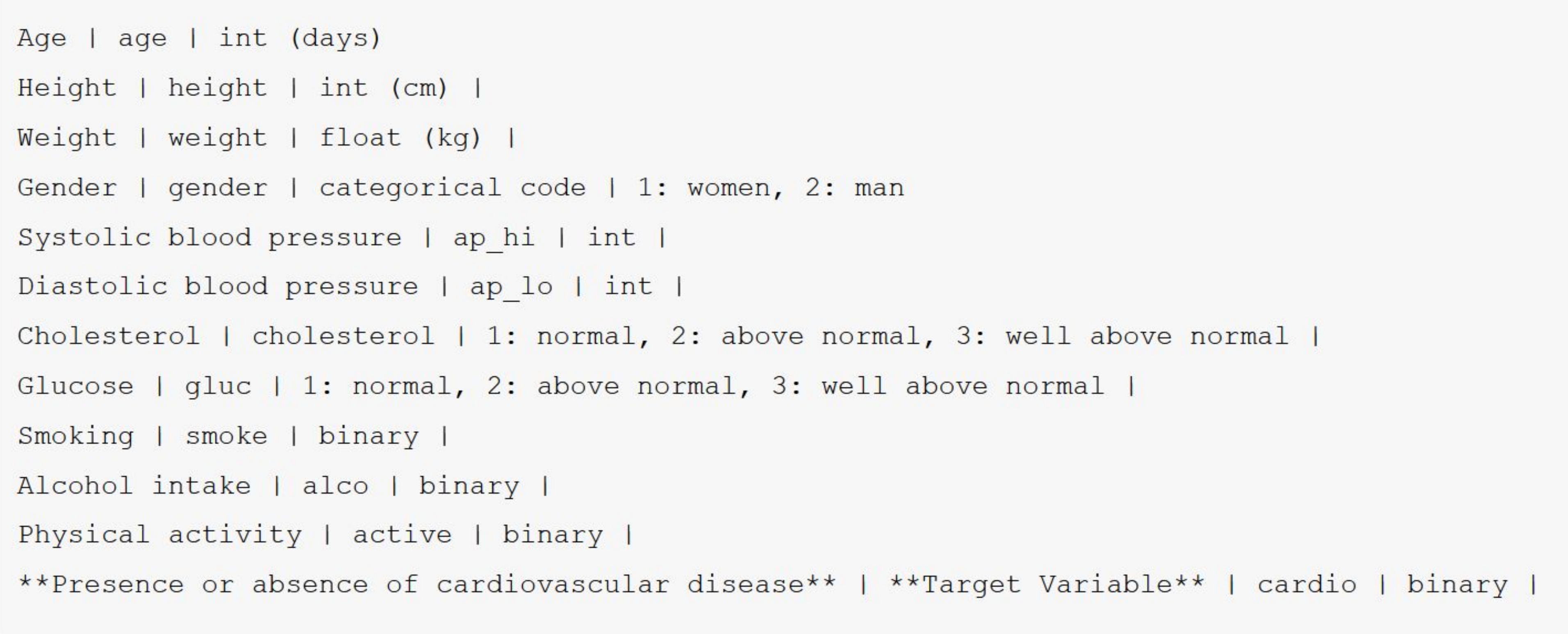
Predicting Cardiovascular Disease

Christopher Singh

Motivation - number one cause of death globally.

The need for a predictive analytical model is crucial because its usage will help determine the presence or the absence of cardiovascular disease.

Dataset from kaggle



```
Age | age | int (days)

Height | height | int (cm) |

Weight | weight | float (kg) |

Gender | gender | categorical code | 1: women, 2: man

Systolic blood pressure | ap_hi | int |

Diastolic blood pressure | ap_lo | int |

Cholesterol | cholesterol | 1: normal, 2: above normal, 3: well above normal |

Glucose | gluc | 1: normal, 2: above normal, 3: well above normal |

Smoking | smoke | binary |

Alcohol intake | alco | binary |

Physical activity | active | binary |

**Presence or absence of cardiovascular disease** | **Target Variable** | cardio | binary |
```

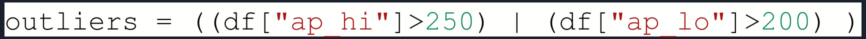
Added BMI Column

```
df['BMI'] = (df['weight'])/((df['height']) * df['height'])
```

Divide the weight of the person by their hieght squared

Outlier Checking

1. Systolic blood pressure cannot be higher than 250
2. Diastolic blood pressure cannot be higher than 200



```
outliers = ((df["ap_hi"]>250) | (df["ap_lo"]>200) )
```

Using this logic, I removed 993 records.

These thresholds were defined by the CDC

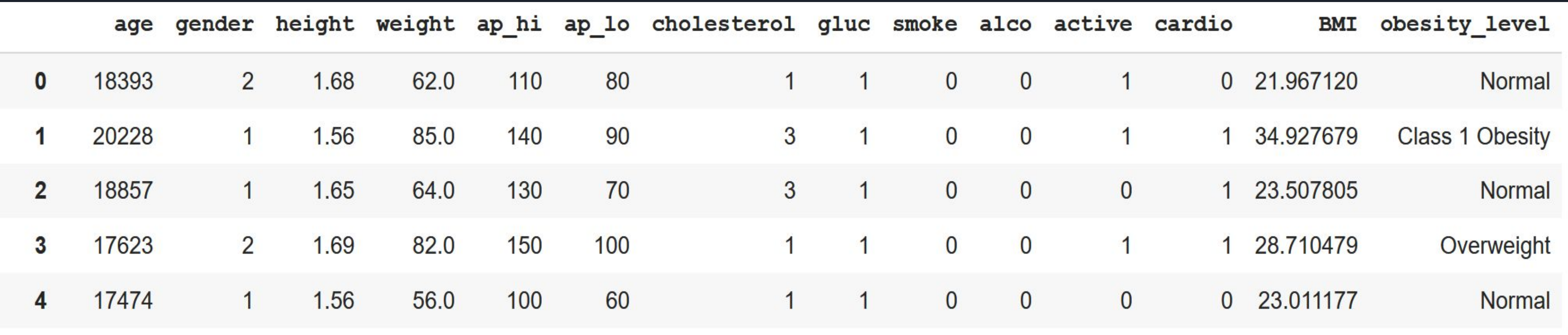
Add level of obesity column

- Appended the obesity level based on the BMI value.
- These thresholds were also defined by the CDC at:
- https://www.cdc.gov/obesity/adult/defining.html

The 6 categories were:

- Underweight
- Normal
- Overweight
- Class 1 Obesity
- Class 2 obesity
- Class 3 obesity

Dataframe head



	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	BMI	obesity_level
0	18393	2	1.68	62.0	110	80	1	1	0	0	1	0	21.967120	Normal
1	20228	1	1.56	85.0	140	90	3	1	0	0	1	1	34.927679	Class 1 Obesity
2	18857	1	1.65	64.0	130	70	3	1	0	0	0	1	23.507805	Normal
3	17623	2	1.69	82.0	150	100	1	1	0	0	1	1	28.710479	Overweight
4	17474	1	1.56	56.0	100	60	1	1	0	0	0	0	23.011177	Normal

Data Analysis

Link:
https://datastudio.google.com/reporting/f532c9df-d1f1-45bc-910a-1bb85294c1a1/page/lwtqB

Quick analysis of the data in terms of gender and obesity level

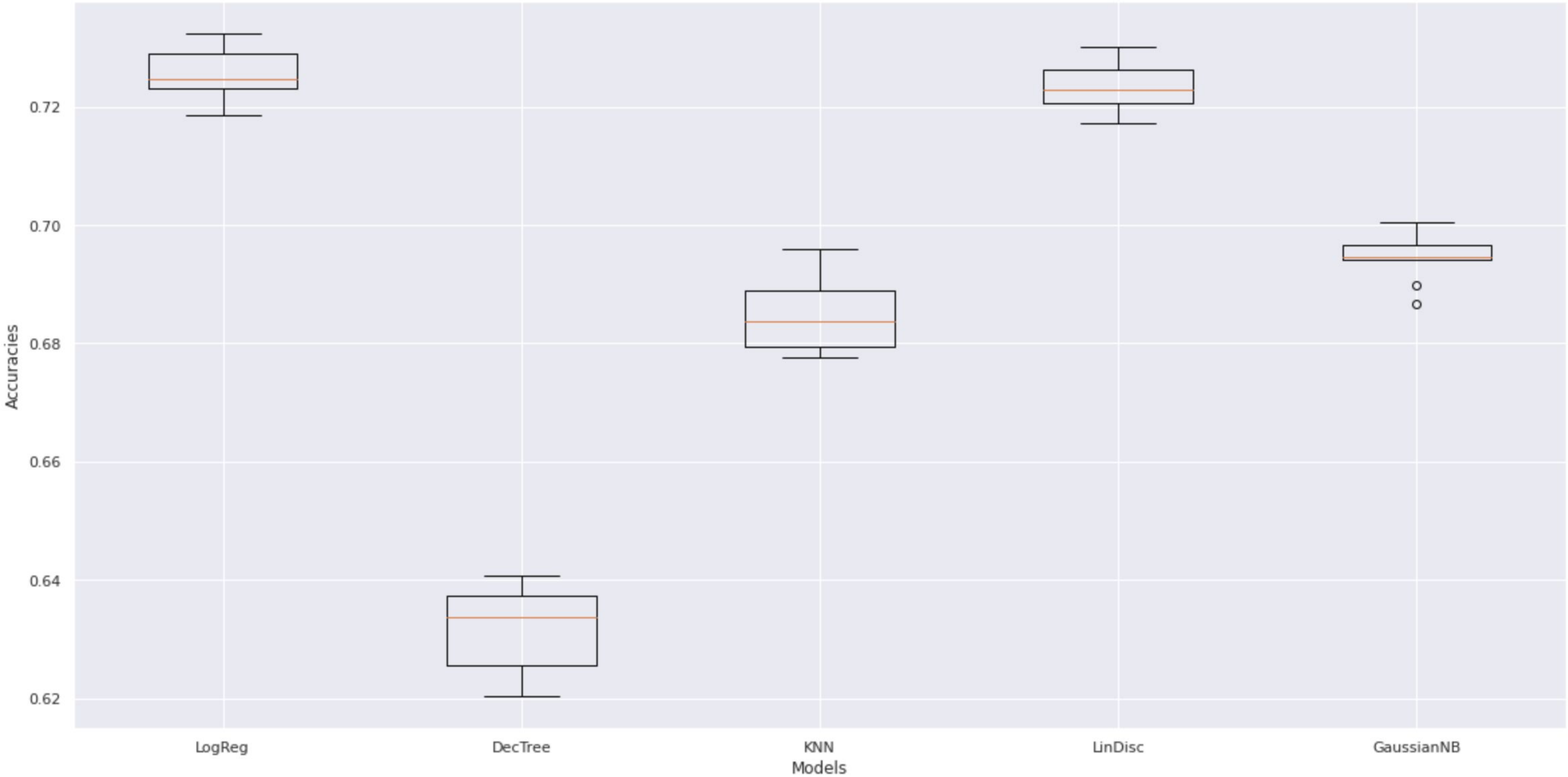
Data Pre-processing

- Convert categorical values into a numerical format. (Gender and obesity level)
- Apply min-max normalization to scale the dataframe into values between 0-1
- Check and remove records that were NA. (0 were found)
- Split the dataset into:
  - 70% training
  - 30% testing

ML Techniques Explored

- Decision Tree
- Naive Bayes Classifiers
- Neural Network
- K-Means Clustering
- Logistical Regression
- Linear Regression

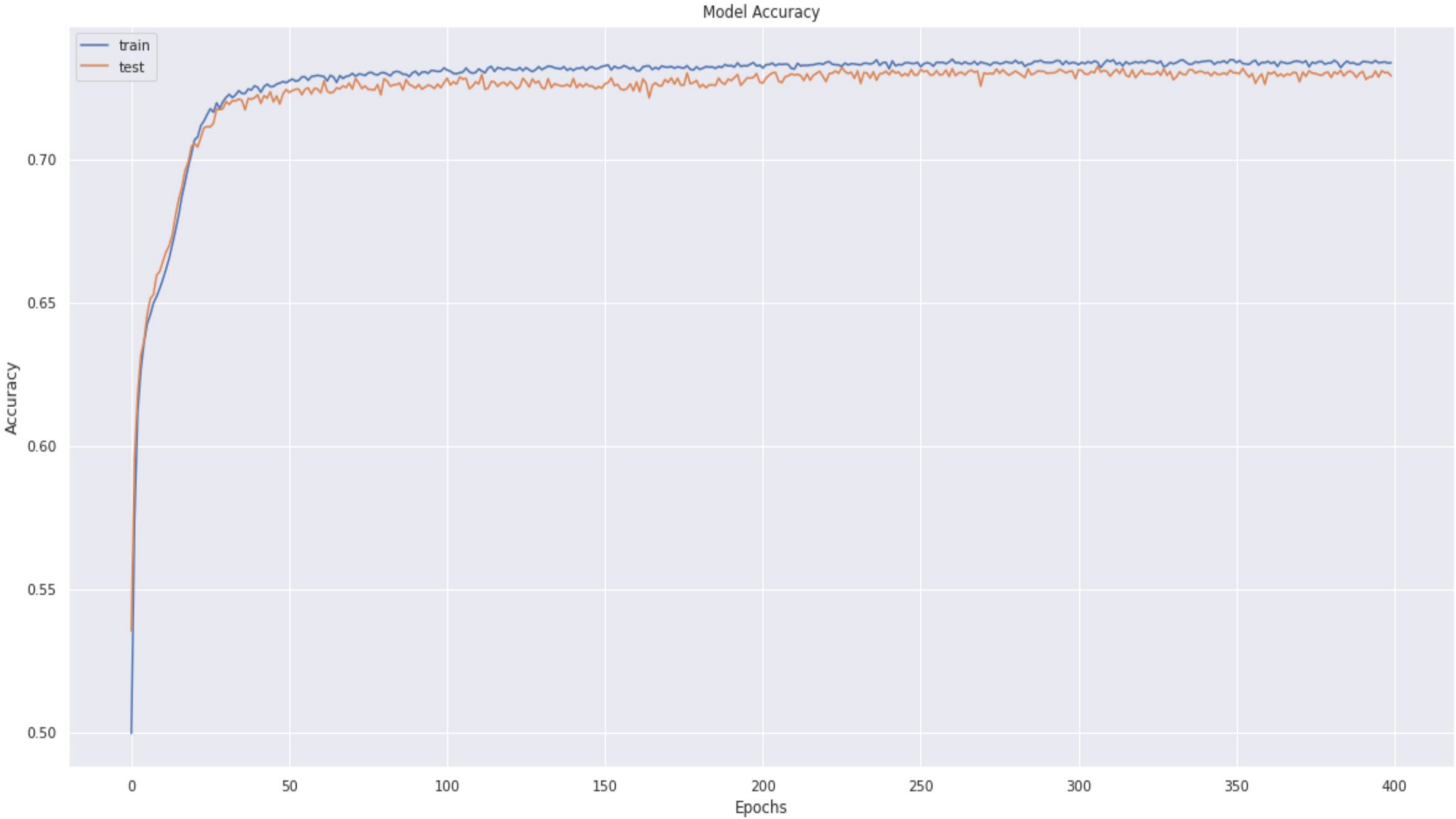
10-Fold Cross Validation To See Best Model

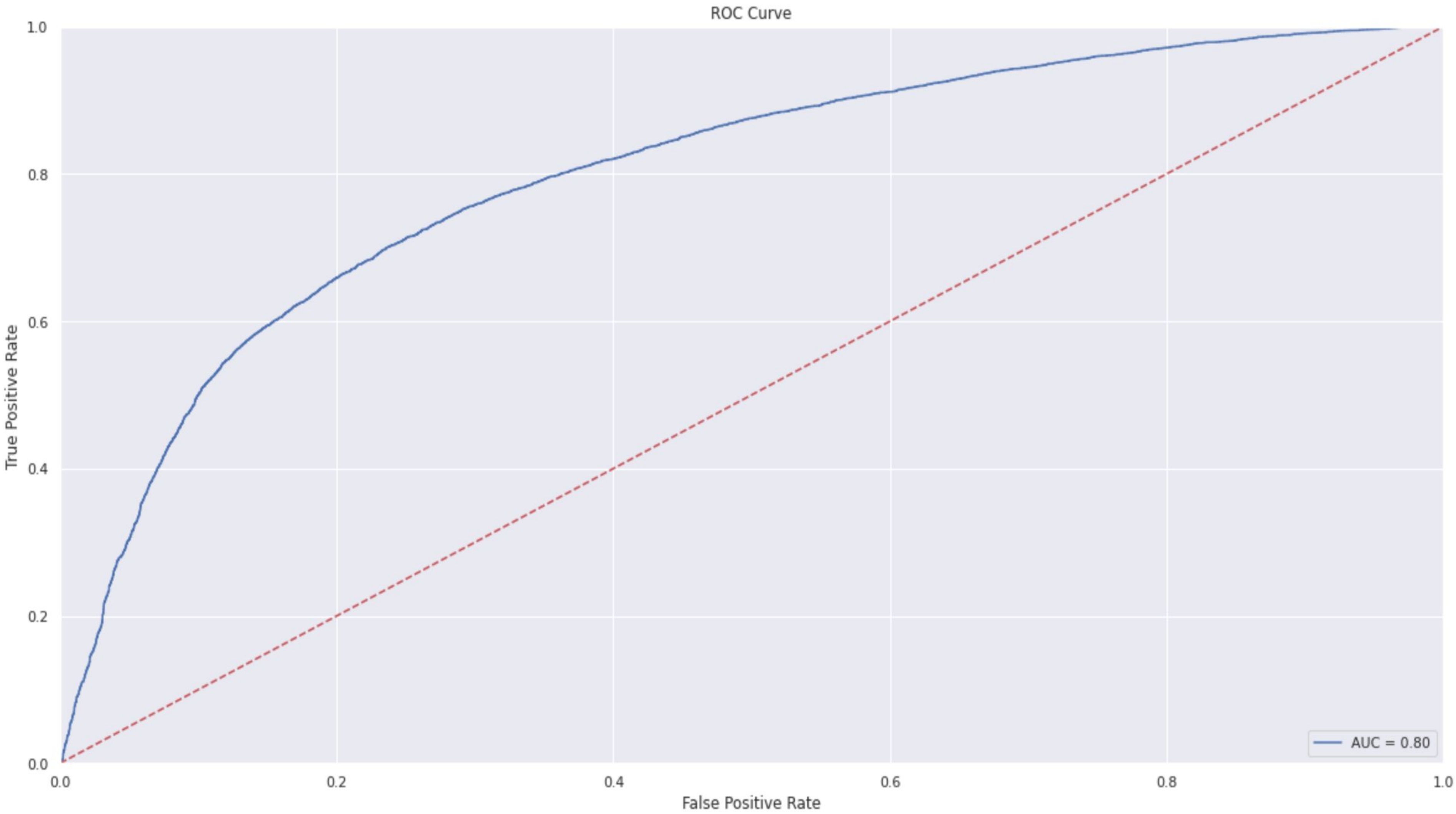


Neural Network Architecture

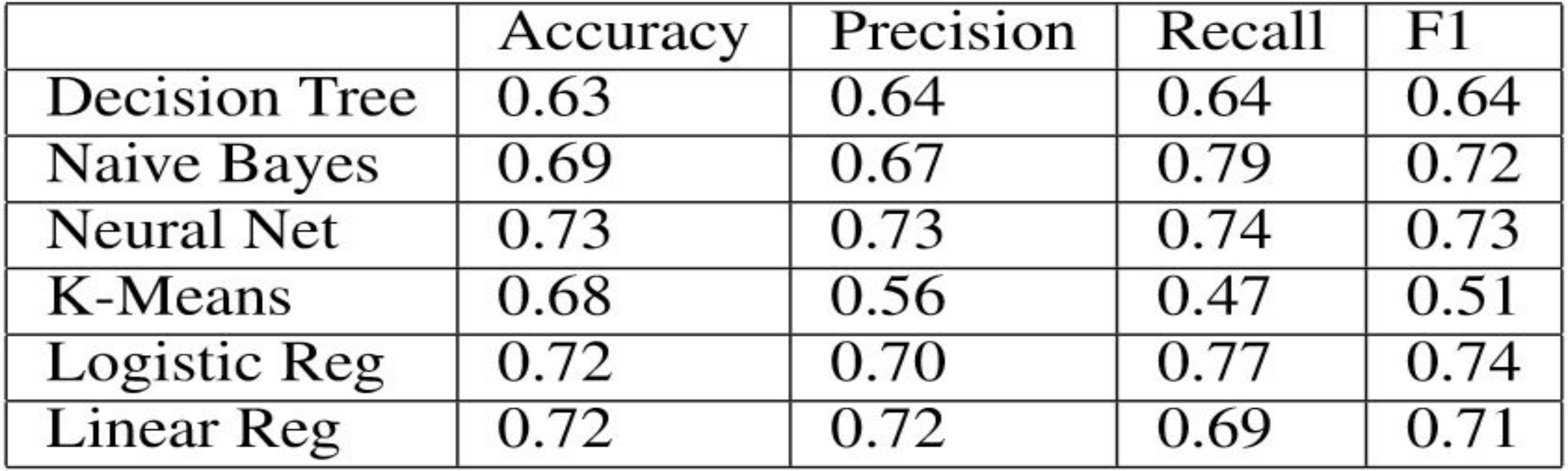
- 1 input layer with 13 neurons
- 2 hidden layers with 7 and 5 neurons respectively
- 1 output layer with 1 neuron

Loss function was evaluated using binary cross entropy (2 possible outcomes)


Model Accuracy
train
test
Accuracy
0.50
0.55
0.60
0.65
0.70
0
50
100
150
200
250
300
350
400
Epochs


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.80

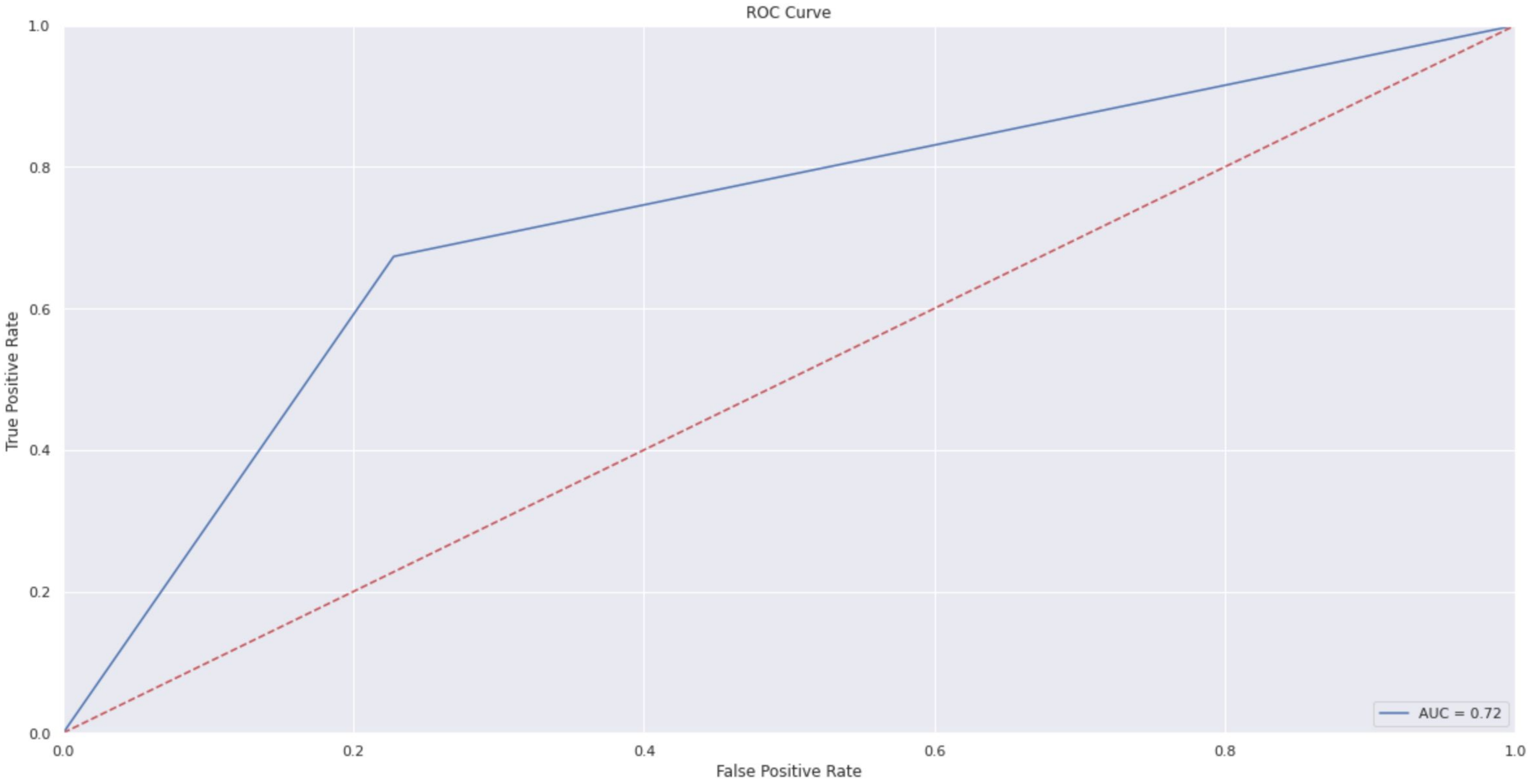
Results of all ML Models



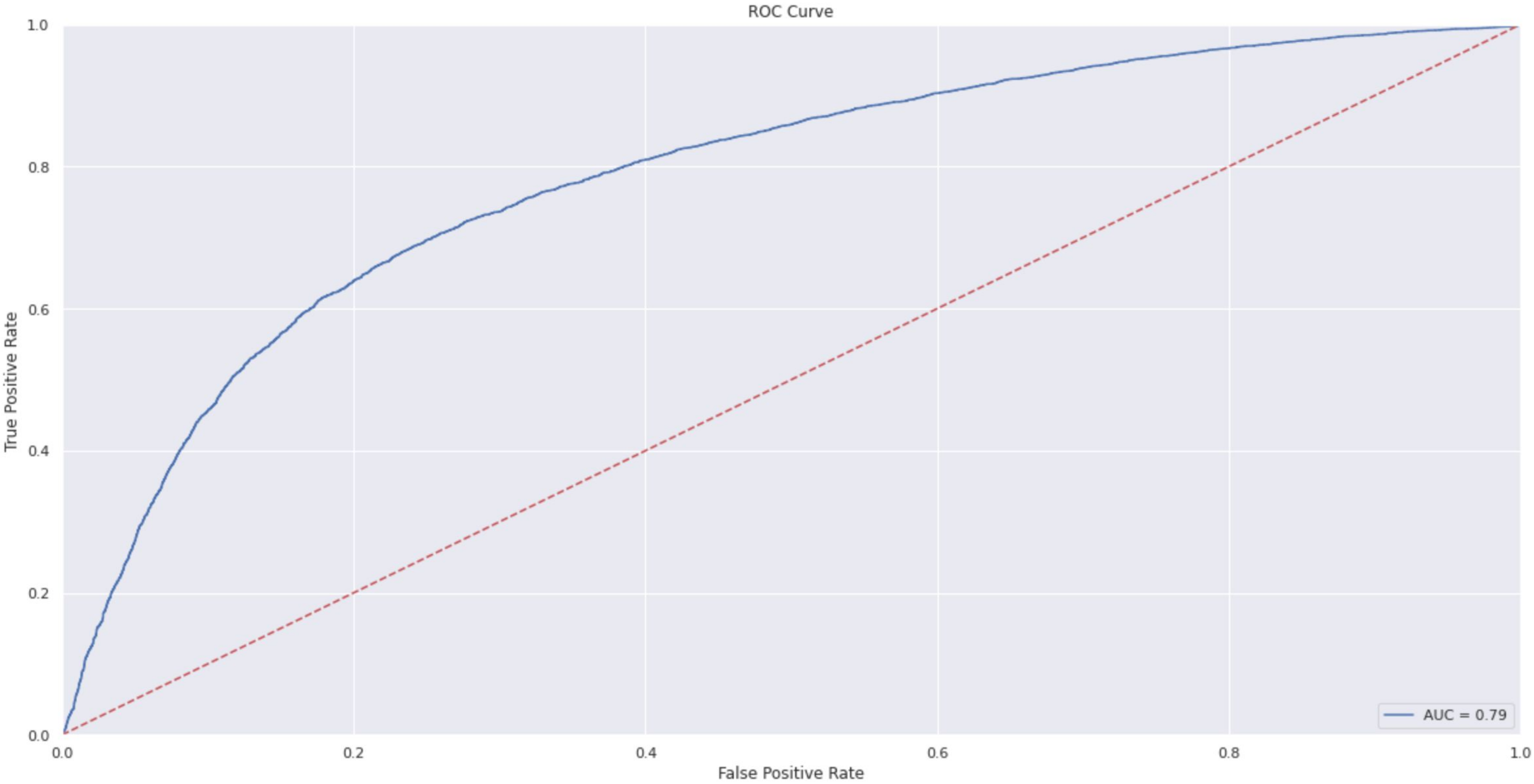
	Accuracy	Precision	Recall	F1
Decision Tree	0.63	0.64	0.64	0.64
Naive Bayes	0.69	0.67	0.79	0.72
Neural Net	0.73	0.73	0.74	0.73
K-Means	0.68	0.56	0.47	0.51
Logistic Reg	0.72	0.70	0.77	0.74
Linear Reg	0.72	0.72	0.69	0.71

Table 1: Evaluation Metrics Of All Machine Learning Models

Logistical Regression ROC



Linear Regression ROC



Conclusions

The ANN outperformed all of the other ML algorithms but not by much. Both linear and logistical regression were close in terms of accuracy.

The decision tree had the worst accuracy.

K-Means, Naive Bayes and decision tree should be thought of as a second alternative to ann, linear and logistical regressions.

Future work

Explore dimensionality reduction to consider only the most important features in the analysis

Thanks!