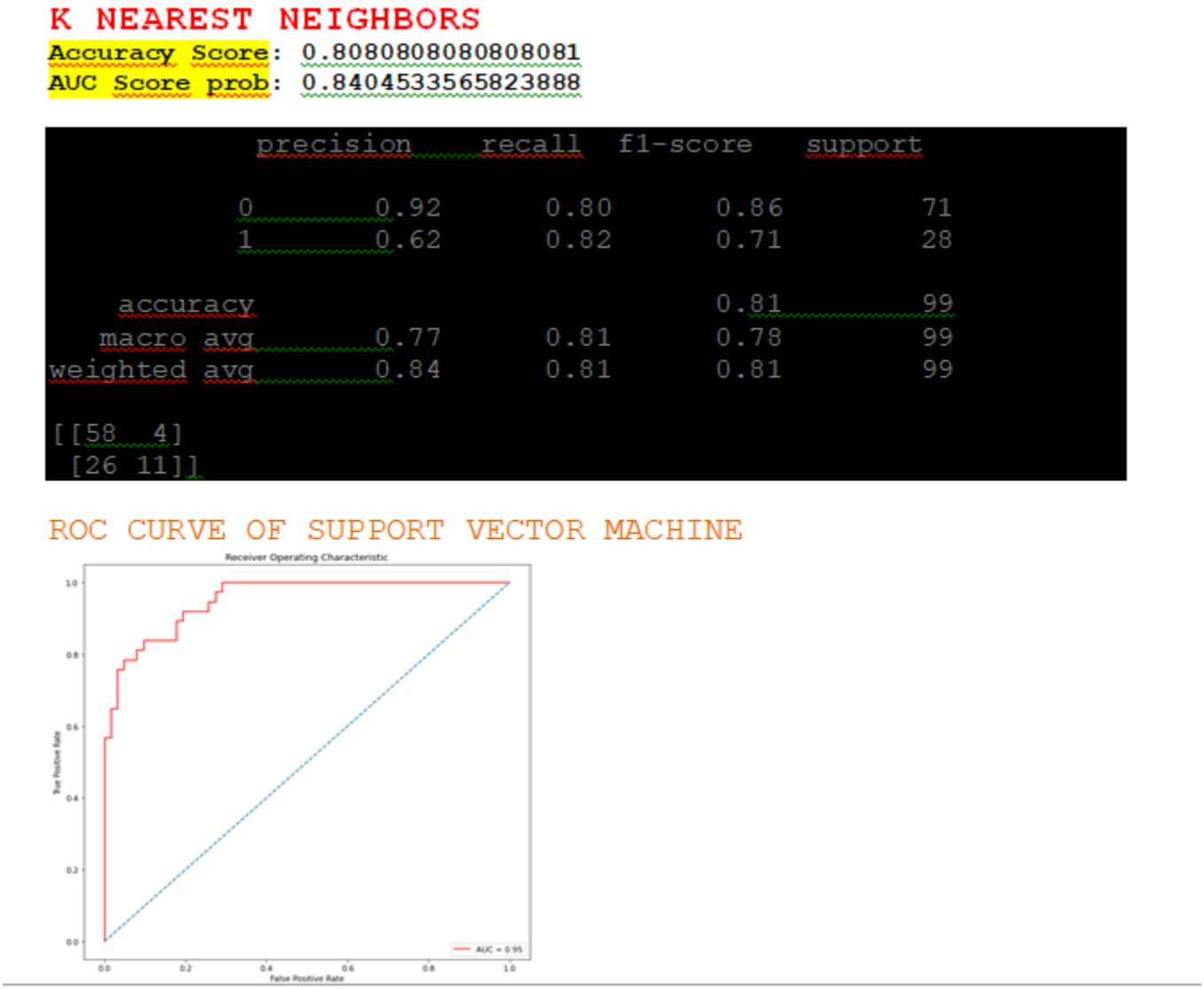


Figure 31: Score of KNN

Test size is 99 so, if we look at our confusion matrix TP = 58 , FP = 4 , FN = 20, TN = 11
Accuracy Score =0.80
precision = TP/(TP+FP) = 0.92
recall = TP/(TP+FN) = 0.80
f1-score, works with precision and recall f1 score is combined version of them.
(2*(precision*recall)/(precision+recall)) f1 score=0.86
AUC SCORE PROB = 0.84 score for probability of classes.

5.5 Prediction of Random Forest

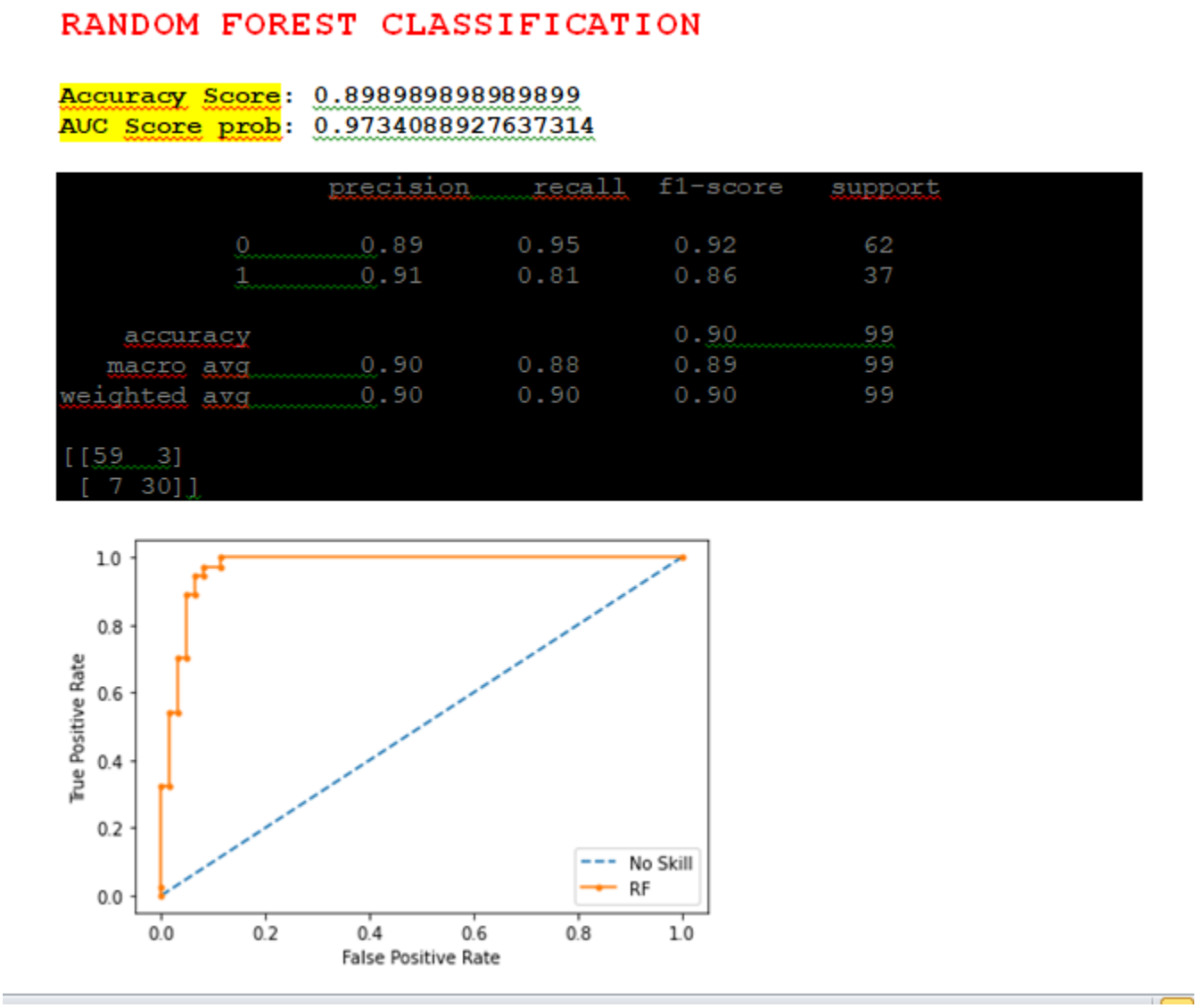


Figure 32: Random Forest Confusion Matrix

Our test size is 99 so, if we look at our confusion matrix TP = 59 , FP = 3 , FN = 7, TN = 30

Means 89 of 99 values are predicted right (TP+FN).

The score of Accuracy with this information 89/99 = 0.8989

precision = TP/(TP+FP) = 0.89

recall = TP/(TP+FN) = 0.95

f1-score, works with precision and recall f1 score is combined version of them.

(2*(precision*recall)/(precision+recall)) f1 score= 0.92

AUC SCORE PROB = 0.97 score for probability of classes.

5.6 Prediction of Logistic Regression

LOGISTIC REGRESSION

```
Accuracy Score: 0.9494949494949495
AUC Score Prob: 0.9734088927637314
              precision    recall  f1-score   support

           0       0.95      0.97      0.96        61
           1       0.95      0.92      0.93        38

    accuracy                           0.95        99
   macro avg       0.95      0.94      0.95        99
weighted avg       0.95      0.95      0.95        99

[[59  3]
 [ 2 35]]
```

ROC CURVE OF LOGISTIC REGRESSION

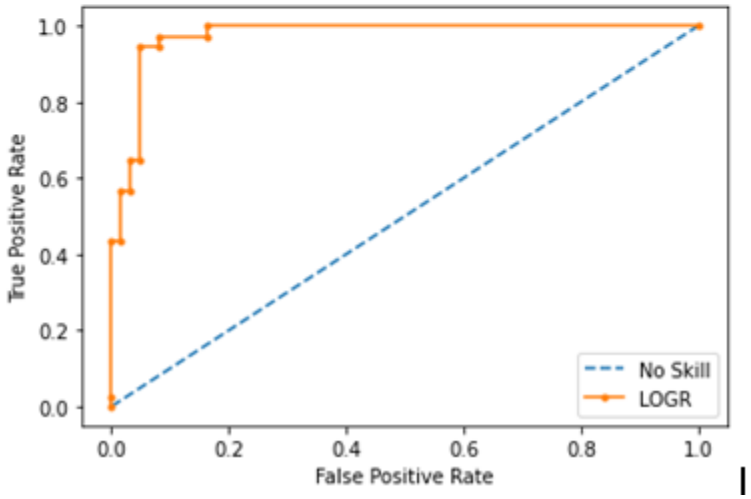


Figure 33: Score of Logistic Regression

Accuracy = 0.94
precision = TP/(TP+FP) = 0.95
recall = TP/(TP+FN) = 0.97
f1-score, works with precision and recall f1 score is combined version of them.
(2*(precision*recall)/(precision+recall)) f1 score= 0.96
AUC SCORE PROB = 0.97 score for probability of classes.
Test size is 99 so, if we look at our confusion matrix TP = 59 , FP = 3 , FN = 4, TN = 33
Means 92 of 99 values are predicted right (TP+FN).
Accuracy = 92/ 99 = 0.929

5.7 Evaluate of Artificial Neural Network

```
ARTIFICIAL NEURAL NETWORK
Accuracy Score: 0.8686868686868687
AUC Score prob: 0.9507410636442895
              precision    recall  f1-score   support

           0       0.87      0.94      0.90        62
           1       0.88      0.76      0.81        37

    accuracy                           0.87        99
   macro avg       0.87      0.85      0.86        99
weighted avg       0.87      0.87      0.87        99

[[58  4]
 [ 9 28]]
```

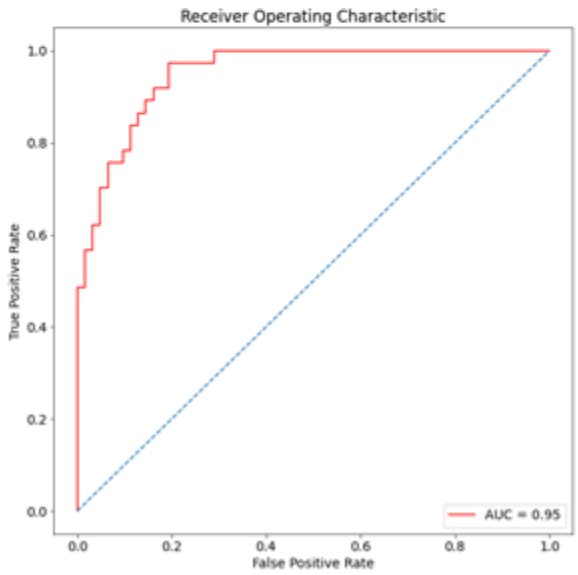


Figure 34: Score of ANN

Accuracy = 0.86
precision = TP/(TP+FP) = 0.87
recall = TP/(TP+FN) = 0.94
f1-score, works with precision and recall f1 score is combined version of them.
(2*(precision*recall)/(precision+recall)) f1 score= 0.90
AUC SCORE PROB = 0.95 score for probability of classes.

5.8 Evaluate of Bagging

BAGGING

Accuracy: 0.9292929292929293
AUC Score prob: 0.9736268526591108

	precision	recall	f1-score	support
0	0.97	0.92	0.94	62
1	0.88	0.95	0.91	37
accuracy			0.93	99
macro avg	0.92	0.93	0.93	99
weighted avg	0.93	0.93	0.93	99

[[57 5]
[2 35]]

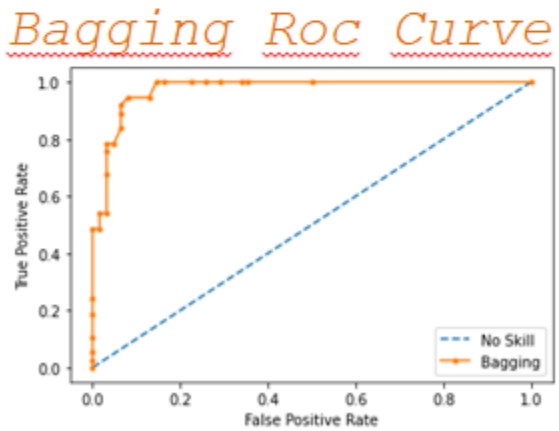


Figure 35: Score of Bagging

Accuracy = 0.92
precision = TP/(TP+FP) = 0.97
recall = TP/(TP+FN) = 0.92
f1-score, works with precision and recall f1 score is combined version of them.
(2*(precision*recall)/(precision+recall)) f1 score= 0.94
AUC SCORE PROB = 0.97 score for probability of classes.

5.9 Evaluate of AdaBoost

ADABOOST

Accuracy Score: 0.8686868686868687
AUC Score prob: 0.9559721011333915

```
AdaBoost Classifier
Accuracy  Score: 0.8686868686868687
AUC Score prob: 0.9559721011333915
              precision    recall  f1-score   support

           0       0.88      0.92      0.90        62
           1       0.85      0.78      0.82        37

    accuracy                           0.87        99
   macro avg       0.86      0.85      0.86        99
weighted avg       0.87      0.87      0.87        99

[[57  5]
 [ 8 29]]
```

ADABOOST ROC CURVE

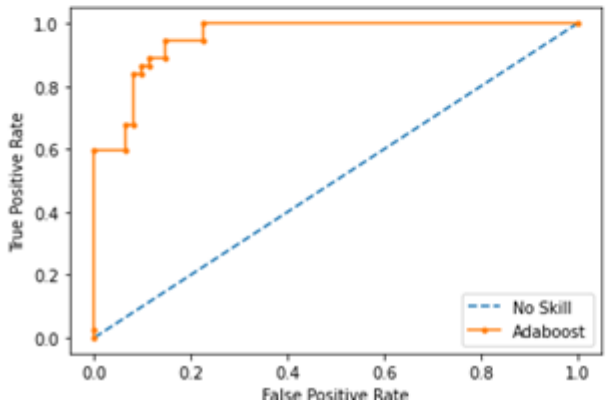


Figure 36: Score of AdaBoost

5.10 Evaluate of XgBoost

XGBOOST

```
Accuracy Score: 0.9090909090909091
AUC Score prob: 0.9760244115082826
              precision    recall  f1-score   support

           0       0.91      0.95      0.93        62
           1       0.91      0.84      0.87        37

    accuracy                           0.91        99
   macro avg       0.91      0.89      0.90        99
weighted avg       0.91      0.91      0.91        99

[[59  3]
 [ 6 31]]
```

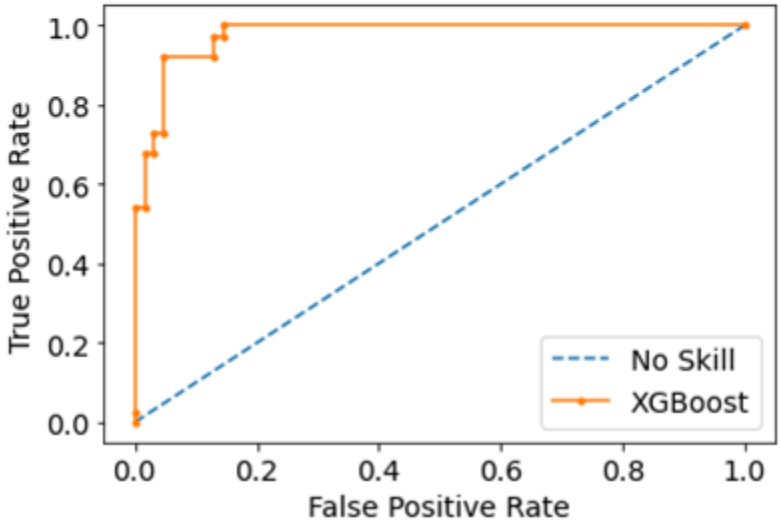


Figure 37: Score of XgBoost

Accuracy = 0.90
precision = TP/(TP+FP) = 0.91
recall = TP/(TP+FN) = 0.95
f1-score, works with precision and recall f1 score is combined version of them.
(2*(precision*recall)/(precision+recall)) f1 score= 0.93
AUC SCORE PROB = 0.97 score for probability of classes.

6 EXPERIMENTS

Experimental Setup:

The dataset is the ready data taken from the Kaggle platform .There are 33 columns in our data and each represents a separate feature. Features includes students' demographic structure, socioeconomic status, health/sports/activity characteristics, attitude questions, and lesson notes. Of the features in our data, 16 are numerical and 17 are categorical.395 students participated in the survey study.There is no missing data in our data. The features in the data set can be extracted, com-pressed, or categorized by applying Scale, Feature selection, LDA, PCA methods, provided that they change with the algorithms used. The data was uploaded to the system in the CSV file and the coding was done by importing it through this file. To take advantage of algorithms. Although IPython makes use of its own functions, open-source libraries are also heavily used. These libraries are Pandas, Numpy, Seaborn, and Matplotlib. Anaconda /Jupyter, Anaconda /Scrapy, and Pycharm were used while writing the codes. The laptop used has an Intel Core i5 processor.

Experimental Results: Both regression and classification models are used in the estimation algorithms. In regression; While multilinear regression, Lidge regression, Lasso regression were successful, Decision tree, Random forest,K Nearest Neighbors and Support Vector Machine are successful in classification. Artificial Neural networks, one of the deep learning methods, have been applied and yielded successful results.Ensemble Learning methods are used succesfully. The evaluation of the algorithms used in our study as classification and regression is as follows.

	ACCURACY SCORE	AUC SCORE	PRECISION	RECALL	F1 SCORE	SUPPORT
XgBOOST	0.90	0.97	0,91	0.95	0.93	62
			0.91	0.84	0.87	37
AdaBOOST	0.86	0.95	0.88	0.92	0.90	62
			0.85	0.78	0.82	37
BAGGING	0.92	0.97	0.97	0.92	0.94	62
			0.88	0.95	0.91	37
RANDOM FOREST	0.89	0.97	0.89	0.95	0.92	62
			0.91	0.81	0.86	37
DECISION TREE	0.89	0.97	0.89	0.95	0.92	62
			0.91	0.81	0.86	37
SUPPORT VECTOR MACHINE	0.92	0.97	0.97	0.92	0.94	62
			0.88	0.95	0.91	37
LOGISTIC REGRESSION	0.94	0.97	0.95	0.97	0.96	59
			0.95	0.92	0.93	40
K NEAREST NEIGHBORS	0.80	0.84	0.80	0.92	0.86	62
			0.82	0.62	0.71	37
ARTIFICIAL NEURAL NETWORK	0.86	0.95	0.87	0.94	0.90	62
			0.88	0.76	0.81	34

Figure 38: Compare of Classification Algorithms

Comparison of Regression

All regressions result in the same scores, but when we look at the margins of error, it is seen that the best performance is the ridge regression.

	Train Score	Test Score	Mean Squared Error
Linear Regression	0.87	0.77	6.40
Ridge Regression	0.87	0.77	6.41
Lasso Regression	0.87	0.77	6.39

Figure 39: Compare of Regression Algorithms

If the regression models are evaluated within themselves and the classification models are evaluated within themselves, the models with the best performance are evaluated. Among the regression models, the training and test results of Multilinear Regression, Lasso Regression, and Ridge Regression models (all three are the same) are 0.87 and 0.77, respectively. When the mean squares error coefficient values are examined, Multilinear Regression 6.40, Ridge Regression 6.41, Lasso Regression 6.39 mean squares error coefficient. When evaluated in regression models, Lasso Regression was the classification model that worked with the best performance by looking at the scores values, not much different from the others. classification models are evaluated within themselves;
When the Decision Tree algorithm is evaluated, Accuracy value:0.89 Value of the area under the Roc curve: 0.97
When the Random Forest algorithm is evaluated, Accuracy value: 0.89 Value of the area under the Roc curve: 0.97
When the Support Vector Machine algorithm is evaluated Accuracy value: 0.92 Value of the area under the Roc curve: 0.97
When the AdaBoost algorithm is evaluated Accuracy value:0.86 Value of the area under the Roc curve:0,95
When the logistic regression algorithm is evaluated Accuracy value: 0.94 Area value under the Roc curve: 0.97
When the XgBoost algorithm is evaluated Accuracy value: 0.90 Value of the area under the Roc curve: 0.97
When the Bagging algorithm is evaluated Accuracy value: 0.92 Value of the area under the Roc curve: 0.97
When the K-nearest neighbors algorithm is evaluated, Accuracy value: 0.80 The area value under the Roc curve: 0.84 results have been reached.
An accuracy value of 0.86 and under the Roc curve is 0.95 is reached in the Artificial Neural Networks model, which is a deep learning algorithm.

7 CONCLUSION AND FUTURE

In this study, by using the factors affecting academic performance, classification, regression, and clustering methods were used, and the degree of influence of the features on each other and the target variable by using estimation algorithms according to the method was predicted, and when the numerical values, which are the evaluation criteria of the models, were examined, significant results were obtained. When "Chapter5" in this study was examined, these numerical values were included in the relevant estimation algorithms.Survey questions including Demographic, Social-economic, Attitude, Health and Sports, Learning Types and Motivation, Social Support / Social Activity, Emotional Intelligence, life Satisfaction, and Academic were used as factors affecting academic performance, and the features were evaluated.

Contributing to education in our country by applying the models and algorithms that give meaningful results to the data set collected from schools in Turkey are the close targets of the study. Factors affecting learning and their positive reflection on students will be studied at different grade levels, in different school types, in different regions, in short, by providing all the support that will enlarge the sample. At the end of this study, it is aimed to reach the data set with a large sample of different school types in Turkey, in different regions, by expanding the features together with the successful algorithms and models.

8 PYTHON CODES

To overview the output of the codes as a pdf, click here. If you want to run any of the codes, you should download the regarding codes in consideration with that those are Jupiter python here .

9 BIBLIOGRAPY

References

[1] D. Boyd and K. Crawford, Information, Commun. Soc. 15, 662 (2012).

[2] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, Harv. Bus. Rev. 90, 61 (2012).

[3] Data Mining Concepts and Techniques Third Edition Jiawei Han University of Illinois at Urbana–Champaign Micheline Kamber Jian Pei Simon Fraser University

[4] Data Mining Concepts and Techniques Third Edition Jiawei Han University of Illinois at Urbana–Champaign Micheline Kamber Jian Pei Simon Fraser University

[5] (https://www.ibm.com/cloud/learn/deep-learning)

[6] (https://www.ibm.com/cloud/learn/what-is-artificial-intelligence)

[7] Journal of Environmental Management,2015

[8] https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selectio

[9] https://www.veribilimiokulu.com/ozellik-olusumu-ve-ozellik-secimifeature-selection-1/

[10] www.ankara.edu.tr.https://www.youtube.com/watch?v=zokkncTmkfUt=50s

[11] Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman

[12] Linear Model Selection And Regularization – Nathan Bastian

[13] http//statquest.org/regularization-part-1-ridge-regression/

[14] British statistician Sir Ronald A. Fisher (1890-1962) in the 1920s(Yrd. Doç. Dr. A. Talha YAL)

[15] https://www.saedsayad.com/logistic$_r$*egression.html*

[16] DATA -Veri Madencili gi Veri Analizi (Haldun Akpınar), Papatya Bilim, 2014

[17] Predicting Academic Achievement with Machine Learning Methods Murat GÖK1, * 1Yalova University, Faculty of Engineering, Department of Computer Engineering, 77100, YALOVA

[18] Paulo Cortez and Alice Silva Dep. Information Systems/Algoritmi RD Centre University of Minho 4800-058 Guimar˜aes, PORTUGAL
P. Cortez and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", pp. 5-12, 2008.

[19] Student academic performance prediction via artificial intelligence using machine learning algorithms / HATİCE NAZLI BASTEM, thesis no: 703366
https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp)

[20] https://veribilimcisi.com/2018/11/28/siniflandirma-metrikleri/