University Entrance Exam Data Science Studies

Berke Eren-Korkut Emre Arslantürk
250206008-250206039
Assist.Prof.Dr. Mehmet Serkan Apaydın
Izmir Institute of Technology
Electronics and Communication Engineering

ABSTRACT

- Placement results data were collected from OSYM and Hacettepe University preference robot. [1], [2]
- Data were subjected to certain filtering processes and studied only on computer engineering data.
- There were 4 different categories of prediction:

Entry threshold score, ranking, quota problem, and number of admitted student.

- The results obtained using different parameters and models were compared.

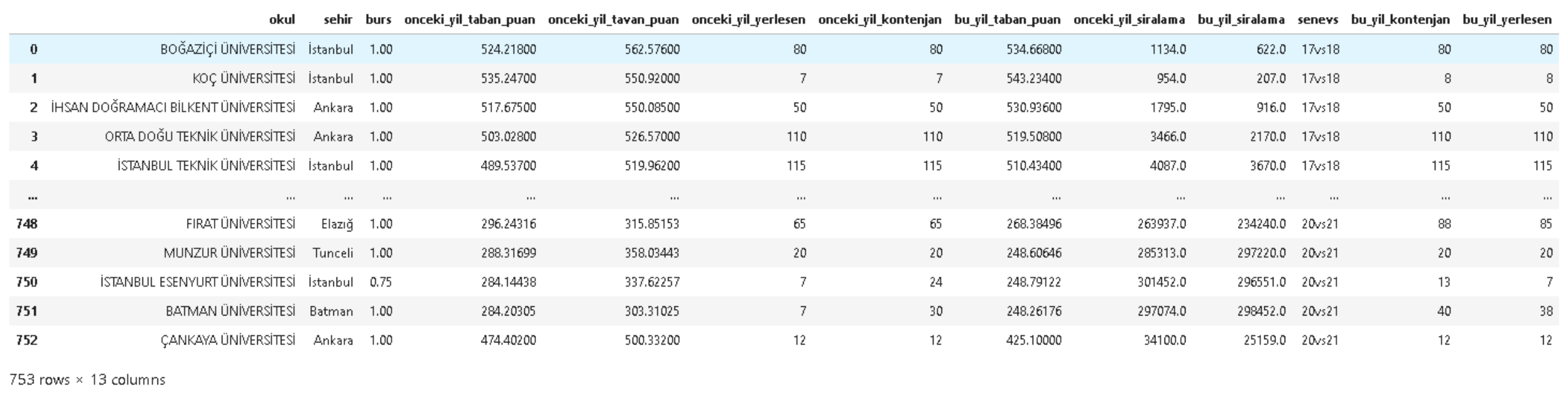
INTRODUCTION

- Every year, both students and universities plan for the next year according to the placement results announced by OSYM.
- Our motivation is to estimate the desired values using extracted features.
- To get the best prediction performance, 4 different methods have been performed.

DATA ANALYSIS

- The university entrance exam placement results were filtered for Computer Engineering.

Dataset

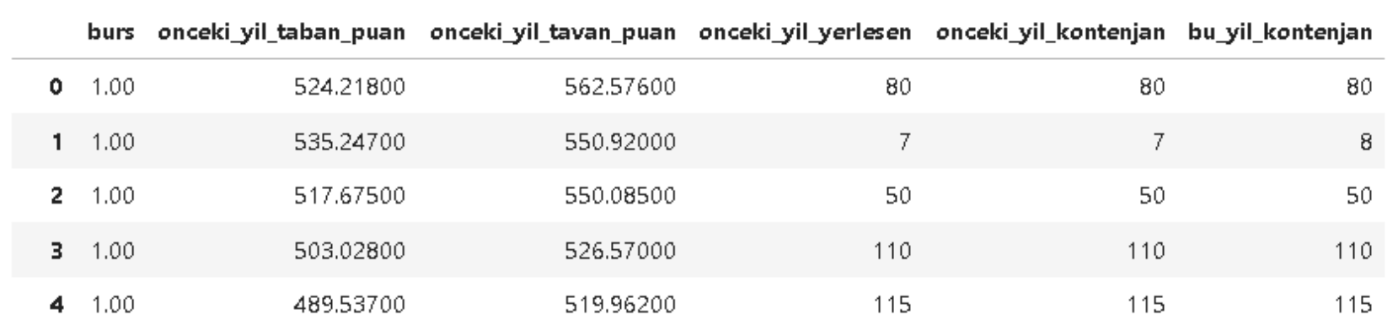


	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_taban_puan	onceki_yil_siralama	bu_yil_siralama	senevs	bu_yil_kontenjan	bu_yil_yerlesen
0	BOĞAZİÇİ ÜNİVERSİTESİ	İstanbul	1.00	524.21800	562.57600	80	80	534.66800	1134.0	622.0	17vs18	80	80
1	KOÇ 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

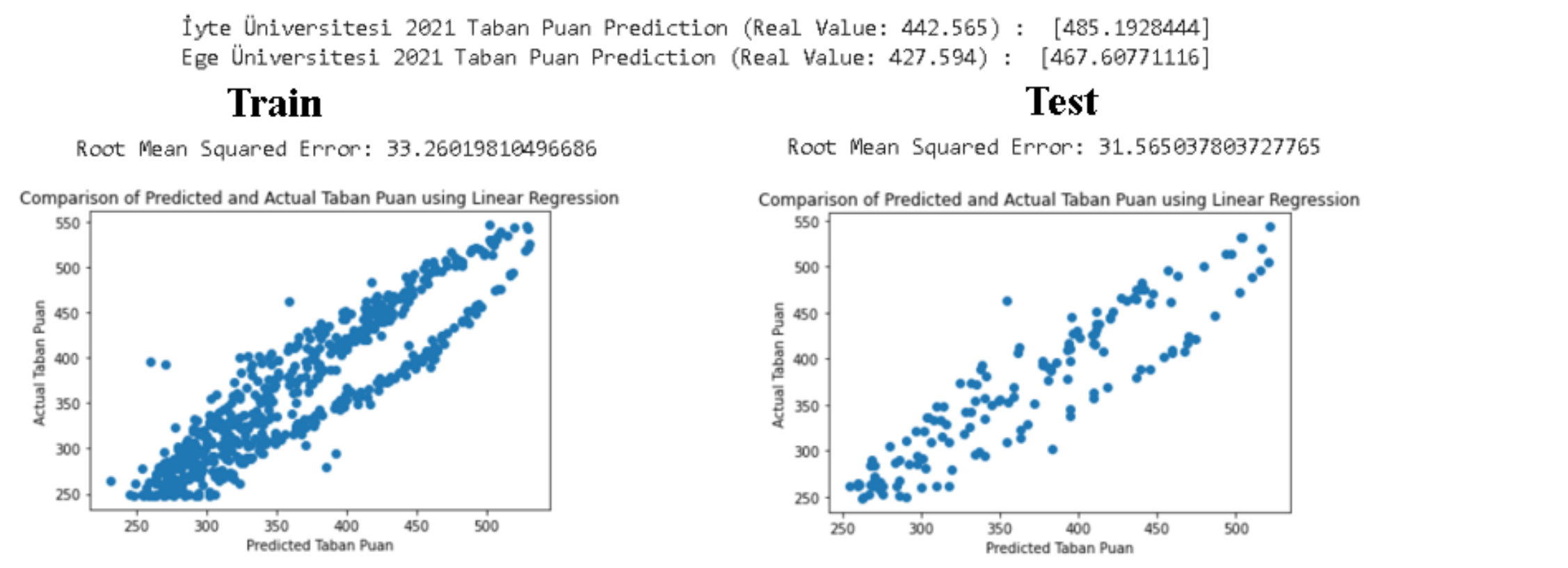
- Then, universities with data between 2016 and 2021 were selected and each year compared with previous year.
- The dataset consists of 753 samples and 13 features.

TABAN PUAN PREDICTION



	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

- The features in figure are given to the model as input.
- 80% of the entire dataset is allocated in a random way for training and 20% for the test set.
- Linear regression utilized is a linear model, a model in which the input variables and the single output variable have a linear relationship.



```
İyte Üniversitesi 2021 Taban Puan Prediction (Real Value: 442.565) :  [485.1928444]
Ege Üniversitesi 2021 Taban Puan Prediction (Real Value: 427.594) :  [467.60771116]
```

Train

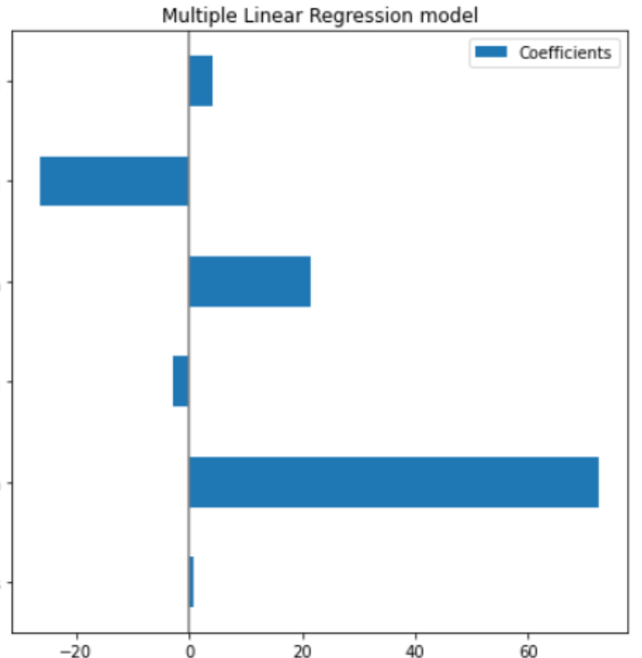
```
Root Mean Squared Error: 33.26019810496686
```

Test

```
Root Mean Squared Error: 31.565037803727765
```

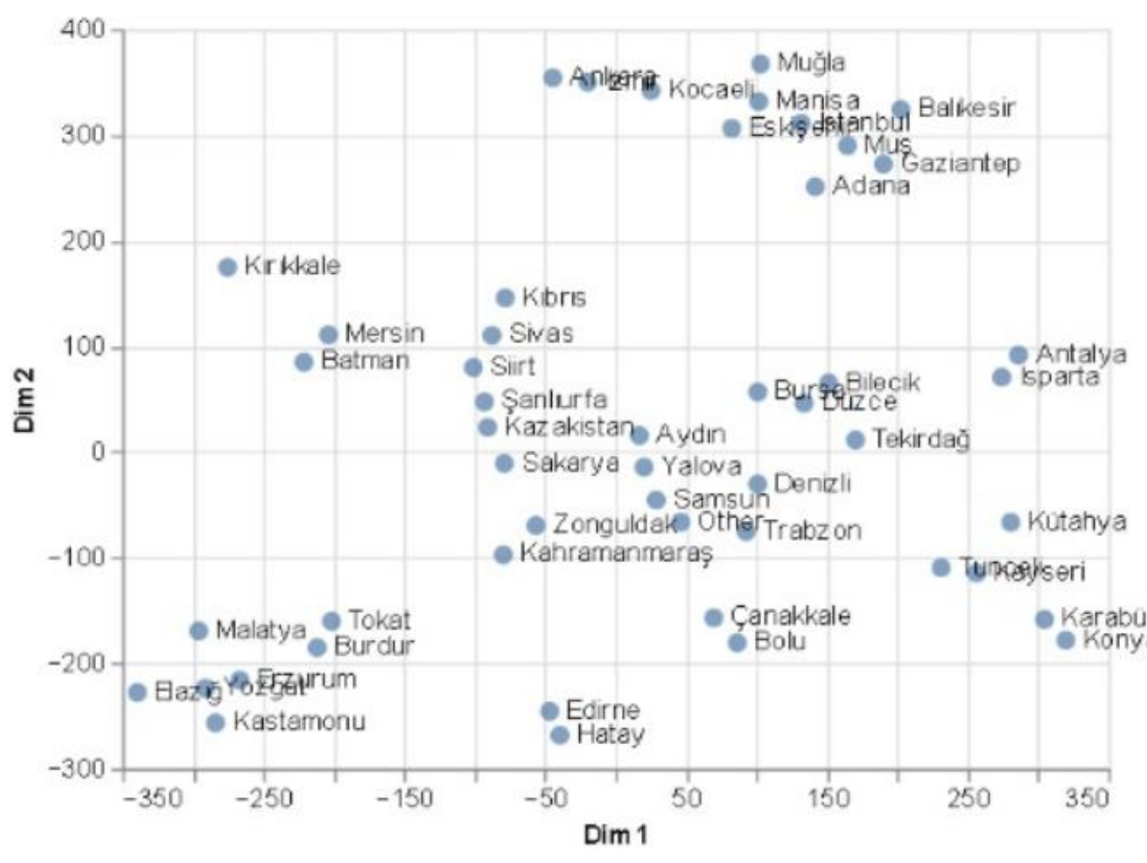
- The root mean squared error values obtained for both the train and the test set were calculated in terms of points.
- Comparison of predicted and actual threshold entry score using linear regression can be seen.

Feature Importance



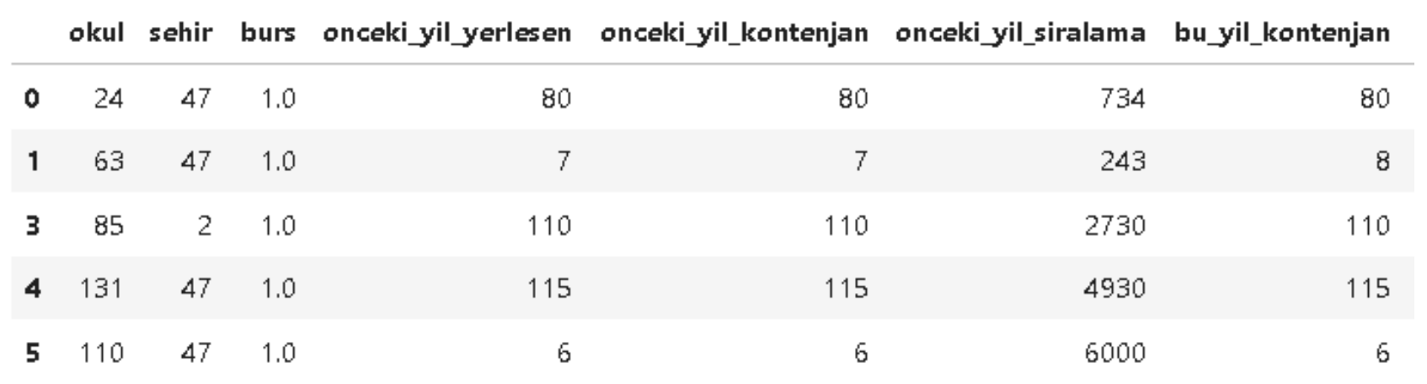
- Techniques that generate a score for all input characteristics for a particular model are referred to as feature importance.

Categorical Embedding



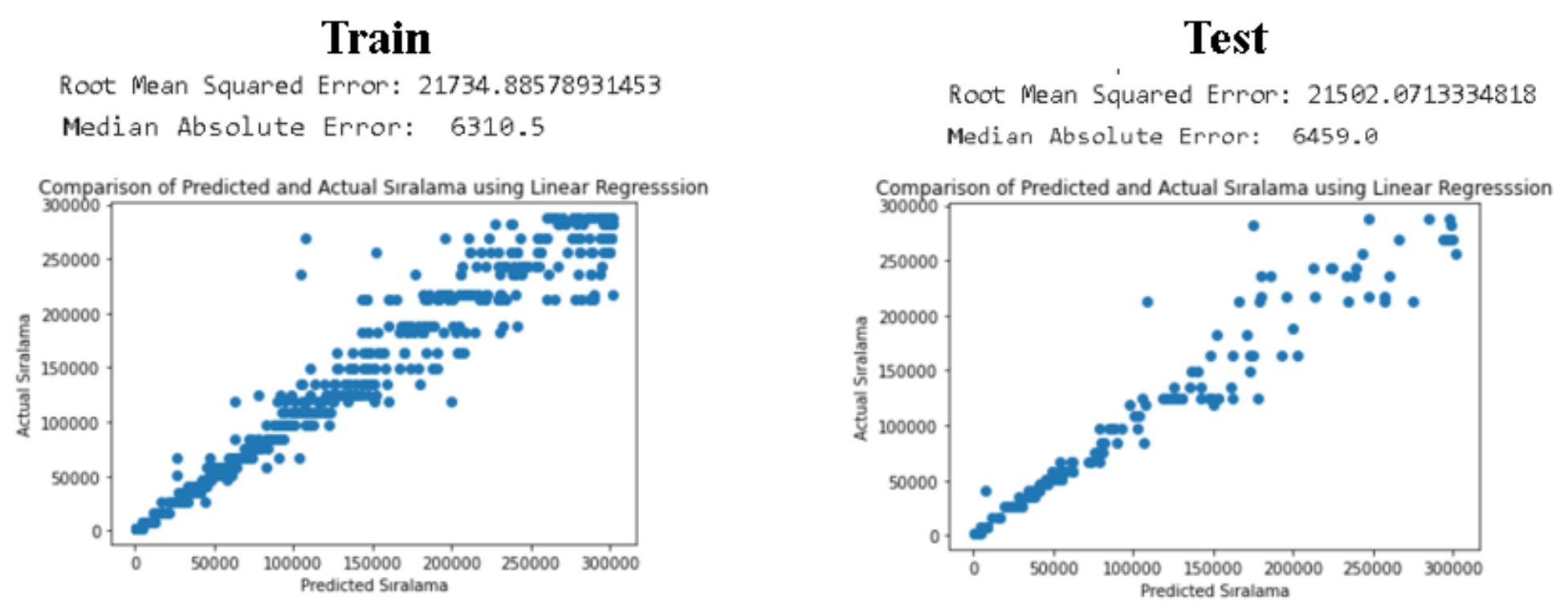
- To use these embedding layers, we first encoded the categorical variable with integer values.
- Each of these integers will then correspond to a vector representation of corresponding category.

SIRALAMA PREDICTION



	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

- The ranking is a parameter regardless of the difficulty of the exam. So, this prediction may give more reliable results.
- ”City” and ”university name" categorical features were added to the model by label encoding.



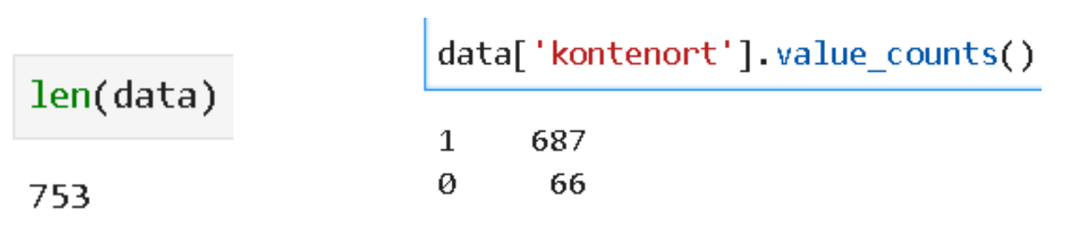
Train

```
Root Mean Squared Error: 21734.88578931453
Median Absolute Error:  6310.5
```

Test

```
Root Mean Squared Error: 21502.0713334818
Median Absolute Error:  6459.0
```

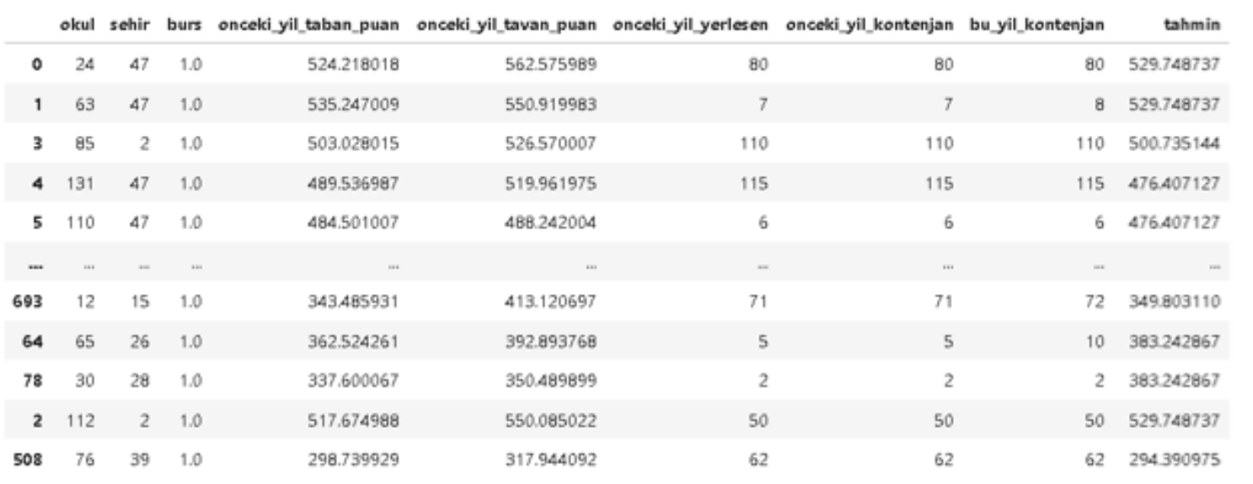
THE QUOTA PROBLEM



```
len(data)

753
```

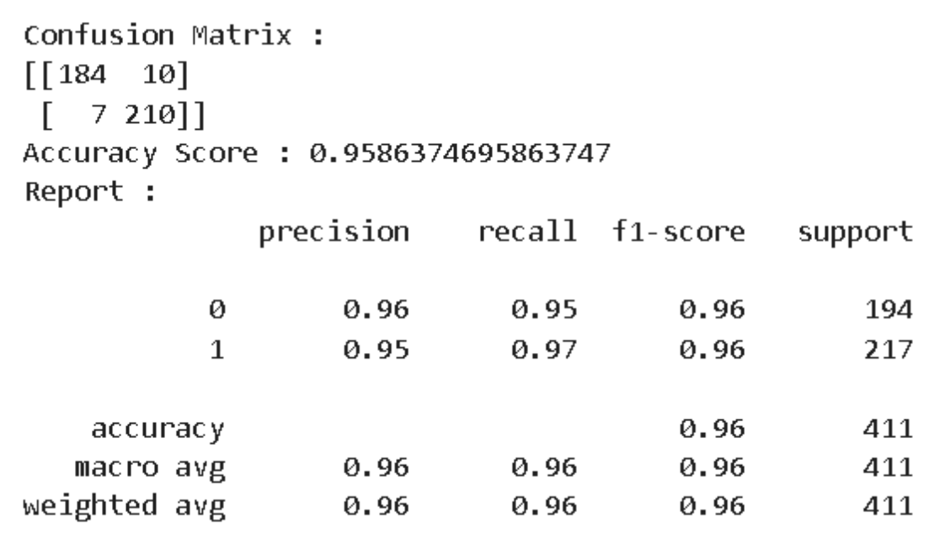
```
data['kontenort'].value_counts()

1     687
0      66
```

- Class 0 indicates that the university cannot fill its quota, while class 1 indicates that it will fill it.
- Unbalanced data set has been balanced by applying oversampling to the data set.



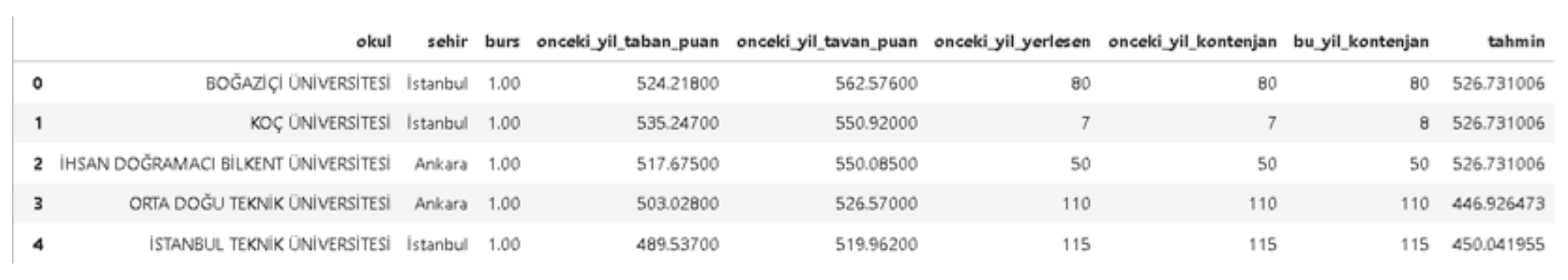
	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan	tahmin
0	24	47	1.0	524.218018	562.575989	80	80	80	529.748737
1	63	47	1.0	535.247009	550.919983	7	7	8	529.748737
3	85	2	1.0	503.028015	526.570007	110	110	110	500.735144
4	131	47	1.0	489.536987	519.961975	115	115	115	476.407127
5	110	47	1.0	484.501007	488.242004	6	6	6	476.407127
...	...	...	...	...	...	...	...	...	...
693	12	15	1.0	343.485931	413.120697	71	71	72	349.803110
64	65	26	1.0	362.524261	392.893768	5	5	10	383.242867
78	30	28	1.0	337.600067	350.489899	2	2	2	383.242867
2	112	2	1.0	517.674988	550.085022	50	50	50	529.748737
508	76	39	1.0	298.739929	317.944092	62	62	62	294.390975

- The results of the score estimation were also given to the model as input.



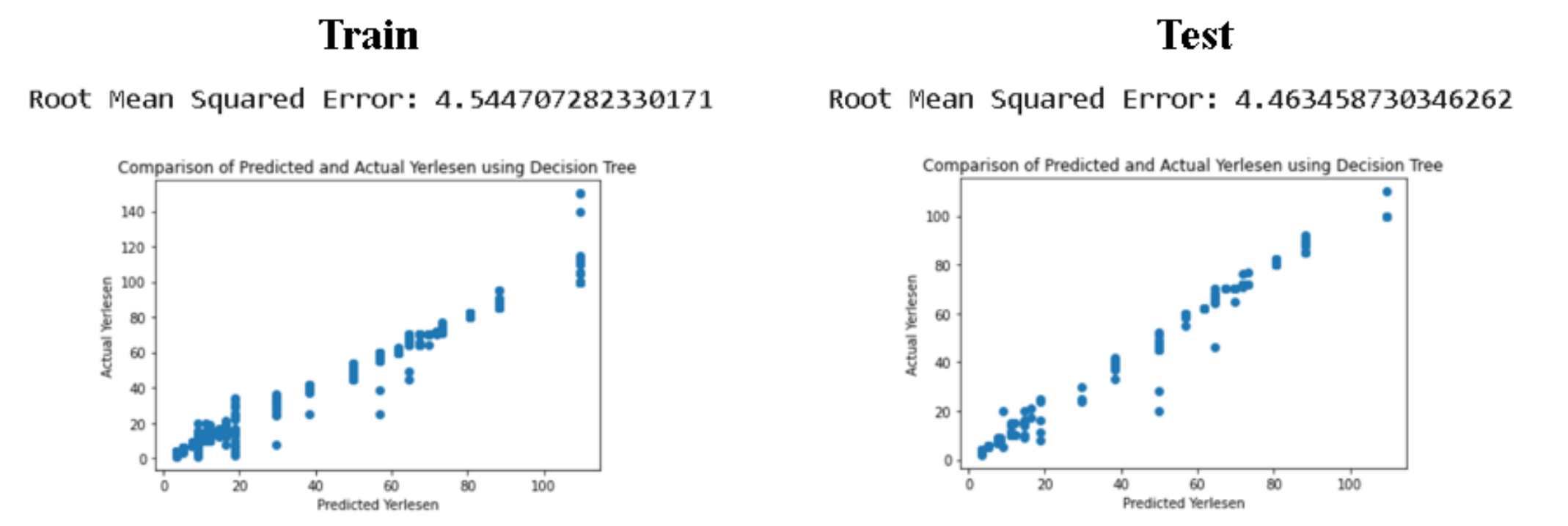
```
Confusion Matrix :
[[184  10]
 [  7 210]]
Accuracy Score : 0.9586374695863747
Report :
              precision    recall  f1-score   support

           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
```

YERLESEN PREDICTION



	okul	sehir	burs	onceki_yil_taban_puan	onceki_yil_tavan_puan	onceki_yil_yerlesen	onceki_yil_kontenjan	bu_yil_kontenjan	tahmin
0	BOĞAZİÇİ ÜNİVERSİTESİ	İstanbul	1.00	524.21800	562.57600	80	80	80	526.731006
1	KOÇ ÜNİVERSİTESİ	İstanbul	1.00	535.24700	550.92000	7	7	8	526.731006
2	İHSAN DOĞRAMACI BİLKENT ÜNİVERSİTESİ	Ankara	1.00	517.67500	550.08500	50	50	50	526.731006
3	ORTA DOĞU TEKNİK ÜNİVERSİTESİ	Ankara	1.00	503.02800	526.57000	110	110	110	446.926473
4	İSTANBUL TEKNİK ÜNİVERSİTESİ	İstanbul	1.00	489.53700	519.96200	115	115	115	450.041955

- A model was developed on calculating whether a university can fill its quota by estimating the number of people who have admitted.



Train

```
Root Mean Squared Error: 4.544707282330171
```

Test

```
Root Mean Squared Error: 4.463458730346262
```

RESULTS

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
Sıralama	-	Common Features,onceki_yil_siralama	Linear Regression	25735	People
Sıralama	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

- For estimation the entry threshold score, the most successful result belongs to linear regression without ”city“ and ”school” categorical features.
- In the prediction of yerlesen, 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 quota was obtained in the model where categorical values were added, oversampling was applied to the dataset and the predicted entry threshold score was added.

CONCLUSION

- It is decided that a trending ranking can be added as a feature by researching at the search numbers of universities and departments in 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 that as a trend.

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/