SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Predict Students Dropout and Academic Success

SIC 603

Supervised by :

Facilitator : Shaimaa Othman

Team Members

Ashraf Saber

Mahmoud Wahban

Salma Sherif

Agenda

- Motivation
- Targeted audience
- About the problem
- Understanding the data
- EDA & Visualization
- Preprocessing
- Modeling
- Final Results
- Future Improvements
- Conclusion

1. Motivation

1. Motivation

Why is it important to Predict Students Dropout and Academic Success ?

Predicting student dropout helps identify at-risk students, enabling timely support from educators. This fosters resilient environments and reduces dropout rates, positively impacting communities. By prioritizing academic success, we inspire lifelong learning and empower students to change their life trajectories.

2.Targeted audience

2.Targeted Audience

جامعة الإسكندرية
ALEXANDRIA
UNIVERSITY

مدينة زويل
ZEWAIL CITY
ESTABLISHED 2000
INAUGURATED 2011

2006
SOHAG UNIVERSITY
جامعة سوهاج

3. About the problem

3. About the problem

In a bustling city, a vibrant university attracted diverse students, each with unique backgrounds and challenges. A rich dataset captured their stories, highlighting factors like marital status, application modes, and courses chosen, reflecting their aspirations. This data revealed patterns of success and struggle, particularly showing that single and evening students faced more obstacles.

4. Understanding the data

4. Understanding the data

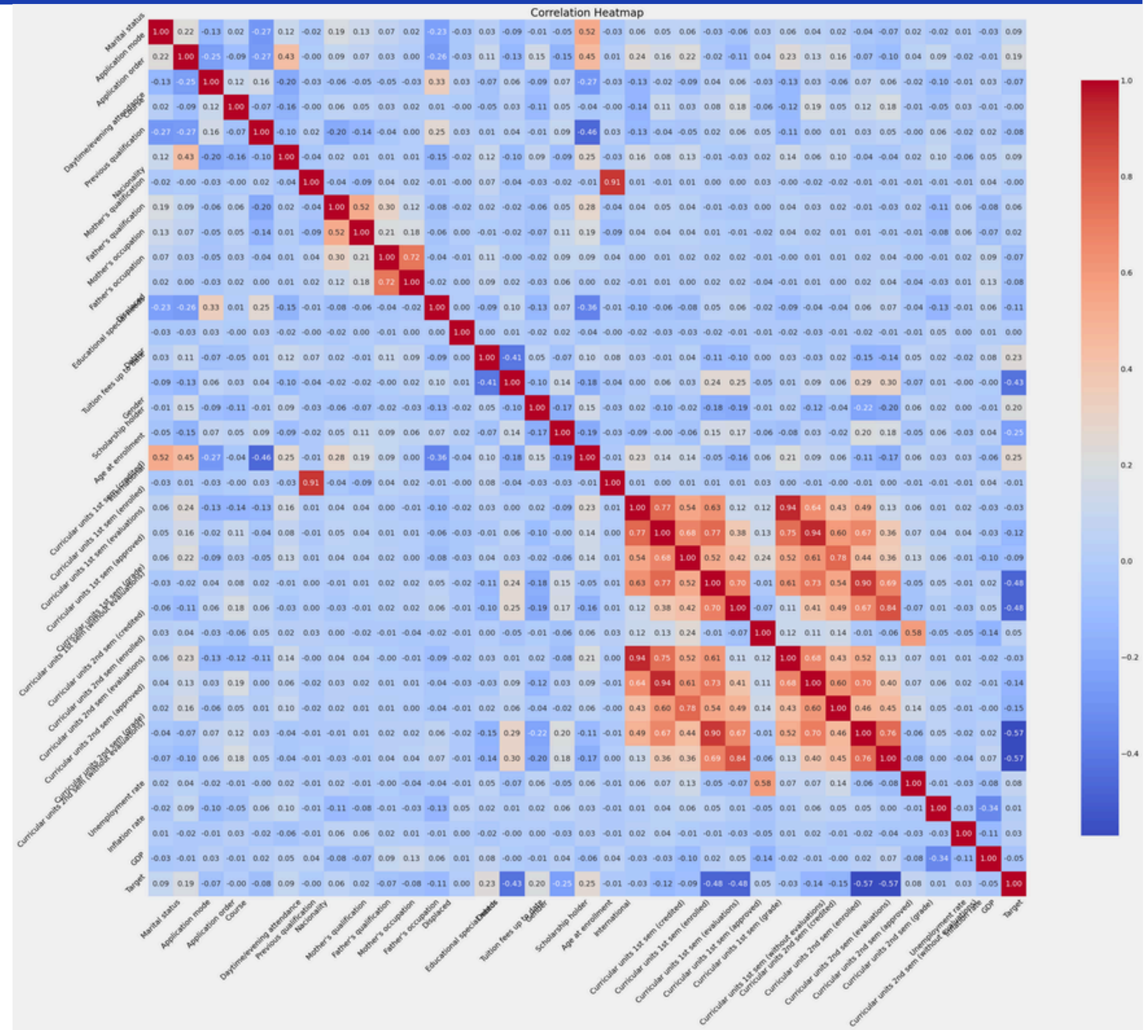
Our data consists of 37 features, here’s the most important ones.

- **Marital Status**: The marital status of the student (e.g., single, married, divorced).
- **Application Mode**: Refers to the mode or type of application the student submitted to enroll in the course.
- **Application Order**: Indicates the order in which the student applied for the course. For example, whether it was the student’s first, second, or third choice.
- **Course**: The course or degree program the student is enrolled in (e.g., Computer Science, Engineering, etc.).
- **Daytime/Evening Attendance**: Specifies whether the student attends the course during the day or in the evening, representing their attendance schedule.
- **Previous Qualification**: The type of academic qualification the student had before enrolling in the course (e.g., high school diploma, vocational training).
- **Previous Qualification (Grade)**: The final grade or score associated with the student's previous qualification.
- **Nationality**: The nationality of the student.
- **Mother's Qualification**: The highest academic qualification attained by the student's mother.
- **Father's Qualification**: The highest academic qualification attained by the student's father.

5. EDA & Visualization

5. EDA & Visualization

Correlation between features

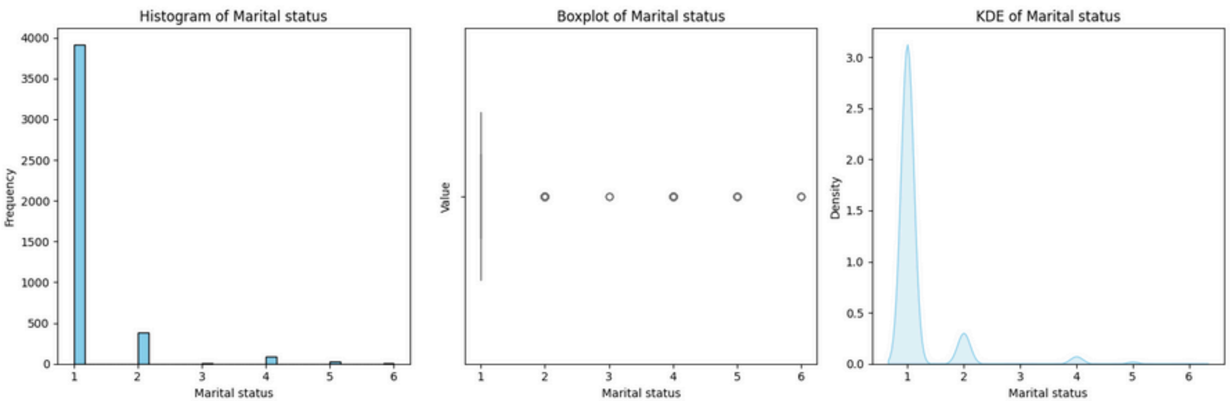


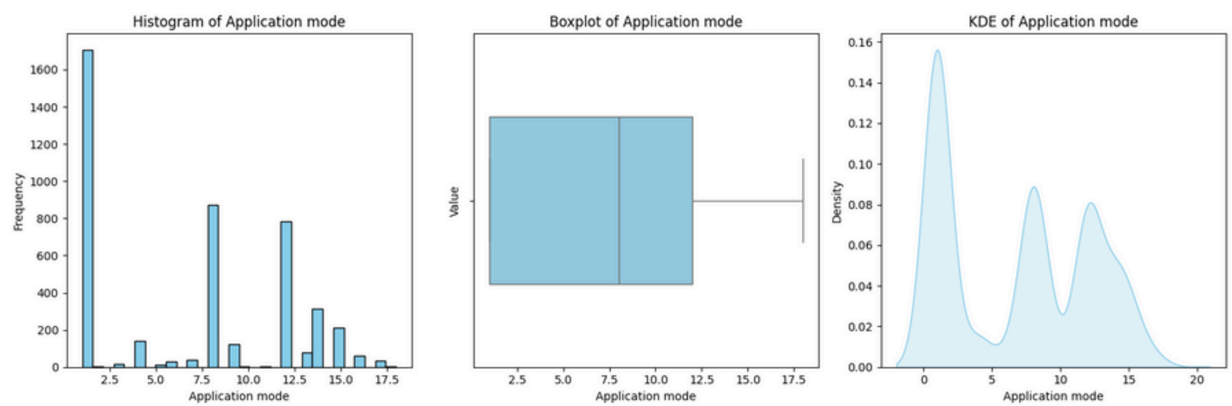
Strongest correlation with target :

Curricular units 2nd sem (grade) 0.571792
Curricular units 2nd sem (approved) 0.569500
Curricular units 1st sem (grade) 0.480669

5. EDA & Visualization

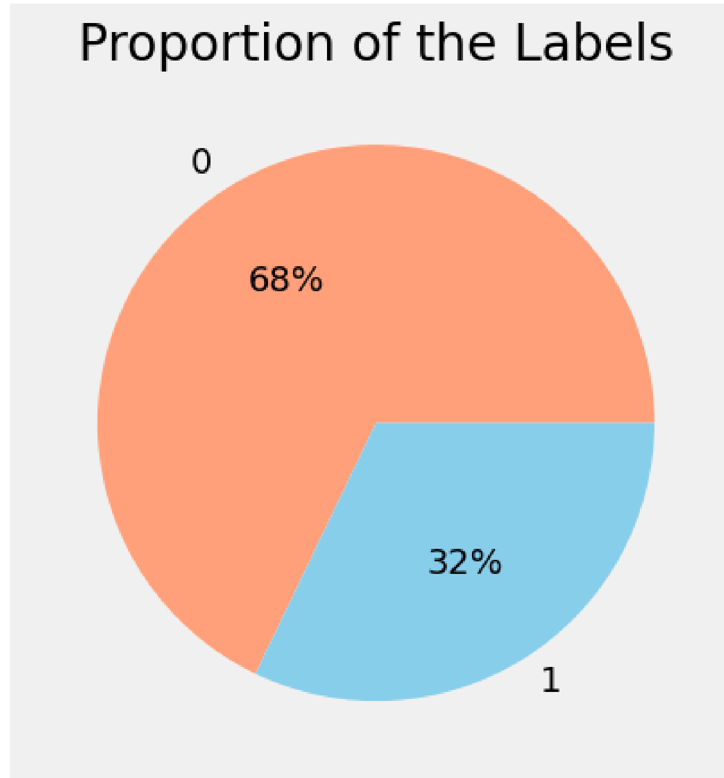
Different types of visuals to check up features





5. EDA & Visualization

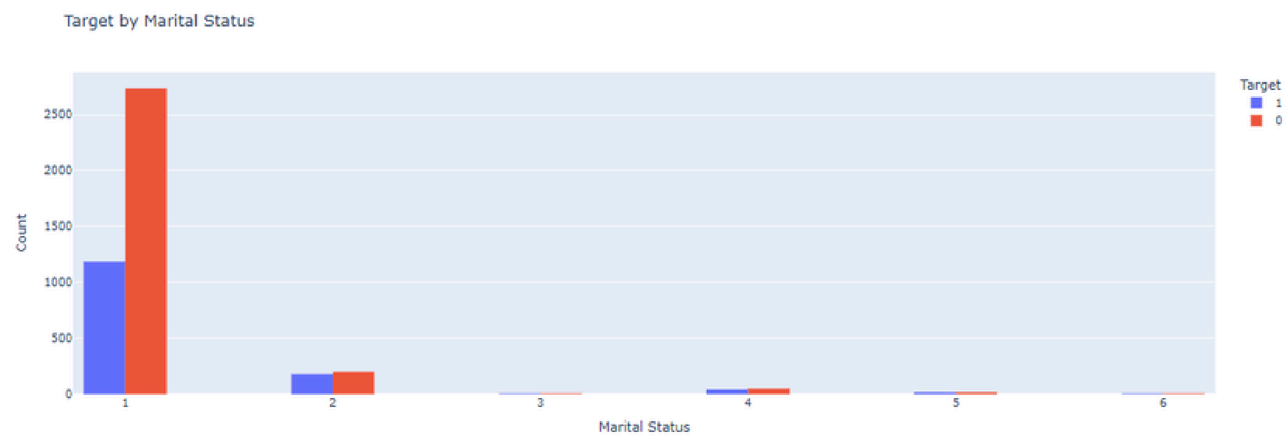
Target column distribution



imbalanced data

5. EDA & Visualization

Target by
Marital Status



Single Individuals are more likely to have dropped out

5. EDA & Visualization

Target by Gender

Females are more likely to have dropped out

6. Preprocessing

6. Preprocessing

Encoding

We did some encoding and mapping to our nominal and categorical columns such as one hot encoding and label encoding.

6. Preprocessing

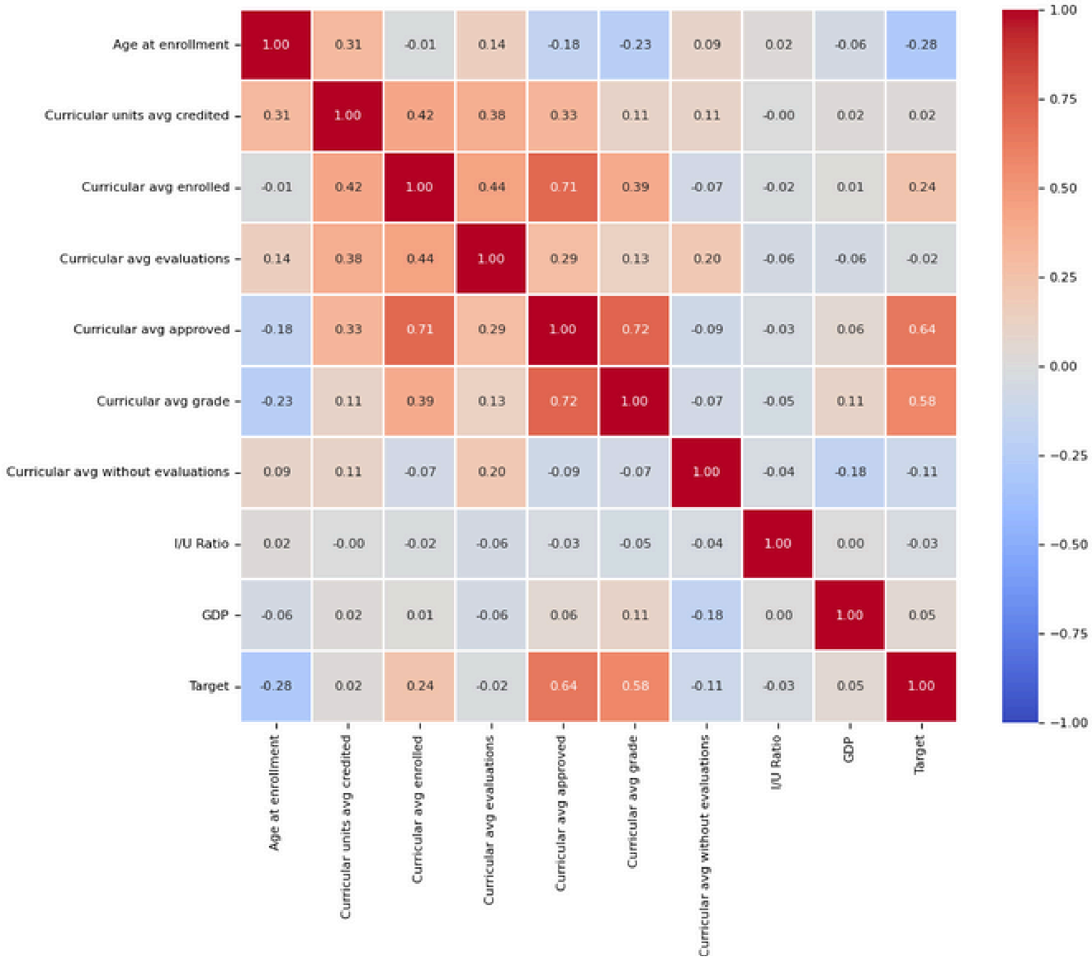
Feature Engineering

We started by dropping some unnecessary columns such as Nacionality, International and Educational special needs

Then, we decided to create a new feature that represents inflation to unemployment ratio instead of the two features.

Finally, we created some features that represents the average of some features such as grades, evaluations and courses enrolled.

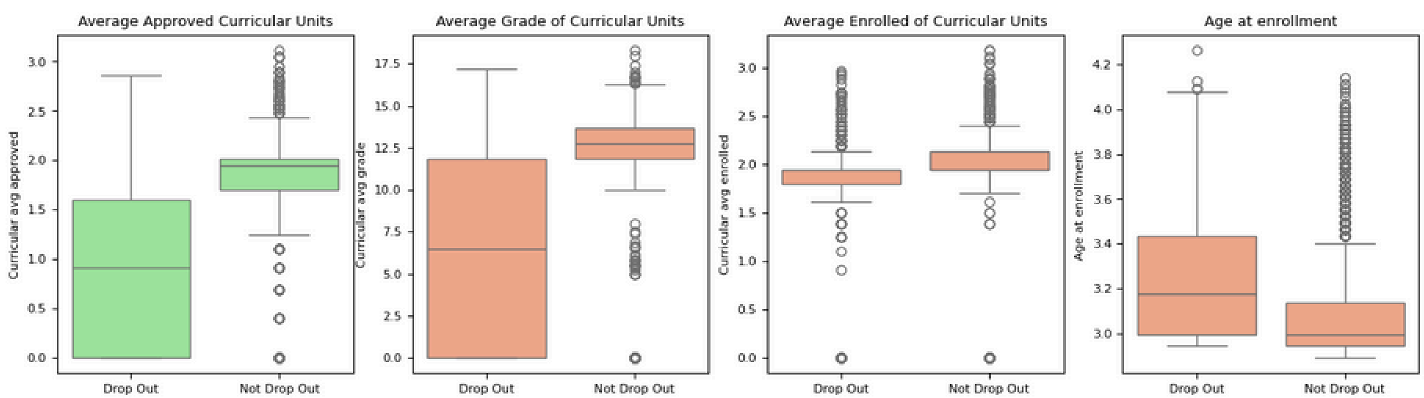
6. Preprocessing



Feature Engineering

6. Preprocessing

Outliers



it’s pretty obvious that no drop out student should have 0 credits, so we should drop them

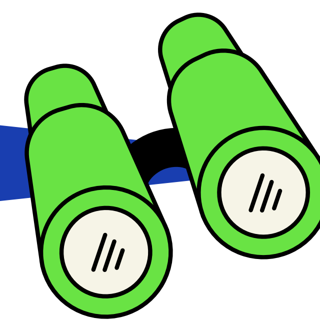
6. Preprocessing

Data Transformation and Scaling

From the EDA part, we could see that we have some skewed features, so we used log transformation to cure skewed features.

Then, we used standard scaling for our classification models, and robust scaling for clustering.

Now, let's dive into our models performance!



7. Modeling

7. Modeling

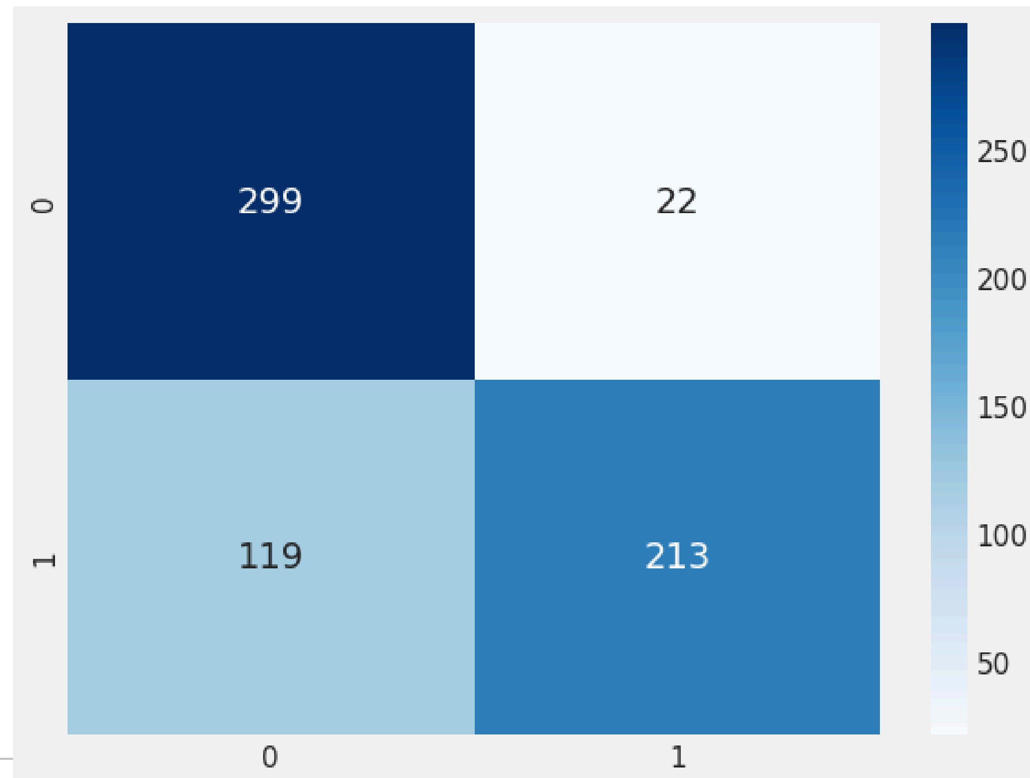
We will discuss these models and compare the results of each :

. Logistic Regression
. KNN
. SVM
. Random Forest
. Naive Bayes
. Decision Tree
. AdaBoost Classifier with GridSearch and without
. GradeintBoost Classifier with GridSearch and without
. XGBoost Classifier with GridSearch and without

7. Modeling

Naive Bayes

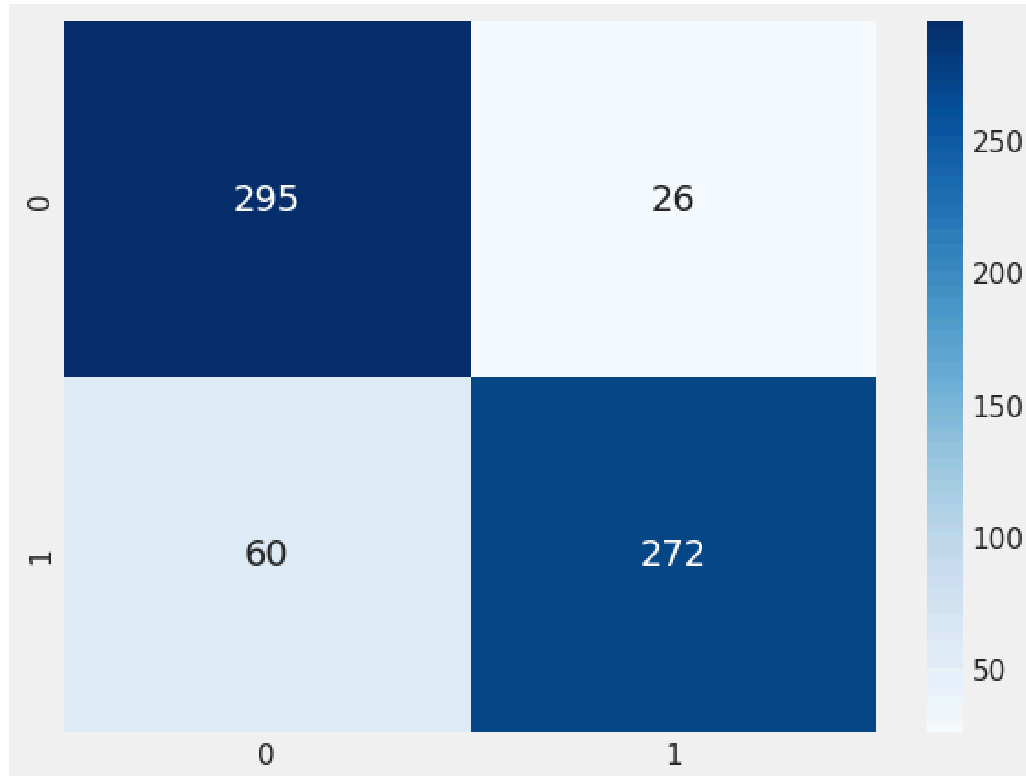
```
Train Accuracy: 0.776281208935611
Validation Accuracy: 0.7484662576687117
Test Accuracy: 0.7840735068912711
==================================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.70      0.94      0.81      1498
         1.0       0.92      0.62      0.74      1546

    accuracy                           0.78      3044
   macro avg       0.81      0.78      0.77      3044
weighted avg       0.81      0.78      0.77      3044

==================================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.67      0.93      0.78       315
         1.0       0.89      0.58      0.71       337

    accuracy                           0.75       652
   macro avg       0.78      0.75      0.74       652
weighted avg       0.79      0.75      0.74       652

==================================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.72      0.93      0.81       321
         1.0       0.91      0.64      0.75       332

    accuracy                           0.78       653
   macro avg       0.81      0.79      0.78       653
weighted avg       0.81      0.78      0.78       653
```



7. Modeling

Logistic Regression

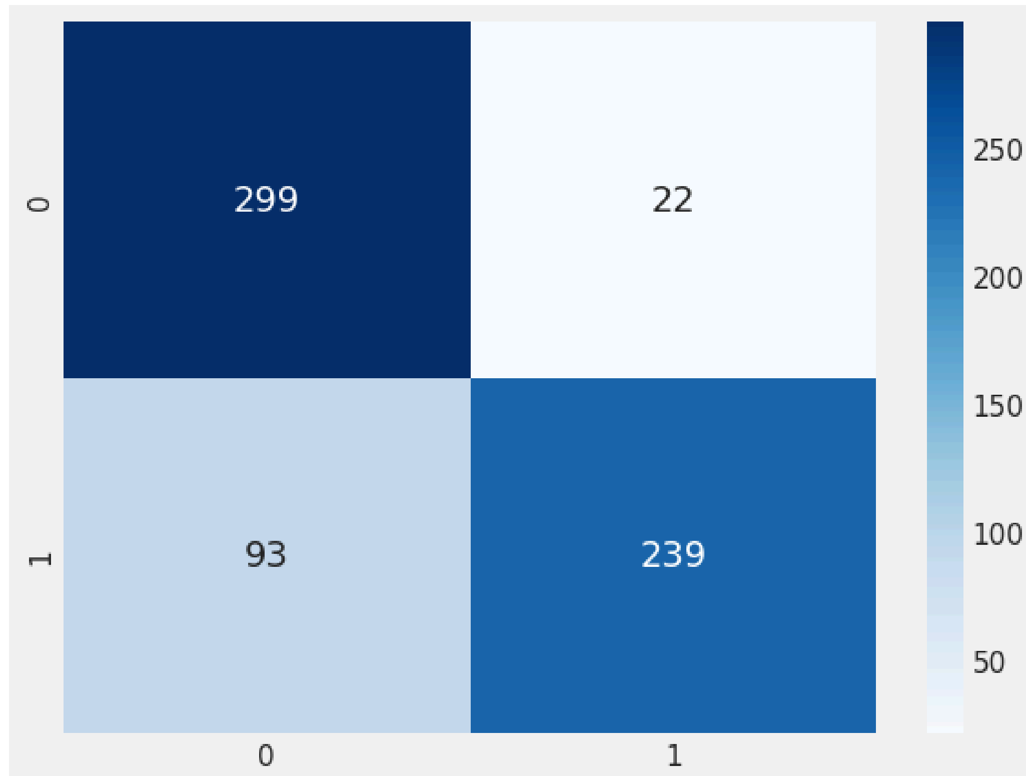
```
Train Accuracy: 0.8610381077529566
Validation Accuracy: 0.8496932515337423
Test Accuracy: 0.8683001531393568
==========================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.83      0.90      0.86      1498
         1.0       0.90      0.82      0.86      1546

    accuracy                           0.86      3044
   macro avg       0.86      0.86      0.86      3044
weighted avg       0.86      0.86      0.86      3044

==========================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.82      0.89      0.85       315
         1.0       0.89      0.81      0.85       337

    accuracy                           0.85       652
   macro avg       0.85      0.85      0.85       652
weighted avg       0.85      0.85      0.85       652

==========================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.83      0.92      0.87       321
         1.0       0.91      0.82      0.86       332

    accuracy                           0.87       653
   macro avg       0.87      0.87      0.87       653
weighted avg       0.87      0.87      0.87       653
```



7. Modeling

KNN

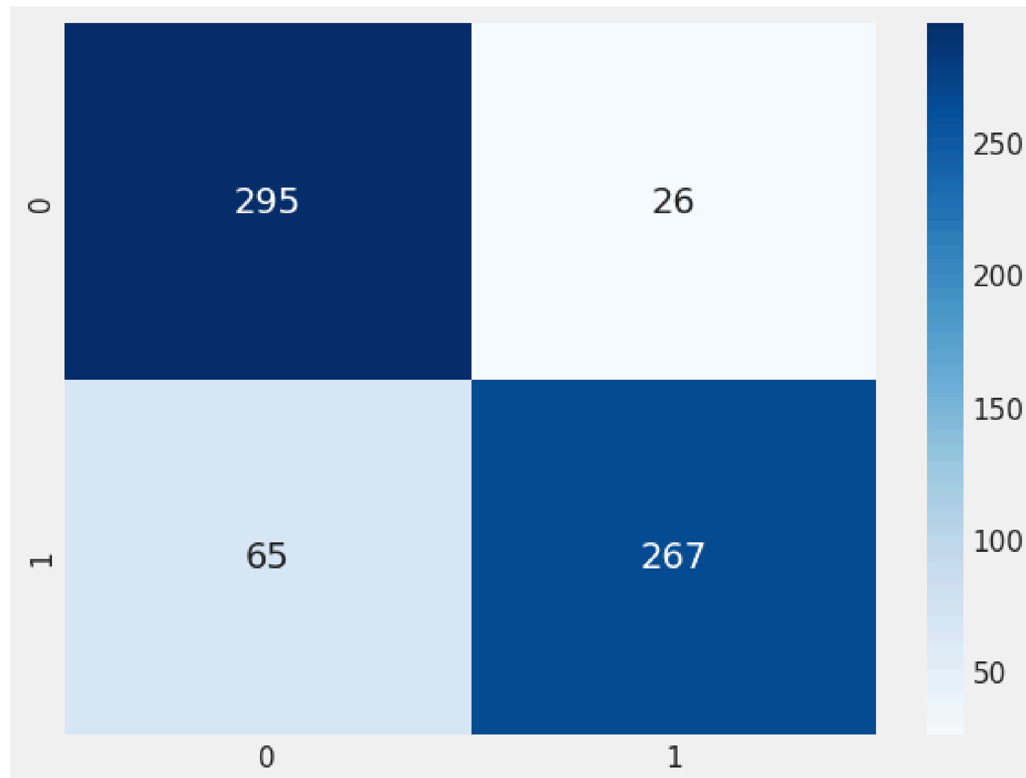
```
Train Accuracy: 0.8350854139290408
Validation Accuracy: 0.799079754601227
Test Accuracy: 0.8238897396630934
==================================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.78      0.93      0.85      1498
         1.0       0.92      0.74      0.82      1546

    accuracy                           0.84      3044
   macro avg       0.85      0.84      0.83      3044
weighted avg       0.85      0.84      0.83      3044

==================================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.74      0.91      0.81       315
         1.0       0.89      0.69      0.78       337

    accuracy                           0.80       652
   macro avg       0.81      0.80      0.80       652
weighted avg       0.82      0.80      0.80       652

==================================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.76      0.93      0.84       321
         1.0       0.92      0.72      0.81       332

    accuracy                           0.82       653
   macro avg       0.84      0.83      0.82       653
weighted avg       0.84      0.82      0.82       653
```



7. Modeling

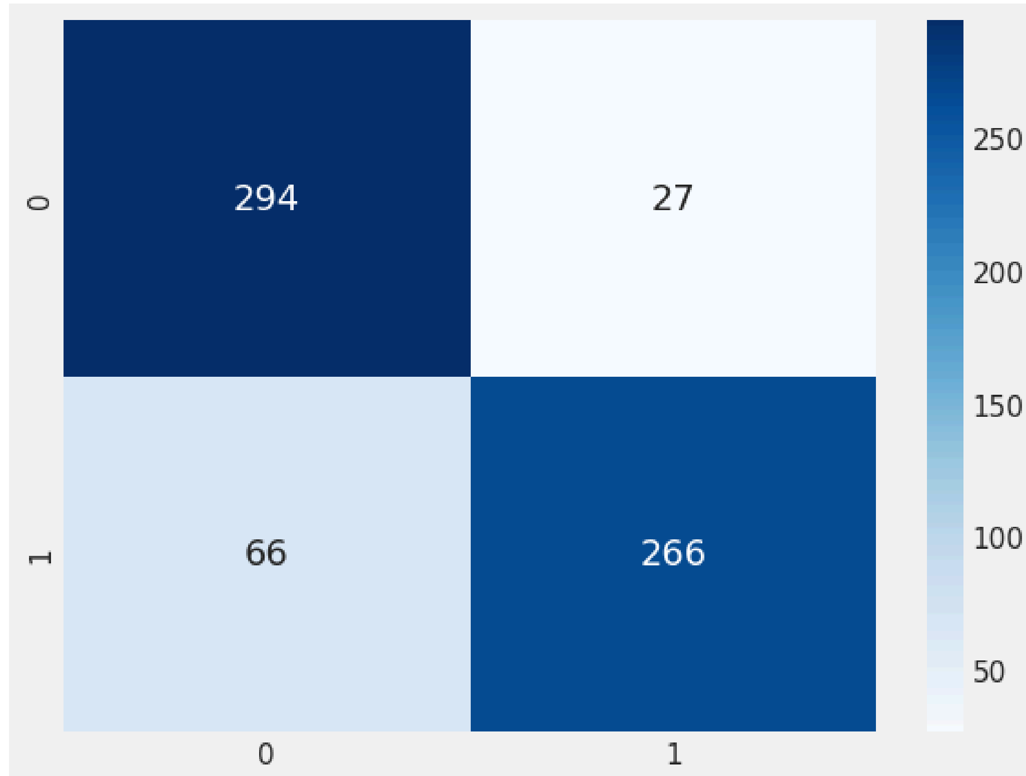
SVM

```
Train Accuracy: 0.864323258869908
Validation Accuracy: 0.8573619631901841
Test Accuracy: 0.8606431852986217
==========================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.83      0.91      0.87      1498
         1.0       0.91      0.82      0.86      1546

    accuracy                           0.86      3044
   macro avg       0.87      0.87      0.86      3044
weighted avg       0.87      0.86      0.86      3044

==========================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.82      0.90      0.86       315
         1.0       0.90      0.81      0.85       337

    accuracy                           0.86       652
   macro avg       0.86      0.86      0.86       652
weighted avg       0.86      0.86      0.86       652

==========================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.82      0.92      0.87       321
         1.0       0.91      0.80      0.85       332

    accuracy                           0.86       653
   macro avg       0.87      0.86      0.86       653
weighted avg       0.87      0.86      0.86       653
```



AdaBoost Classifier

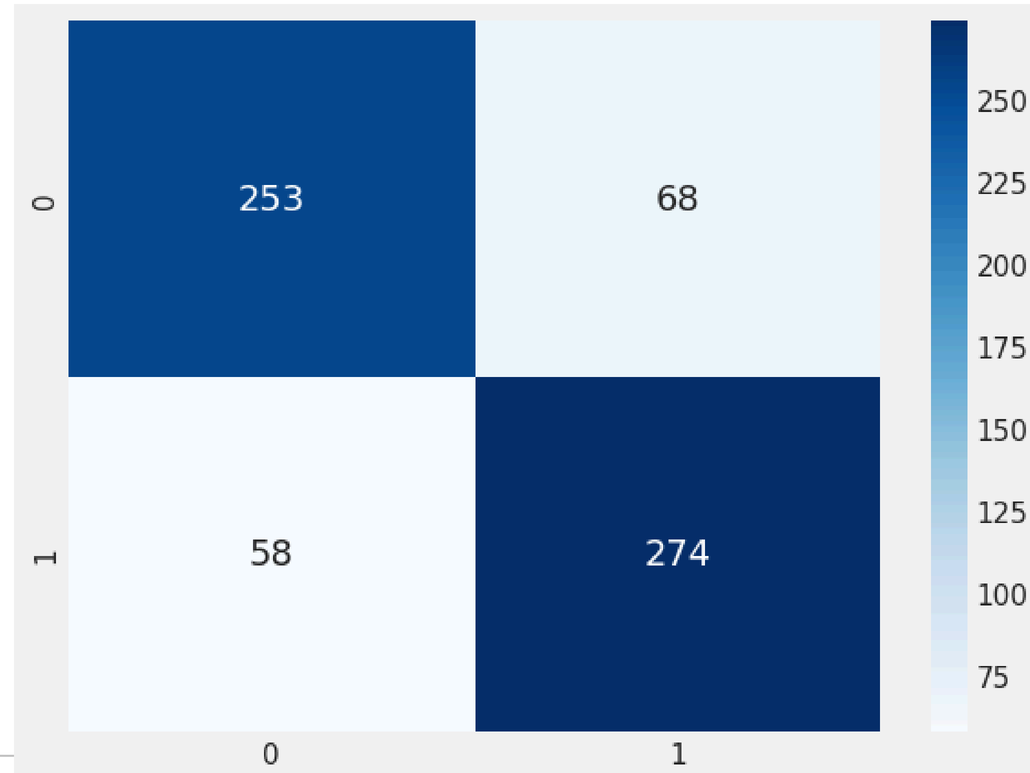
```
Train Accuracy: 0.8488830486202366
Validation Accuracy: 0.8466257668711656
Test Accuracy: 0.8575803981623277
==================================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.80      0.91      0.86      1498
         1.0       0.90      0.79      0.84      1546

    accuracy                           0.85      3044
   macro avg       0.85      0.85      0.85      3044
weighted avg       0.86      0.85      0.85      3044

==================================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.80      0.92      0.85       315
         1.0       0.91      0.78      0.84       337

    accuracy                           0.85       652
   macro avg       0.85      0.85      0.85       652
weighted avg       0.86      0.85      0.85       652

==================================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.82      0.92      0.86       321
         1.0       0.91      0.80      0.85       332

    accuracy                           0.86       653
   macro avg       0.86      0.86      0.86       653
weighted avg       0.86      0.86      0.86       653
```



7. Modeling

Decision Tree

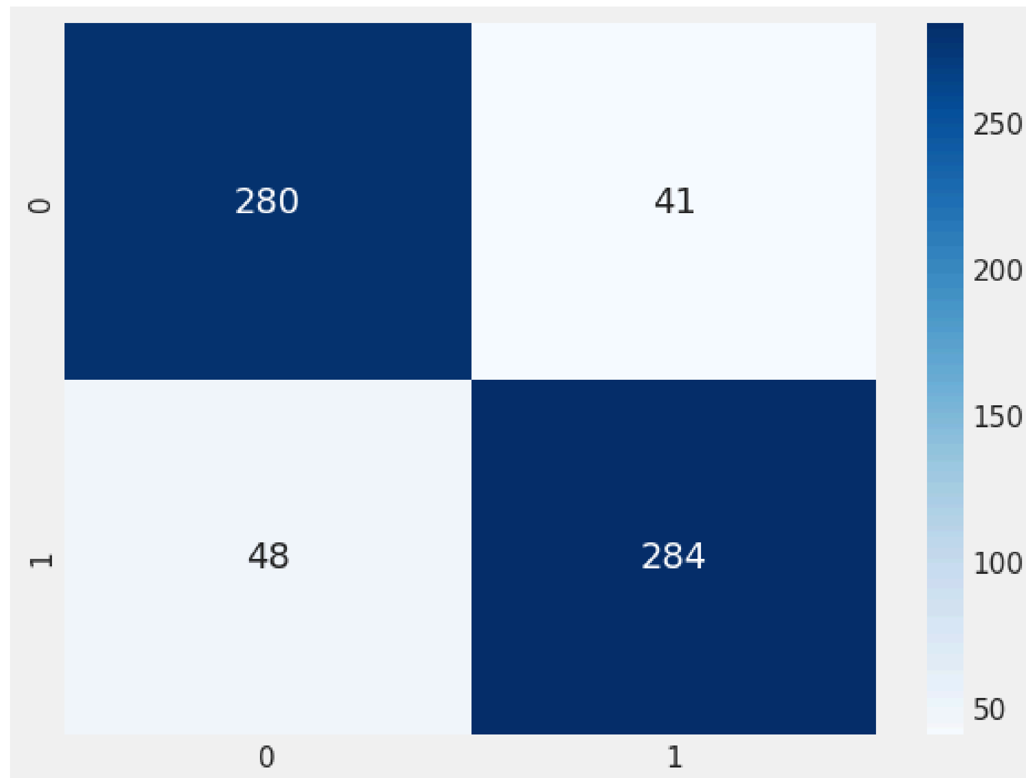
```
Train Accuracy: 1.0
Validation Accuracy: 0.7837423312883436
Test Accuracy: 0.8070444104134763
==================================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1498
         1.0       1.00      1.00      1.00      1546

    accuracy                           1.00      3044
   macro avg       1.00      1.00      1.00      3044
weighted avg       1.00      1.00      1.00      3044

==================================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.77      0.79      0.78       315
         1.0       0.80      0.78      0.79       337

    accuracy                           0.78       652
   macro avg       0.78      0.78      0.78       652
weighted avg       0.78      0.78      0.78       652

==================================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.81      0.79      0.80       321
         1.0       0.80      0.83      0.81       332

    accuracy                           0.81       653
   macro avg       0.81      0.81      0.81       653
weighted avg       0.81      0.81      0.81       653
```



7. Modeling

GradientBoost Classifier

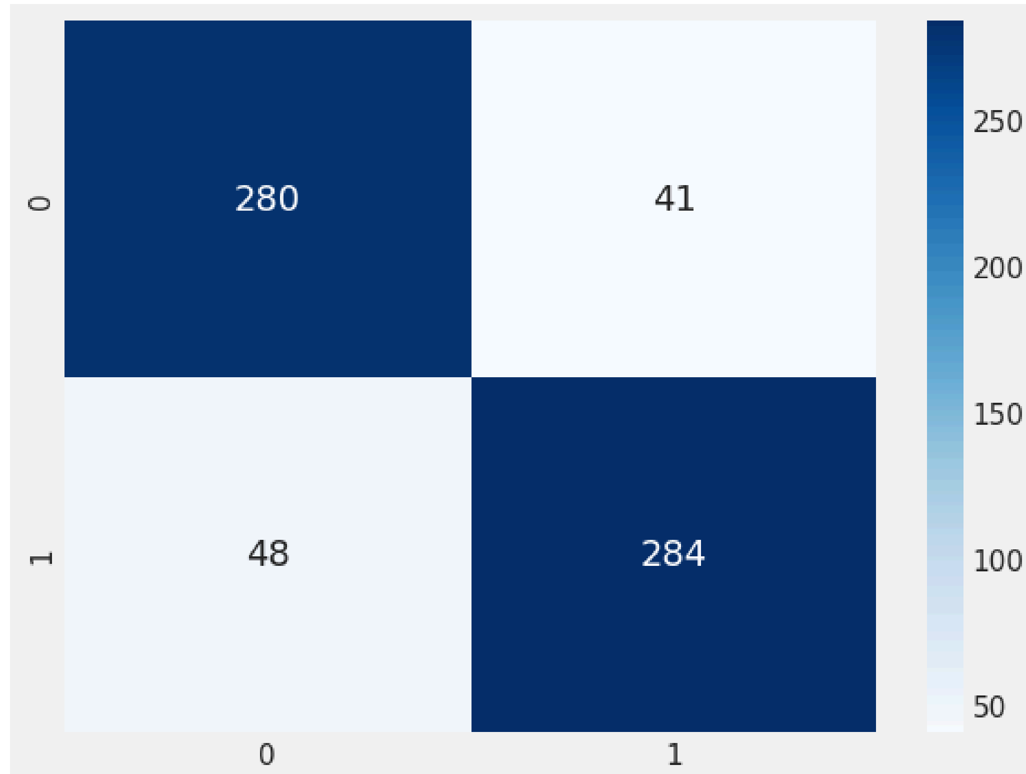
```
Train Accuracy: 0.8837056504599211
Validation Accuracy: 0.843558282208589
Test Accuracy: 0.8637059724349158
==============================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.85      0.92      0.89      1498
         1.0       0.92      0.85      0.88      1546

    accuracy                           0.88      3044
   macro avg       0.89      0.88      0.88      3044
weighted avg       0.89      0.88      0.88      3044

==============================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.81      0.89      0.85       315
         1.0       0.89      0.80      0.84       337

    accuracy                           0.84       652
   macro avg       0.85      0.85      0.84       652
weighted avg       0.85      0.84      0.84       652

==============================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.85      0.87      0.86       321
         1.0       0.87      0.86      0.86       332

    accuracy                           0.86       653
   macro avg       0.86      0.86      0.86       653
weighted avg       0.86      0.86      0.86       653
```



7. Modeling

AdaBoost Classifier with GridSearch

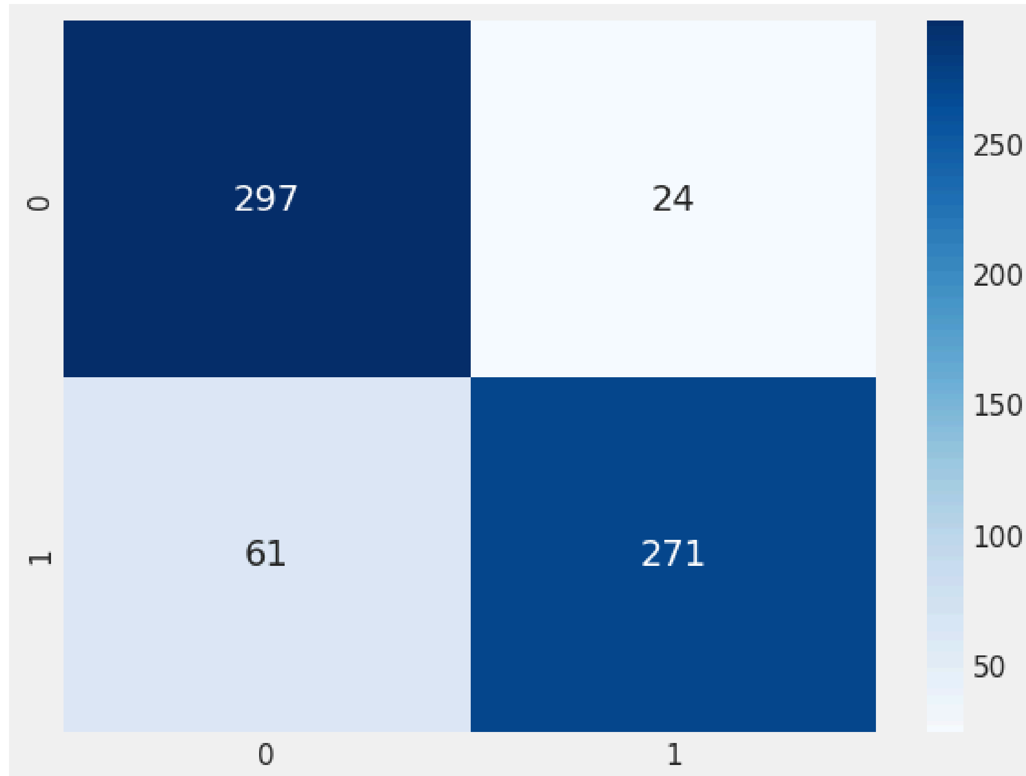
```
Best AdaBoost Classifier parameters: {'algorithm': 'SAMME.R', 'n_estimators': 200}
Train Accuracy: 0.8837056504599211
Validation Accuracy: 0.843558282208589
Test Accuracy: 0.8637059724349158
==========================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.85      0.92      0.89      1498
         1.0       0.92      0.85      0.88      1546

    accuracy                           0.88      3044
   macro avg       0.89      0.88      0.88      3044
weighted avg       0.89      0.88      0.88      3044

==========================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.81      0.89      0.85       315
         1.0       0.89      0.80      0.84       337

    accuracy                           0.84       652
   macro avg       0.85      0.85      0.84       652
weighted avg       0.85      0.84      0.84       652

==========================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.85      0.87      0.86       321
         1.0       0.87      0.86      0.86       332

    accuracy                           0.86       653
   macro avg       0.86      0.86      0.86       653
weighted avg       0.86      0.86      0.86       653
```



7. Modeling

XGBoost Classifier

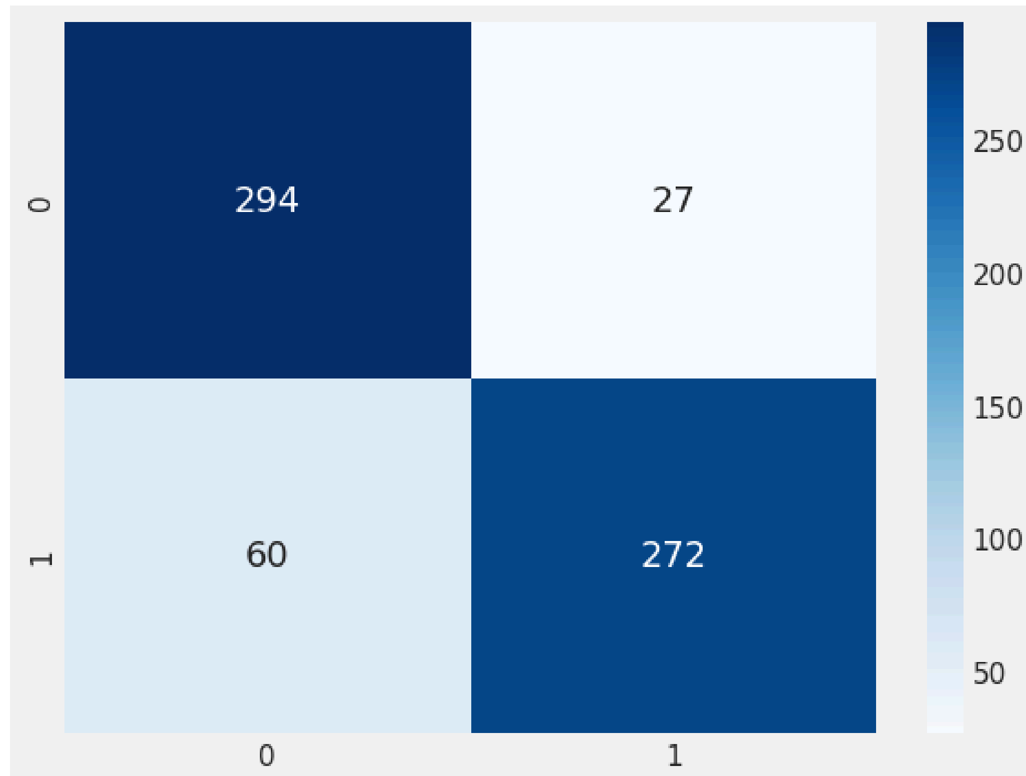
```
Train Accuracy: 0.8912614980289093
Validation Accuracy: 0.8374233128834356
Test Accuracy: 0.8698315467075038
==============================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.85      0.94      0.90      1498
         1.0       0.94      0.84      0.89      1546

    accuracy                           0.89      3044
   macro avg       0.90      0.89      0.89      3044
weighted avg       0.90      0.89      0.89      3044

==============================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.85      0.80      0.83       315
         1.0       0.83      0.87      0.85       337

    accuracy                           0.84       652
   macro avg       0.84      0.84      0.84       652
weighted avg       0.84      0.84      0.84       652

==============================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.83      0.93      0.87       321
         1.0       0.92      0.82      0.86       332

    accuracy                           0.87       653
   macro avg       0.87      0.87      0.87       653
weighted avg       0.87      0.87      0.87       653
```



7. Modeling

XGBoostClassifier with GridSearch

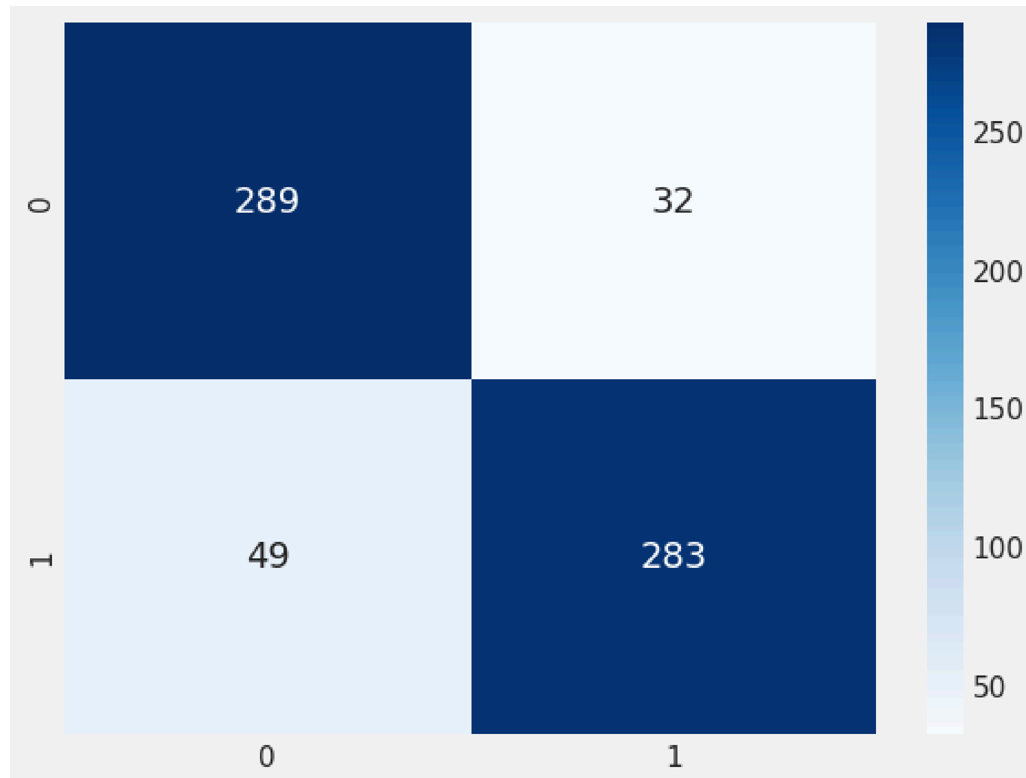
```
Best XGBoost parameters: {'gamma': 0.1, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}
Train Accuracy: 0.8915900131406045
Validation Accuracy: 0.8358895705521472
Test Accuracy: 0.8667687595712098
==========================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.85      0.94      0.90      1498
         1.0       0.94      0.84      0.89      1546

    accuracy                           0.89      3044
   macro avg       0.90      0.89      0.89      3044
weighted avg       0.90      0.89      0.89      3044

==========================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.85      0.80      0.82       315
         1.0       0.82      0.87      0.85       337

    accuracy                           0.84       652
   macro avg       0.84      0.83      0.84       652
weighted avg       0.84      0.84      0.84       652

==========================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.83      0.92      0.87       321
         1.0       0.91      0.82      0.86       332

    accuracy                           0.87       653
   macro avg       0.87      0.87      0.87       653
weighted avg       0.87      0.87      0.87       653
```



7. Modeling

GradientBoost Classifier with GridSearch

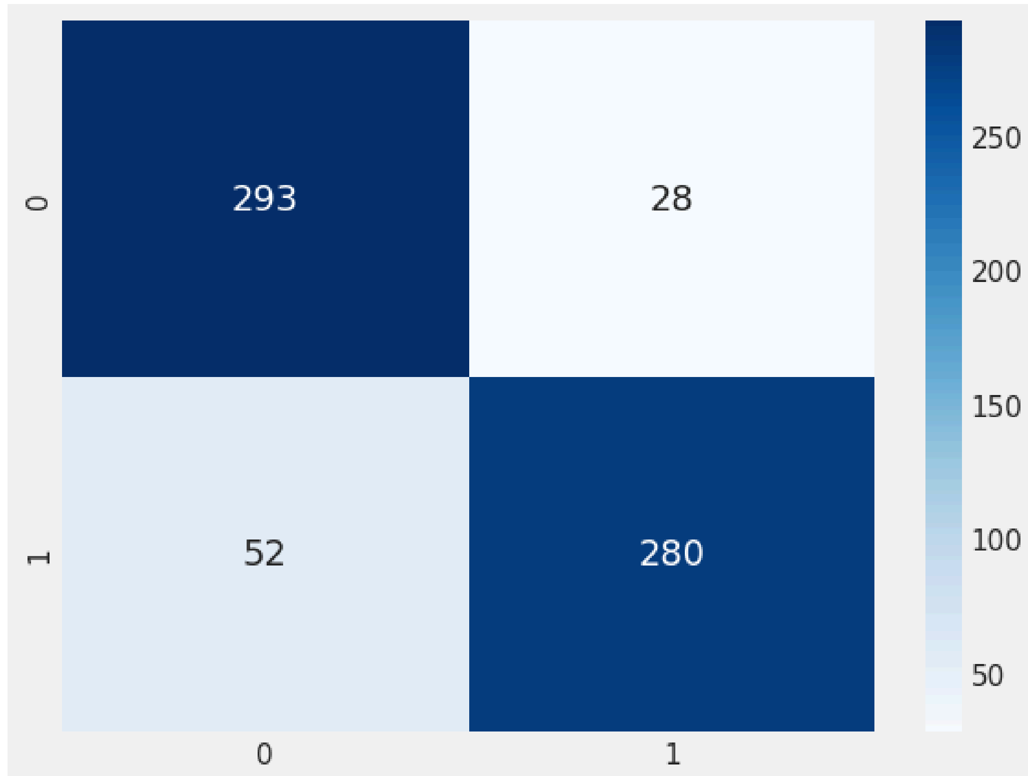
```
Best Gradient Boosting parameters: {'max_depth': 3, 'n_estimators': 200}
Train Accuracy: 0.9250985545335085
Validation Accuracy: 0.838957055214724
Test Accuracy: 0.8759571209800919
==========================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       0.90      0.96      0.93      1498
         1.0       0.96      0.89      0.92      1546

    accuracy                           0.93      3044
   macro avg       0.93      0.93      0.93      3044
weighted avg       0.93      0.93      0.93      3044

==========================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.80      0.89      0.84       315
         1.0       0.89      0.79      0.84       337

    accuracy                           0.84       652
   macro avg       0.84      0.84      0.84       652
weighted avg       0.84      0.84      0.84       652

==========================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.86      0.90      0.88       321
         1.0       0.90      0.85      0.87       332

    accuracy                           0.88       653
   macro avg       0.88      0.88      0.88       653
weighted avg       0.88      0.88      0.88       653
```



7. Modeling

Random Forest

```
Train Accuracy: 1.0
Validation Accuracy: 0.8542944785276073
Test Accuracy: 0.877488514548239
==============================================================================================
Classification Report for Train Set:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1498
         1.0       1.00      1.00      1.00      1546

    accuracy                           1.00      3044
   macro avg       1.00      1.00      1.00      3044
weighted avg       1.00      1.00      1.00      3044

==============================================================================================
Classification Report for Validation Set:
              precision    recall  f1-score   support

         0.0       0.81      0.92      0.86       315
         1.0       0.91      0.80      0.85       337

    accuracy                           0.85       652
   macro avg       0.86      0.86      0.85       652
weighted avg       0.86      0.85      0.85       652

==============================================================================================
Classification Report for Test Set:
              precision    recall  f1-score   support

         0.0       0.85      0.91      0.88       321
         1.0       0.91      0.84      0.88       332

    accuracy                           0.88       653
   macro avg       0.88      0.88      0.88       653
weighted avg       0.88      0.88      0.88       653
```

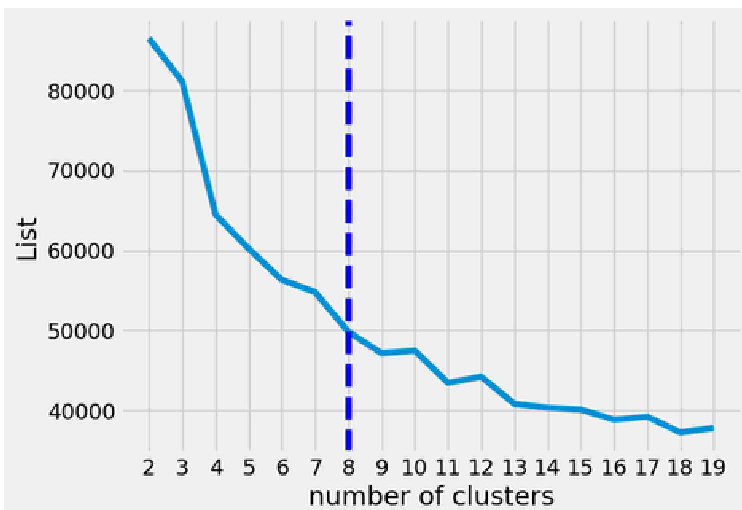


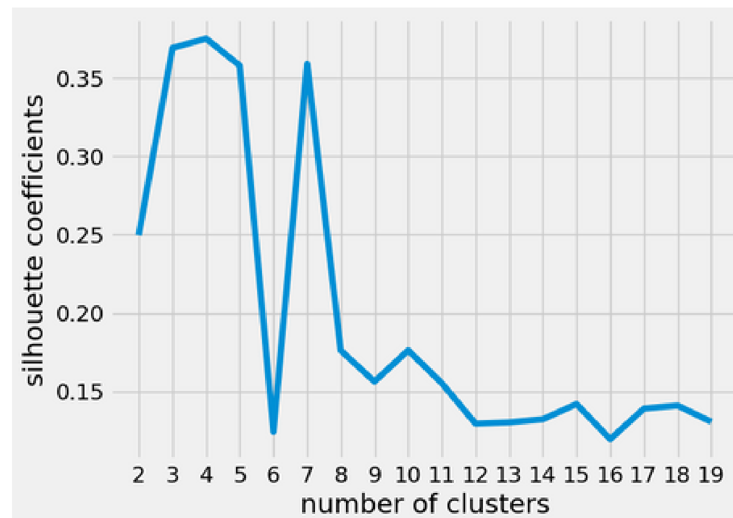
7. Modeling

Clustering

We preferred to work with robust scaled data in this part.

1.KMeans



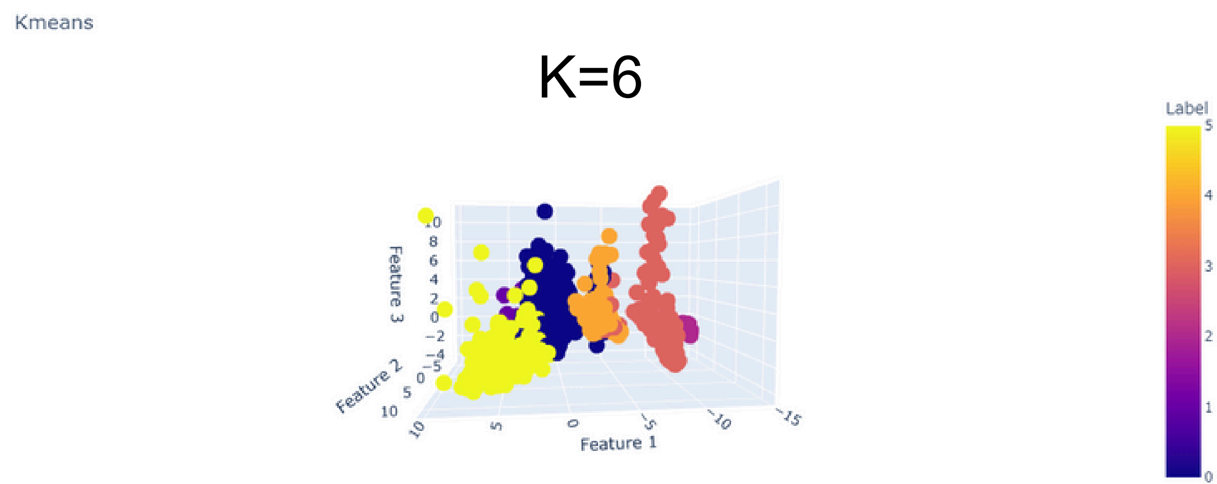


7. Modeling

Clustering

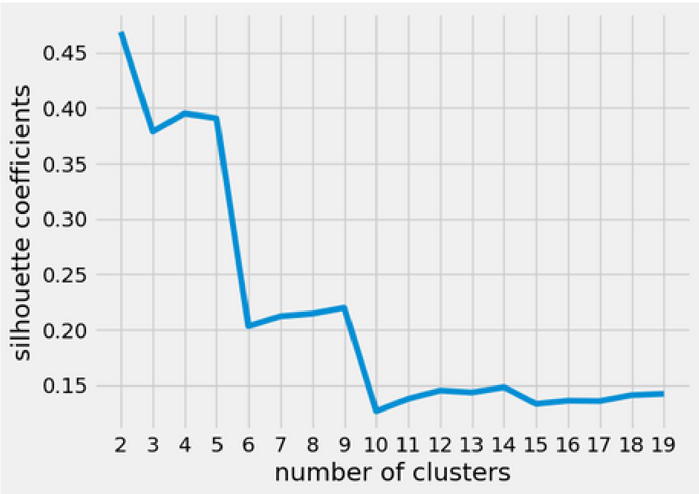
1.KMeans

K=6

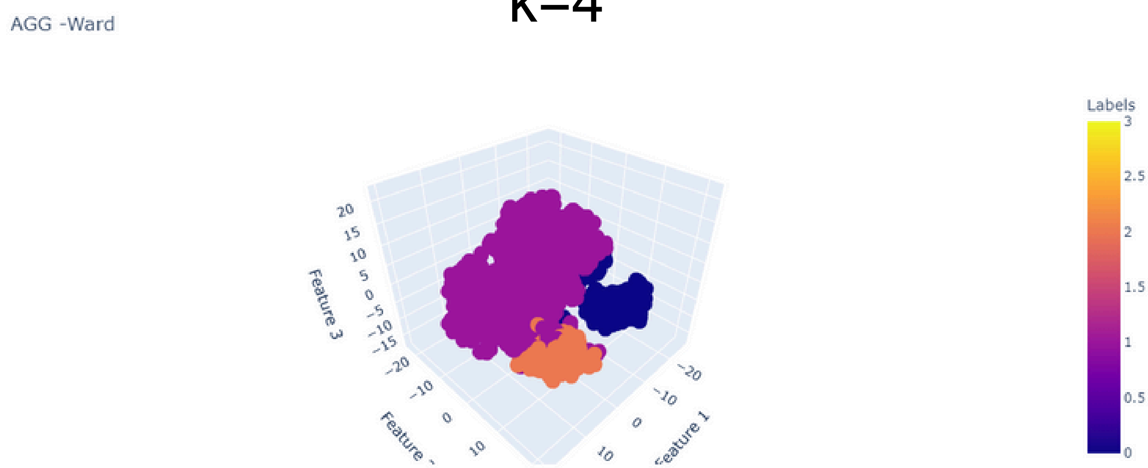


7. Modeling

2.Agglomerative Clustering Ward

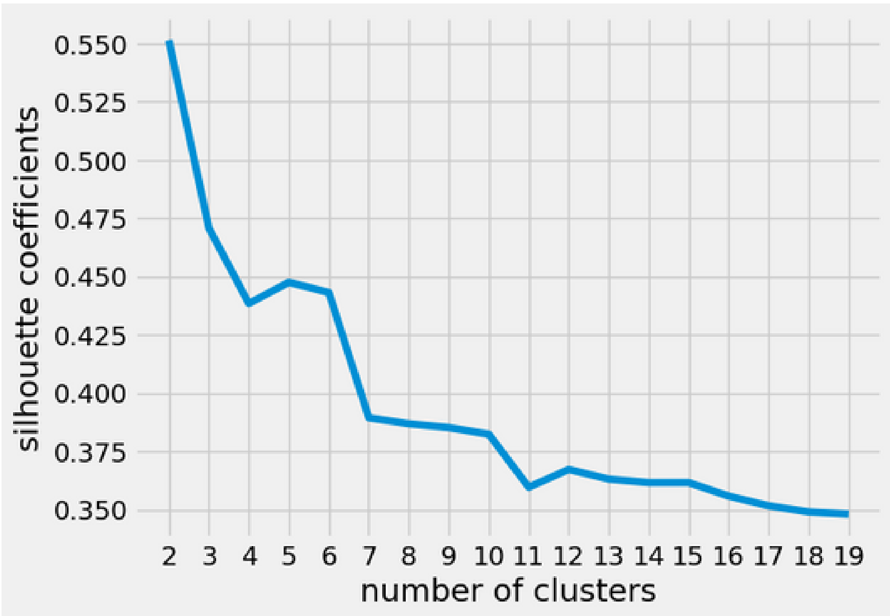


k=4

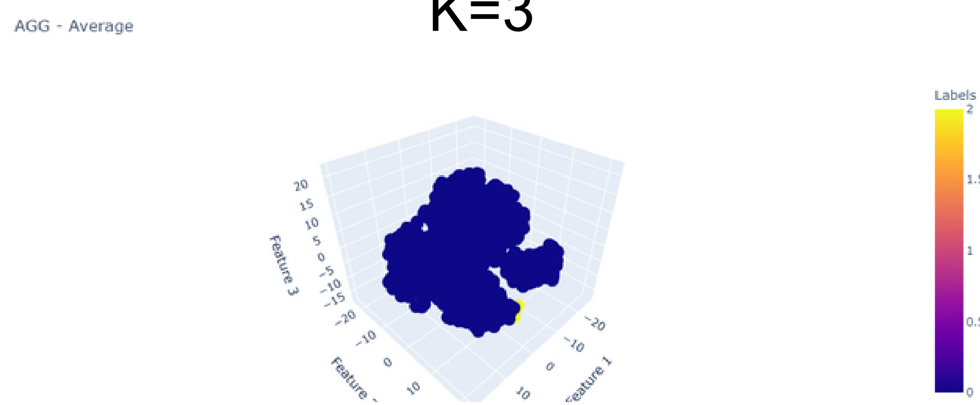


7. Modeling

2.Agglomerative Clustering
Average

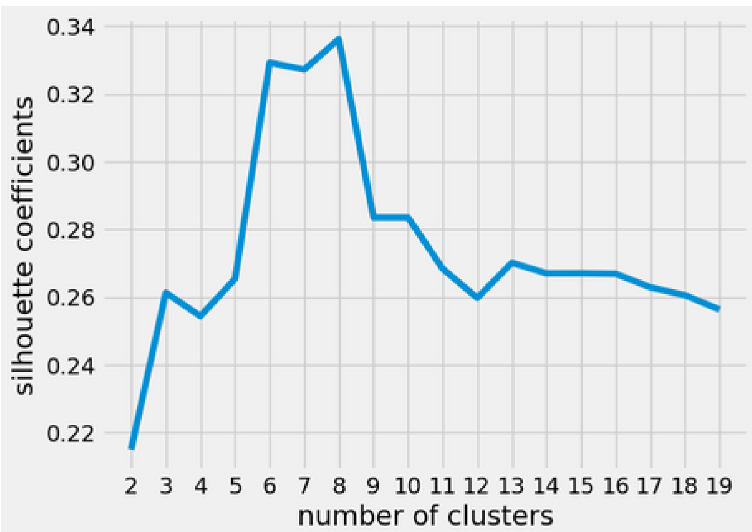


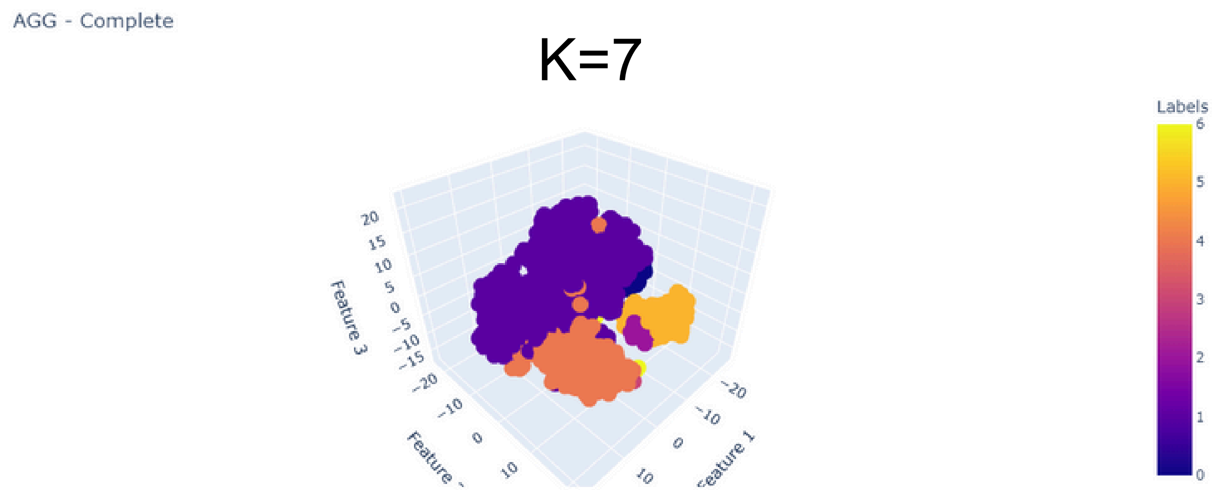
K=3



7. Modeling

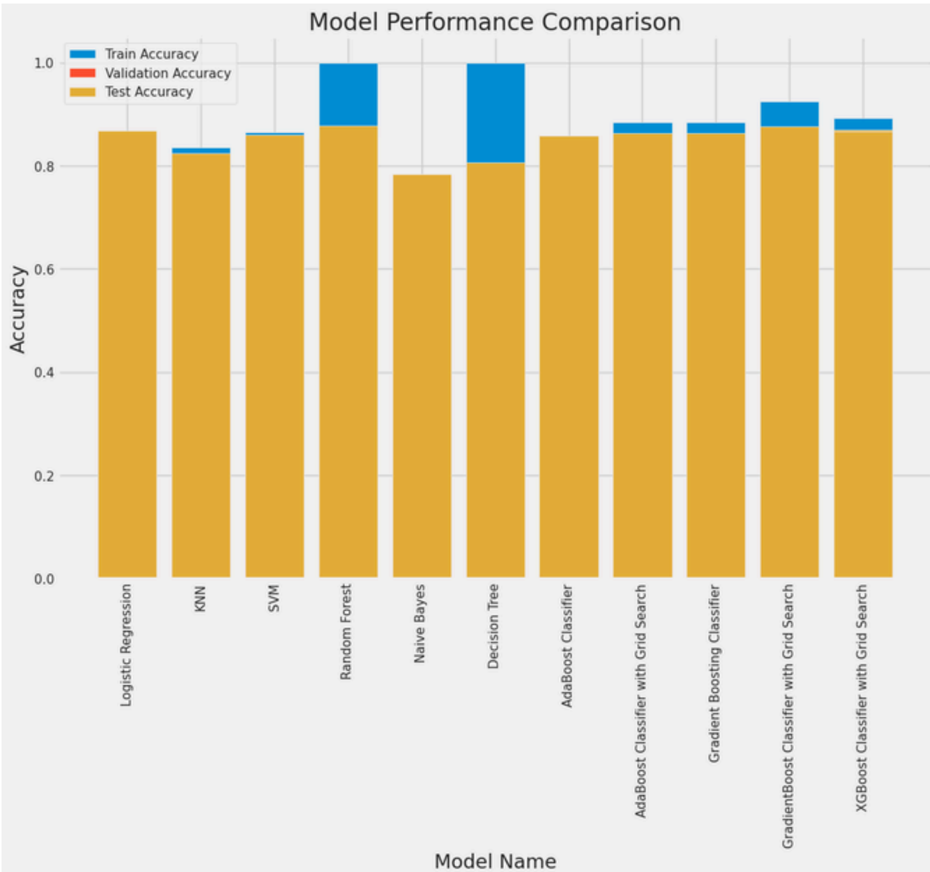
2.Agglomerative Clustering Complete





8. Final results

8. Final Results



Best Model is :
Random Forest
Classifier

9. Conclusion

9. Conclusion

Early Detection of Student Dropout Trends

- The analysis of the student performance dataset aims to help universities and schools **identify students at risk** of academic dropout. By predicting student trends early, educational institutions can intervene with personalized solutions to improve academic outcomes.
-
- In this project, several machine learning models were explored, with **Random Forest** being the most effective model. Despite a slight overfitting issue, the model achieved an accuracy of **87.75%**, making it a reliable tool for predicting students likely to drop out.

9. Conclusion

Early Detection of Student Dropout Trends

Key takeaways include:

- Early intervention is crucial for helping at-risk students.
- Data driven insights enable educational institutions to develop tailored support programs.
- Random Forest can serve as the primary model for student performance prediction with continuous tuning to mitigate overfitting.

By leveraging such models, schools and universities can improve retention rates and provide timely support to students who may need additional academic assistance.

10.Future imporvments

10.Future improvements

We could enhance our deployment by making the student only enter their ID

The clustering part could be much better with some optimization.

Any Questions ?

Team work

Name	Contribution
Salma Sherif	StoryTelling, introduction, Second part of Modeling
Ashraf Saber	EDA, Preprocessing, Conclusion
Mahmoud Wahban	First part of Modeling, Clustering

SAMSUNG

Together for Tomorrow! Enabling People
Education for Future Generations

Thank you

Together for Tomorrow!
Enabling People

Education for Future Generations

ⓒ2020 SAMSUNG. All rights reserved.
Samsung Electronics Corporate Citizenship Office holds the copyright of book.
This book is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.
To use this book other than the curriculum of Samsung innovation Campus or to use the entire or part of this book, you must receive written consent from copyright holder.