DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

EE492 PROJECT REPORT

University Entrance Exam
Data Science Studies

Berke Eren **250206008**

Korkut Emre Arslantürk **250206039**

Assist.Prof.Dr. Mehmet Serkan Apaydın

DATE: 05/06/2022

ABSTRACT

In that project, using data collected from the university admission exam, it was estimated the entry threshold score, threshold order, number of persons admitted, and whether the department's quota was filled or not.

The data were taken from OSYM and Hacettepe University preference robot [1][2]. These data were subjected to certain filtering processes and studied only on computer engineering data. Selected data were given as input to the models, to be used in the prediction.

Many machine learning techniques were used to make these predictions such as Random Forest, Neural Network, Decision Tree, and Linear Regression. Apart from numerical features, categorical features are given to models using machine learning techniques as input like school name and city. The fastai library was used while using these techniques, which is a deep learning library that provides practitioners with high-level components in standard deep learning domains.

Models, which were created with different techniques and features, were compared with each other. As a result, an estimation was realized that could be useful to both students and university administrations.

TABLE OF CONTENTS

ABBREVIATIONS

OSYM: Student Selection and Placement Centre

IZTECH: Izmir Institute of Technology

RMSE: Root Mean Squared Error

T-SNE: t-Distributed Stochastic Neighbour Embedding

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The university exam is an exam held every year and affects the lives of millions of students. Every year, both students and universities plan for the next year according to the placement results announced by OSYM.

In this project, our motivation is to estimate the desired values using extracted features. OSYM's previous university entrance exam data were analyzed and the entry threshold score and threshold ranking for the students, the quota problem and the number of admitted students for the institutions were estimated with different ensemble learning methods.

The fastai library was used while making this estimation. The purpose of using this library is that it contains the functions of the applied machine learning techniques and these functions are easy to use. To get the best prediction performance, 4 different methods have been tried. These are random forest, linear regression, decision tree, and neural network. In addition to these methods, the effect of the city and name information of the universities as input on the models was examined. Thus, the suitability of the models created with the techniques used for the data set was tested. It was observed that the resulting estimates were close to actual results. In this project, the performance of machine learning techniques and how categorical data are added to the models are observed.

2. PROBLEM DEFINITON

It is a vital question for students which university they can admitted in with the exam score and ranking they have. Every student wants to know which university they can admit with the exam score or ranking they have. Also, universities would like to take an action by predicting whether they will be able to fill their quotas or not. The most important reason for this is that the entry threshold score cannot be determined if the quota is not filled.

Many foundation universities, such as Özyeğin University and Koç University, have many data analysts on-site to review the university entrance exam data they have obtained from previous years. We have also done a similar data science project based on open-source data.

To meet this need, we developed a project that addresses 4 different problems, and this project can be generalized into a product. Thus, we have established a prediction mechanism with certain parameters for both students and institutions.

3. PROPOSED SOLUTION

3.1. Data Analysis

First, the university entrance exam placement results announced by OSYM for the past years were researched. Then these data were filtered for Computer Engineering. Then, universities with data between 2016 and 2021 were selected.

	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_taban_puan	onceki_yil_siralama	bu_yil_siralama	senevs	bu_yil_kontenjan	bu_yil_yerlesen
0	BOĞAZİÇİ ÜNİVERSİTESİ	İstanbul	1.00	524.21800	562.57600	80	80	534.66800	1134.0	622.0	17vs18	80	80
1	KOÇ ÜNİVERSİTESİ	İstanbul	1.00	535.24700	550.92000	7	7	543.23400	954.0	207.0	17vs18	8	8
2	İHSAN DOĞRAMACI BİLKENT ÜNİVERSİTESİ	Ankara	1.00	517.67500	550.08500	50	50	530.93600	1795.0	916.0	17vs18	50	50
3	ORTA DOĞU TEKNİK ÜNİVERSİTESİ	Ankara	1.00	503.02800	526.57000	110	110	519.50800	3466.0	2170.0	17vs18	110	110
4	İSTANBUL TEKNİK ÜNİVERSİTESİ	İstanbul	1.00	489.53700	519.96200	115	115	510.43400	4087.0	3670.0	17vs18	115	115
...	...	...	...	...	...	...	...	...	...	...	...	...	...
748	FIRAT ÜNİVERSİTESİ	Elazığ	1.00	296.24316	315.85153	65	65	268.38496	263937.0	234240.0	20vs21	88	85
749	MUNZUR ÜNİVERSİTESİ	Tunceli	1.00	288.31699	358.03443	20	20	248.60646	285313.0	297220.0	20vs21	20	20
750	İSTANBUL ESENYURT ÜNİVERSİTESİ	İstanbul	0.75	284.14438	337.62257	7	24	248.79122	301452.0	296551.0	20vs21	13	7
751	BATMAN ÜNİVERSİTESİ	Batman	1.00	284.20305	303.31025	7	30	248.26176	297074.0	298452.0	20vs21	40	38
752	ÇANKAYA ÜNİVERSİTESİ	Ankara	1.00	474.40200	500.33200	12	12	425.10000	34100.0	25159.0	20vs21	12	12

753 rows × 13 columns

Figure 1: Dataset

Then, the data set seen in Figure 1 was created by comparing each year with the previous year. This data set consists of the name of the school, the city where it is located, the amount of scholarships, entry threshold score and the highest score in the previous year, the number of people who admitted in the previous year, the number of quotas in the previous year, the base score in this year, the previous year's ranking, the ranking in this year, the number of quotas in this year and the number of people who admitted in this year. There are two different types of universities in the database, state and foundation universities. When creating the scholarship feature, data were collected according to the different scholarship rates of the foundation universities, and one hundred percent of the scholarships were registered for each state university. The dataset consists of 753 samples and 13 features.

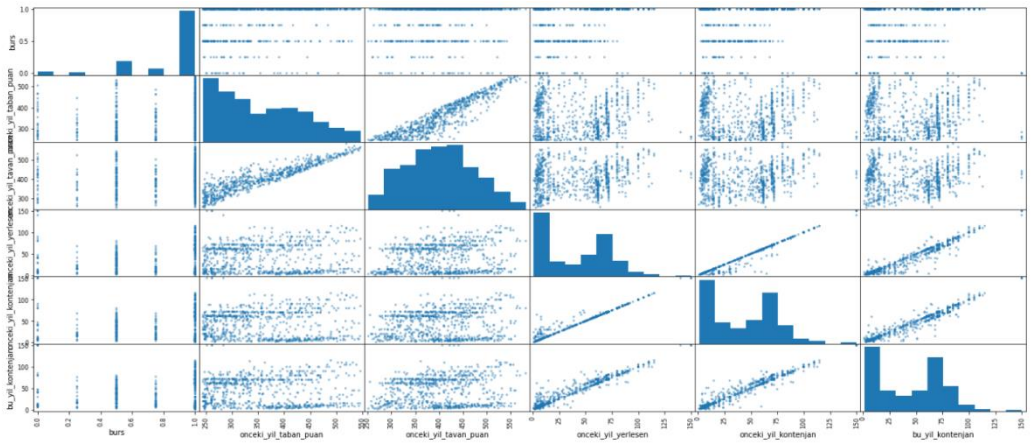

Figure 2: Distribution Matrix of Some Features

A distribution matrix compactly plots the numerical variables we have in a dataset against each other. It allows to simultaneously visualize the relationship between multiple variables in a dataset. The linear graphs of the properties that are related to each other can be seen in Figure 2.

3.2. Taban Puan Prediction

3.2.1. Taban Puan Prediction using Linear Regression

Linear regression is a model in which the many input variables and the single output variable have a linear relationship. When predicting the threshold base score using the linear regression model, the features seen in Figure 3 which were given to the model as input.

	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan
0	1.00	524.21800	562.57600	80	80	80
1	1.00	535.24700	550.92000	7	7	8
2	1.00	517.67500	550.08500	50	50	50
3	1.00	503.02800	526.57000	110	110	110
4	1.00	489.53700	519.96200	115	115	115

Figure 3: Features which are used for Taban Puan Prediction

80 Percent of the entire dataset is allocated in a random way for training and 20 percent for the test set.

As an example, the predicted and actual entry threshold score values of IZTECH and EGE University have been printed. After that, the root mean squared error values obtained for both the train and the test set were calculated in terms of points.

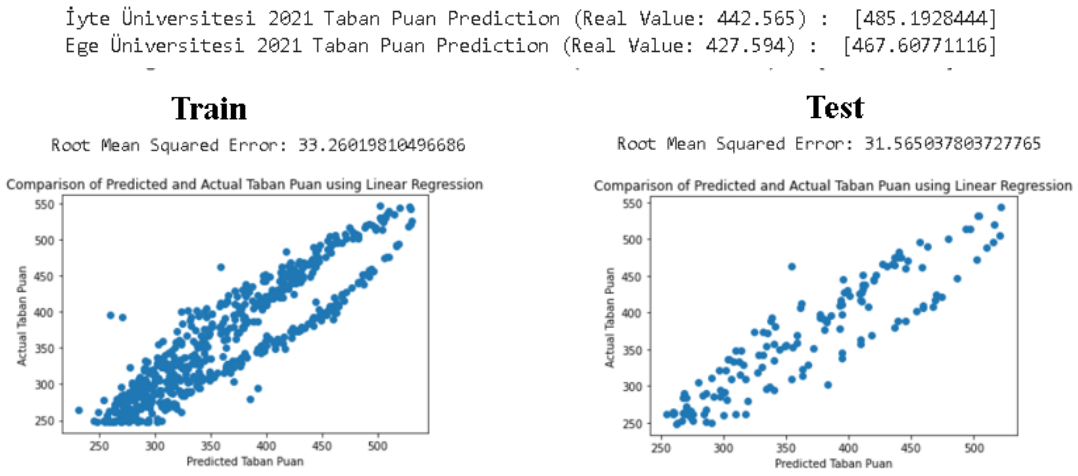

Figure 4: Comparison of Predicted and Actual Taban Puan using Linear Regression for Train and Test Sets.

A scatter plot is a form of plot or mathematical diagram that displays values for typically two variables for a collection of data using cartesian coordinates. In figure 4 comparison of predicted and actual threshold entries score using linear regression can be seen for train and test sets.

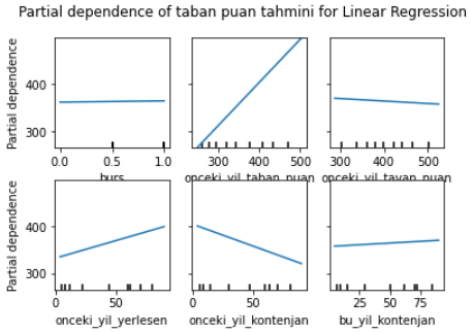

Figure 5: Partial Dependence of Threshold Base Score Estimation

The partial dependency plot shows the functional relationship between input variables and the predicted entry threshold score. It shows how the estimates vary according to each of the relevant input variables. In Figure 5, it can be seen that the entry threshold score of the previous year greatly affects the estimation.

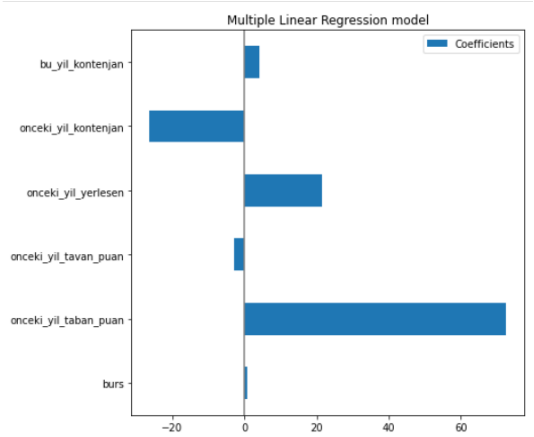

Figure 6: Feature Importance Graphic

Feature importance technique generates a score for all of a model's input attributes. In a summary, scores denote the "importance" of each feature. As can be seen in figure 6, the effect of the "previous year threshold score" is high as positive for the prediction, while the "previous year quota" feature has big negative effect. It means that if the entry threshold score was too high in the previous year, the entry threshold score in this year will also be too high. If the quota increased in the previous year, it causes the entry threshold score to decrease this year.

3.2.2. Taban Puan Prediction using Decision Tree

Decision Trees are a non-parametric supervised learning technique for classification and regression. The goal is to learn basic decision principles and develop a model that predicts the value of a target variable using data properties. If the conditions, which is created randomly, is met, a condition lower than it is checked by continuing from the left side. However, if the first condition is not met, the maximum score is checked at regular intervals up to 360 points. The total number of samples that meet these conditions is seen in the blocks. Thus, data classification is made.

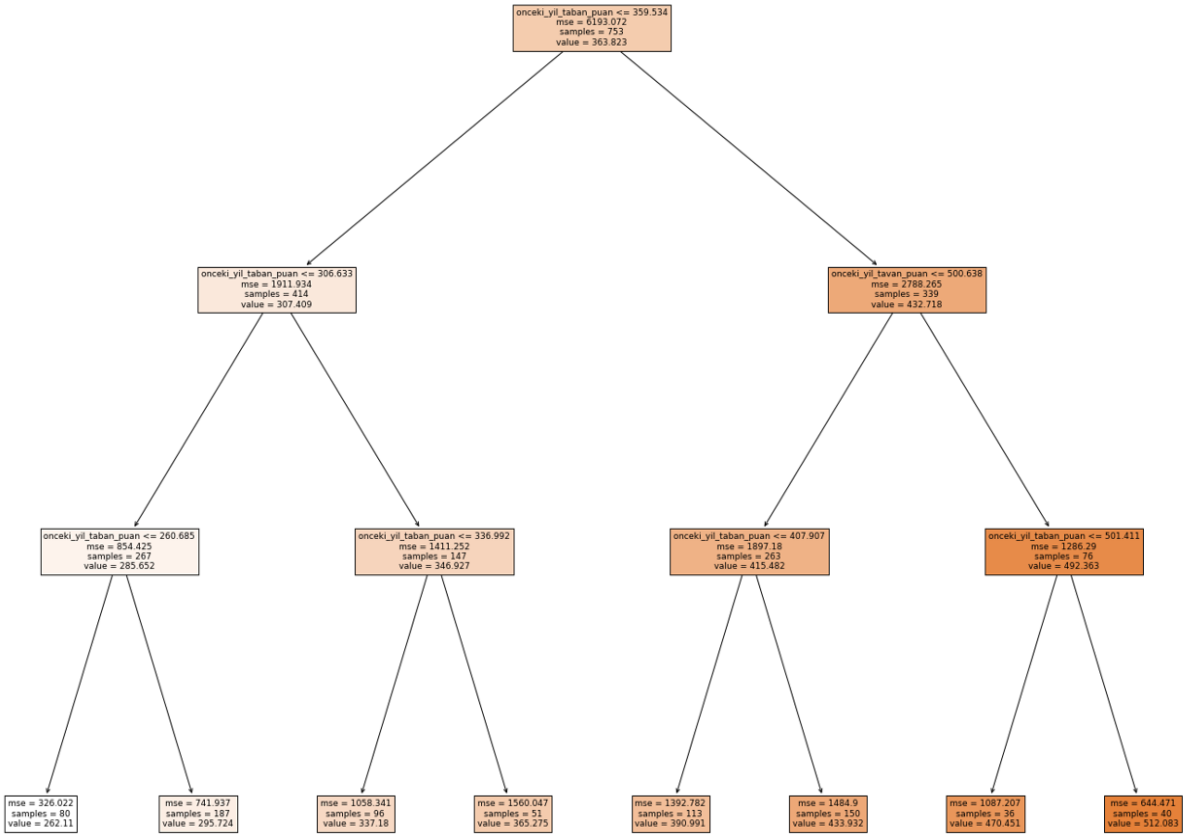

Figure 7: Decision Mechanism of Tree

3.2.3. Taban Puan Prediction using Random Forest

	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan
0	1.00	524.21800	562.57600	80	80	80
1	1.00	535.24700	550.92000	7	7	8
2	1.00	517.67500	550.08500	50	50	50
3	1.00	503.02800	526.57000	110	110	110
4	1.00	489.53700	519.96200	115	115	115

Figure 8: Features which are used in model.

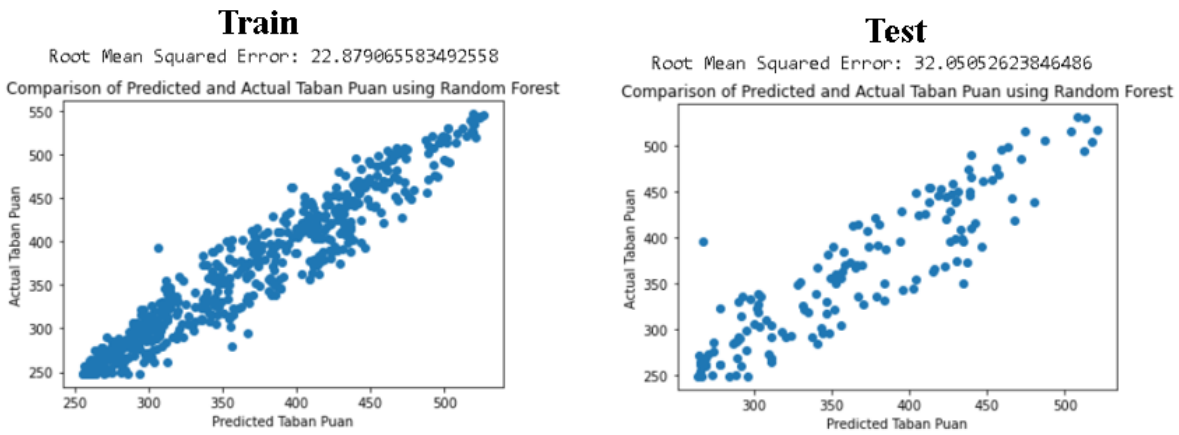

Figure 9: Comparison of Predicted and Actual Taban Puan using Random Forest for Train and Test Sets.

Random forests is a classification, regression, and other problems ensemble learning method that works by training a huge number of decision trees. The features used were kept constant and the training was carried out using the Random Forest model. Similar results were obtained for the test set, while the root mean squared error decreased by 10 points in the Train set if we compare that model with Linear Regression. In addition, the scatter plots of the train and test sets can be seen in Figure 9.

3.2.4. Label Encoding for City of where University is Located

	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan
0	47	1.0	524.218018	562.575989	80	80
1	47	1.0	535.247009	550.919983	7	7
3	2	1.0	503.028015	526.570007	110	110

Figure 10: Features which are used in model.

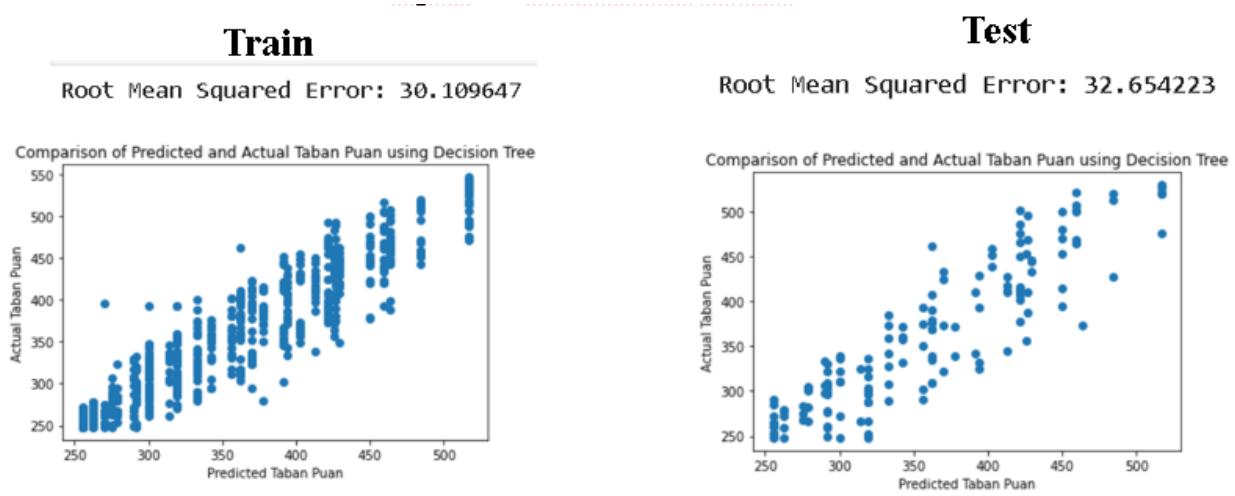

Figure 11: Comparison of Predicted and Actual Taban Puan using Decision Tree for Train and Test Sets.

In tabular data, some columns may contain numeric data, while others may contain string values. String values need to be converted to numbers while numeric data can be fed directly into the model. Label encoding method was used while converting these string values to numerical values. Label Encoding is the process of assigning a number to each category and labelling them into numerical values[3]. First, city information was added to the model as a feature. As a result, the relationship between the RMSE values obtained and the actual score and the predicted score were shown on both test and train sets.

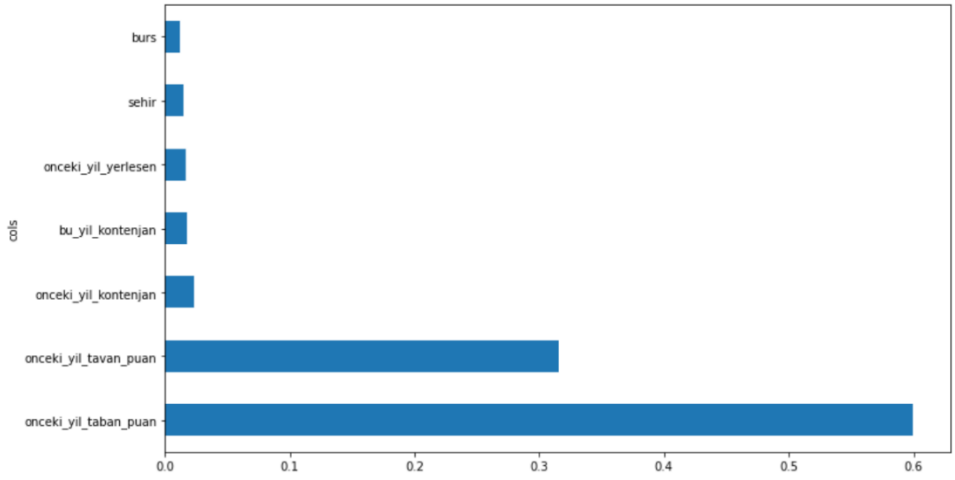

Figure 12: Feature Importance Graph

Removed Feature	RMSE
Original Model	33.09
onceki_yil_tavan_puan	32.66
onceki_yil_taban_puan	40.62
sehir	32.57
bu_yil_kontenjan	32.94
onceki_yil_yerlesen	33.43
onceki_yil_kontenjan	32.93

Table 1: Calculated RMSE Values by Removing Features

After the city information was categorically added to the features, there were changes in the partial dependence of the features. As a result, the effects can be observed in figure 12. As can be seen, the effect of the "previous year threshold score" and “previous year highest score” is high for the prediction. Table 1 shows how the removal of these features changes the RMSE value obtained as a result of the estimation. It is observed that the added city feature affects the prediction worse.

3.2.5. Taban Puan Prediction using a Neural Network

	sehir	onceki_yil_taban_puan	burs	bu_yil_kontenjan	onceki_yil_kontenjan	onceki_yil_yerlesen	bu_yil_taban_puan	bu_yil_taban_puan_pred
0	15.0	-1.076155	0.535466	0.566689	0.600447	0.627793	283.057495	306.000122
1	30.0	-1.438865	0.535466	0.822830	0.957851	0.913300	267.064240	288.492462
2	30.0	-0.908257	0.535466	1.207042	1.282763	1.293977	336.828735	326.631805
3	47.0	-0.940531	-1.389962	-0.297787	-0.601728	-0.545961	359.308167	306.085693
4	47.0	-1.401312	-1.389962	0.246513	0.697921	0.722962	279.049255	296.468018
5	30.0	-1.403912	0.535466	0.886865	0.632939	0.659516	250.303116	295.443237
6	30.0	-1.115434	0.535466	1.207042	0.925360	0.945023	292.673218	312.049988
7	3.0	-1.158758	-1.389962	-0.073664	-0.471763	-0.419069	332.808716	294.527283
8	29.0	-0.886864	0.535466	0.566689	0.632939	0.659516	336.835327	319.232635

Figure 13: Used Features and Comparison of Results.

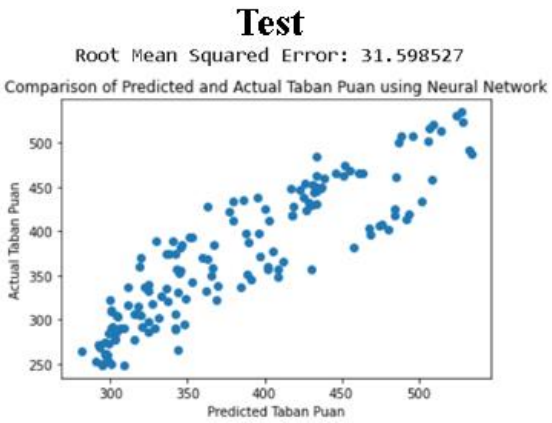

Figure 14: Comparison of Predicted and Actual Taban Puan using Neural Network for Test Set

When we hear the name Neural Network, we feel that it consist of many and many hidden layers but there is a type of neural network with a few numbers of hidden layers. Shallow neural networks consist of only 1 or 2 hidden layers. The neural network we created has 2 hidden layers, one of which is made up of 100 and the other 200 neurons. The features seen in Figure 13 are used for this model. Similar results were obtained for the test se. In addition, the scatter plots of test set can be seen in Figure 14.

3.2.6. Categorical Embedding

Categorical embedding is implemented as lookup tables that map integer values to floating point vectors. To use these embedding layers, we first encoded the categorical variable with integer values. Each of these integers will then correspond to a vector representation of the corresponding category. The size of this vector is automatically suggested by Fast AI. These high-dimensional vectors created for each city have been reduced to 2 dimensions to make them more meaningful, as seen in the figure 15. t-Distributed Stochastic Neighbour Embedding (t-SNE) is a probabilistic technique we use for dimensionality reduction [4].

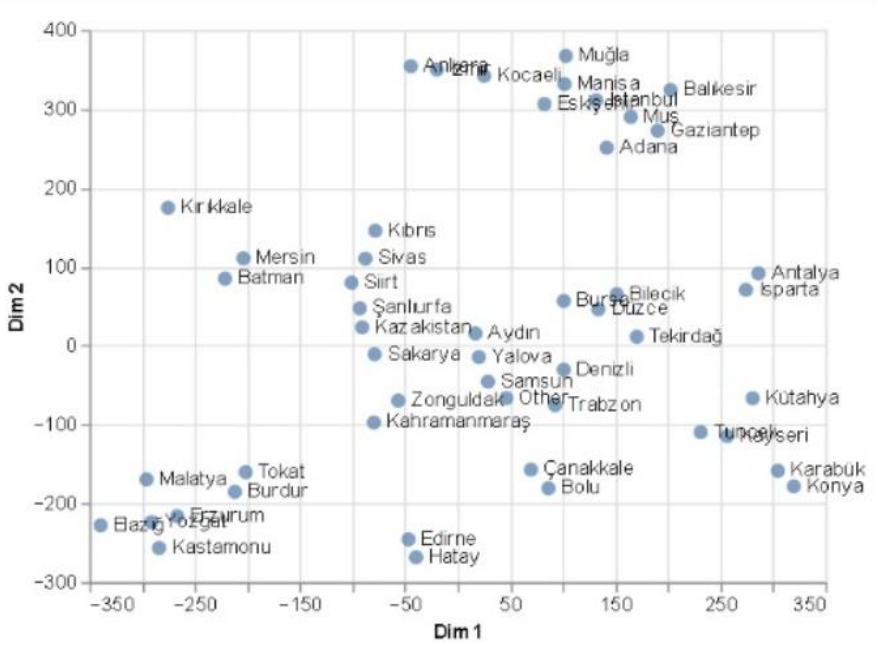

Figure 15: Visualization of City feature to observe how they are close to each other.

As can be seen in Figure 15, most of the cities expected to be close in terms of features could be grouped with each other. For example, cities such as Ankara, İzmir, Istanbul on the upper right and Elazığ, Tokat and Yozgat on the lower left are close to each other in terms of their characteristics. It means that, cities which are close to each other, have similar effect.

3.2.7. Label Encoding for City and University Name

	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan
0	24	47	1.00	524.218018	562.575989	80	80	80
1	63	47	1.00	535.247009	550.919983	7	7	8
2	112	2	1.00	517.674988	550.085022	50	50	50
5	110	47	1.00	484.501007	488.242004	6	6	6
7	112	2	0.50	489.721008	515.463989	20	20	25
...	...	...	...	...	...	...	...	...
745	64	26	0.50	284.761810	388.166687	42	42	38
749	73	42	1.00	288.316986	358.034424	20	20	20
750	122	47	0.75	284.144379	337.622559	7	24	13
751	19	6	1.00	284.203064	303.310242	7	30	40
752	108	2	1.00	474.402008	500.332001	12	12	12

Figure 16: Features which are used in model.

Finally, the categorical feature containing the university names was added to the model using Decision Tree Algorithm while other features keep constant as shown in Figure 16.

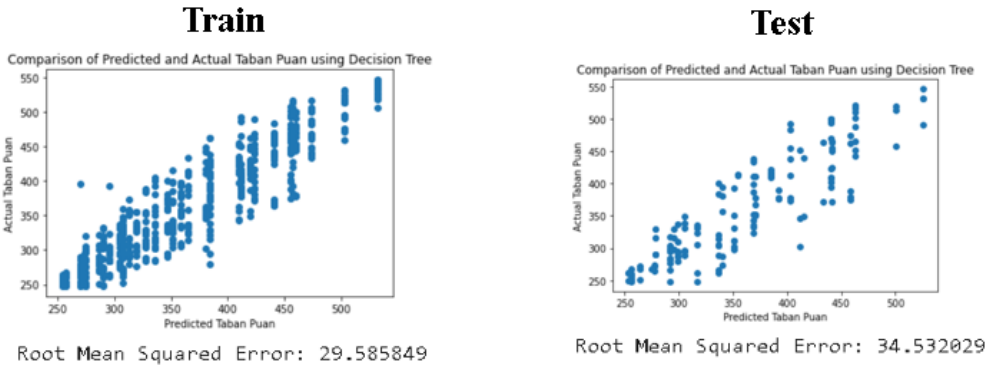

Figure 17: Comparison of Predicted and Actual Taban Puan using Decision Tree for Train and Test Sets.

Then the RMSE value are calculated and the actual and predictive score were compared. As a result, the categorical features we added in general increased the error in the model as can be seen in Figure 17.

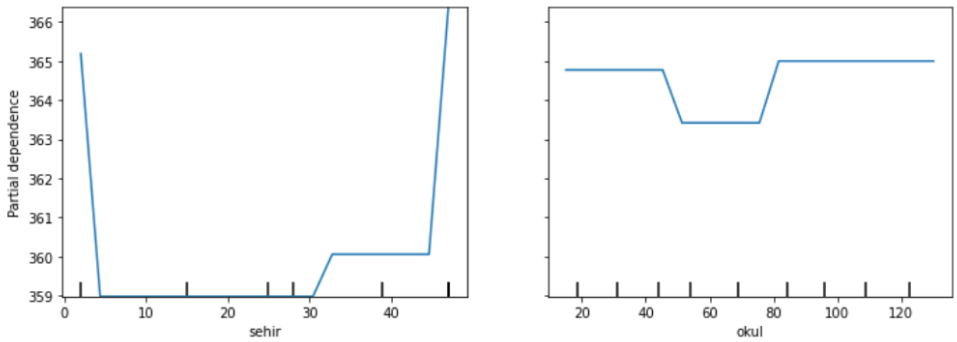

Figure 18: Partial Dependence of Sehir and Okul features.

sehir_isim	sehir_index
İstanbul	47
Ankara	2
İzmir	48
Kocaeli	25
Kıbrıs	28
Sakarya	35
Konya	26
Muğla	33
Trabzon	41
Çanakkale	46
Aydın	4
Samsun	36
Bolu	8

Okul_isim	Okul_index
BOĞAZİÇİ ÜNİVERSİTESİ	24
İHSAN DOĞRAMACI BİLKENT ÜNİVERSİTESİ	112
İSTANBUL TEKNİK ÜNİVERSİTESİ	131
ÖZYEĞİN ÜNİVERSİTESİ	110
BAHÇEŞEHİR ÜNİVERSİTESİ	17
ÇUKUROVA ÜNİVERSİTESİ	109
İSTANBUL RUMELİ ÜNİVERSİTESİ	128
KIRIKKALE ÜNİVERSİTESİ	60
ODTU KUZEY KIBRIS KAMPUSU	81
GAZİ ÜNİVERSİTESİ	44
NECMETTİN ERBAKAN ÜNİVERSİTESİ	78
SELÇUK ÜNİVERSİTESİ	89
KARABÜK ÜNİVERSİTESİ	57

Table 2: Determined Index Values

The partial dependence of the school names and city on the treshold score estimation can be seen in figure 18. It is seen which indexes the names of different cities and schools correspond to. Each feature is kept constant, and the impact of the categorical value is observed part by part. It is observed that cities such as Istanbul and Ankara have a high impact, and cities such as Kocaeli and Konya have a low impact as can be observed in Figure 18 and Table 2.

3.3. Sıralama Prediction

3.3.1. Sıralama Prediction Using Linear Regression

Although score estimation engines are popular among students, the difficulty of the exam cannot be considering while estimating the score. The entry threshold score is a factor that may vary depending on the difficulty of the exam and is a feature that cannot be considering while training the model. However, the ranking is a parameter that varies depending on the trends of students in the selection of departments, regardless of the difficulty of the exam. For all these reasons, the ranking prediction can give more realistic results for both students and institutions.

	burs	onceki_yil_yerlesen	onceki_yil_kontenjan	onceki_yil_siralama	bu_yil_kontenjan
0	1.00	80	80	734	80
1	1.00	7	7	243	8
2	1.00	50	50	1210	50
3	1.00	110	110	2730	110
4	1.00	115	115	4930	115
...	...	...	...	...	...
748	1.00	65	65	263937	88
749	1.00	20	20	285313	20
750	0.75	7	24	301452	13
751	1.00	7	30	297074	40
752	1.00	12	12	34100	12

Figure 19: Features which are used in model.

Features which were given as input to the model are shown in Figure 19. 20 percent of data was reserved for testing randomly, while the rest was used for training. The root mean squared error values were calculated for both the train and the test set in terms of number of people.

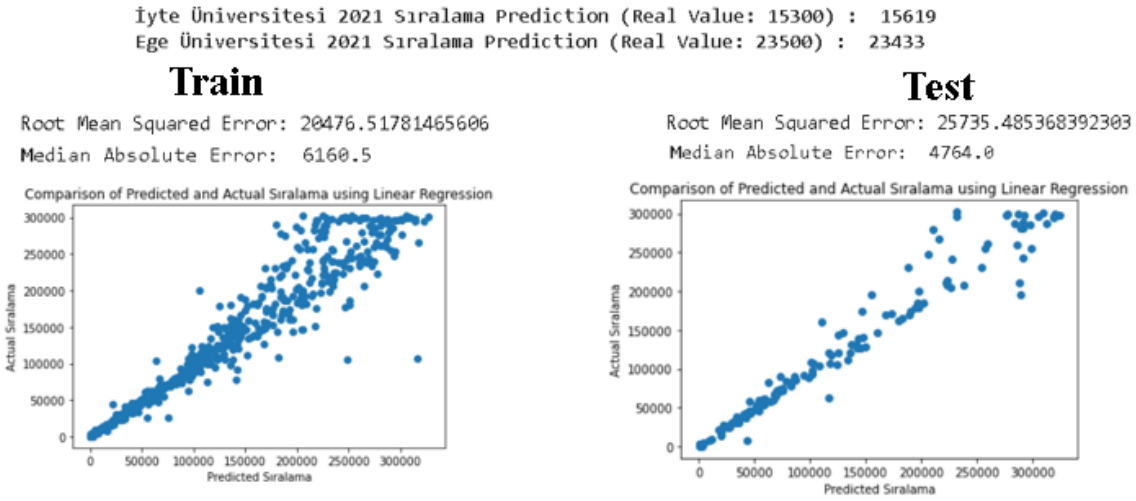

Figure 20: Comparison of Predicted and Actual Sıralama using Linear Regression for Train and Test Sets.

As can be seen in Figure 20, the ranking estimates are correct up to the first 200 thousand, and then deteriorate. This is due to the fact that the student rankings accepted by the schools that usually receive the first 200 thousand do not change much, while other schools change so much in a year. So, the median absolute error is calculated. It has been observed that the median absolute error is only 6500.

3.3.2. Label Encoding for City and University Name

After that, for ranking prediction," city" and" university name" categorical features were added to the model by label encoding as can be seen in Figure 21.

	okul	sehir	burs	onceki_yil_yerlesen	onceki_yil_kontenjan	onceki_yil_siralama	bu_yil_kontenjan
0	24	47	1.0	80	80	734	80
1	63	47	1.0	7	7	243	8
3	85	2	1.0	110	110	2730	110
4	131	47	1.0	115	115	4930	115
5	110	47	1.0	6	6	6000	6

Figure 21: Features which are used in model.

It has been observed that the root mean squared error and median absolute error in the test set has decreased if we compared with the previous model.

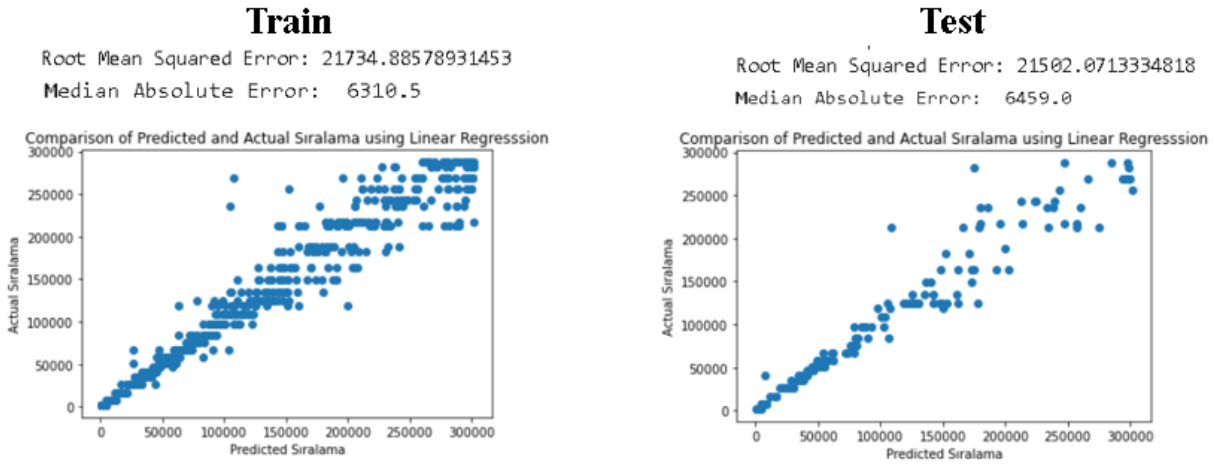

Figure 22: Comparison of Predicted and Actual Sıralama using Linear Regression for Train and Test Sets.

3.4. Prediction of the quota be completed or not

3.4.1. Imbalanced Data

Whether or not to complete the quota is an important problem for the prestige of universities. Therefore, universities want to know in advance whether they can complete their quotas and take action accordingly. It was considered as a class problem whether the university could fill its quota or not. Class 0 indicates that the university cannot fill its quota, while class 1 indicates that it will fill it.

	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	onceki_yil_siralama	bu_yil_kontenjan
0	1.00	524.21800	562.57600	80	80	734	80
1	1.00	535.24700	550.92000	7	7	243	8
2	1.00	517.67500	550.08500	50	50	1210	50
3	1.00	503.02800	526.57000	110	110	2730	110
4	1.00	489.53700	519.96200	115	115	4930	115
...	...	...	...	...	...	...	...
748	1.00	296.24316	315.85153	65	65	263937	88
749	1.00	288.31699	358.03443	20	20	285313	20
750	0.75	284.14438	337.62257	7	24	301452	13
751	1.00	284.20305	303.31025	7	30	297074	40
752	1.00	474.40200	500.33200	12	12	34100	12

Figure 23: Features which are used in model**.**

Features which are given as an input to the model are shown in Figure 23. Of the 753 data in the database, 687 of the quotas were filled, while only 66 of the quotas were not filled. This caused an unbalanced training set to be created because 20 percent of the total data was randomly allocated for testing and the rest was allocated for training.

```
len(data)
```

```
753
```

```
data['kontenort'].value_counts()
```

```
1    687
0     66
```

Figure 24: Amount of Classes.

```
Confusion Matrix :
[[ 12   9]
 [  4 201]]
Accuracy Score : 0.9424778761061947
Report :
              precision    recall  f1-score   support

           0       0.75      0.57      0.65        21
           1       0.96      0.98      0.97       205

    accuracy                           0.94       226
   macro avg       0.85      0.78      0.81       226
weighted avg       0.94      0.94      0.94       226
```

Figure 25: Confusion Matrix of model with RandomForestClassifier

When we examine the Confusion matrix, we can see that it was correctly estimated that 12 universities could not fill their quota. Also, it has been estimated that 9 universities have filled their quota although have not been able to fill it in actual. Furthermore, it has been estimated correctly that 201 universities have filled their quota and 4 universities estimated filled their quota even though they could not fill it actual.

When we review the report, precision gives us the correct estimated number/total estimated number. Recall gives the correct predicted number / actual number. The F-1 Score is the geometric average of these two values. The important parameter is the f-1 score of universities that cannot fill the quota for us.

Since the data set is not balanced and contains quite a few universities that cannot fill their quota, the f-1 score of the universities that can fill the quota is 97 percent, while the f1 score of the universities that cannot fill their quota is only 65 percent as can be seen in Figure 25.

3.4.2. Oversampling

Unbalanced data set has been balanced by applying oversampling to the data set. Oversampling involves the selection of samples from the minority class by displacement and their addition to the training set.

As a result of this, the total data set increased to 1379 data, while 692 universities that could not fill their quota and 687 universities that could fill it were obtained and the data set was balanced.

```
Confusion Matrix :
[[178  11]
 [  9 216]]
Accuracy Score : 0.9516908212560387
Report :
              precision    recall  f1-score   support

           0       0.95      0.94      0.95       189
           1       0.95      0.96      0.96       225

    accuracy                           0.95       414
   macro avg       0.95      0.95      0.95       414
weighted avg       0.95      0.95      0.95       414
```

Figure 26: Confusion Matrix of model with RandomForestClassifier.

As a result, the f1 score increased from 65 percent to 95 percent for class 0 according to the Figure 26. The f1 score just decreased 1 percent for class 1.

3.4.3. Using Categorical Features

After that, it was aimed to increase the accuracy by adding categorical features of city and university names to the model. Other features are kept constant and only city and university features are included in the model with the label encoding. Then the oversampling process was performed, and the model was trained.

```
Confusion Matrix :
[[185   8]
 [ 14 205]]
Accuracy Score : 0.9466019417475728
Report :
              precision    recall  f1-score   support

           0       0.93      0.96      0.94       193
           1       0.96      0.94      0.95       219

    accuracy                           0.95       412
   macro avg       0.95      0.95      0.95       412
weighted avg       0.95      0.95      0.95       412
```

Figure 27: Confusion Matrix of model with RandomForestClassifier.

It was observed that the f-1 scores obtained for both class 0 and class 1 were slightly reduced compared to the previous model as can be seen in Figure 27.

3.4.4. Using Result of Taban Puan Prediction

	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan
0	24	47	1.0	524.218018	562.575989	80	80	80
1	63	47	1.0	535.247009	550.919983	7	7	8
3	85	2	1.0	503.028015	526.570007	110	110	110
4	131	47	1.0	489.536987	519.961975	115	115	115
5	110	47	1.0	484.501007	488.242004	6	6	6
...	...	...	...	...	...	...	...	...
693	12	15	1.0	343.485931	413.120697	71	71	72
64	65	26	1.0	362.524261	392.893768	5	5	10
78	30	28	1.0	337.600067	350.489899	2	2	2
2	112	2	1.0	517.674988	550.085022	50	50	50
508	76	39	1.0	298.739929	317.944092	62	62	62

Figure 28: Features which are used in model**.**

Finally, with the previous model, all the features were kept constant and the results of the score estimation, which is the first subject of the project, were also given to the model as input and its effect on the results was observed as can be seen in Figure 28.

```
Confusion Matrix :
[[184  10]
 [  7 210]]
Accuracy Score : 0.9586374695863747
Report :
              precision    recall  f1-score   support

           0       0.96      0.95      0.96       194
           1       0.95      0.97      0.96       217

    accuracy                           0.96       411
   macro avg       0.96      0.96      0.96       411
weighted avg       0.96      0.96      0.96       411
```

Figure 29: Confusion Matrix of model with RandomForestClassifier.

According to the Figure 29, compared to the previous results, it was observed that the f-1 scores obtained in the university prediction, which can both fill in and not fill out its quota, increased from 95 percent to 96 percent.

3.5. Yerlesen Prediction

After that, a model was developed on calculating whether a university can fill its quota not as a class problem, but by estimating the number of people who have admitted.

In this way, it was possible for institutions to test how many people will be admitted if they change their quotas. Futures which are given to the model as an input are shown in Figure 30. Using the Decision Tree Algorithm, the number of people admitted was estimated.

3.5.1. Yerlesen Prediction using Decision Tree

	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan
0	BOĞAZİÇİ ÜNİVERSİTESİ	İstanbul	1.00	524.21800	562.57600	80	80	80
1	KOÇ ÜNİVERSİTESİ	İstanbul	1.00	535.24700	550.92000	7	7	8
2	İHSAN DOĞRAMACI BİLKENT ÜNİVERSİTESİ	Ankara	1.00	517.67500	550.08500	50	50	50
3	ORTA DOĞU TEKNİK ÜNİVERSİTESİ	Ankara	1.00	503.02800	526.57000	110	110	110
4	İSTANBUL TEKNİK ÜNİVERSİTESİ	İstanbul	1.00	489.53700	519.96200	115	115	115

Figure 30: Features which are used in model**.**

The root mean squared error was calculated less than 6 people for the test set, and less than 5 people for the training set. Also, visualization of the obtained for both the test and the training set was performed as can be seen in Figure 31.

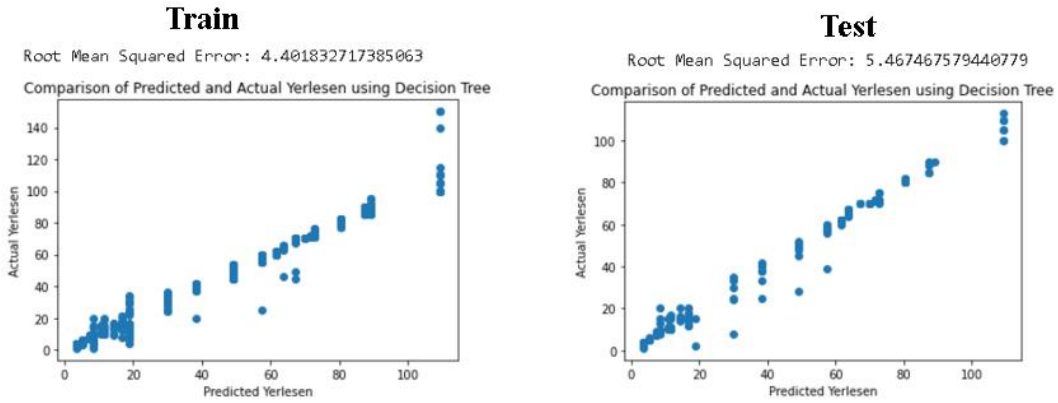

Figure 31: Comparison of Predicted and Actual Yerlesen using Decision Tree for Train and Test Sets.

3.5.2. Using Result of Taban Puan Prediction

Then, with the previous model, all the features were kept constant and the results of the entry threshold score estimation, which is the first subject of the project, were added to model like can be seen in Figure 32.

	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan
0	BOĞAZİÇİ ÜNİVERSİTESİ	İstanbul	1.00	524.21800	562.57600	80	80	80
1	KOÇ ÜNİVERSİTESİ	İstanbul	1.00	535.24700	550.92000	7	7	8
2	İHSAN DOĞRAMACI BİLKENT ÜNİVERSİTESİ	Ankara	1.00	517.67500	550.08500	50	50	50
3	ORTA DOĞU TEKNİK ÜNİVERSİTESİ	Ankara	1.00	503.02800	526.57000	110	110	110
4	İSTANBUL TEKNİK ÜNİVERSİTESİ	İstanbul	1.00	489.53700	519.96200	115	115	115

Figure 32: Features which are used in model**.**

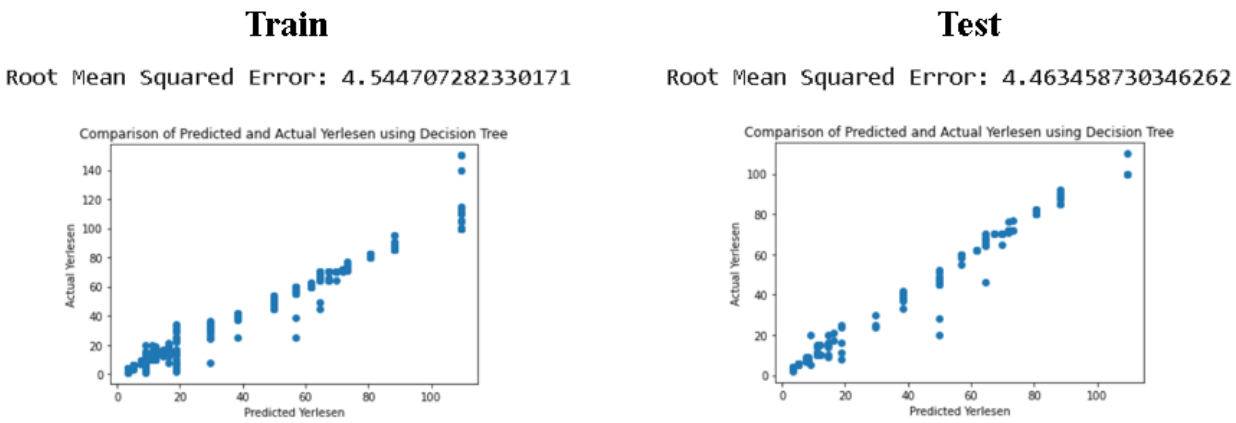

Figure 33: Comparison of Predicted and Actual Yerlesen using Decision Tree for Train and Test Sets.

According to the Figure 33, it can be observed that, although the root mean squared error remained constant in the training set, it was observed that it decreased from 5.46 to 4.46 for the test set.

4. RESULTS AND DISCUSSIONS

Prediction	Categorical Features	Numerical Features	Model	Test RMSE	In terms of
Taban Puan	-	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan	Linear Regression	31.56	Points
Taban Puan	-	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan	Random Forest	32.05	Points
Taban Puan	sehir	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan	Decision Tree	32.65	Points
Taban Puan	sehir, okul	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan	Decision Tree	34.53	Points
Taban Puan	sehir, okul	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan	Neural Network	31.59	Points
Siralama	-	Common Features,onceki_yil_siralama	Linear Regression	25735	People
Siralama	sehir, okul	Common Features,onceki_yil_siralama	Linear Regression	21502	People
Yerlesen	sehir, okul	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan	Decision Tree	5.46	People
Yerlesen	sehir, okul	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan,tahmini_taban_puan	Decision Tree	4.46	People

Prediction	Categorical Features	Numerical Features	Oversampling	F-1 Score(for 0)
Kontenjan	-	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan,onceki_yil_siralama	-	0.65
Kontenjan	-	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan,onceki_yil_siralama	+	0.95
Kontenjan	+	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan,onceki_yil_siralama	+	0.94
Kontenjan	+	Common Features,onceki_yil_taban_puan, onceki_yil_tavan_puan,onceki_yil_siralama,tahmini_taban_puan	+	0.96

Common Features: burs, onceki_yil_yerlesen, onceki_yil_kontenjan, bu_yil_kontenjan

Table 3: Comparison of Observed Results

According to the Table 3, when we compare the predictions, it was observed that the two most successful results in estimating the entry threshold score were given by the Linear Regression Model and the Neural Network. The Neural Network model was given the ”city“ and ”school" features, while that features are not given to the linear regression model.

Categorical “school name” and “city” features for entry threshold ranking prediction were observed to improve the prediction.

In the prediction of the number of people admitted, it was observed that the estimated entry threshold score added to the model improved the result.

It was observed that the best result in the problem of whether the quota should be filled or not was obtained in the model where categorical values were added, oversampling was applied to the dataset and the predicted entry threshold score was added.

5. CONCLUSIONS

In Conclusion, 4 different predictions were made, including the threshold score prediction, the threshold ranking prediction, the prediction of whether the university will be able to fill the quota, and the prediction of the number of people who have been admitted. It was decided that the ranking estimation would give a more reliable result because the threshold score estimation was also a variable that could not be considered the exam difficulty. In addition, the model closest to reality was determined with 4 different predictions. Thus, the performance of the machine learning techniques used and optionally added categorical data were observed.

It is decided that a trending ranking can be added as a feature by researching the search numbers of universities and departments in the Google search engine. It was decided to enlarge the dataset by adding it to the other departments as well as computer engineering. When creating a dataset, instead of making a comparison with just a year ago, it will be observed as a trend.

In this project, machine learning techniques were comprehended. In addition, the results of these techniques were analyzed. The use and effects of data were observed. Finally, the use of categorical data used in machine learning was learned.

REFERENCES

[1] *Yükseköğretim Kurumları Sınavı*. 2021. (n.d.). Retrieved June 5, 2022, from https://www.osym.gov.tr/TR,21232/2021.html

[2] *Osys Sonu Analalizor*. (n.d.). Retrieved June 5, 2022, from http://yks.ee.hacettepe.edu.tr/

[3] Géron Aurélien. (2022). *Hands-on machine learning with scikit-learn, Keras, and tensorflow concepts, tools, and techniques to build Intelligent Systems*. O'Reilly.

[4] Making neural nets uncool again. (n.d.). Retrieved June 5, 2022, from https://www.fast.ai/