GROUP 19

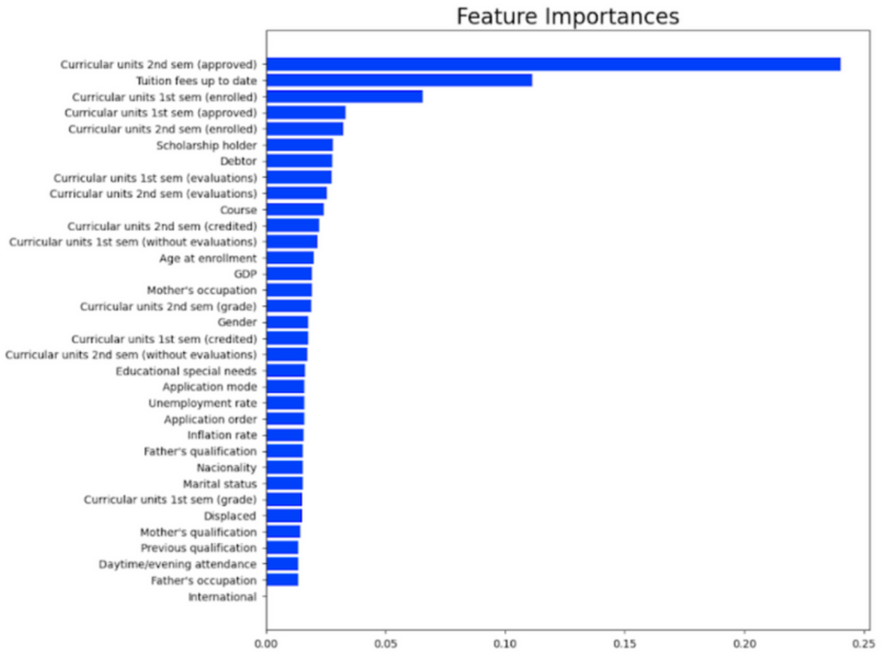
FINAL PRESENTATION - ACADEMIC SUCCESS PREDICTOR

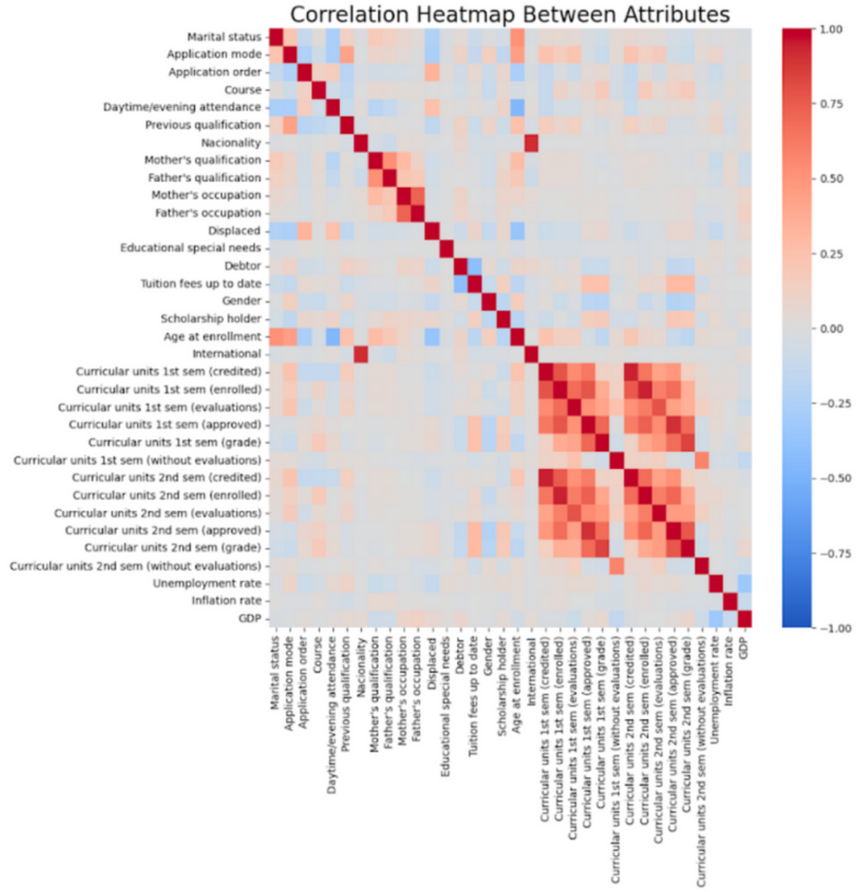
https://github.com/ranapp/ECS171Project

INTRODUCTION

College education is an important part of many people's lives. To many people it is one of the biggest opportunities that can jumpstart someone's career and livelihood. However, everyone that goes to college comes from different backgrounds and live in different ways while at college. Our machine learning model trains data to predict the graduation results of students from various backgrounds.

EXPLORATORY DATA ANALYSIS





Heatmap and feature importance

DATA PREPROCESSING

- One hot encoding for categorical attributes
- Normalization for numerical attributes
- Dropped 4 columns based on Feature Importance
- Dropped confusing results -- "Enrolled"

MODELS

We tried 6 models for our project:

- Polynomial Regression with Sigmoid Function
- Logistic Regression with K-fold Cross Validation
- Decision Tree
- Neural Network
- Naive Bayes
- Support Vector Machine (SVM)

The best two performing models are logistic regression with K-fold cross-validation and support vector machines (SVM) with linear kernel. We will focus on presenting these two.

SPLITTING THE DATA

We split the datasets into training and test sets with an 80:20 ratio, ensuring consistency across all models. By setting the random seed to 42, we guarantee that the same datasets are used for training and testing each model. **This approach facilitates accurate and equitable comparisons based on model accuracy!!!**

EXPERIMENT RESULTS

Because this is a fundamental classification problem, we can use several traditional models that are commonly used for such scenarios. The best-performing models is logistic regression with K-fold cross-validation. It have about 91% accuracy!!!

```
Accuracy scores: [0.9146005509641874, 0.9104683195592287, 0.9077134986225895, 0.9008264462809917, 0.9104683195592287]
Average accuracy: 0.9088154269972453

Confusion Matrix:
[[251  52]
 [ 13 410]]
```

Testing result of logistic regression with k-fold cross validation

EXPERIMENT RESULTS

We constructed multiple SVM models with two kernel options. Linear kernel achieved the highest accuracy of 90% in SVM. This is the second-highest accuracy rate based on a comparison with other models. We think SVM is well suited to solve our data.

	precision	recall	f1-score	support
0	0.93	0.82	0.87	277
1	0.89	0.96	0.93	449
accuracy			0.90	726
macro avg	0.91	0.89	0.90	726
weighted avg	0.91	0.90	0.90	726

Confusion Matrix:
[[226 51]
[18 431]]

Testing result of SVM with linear kernel

EXPERIMENT RESULTS

The average accuracy of the other 4 models is about 85%. They are relatively low by comparing to the first two models. They might not be ideal models for our dataset.

```
Case 1: logistic 100 batch size
Accuracy :  0.8512396694214877
Mean Square Error :  0.1487603305785124
[[1 0]
 [1 0]
 [0 1]
 [0 1]
 [1 0]]
Confusion Matrix for each label :
[[[188  27]
  [ 27 121]]

 [[121  27]
  [ 27 188]]]
Classification Report :
              precision    recall  f1-score   support

           0       0.82      0.82      0.82       148
           1       0.87      0.87      0.87       215

   micro avg       0.85      0.85      0.85       363
   macro avg       0.85      0.85      0.85       363
weighted avg       0.85      0.85      0.85       363
 samples avg       0.85      0.85      0.85       363
```

Testing result of neural network

```
CLASSIFICATION REPORT:
              precision    recall  f1-score   support

         0.0       0.69      0.64      0.66       277
         1.0       0.79      0.82      0.80       449

    accuracy                           0.75       726
   macro avg       0.74      0.73      0.73       726
weighted avg       0.75      0.75      0.75       726


CLASSIFICATION REPORT:
              precision    recall  f1-score   support

         0.0       0.82      0.76      0.79       277
         1.0       0.86      0.90      0.88       449

    accuracy                           0.84       726
   macro avg       0.84      0.83      0.83       726
weighted avg       0.84      0.84      0.84       726
```

Testing result of naive bayes

```
              precision    recall  f1-score   support

           0       0.96      0.75      0.84       296
           1       0.85      0.98      0.91       430

    accuracy                           0.89       726
   macro avg       0.90      0.87      0.88       726
weighted avg       0.89      0.89      0.88       726

Confusion Matrix:
[[223  73]
 [ 10 420]]
```

Testing result of polynomial regression with sigmoid

```
              precision    recall  f1-score   support

           0       0.78      0.83      0.81       277
           1       0.89      0.86      0.87       449

    accuracy                           0.85       726
   macro avg       0.84      0.84      0.84       726
weighted avg       0.85      0.85      0.85       726

Confusion Matrix:
[[230  47]
 [ 64 385]]
```

Testing result of decision tree

THANK'S FOR WATCHING