

**ENGINNERING FACULTY - COMPUTER ENGINEERING DEPARTMENT**

**MACHINE LEARNING**

**2022-2023 SPRING**

**FINAL PROJECT REPORT**

**INSTRUCTOR** : Assoc. Prof. Dr. Ahmet Çağdaş SEÇKİN

**STUDENT NAME** : …………………………………………………………..

**STUDENT ID** : …………………………………………………………..

**OTHER GROUP MEMBERS**

……………………………………………………………

……………………………………………………………

……………………………………………………………

……………………………………………………………

……………………………………………………………

# INSTRUCTIONS AND TABLE OF CONTENTS

Two main tasks will be given to the students to provide a common and unique/original dataset. The common dataset has a unique structure that is not available in known databases. The purpose of this dataset is to enable students to compete, and a ranking will be made based on the performance metric with a scoring ranging from a minimum of 5 points to a maximum of 20 points. In the unique/original dataset problem, the aim is to perform regression, classification, and clustering problems using a dataset obtained in a unique or original way. One or more datasets can be selected. Using a ready to use dataset results in a 5 points penalty. Machine learning final project instructions:

1. The project report will be prepared using the attached format. Do not forget to update the table of contents. Please do not deform the template file.
2. The report cannot exceed 20 pages (exclude cover page). Write short and concise descriptions. Points will be deducted for any content that is difficult to understand.
3. The report submission date will not be extended. Use only remote platform to submit project reports. Project report will not be delivered by mail or e-mail. No need for a printed copy.
4. Each student must submit a project report. Submitting the project report by a member of your group will mean that only that member gets points.
5. Coding will be done using python language. The submission of project codes will be accepted as a single submission for each group. Compress the project codes and your data set together as a zip and share them on the drive. Your codes should be ready to run when downloaded. During the evaluation, only your code will be run, and no settings/adjustment will be made.
6. If you have completed all the conditions above, your project will be evaluated according to the following criteria:
   1. Online exam :20 points
   2. Project and code format :15 points
   3. Common dataset project :[5, 20] points
   4. Regression application :15 points
   5. Classification application :15 points
   6. Clustering application :15 points

## ***TABLE OF CONTENTS***

[1 INSTRUCTIONS AND TABLE OF CONTENTS 1](#_Toc131877475)

[1.1 TABLE OF CONTENTS 1](#_Toc131877476)

[2 COMMON DATASET PROJECT 3](#_Toc131877477)

[2.1 DATASET AND PREPROCESSING 3](#_Toc131877478)

[2.2 COMMON DATASET RESULTS AND MODEL SELECTION 3](#_Toc131877479)

[3 REGRESSION PROJECT 4](#_Toc131877480)

[3.1 DATASET AND PREPROCESSING 4](#_Toc131877481)

[3.2 REGRESSION STEPS 4](#_Toc131877482)

[3.3 REGRESSION RESULTS AND MODEL SELECTION 4](#_Toc131877483)

[3.4 REGRESSION HYPERPARAMTER OPTIMIZATION RESULTS 4](#_Toc131877484)

[4 CLASSIFICATION PROJECT 5](#_Toc131877485)

[4.1 DATASET AND PREPROCESSING 5](#_Toc131877486)

[4.2 CLASSIFICATION STEPS 5](#_Toc131877487)

[4.3 CLASSIFICATION RESULTS AND MODEL SELECTION 5](#_Toc131877488)

[4.4 CLASSIFICATION HYPERPARAMTER OPTIMIZATION RESULTS 5](#_Toc131877489)

[5 CLUSTERING PROJECT 6](#_Toc131877490)

[5.1 DATASET AND PREPROCESSING 6](#_Toc131877491)

[5.2 CLUSTERING STEPS 6](#_Toc131877492)

[5.3 CLUSTERING RESULTS AND MODEL SELECTION 6](#_Toc131877493)

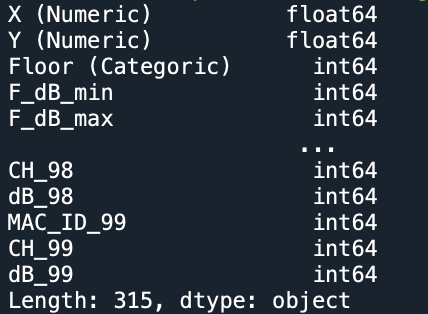
[5.4 CLUSTERING HYPERPARAMTER OPTIMIZATION RESULTS 6](#_Toc131877494)

# COMMON DATASET PROJECT

## DATASET AND PREPROCESSING

* **Data Source:** Doç.Dr.Ahmet Çağdaş Seçkin
* **Data Description:**

These are our columns.



Null count: 0

Number of columns and rows:

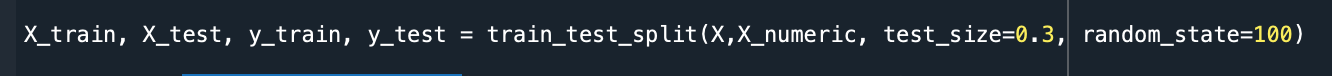


'X (Numeric)', 'Y (Numerical)', 'Floor (Categorical)' these are the columns we will guess. The remaining columns are our attributes.

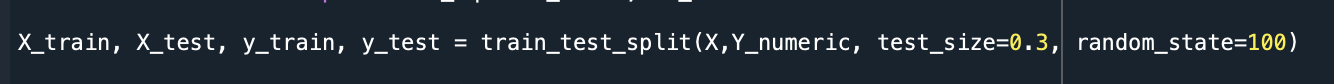
* **Data Split:**

A picture containing text, screenshot, font

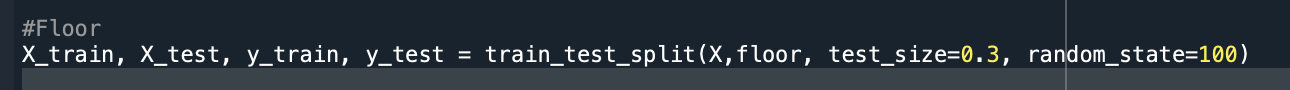
Description automatically generated

Distinction for this X numeric value

Distinction for this Y numeric value

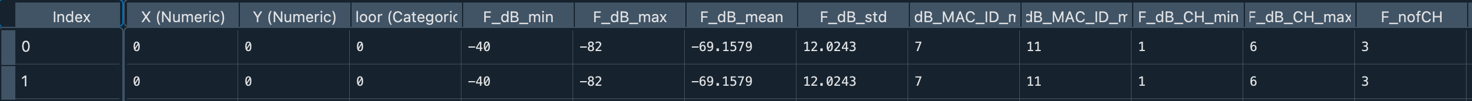


Distinction for this Floor value



* **Data Exploration:**

First two columns:



Here we can see the numerical values of the columns:

A picture containing text, screenshot, menu, font

Description automatically generated

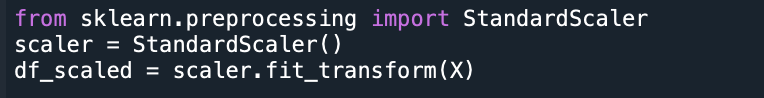
* **Data Preprocessing:**

Since there is a space at the beginning of the columns, I first deleted these spaces.



There are no null values.

There are no duplicate values.

We brought the features closer together with the scale operation.

* **Data Usage:**

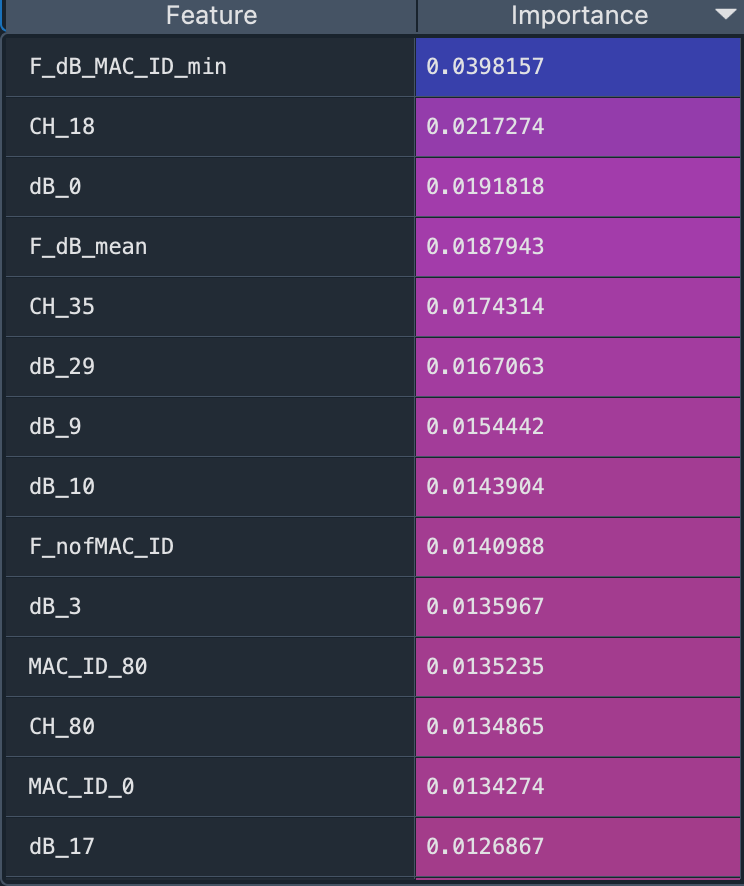
X and Y positions will be calculated with random forest, linear and knn and the best values will be selected according to r2 scores.

Floor classification will be trained with randomforest, adaboost and knn algorithms, and the best model will be found according to the accuracy value.

In this way, the best position where the device can be placed will be determined.

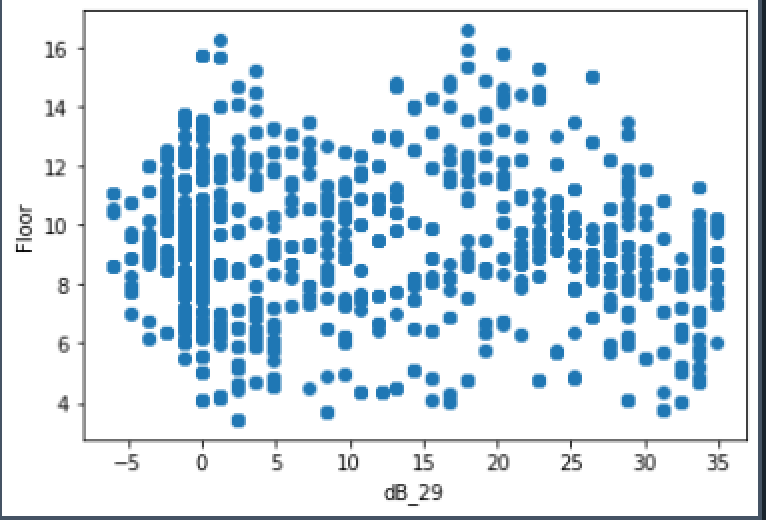
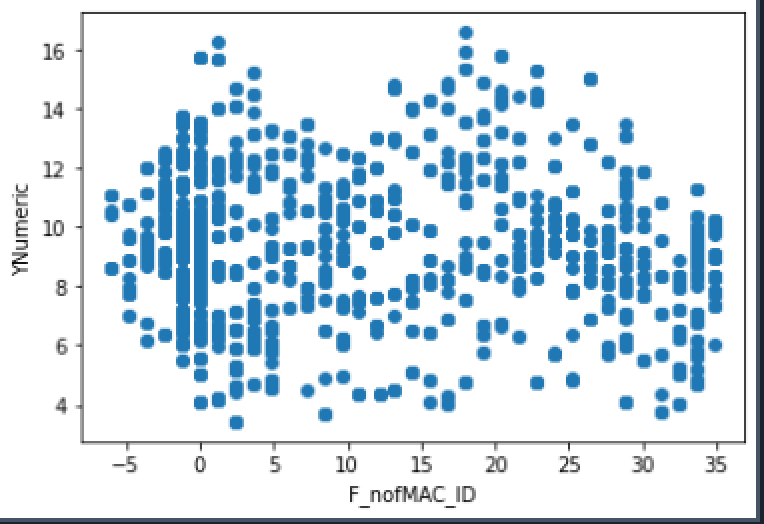
* **Feature engineering:**

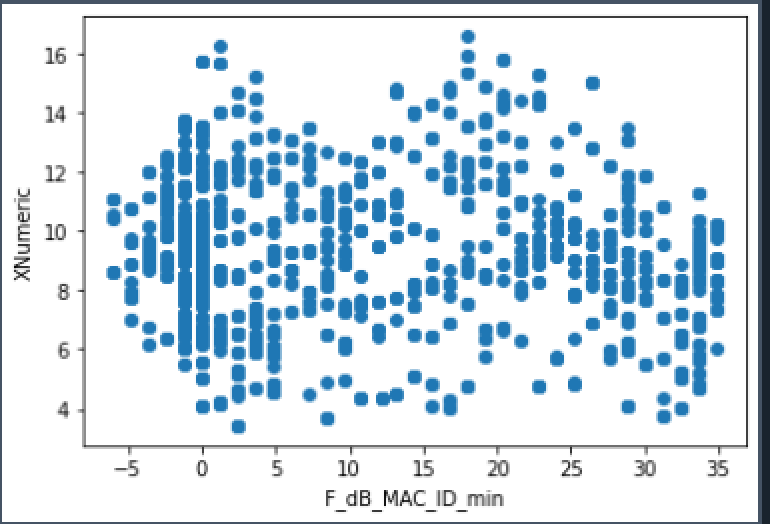
Since there are 315 columns, it is not possible to show the effects of all features, so it would be more accurate to give the weights of 10 features that affect the best model, the random forest.



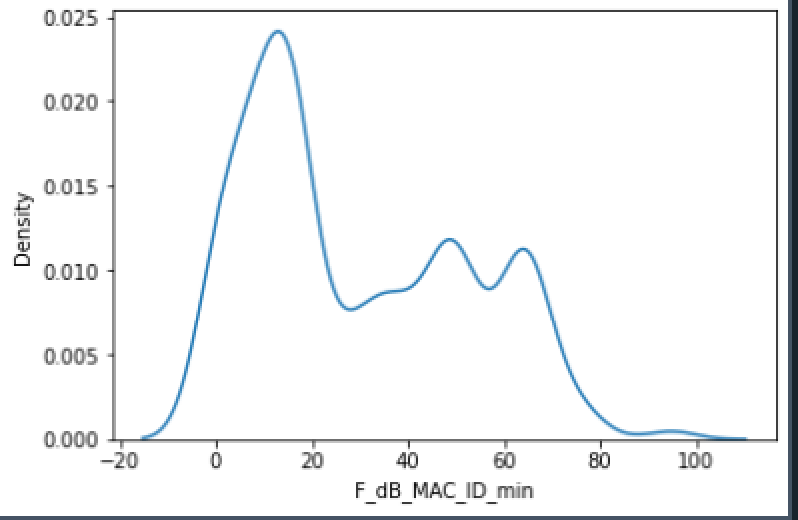
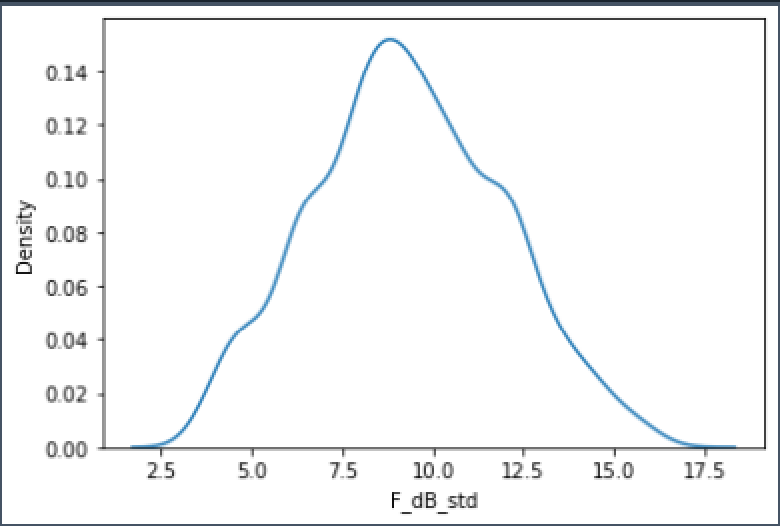
* **Dataset visualization:**

Scatter plots showing the relationship of the two features;

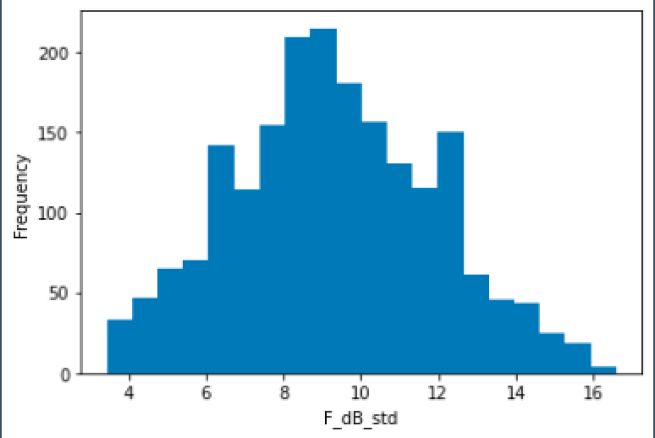
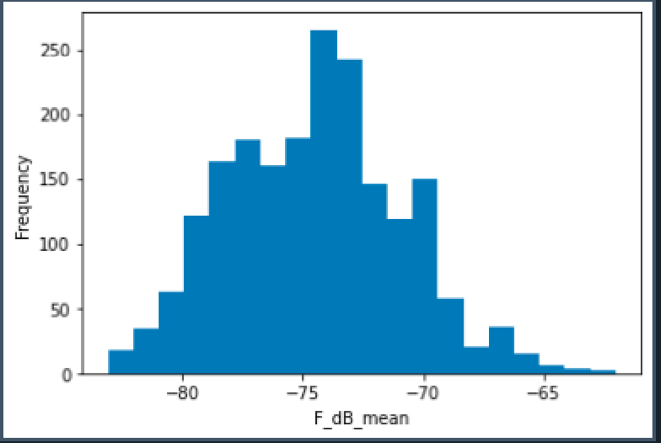




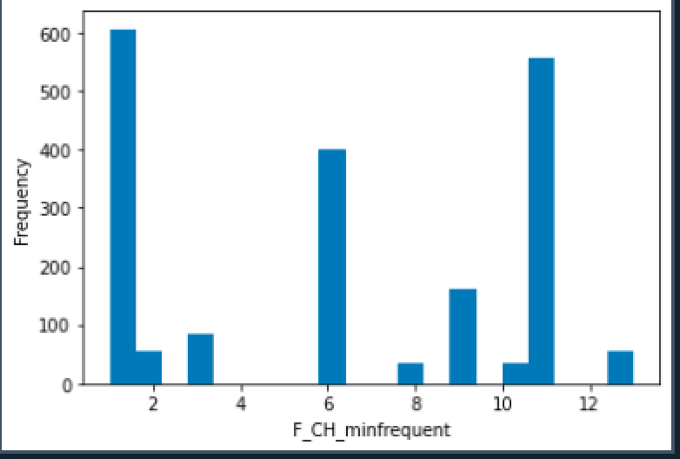
Density representations of the two selected features;

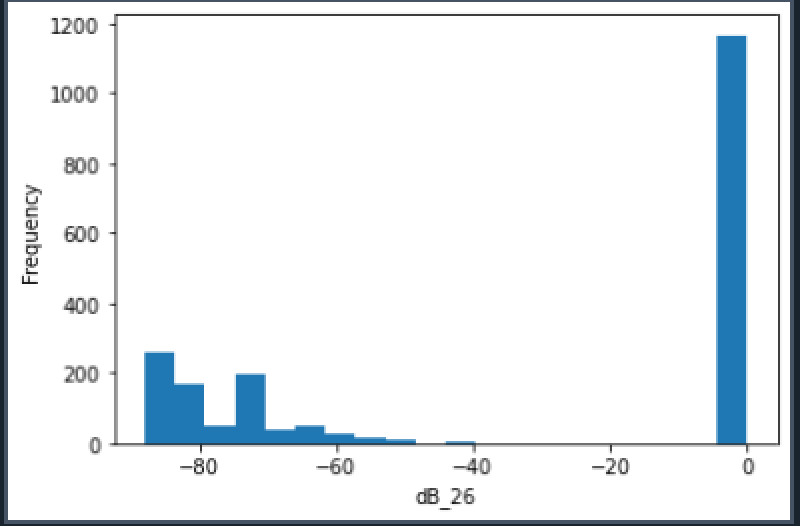
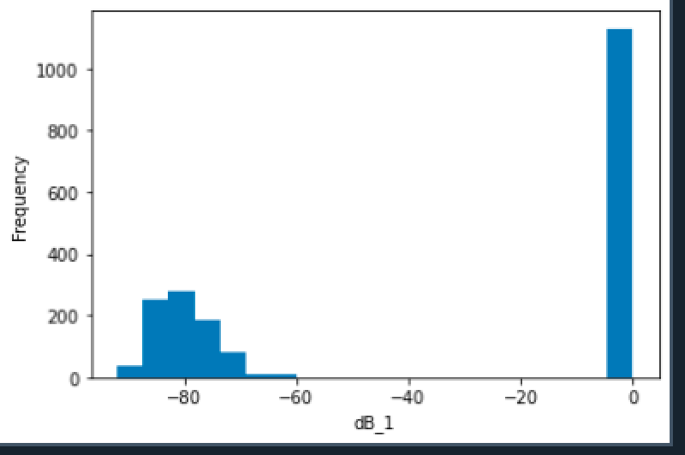


Graph showing the distributions of some selected features;



A picture containing text, screenshot, number, diagram

Description automatically generated



## COMMON DATASET RESULTS AND MODEL SELECTION

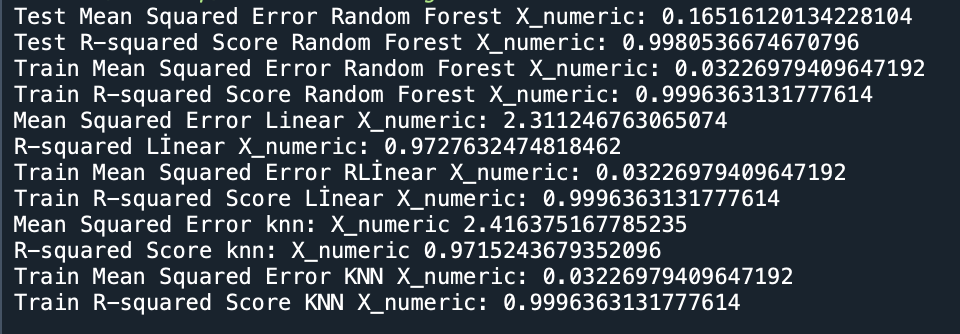
* Use multiple algorithms to get the best performance.

Three different algorithms were used for regression; RandomForest, Linear Regression and KNN

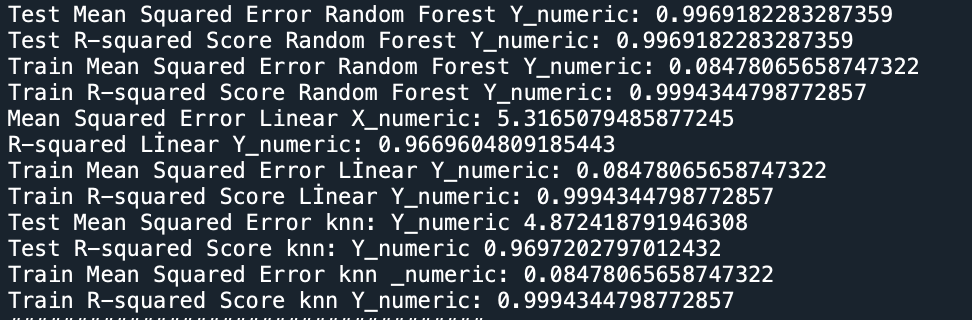
For classification, 3 different algorithms were used, these are random forest, adabood and knn algorithms.

The values of these algorithms according to 3 different labels are given below. According to R2 values and mse values

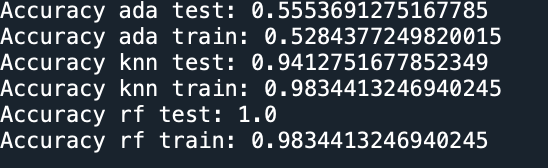
For the value of X



For the value of Y

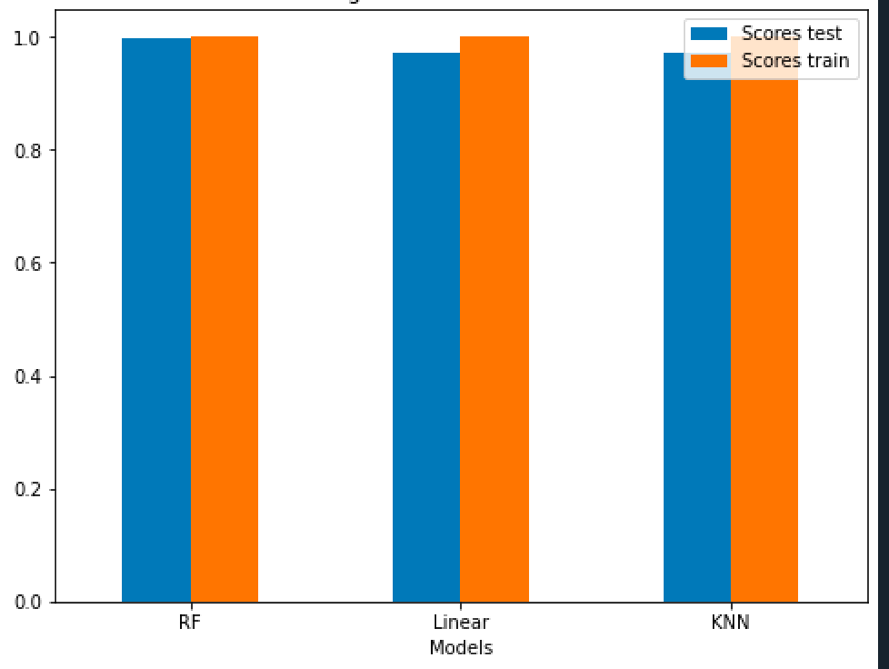


For the value of Floor

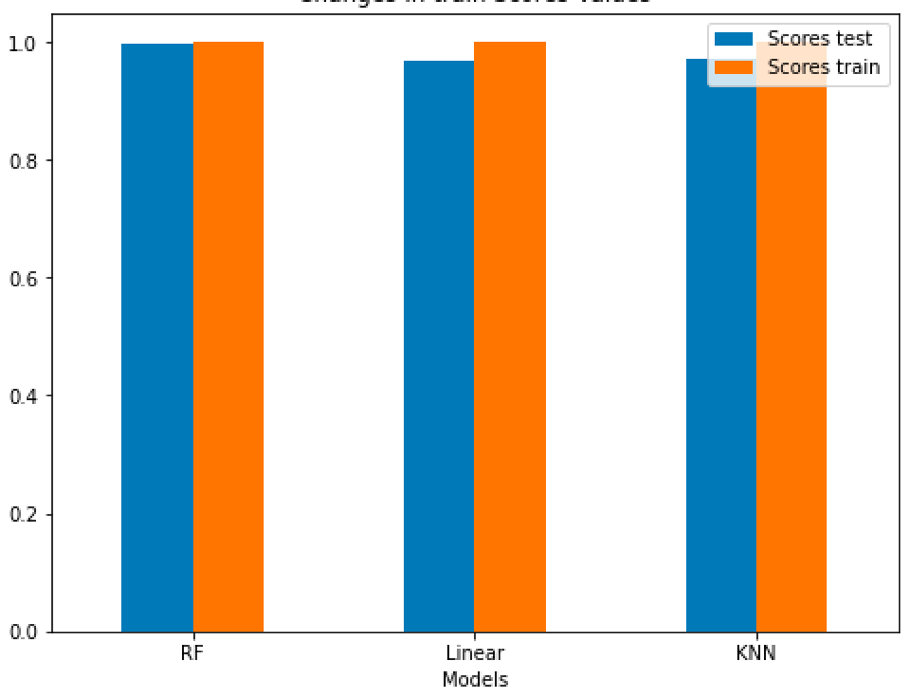


* Compare training results. Create a comparison table. And Compare test results. Create a comparison table.

Visualization of train and test results for X value



Visualization of train and test results for Y value



Visualization of train and test results for Floor valueA picture containing text, screenshot, software, plot

Description automatically generated

* Select the best training algorithm.

For the X value that gives the best r2 value in the regression according to the test results: Linear regression with 0.9980

For Y value: knn algorithm with 0.9969

We look at the one with the highest accuracy value for the Classification.

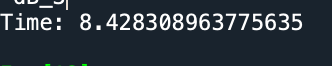
For floor value: It is a random forest model with 1.0.

ACHIEVEMENT VALUE = 0.9980 \* 0.9969 \* 1

= 0.9949062

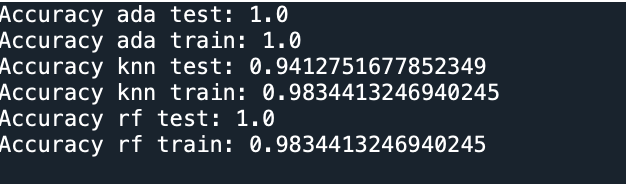
* Try to reduce test and training time.

Training time + test time =



* Try to improve performance metric.

When we made a Standardscaler, it was seen that the classification model gave better results.

The incremental results are below.

# ORIGINAL DATASET REGRESSION PROJECT

## DATASET AND PREPROCESSING

* **Data Source:** The origin of the dataset, including the organization or website where it was obtained. -5 POINTS IF IT IS AN ORDINARY DATASET.
* **Data Description:** A brief summary of the dataset, including the number of instances, features, and labels.
* **ML Problem definition:** Define the regression problem according to dataset.
* **Data Split:** The division of the dataset into training, validation, and testing sets.
* **Data Exploration:** The exploratory data analysis performed on the dataset, including any visualizations or statistical summaries used to understand the data.
* **Data Preprocessing:** The preprocessing steps applied to the dataset, including any missing value imputation, feature scaling, or feature selection techniques.
* **Data Usage:** A description of how the dataset will be used in the final project, including which machine learning algorithms will be applied and how the results will be evaluated.
* **Feature engineering:** show effect of each feature on output. Try to explain effects.
* **Dataset visualization:** try to explain dataset with histograms, plots, cross-correlation tables, scatter plots, etc.

It is important to provide sufficient information about the dataset to ensure that others can reproduce the results and understand the limitations of the data.

## REGRESSION STEPS

* A brief summary of each stage in regression.

## REGRESSION RESULTS AND MODEL SELECTION

* Use multiple regression algorithms to get best performance on MAE and R2.
* Compare training results. Create a comparison table.
* Compare test results. Create a comparison table.
* Select best training algorithm. Plot residual scatter plot.

## REGRESSION HYPERPARAMTER OPTIMIZATION RESULTS

* Use best training algorithm and try to tune hyperparameters to get better performance metrics.
* Try to reduce test and training time.
* Try to improve MAE metric.

# ORIGINAL DATASET CLASSIFICATION PROJECT

## DATASET AND PREPROCESSING

* **Data Source:** The origin of the dataset, including the organization or website where it was obtained. -5 POINTS IF IT IS AN ORDINARY DATASET.
* **Data Description:** A brief summary of the dataset, including the number of instances, features, and labels.
* **ML Problem definition:** Define the classification problem according to dataset.
* **Data Split:** The division of the dataset into training, validation, and testing sets.
* **Data Exploration:** The exploratory data analysis performed on the dataset, including any visualizations or statistical summaries used to understand the data.
* **Data Preprocessing:** The preprocessing steps applied to the dataset, including any missing value imputation, feature scaling, or feature selection techniques.
* **Data Usage:** A description of how the dataset will be used in the final project, including which machine learning algorithms will be applied and how the results will be evaluated.
* **Feature engineering:** show effect of each feature on classes. Try to explain effects.
* **Dataset visualization:** try to explain dataset with histograms, plots, cross-correlation tables, scatter plots, etc.

It is important to provide sufficient information about the dataset to ensure that others can reproduce the results and understand the limitations of the data.

## CLASSIFICATION STEPS

* A brief summary of each stage in classification.

## CLASSIFICATION RESULTS AND MODEL SELECTION

* Use multiple classification algorithms to get best performance.
* Compare training results. Create a comparison table.
* Compare test results. Create a comparison table.
* Select best training algorithm. Show confusion matrix.

## CLASSIFICATION HYPERPARAMTER OPTIMIZATION RESULTS

* Use best training algorithm and try to tune hyperparameters to get better performance metrics.
* Try to reduce test and training time.
* Try to improve performance metric.

# ORIGINAL DATASET CLUSTERING PROJECT

## DATASET AND PREPROCESSING

* **Data Source:** www.twitter.com
* **Data Description:**

A screenshot of a computer

Description automatically generated

A dataset with 284 rows and 19 columns. The final state of this data after preprocessing

Like Count = Shows how many likes the tweet received

Comment Count = Shows how many comments the tweet received

Retweet Count = Shows how many retweet the tweet received

View Count = Shows the number of times the tweet was viewed

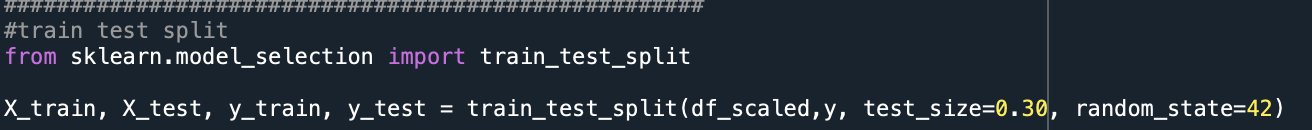
countOfWords = Shows how many words the tweet consists of

countOfPositive = Returns the number of positive words in the tweet. This number was determined according to the words in the data set we prepared ourselves.

countOfNegative = Returns the number of negative words in the tweet. This number was determined according to the words in the data set we prepared ourselves.

Art,Health,Politics,Science,Sport = These columns show which sector the tweet is in, for example, if the tweet is in the Arts sector, it would say 1 if not 0.

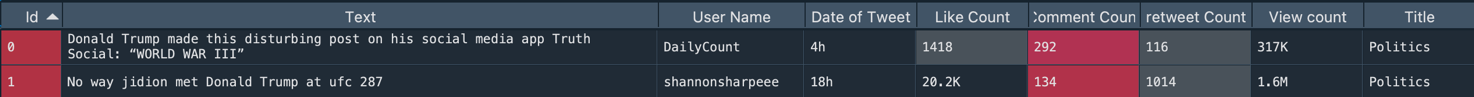
C\_Text\_\_mysteries,C\_Text\_\_replying,T\_Text\_\_replying,T\_Text\_\_manchester,C\_Text\_\_manchester,T\_Text\_\_believe,T\_Text\_\_mark = These columns are the columns that make the biggest contribution to the data as a result of the Tf-idf and count vectorization processes of the tweet content.

* **ML Problem definition:** The purpose of this cluster method is to find outlier values in the data set and to determine how much these values increase or decrease the accuracy of the data set.
* **Data Split:** The division of the dataset into training, validation, and testing sets.

70% of the dataset is reserved for train and 30% for testing.

* **Data Exploration:**

The raw states of the Data Set after we took it ourselves.

First 2 rows of dataset 

Data size is 286 rows and 9 columns.



Data types of columns;

A screenshot of a computer program

Description automatically generated with low confidence

There are 2 numeric values in the columns. The “m”, “k” texts in the like, retweet, view columns should be removed and made numerical in them.

Hist values;

Since there are only 2 numeric values, two figures appear. Here, we can reach the information that a maximum of 0-2000 comments were made.A picture containing text, screenshot, diagram, line

Description automatically generated

When we look at the unique values of the title column, we see 5 different title values.

* **Data Preprocessing:**

**Steps:**

**1)** Instead of the K and M values in the comment, retweet and view columns, we convert those columns into numerical data by typing their numerical equivalents.

A picture containing text, screenshot, font, software

Description automatically generated

**2)** We start to infer meaning in the text column. These text-mining steps

* + - 1. We calculate the number of words in the text content and save it to the countOfWords columnA screen shot of a computer code

         Description automatically generated with low confidence
      2. We find the number of positive and negative words in the text content, based on the data set we prepared ourselves last term. And we save them in countOfNegative and countOfPositive columns.A picture containing text, screenshot, font

         Description automatically generated
      3. We perform text-mining by applying TfIdf and count vectorixation operations to the text column. Since there are about 4 thousand columns here, we make feature importance on Orange and add the ones that give the best results according to r2 and MSE values to the data set.A screen shot of a computer code

         Description automatically generated with low confidence

The values in Orange;

MSE

A screenshot of a computer

Description automatically generated with medium confidence

R2 score: A screenshot of a computer

Description automatically generated

**3)** We are deleting the user\_name and date of tweet columns that do not contribute to the data set. Since we have completed our operations in the Text column, we delete it as well.

A picture containing text, font, screenshot

Description automatically generated

**4)** We apply one-hot-encoding operations to the title column, because our goal is to find outlier values, so object values will not work for us.

A picture containing text, screenshot, font

Description automatically generated

**5)** Finally, if there are duplicate values, we delete them and remove the null values from the data set.

A picture containing font, text, screenshot, graphics

Description automatically generated



* **Data Usage:**

DBSCN, K-MEAN and GaussianMixture methods will be used and after outlier extraction with them, the change of outlier extracted data set and outlier unallocated data set will be observed. Evaluation scores will be silhoutte and Davies scores.

* **Feature engineering:**

A picture containing screenshot, text, software

Description automatically generated A picture containing screenshot, text

Description automatically generatedA picture containing screenshot

Description automatically generated

A picture containing screenshot, text, plot, line

Description automatically generatedA picture containing screenshot, text, line, plot

Description automatically generatedA picture containing screenshot, text, plot, number

Description automatically generated

A picture containing screenshot, text, line, plot

Description automatically generatedA picture containing screenshot, text, software, line

Description automatically generatedA graph with purple and yellow dots

Description automatically generated with low confidence

A picture containing screenshot, line, text

Description automatically generatedA picture containing screenshot, text, display, software

Description automatically generatedA picture containing screenshot, text, line

Description automatically generated

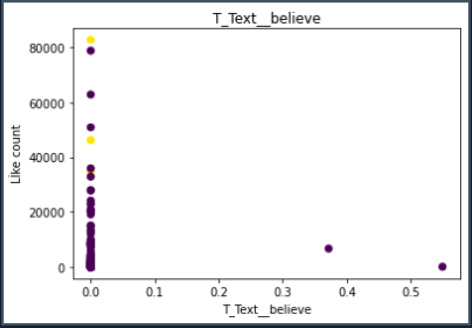
A picture containing screenshot, text, software, number

Description automatically generatedA picture containing screenshot, text, line

Description automatically generatedA picture containing screenshot, text

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generatedA picture containing screenshot, text, software, number

Description automatically generated

In these graphs, we see the effect of the features on outlier extraction with the K-Means algorithm. For example, as the value in the View Count column increases, the outlier values increase. Or, when the number of words in the countOfWords column is between 10 and 30, the outlier values show redundancy here. If there are no negative words in the CountOfNegative column, we can make inferences that outlier values are found here.

* **Dataset visualization:**

Correlation graph, with this graph, it is seen that the columns that affect each other the most are like count, comment column, retweet column and view column. At the same time, the columns obtained as a result of text-mining affect each other.A screenshot of a computer

Description automatically generated with medium confidence

Density map of the Like Count;

A picture containing text, screenshot, plot, line

Description automatically generated

Here, it can be seen that the density of Likes is between 0 and 20000.

Heat map and plot display between Like count and Commet count, where the correlation between them is high;

A picture containing text, screenshot, diagram

Description automatically generated

Here we can see that low number of likes get more comment, this like count is valid for 1000 and below.

The violinplot representation between the number of Likes and the number of views, where the correlation between them is high;A picture containing diagram, text, line, screenshot

Description automatically generated

This chart visualizes the distribution of the "CommentCount" and "LikeCount" properties over samples containing the top 5 most frequent values of the "CommentCount" property..

## CLUSTERING STEPS

1. First of all, a standard scaler operation was applied to the dataset because the data distribution was very common.
2. The dataset is split into train test (70-30). Here, the value of y is the number of Likes, the x dataset contains other features apart from the number of likes.
3. DBSCN, K-Means and GMM algorithms were set up to try different algorithms.
4. Best parameters for DBSCN found via Orange eps=2.246 min\_samples=2 (2 selected to find outlier)
5. Since those that do not belong to any cluster for DBSCN are shown with -1, the rows with the value of -1 are removed from the dataset and a data set with outlier values is obtained for dbsnc.
6. n\_cluster=2 was selected for the K-Means algorithm (to find outlier values)
7. The normal distribution threshold was performed to find outliers for K-Means. If the distance value for each value was greater than the threshold value, the outlier value was accepted. By subtracting these values from the data set, a data set with outlier values for k-Mean was obtained.
8. Silhouette scores of different parameters for GMM were calculated and the parameters that gave the best results were selected.
9. Selected n\_components=2 covarience\_type=” spherical” for GMM.
10. Normal distiribution was applied to find outlier values in GMM.
11. Outliers above threshold 3 are accepted for GMM. (Z-Score). By subtracting these values, a data set without outliers was obtained for GMM.
12. Outliers extracted data sets were fitted and silhouette and Davies scores were calculated.
13. In the data set (X\_test) without outlier values, fit and silhouette and Davies scores were calculated.
14. Evaluations were made in the clusters prepared with 2 different methods of these 3 different cluster methods.

## CLUSTERING RESULTS AND MODEL SELECTION

Three different algorithms were used for the cluster: DBSCN, K-Means and GaussianMixture algorithms. The location and visualization of the outlier values found by these algorithms are given below.

DBSCN;

A graph with red and blue dots

Description automatically generated with low confidence

K-Means;

A graph with red and blue dots

Description automatically generated with low confidence

GMM;

A picture containing screenshot, text, display, software

Description automatically generated

Their performance difference between outlier and non-outlier datasets; Here train dataset is not outlier, test dataset is outlier dataset.

A screenshot of a computer program

Description automatically generated with medium confidence

Visualization of their comparison;A picture containing text, screenshot, line, plot

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated

As we can see from these graphs, when we subtract outlier values in the DBSCN algorithm, it gives better results.

But the GMM algorithm gives the best result with 0.77 without removing the outlier values.

## CLUSTERING HYPERPARAMTER OPTIMIZATION RESULTS

The best working algorithm is the GMM algorithm. The silhouette scores that change when we change the parameters of this algorithm sequentially are given below. Accordingly, since we are scanning outliers, the best covariance\_type is "spherical" with n\_components=2.A screenshot of a computer

Description automatically generated with medium confidence