Predictive Analytics for Financial Services

Team: Nu

Koushik Srivastav Lala
Sahithi Bathini
Vaishnavi Jangili

CS5831 Project Presentation
Michigan Technological University
04/17/2023

INTRODUCTION

- The recent expansion of the credit industry has made credit scoring a very important problem, so the bank's credit section deals with a lot of credit data.

- The motto of our project is to predict if the customer is a good or bad customer based on the risk involved.

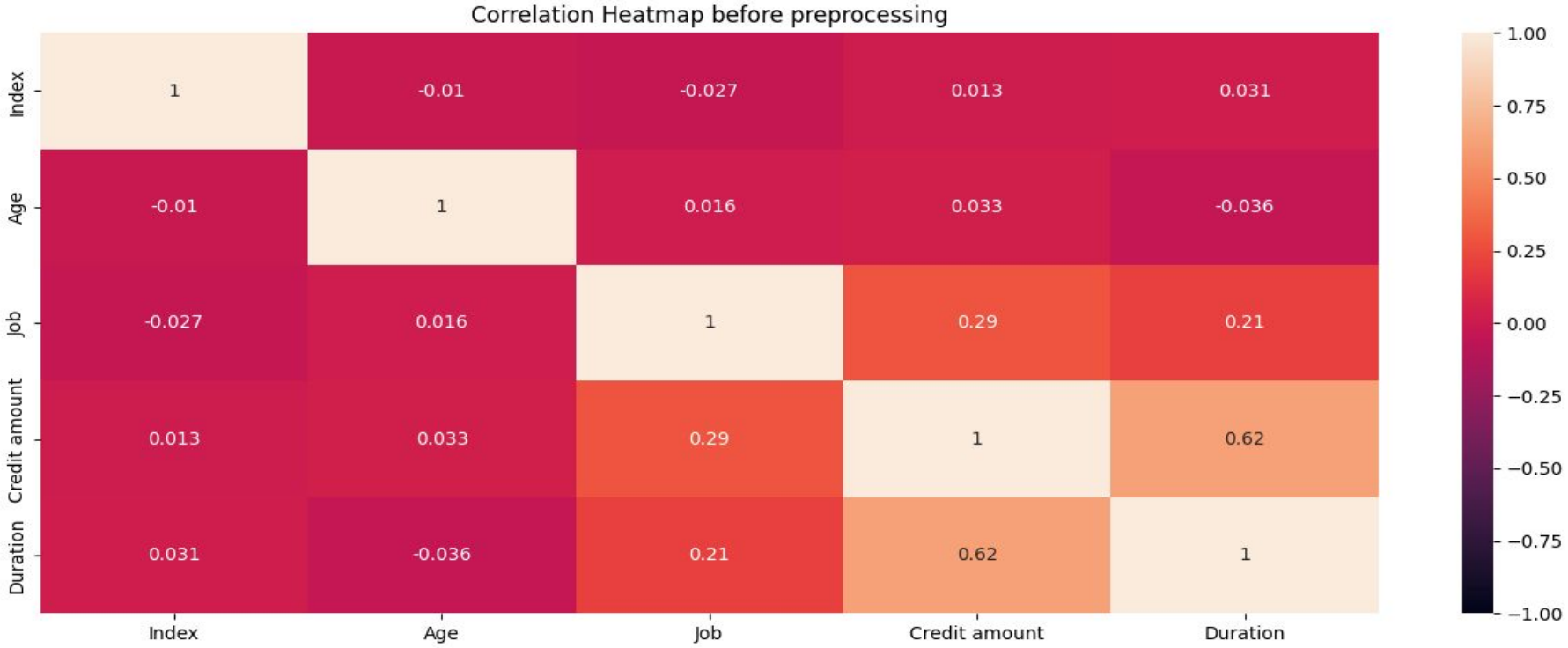
DATA SET

Summary Of the Dataset

- 1000 Records (Kaggle Dataset)
- 11 Features
- One redundant column (index)
- Dataset can be found at: https://www.kaggle.com/datasets/uciml/german-credit

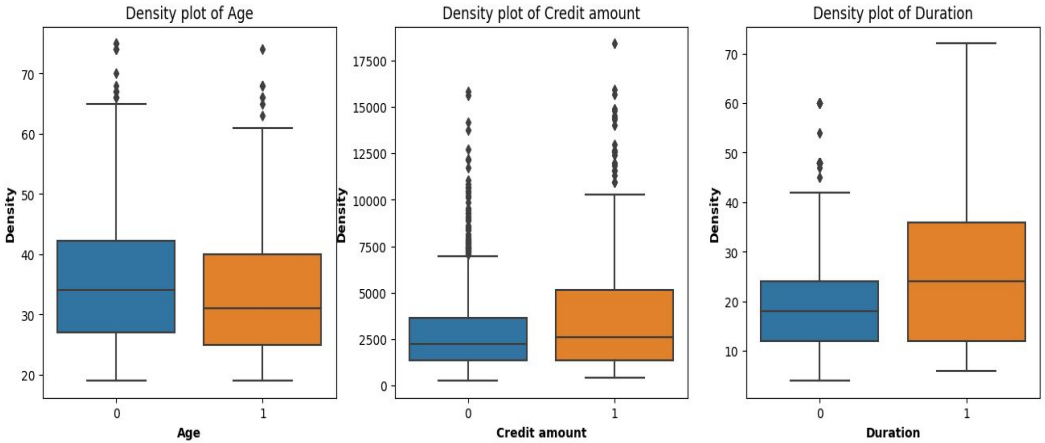
Features	Data Type
Index	Continuous
Age	Continuous
Sex	Categorical
Job	Categorical
Housing	Categorical
Savings accounts	Categorical
Checking account	Categorical
Credit Amount	Continuous
Duration	Continuous
Purpose	Categorical
Risk	**Categorical**

CORRELATION ANALYSIS



DATA PREPROCESSING

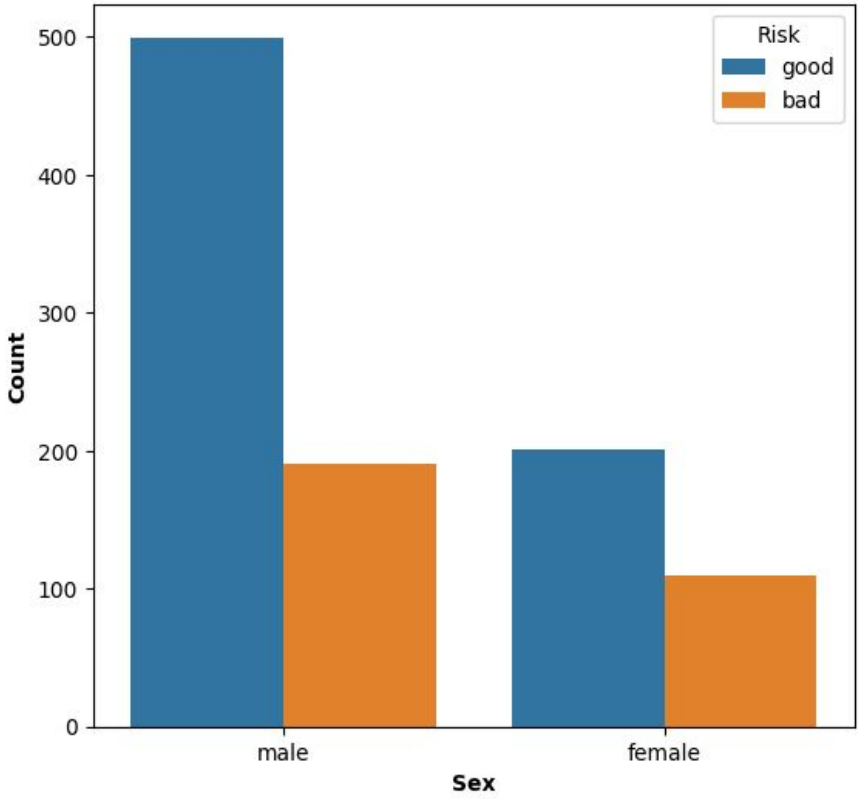
- Removing the redundant attributes
- Removal of Null values (savings account and checking account)
- One-hot-encoding of categorical variables.

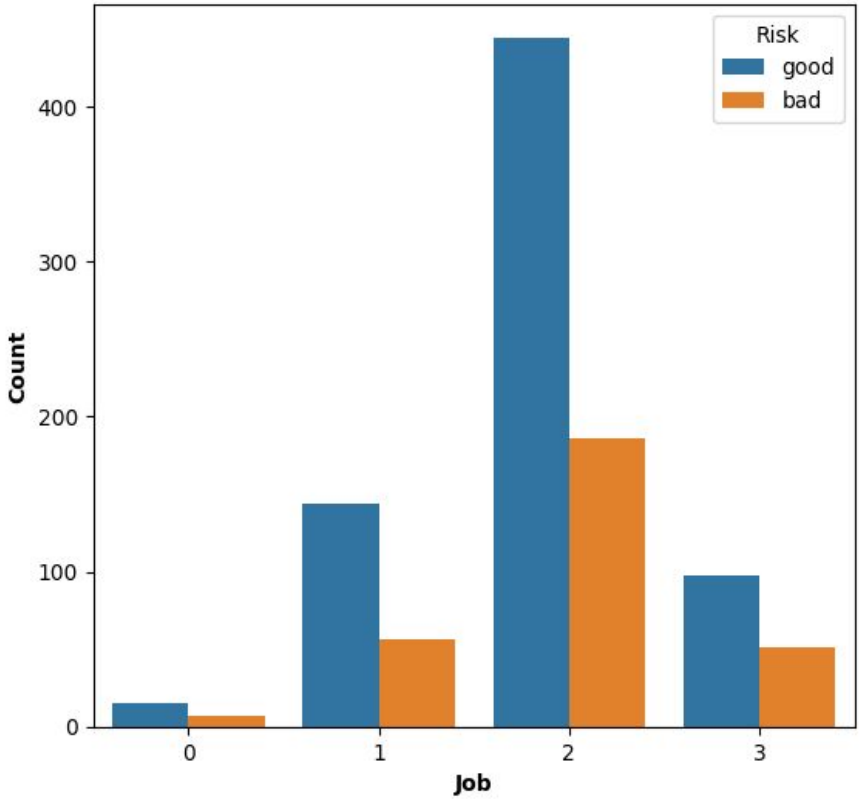



```
credit.isna().sum()
```

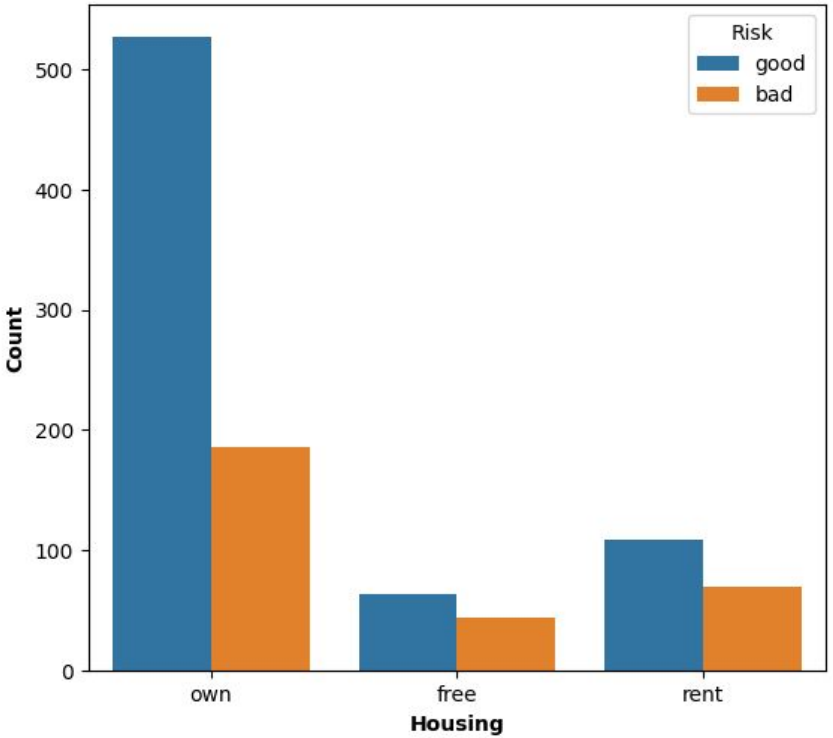
```
Index                0
Age                  0
Sex                  0
Job                  0
Housing              0
Saving accounts    183
Checking account   394
Credit amount        0
Duration             0
Purpose              0
Risk                 0
dtype: int64
```

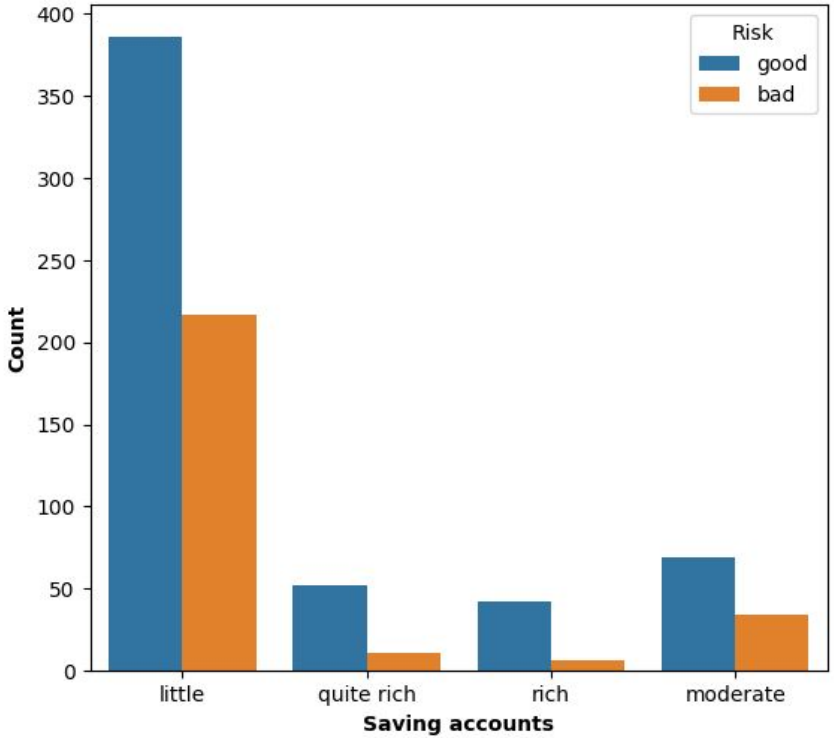
DATA EXPLORATION



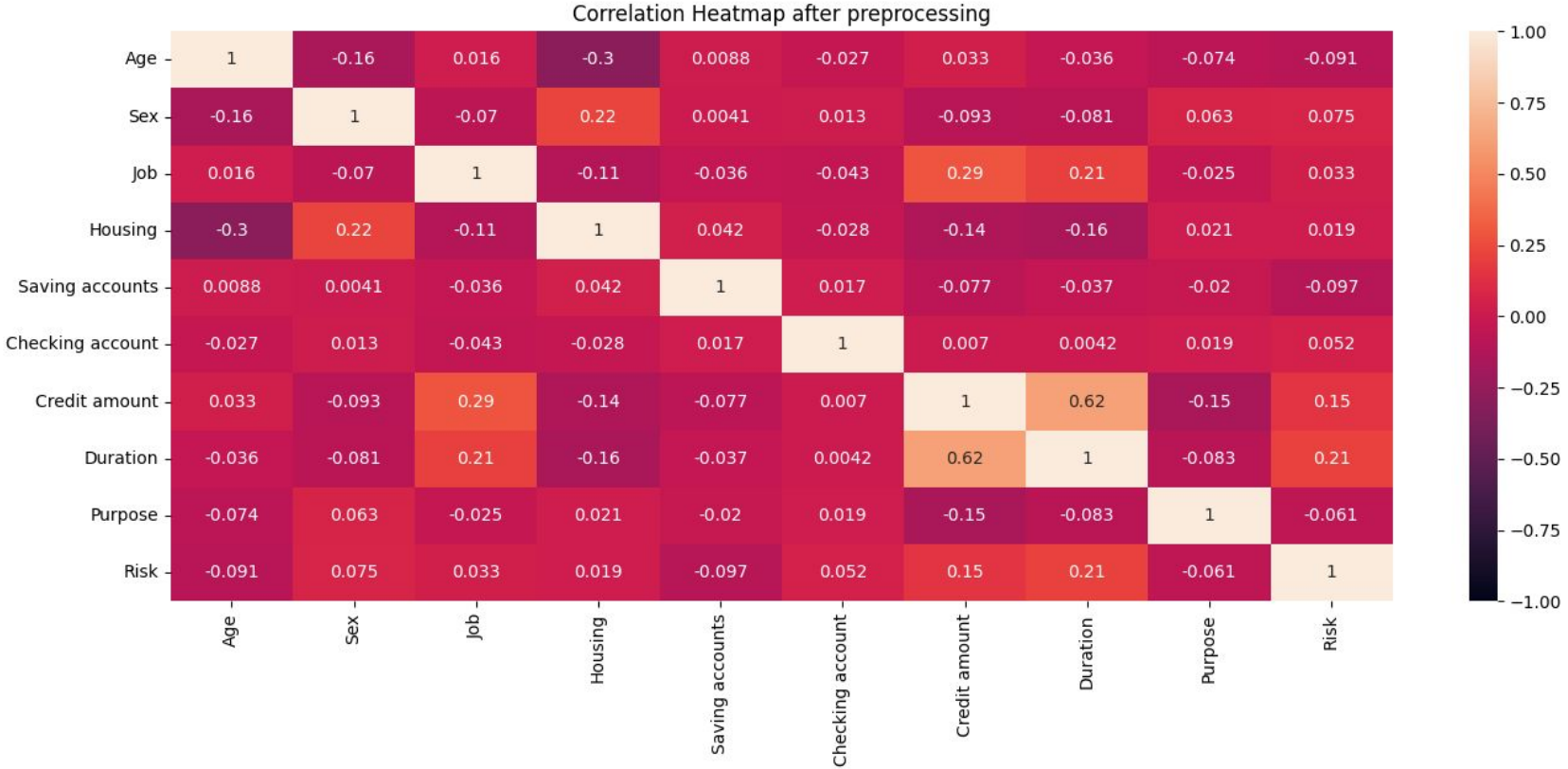


Cont…


Risk
good
bad
Count
0
100
200
300
400
500
own
free
rent
Housing

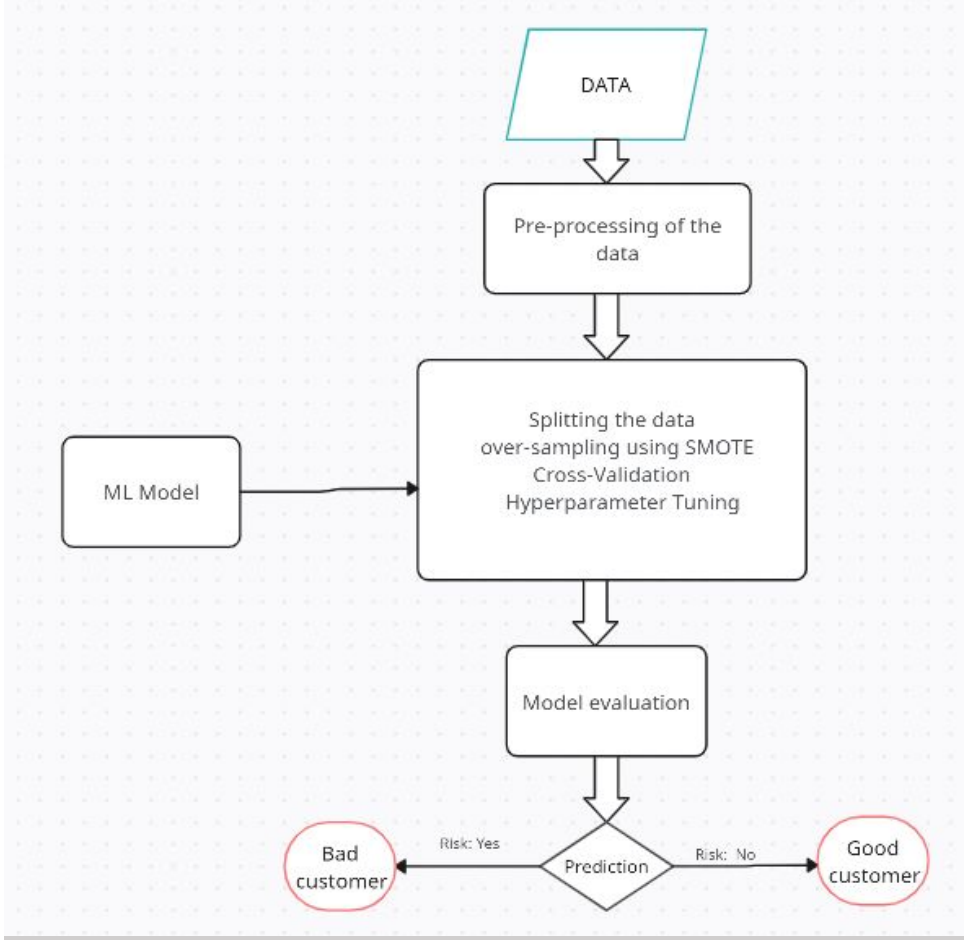

Risk
good
bad
Count
0
50
100
150
200
250
300
350
400
little
quite rich
rich
moderate
Saving accounts

CORRELATION AFTER PREPROCESSING



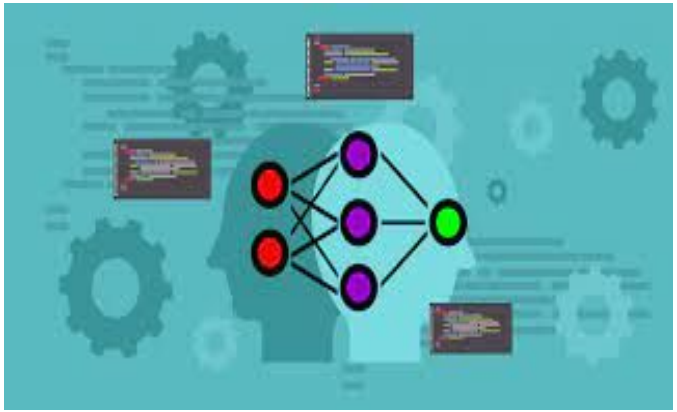
EXPERIMENTAL DESIGN

- Split: 70% training/validation, 30% test.
- Scaler: Standard scaler
- SMOTE: Synthesizing new examples.
- GridSearchCV for tuning models.



CLASSIFICATION MODELS & EVALUATION

In the project we have tried and tested few classification models like



- Random forest
- K-Nearest Neighbors
- SVM
- Decision Tree

Accuracy, Precision, Recall were used to evaluate the models

BEST HYPERPARAMETERS

Random Forest:

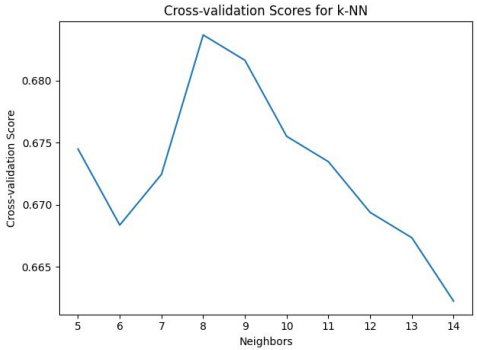
- 'max_depth': None
- 'max_features': 'auto'
- 'min_samples_leaf': 1
- 'min_samples_split': 4
- 'n_estimators': 162

SVM:

- 'C': 10
- 'Kernel': 'rbf'

KNN:

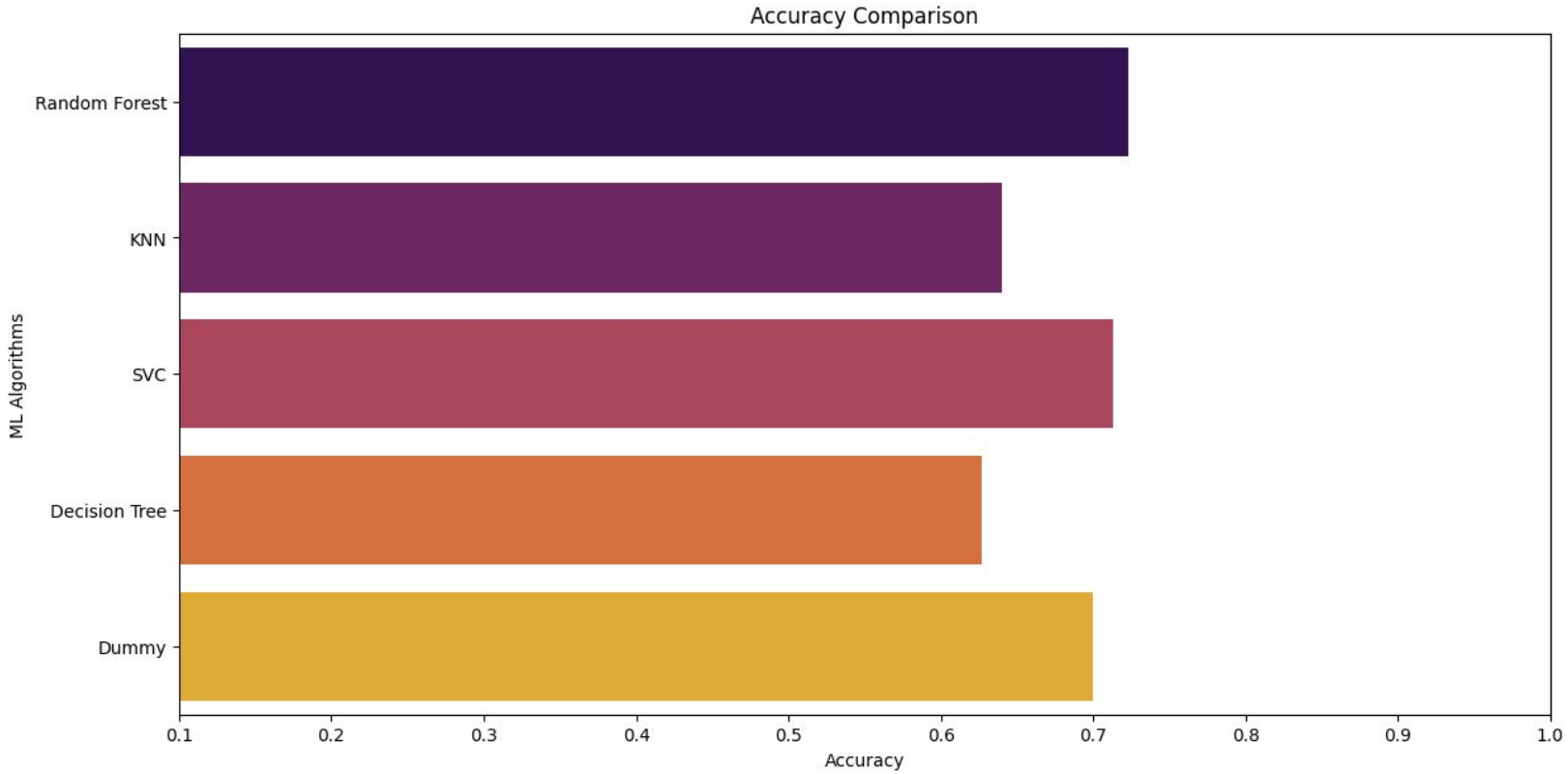
- Best Hyperparameters: {'n_neighbors': 1}
- Best Accuracy Score: 0.686734693877551



DECISION TREE:

- 'criterion': 'gini'
- 'max_depth': 20
- 'min_samples_leaf': 1
- 'min_samples_split': 2

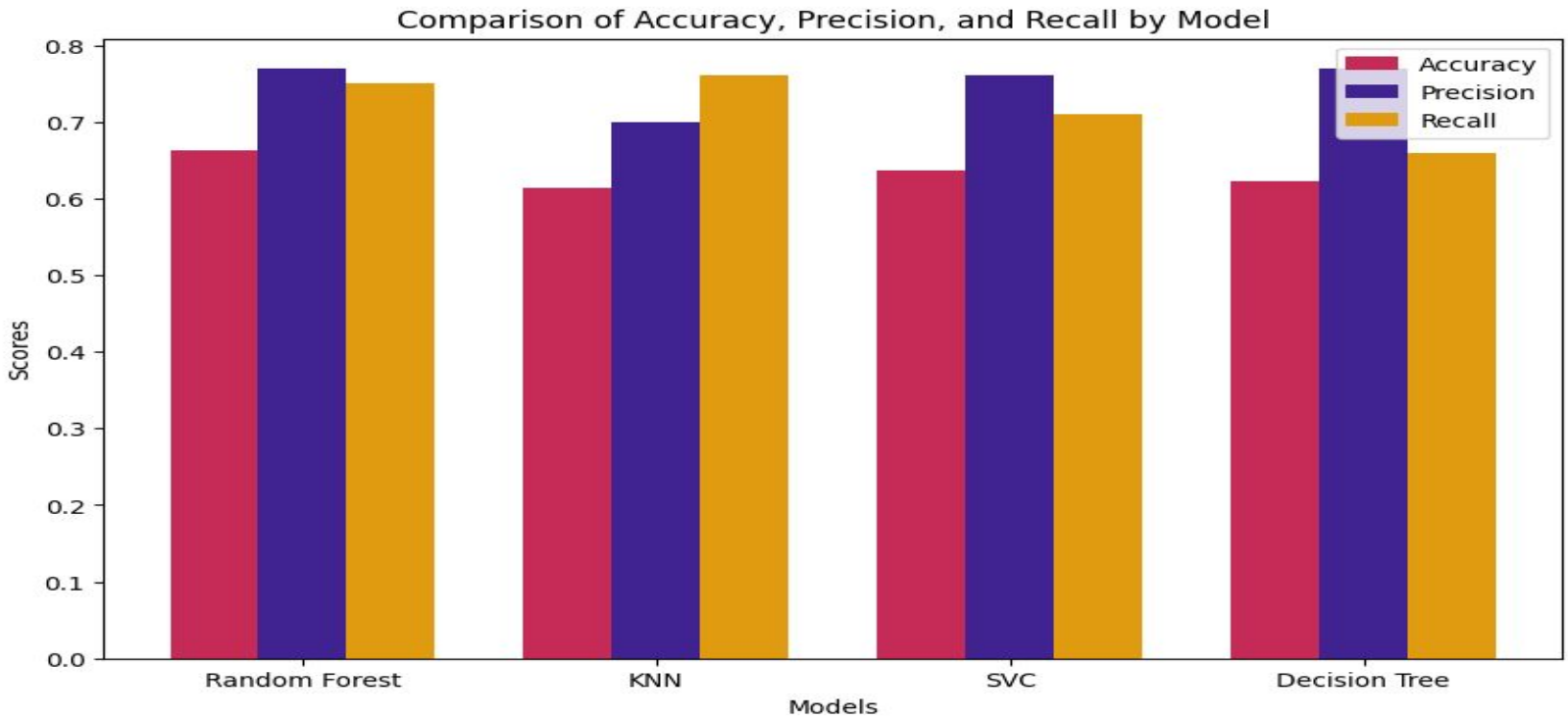
RESULTS AND DISCUSSION



RESULTS AND DISCUSSION

Models	Accuracy of original Data	Recall of Original Data	Precision of original Data	Accuracy of SMOTE Data	Recall of SMOTE Data	Precision of SMOTE Data
Random Forest	0.72	0.91	0.74	0.69	0.78	0.77
KNN	0.69	0.89	0.70	0.61	0.76	0.71
SVM	0.70	0.95	0.72	0.73	0.71	0.76
Decision Tree	0.62	0.70	0.74	0.70	0.66	0.77

RESULTS AND DISCUSSION



FUTURE SCOPE

- In the future, we are going to use classification models like AdaBoost, Naive Bayes, Gradient Boost.
- We would also like to consider additional attributes to improve the accuracy of the prediction.

TO SUMMARIZE…

- The project has developed a model to classify customers as good or bad based on the various attributes related to their occupation, credit amount etc.
- The models are evaluated on basis of accuracy values and best model is selected (Random Forest).