CS 5805 Machine Learning 1
Final Term Project

Jyothi Sevakula

Introduction

VIRGINIA TECH

In this project, we explore a comprehensive airlines dataset to predict flight delays, providing a detailed analysis of factors influencing these delays. Our primary objective is to empower airlines and stakeholders with data-driven insights to improve operational efficiency, minimize disruptions, and enhance passenger satisfaction.

We employ various dimensionality reduction techniques, including Random Forest, PCA, SVD, and Variance Inflation Factor (VIF), alongside statistical analyses like T-tests and F-tests.

Machine learning methodologies such as Linear Regression, Backward Stepwise Regression,Decision Trees, SVM, KNN, Naive Bayes, and Multi-Layer Perceptron are utilized for predictive modeling.

Clustering methods like KMeans and DBSCAN, along with association rule mining using the Apriori algorithm, provide valuable insights into flight patterns and feature relationships.

Description of Dataset

The Airlines dataset consists of 539,383 records with 9 features aimed at predicting flight delays.

Categorical features

- Airline
- Flight
- AirportFrom
- AirportTo
- DayOfWeek
- Delay
- Id

Numerical features

- Time
- Length

Data Preprocessing

1. The dataset is inherently clean with no missing or NaN values.
2. There are no duplicate records present in the dataset.
3. Removed outliers present in data by IQR method
4. Performed one-hot encoding on categorical features
5. After downsampling total number of observations are 37,570

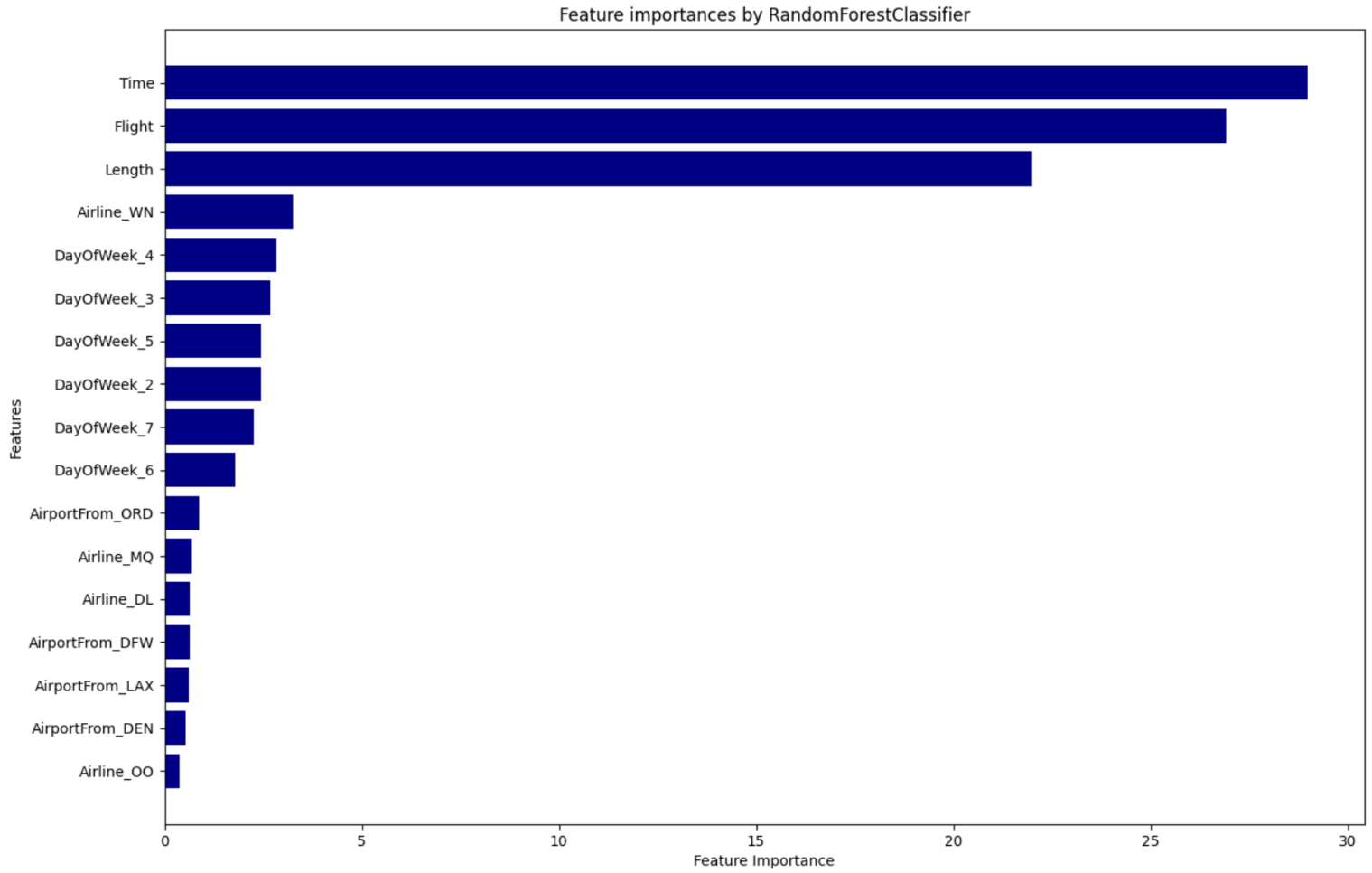
The dataset includes:

1. 18 categories in *Airline*.
2. 293 categories in *AirportFrom*.
3. 293 categories in *AirportTo*.
4. 7 categories in *DayOfWeek*.

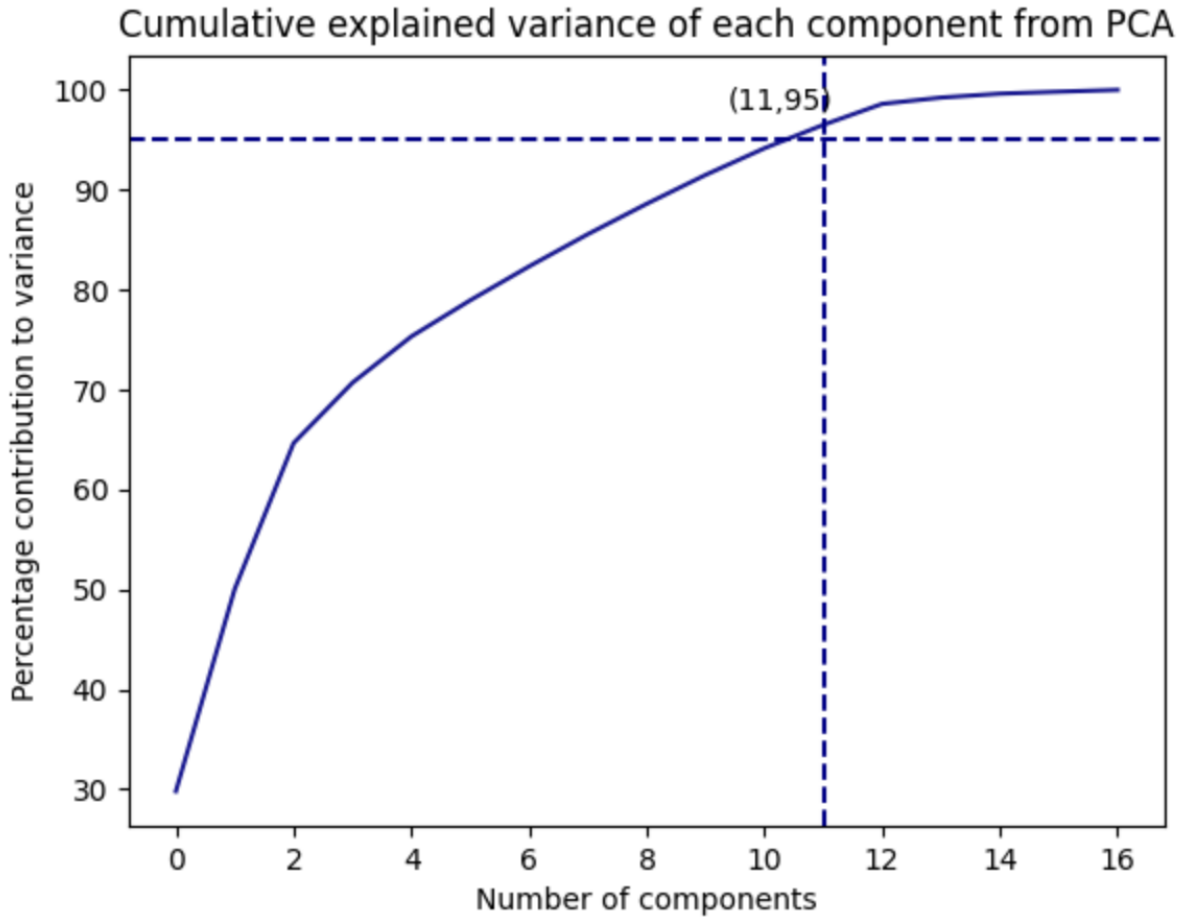
Down Sampling:

- Focused on predicting delays related to the top 5 source airports (*AirportFrom*).
- Dropped the *AirportTo, id* feature as it is not essential for the current analysis.
- Considered only the top 5 airlines, as other airlines have significantly fewer observations.

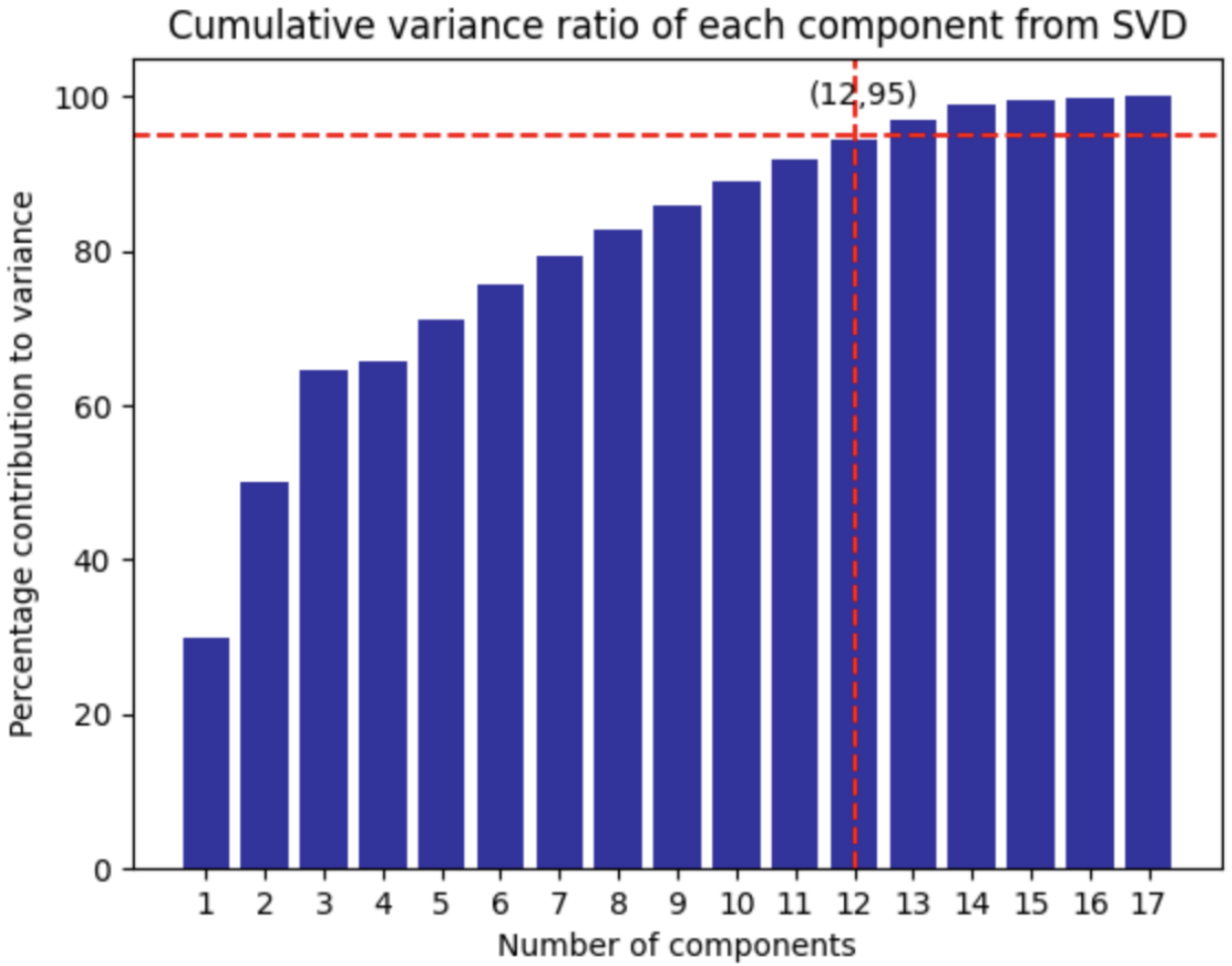
Dimensionality Reduction – Random Forest



Dimensionality Reduction - PCA



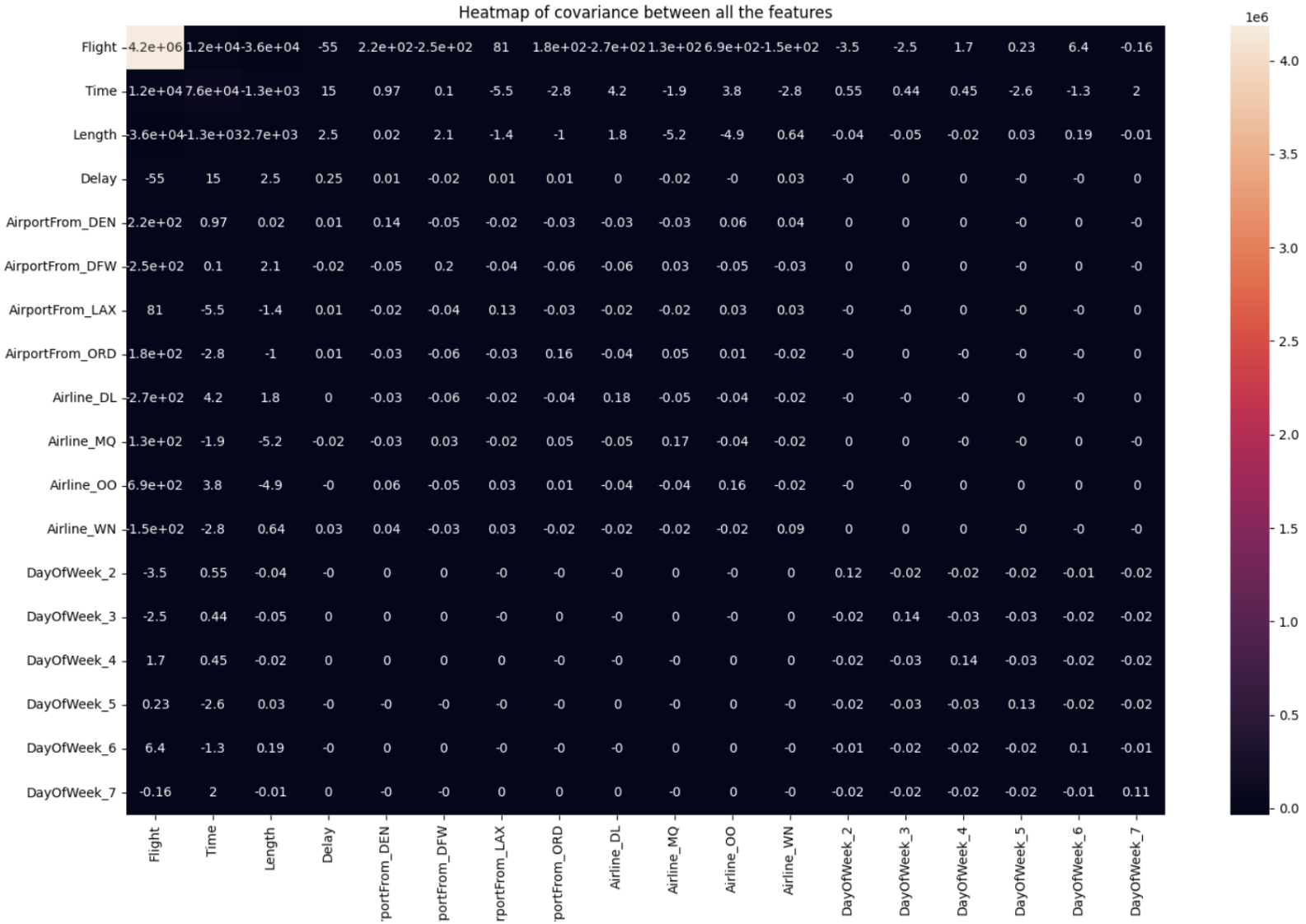
Dimensionality Reduction - SVD



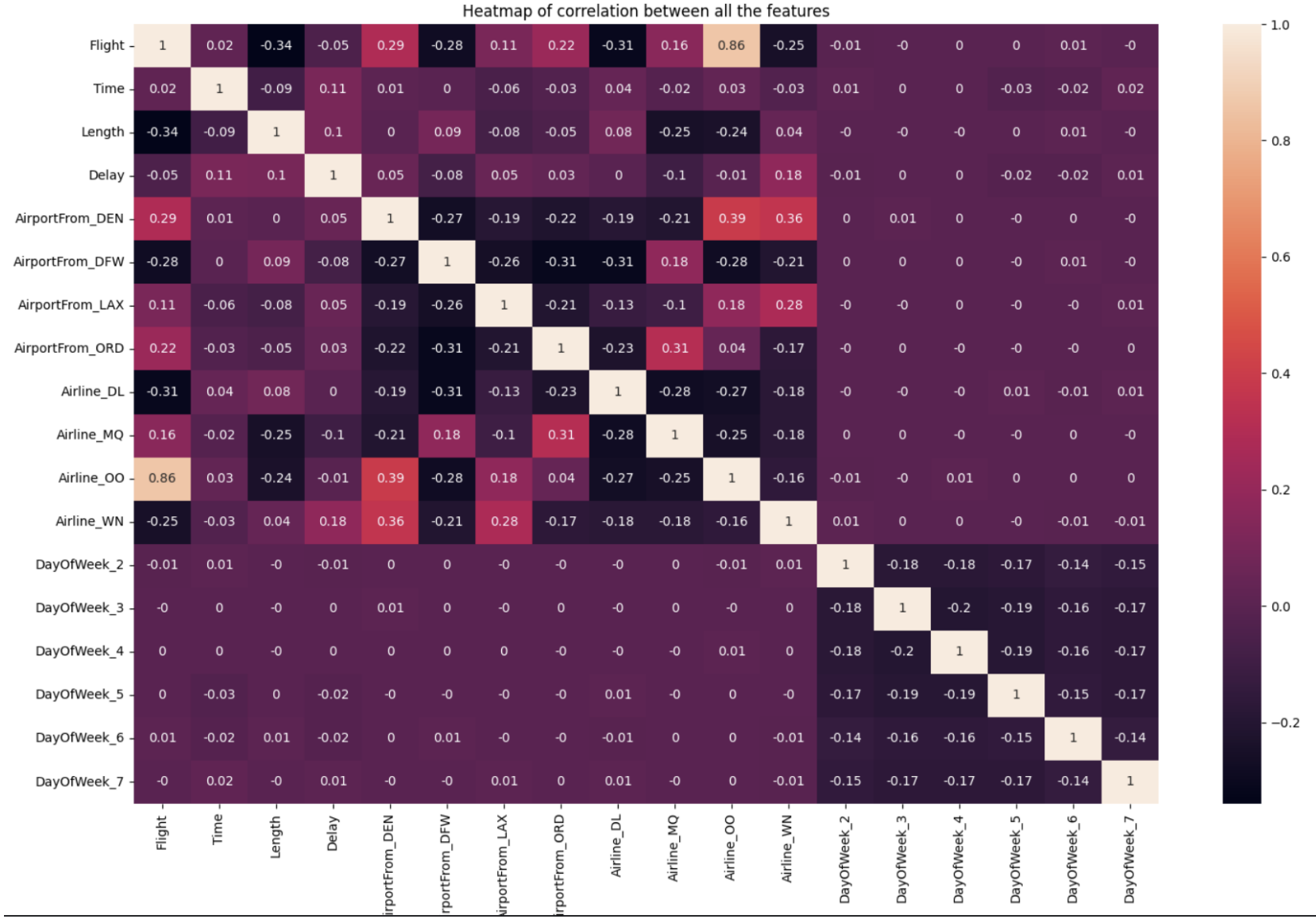
Dimensionality Reduction - VIF

```
VIF analysis
Flight 8.57
Time 1.02
Length 1.28
Delay 3.57
AirportFrom_DEN 3.56
AirportFrom_DFW 2.71
AirportFrom_LAX 2.7
AirportFrom_ORD 2.24
Airline_DL 3.57
Airline_MQ 11.52
Airline_OO 2.4
Airline_WN 1.8
DayOfWeek_2 2.02
DayOfWeek_3 2.02
DayOfWeek_4 1.93
DayOfWeek_5 1.66
DayOfWeek_6 1.8
```

Covariance Matrix



Correlation Matrix



Collinearity Analysis

The features eliminated from random forest analysis are Airline_MQ, AirportFrom_DFW, AirportFrom_LAX, Airline_DL, AirportFrom_DEN, and Airline_OO and leaves us with 11 dimensions which is the least of all the dimensionality reduction techniques. Hence, I chose to follow random forest feature importances.

Linear Regression Model

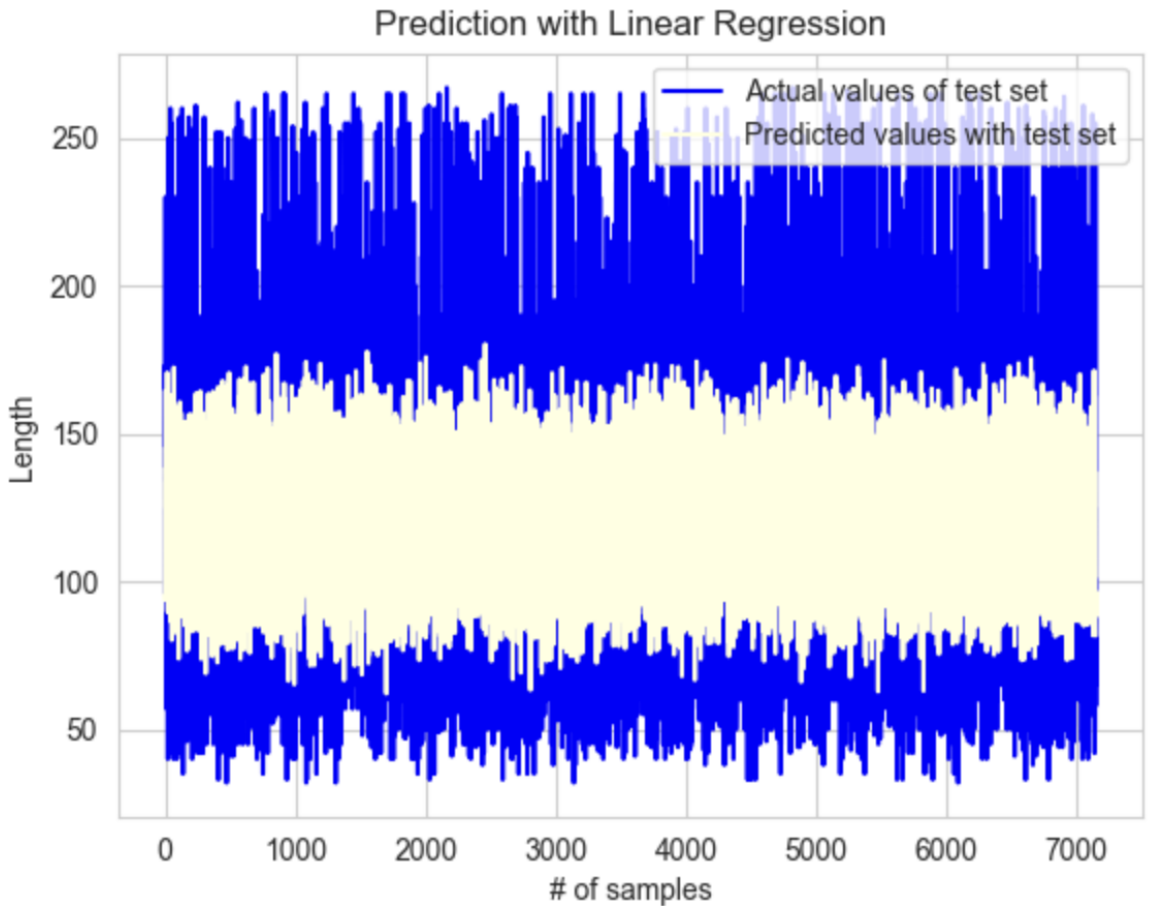
VIRGINIA TECH

```
The coefficients of the Linear Regression model
Feature Coefficient
const 0.0
Flight -0.02
Time -0.09
Delay 0.16
AirportFrom_DEN 0.8
AirportFrom_DFW 0.35
AirportFrom_LAX 0.43
AirportFrom_ORD 0.51
Airline_DL -0.09
Airline_MQ -1.04
Airline_OO -1.22
Airline_WN -0.74
DayOfWeek_2 0.01
DayOfWeek_3 -0.0
DayOfWeek_4 0.0
DayOfWeek_5 0.01
DayOfWeek_6 0.03
DayOfWeek_7 0.02
Mean Squared Error for training set (MSE): 2054.446
```

Model	R-squared	Adj. R-squared	AIC	BIC	Mean Squared Error
Linear Regression	0.23	0.22	218035.43	218175.86	2054.45

Target – ‘Length’

Linear Regression Model



T-test and F-test analysis

t-test analysis

Feature	t-statistic	p-value
const	2.22	0.03
Flight	-1.52	0.13
Time	-16.24	0.0
Delay	15.11	0.0
AirportFrom_DEN	27.45	0.0
AirportFrom_DFW	12.54	0.0
AirportFrom_LAX	15.57	0.0
AirportFrom_ORD	18.46	0.0
Airline_DL	-3.52	0.0
Airline_MQ	-45.77	0.0
Airline_OO	-27.38	0.0
Airline_WN	-28.71	0.0
DayOfWeek_2	0.45	0.65
DayOfWeek_3	-0.17	0.86
DayOfWeek_4	0.25	0.8
DayOfWeek_5	0.43	0.67
DayOfWeek_6	1.48	0.14
DayOfWeek_7	0.75	0.45

Feature	F-test score	p-value
Flight	89.3	0.0
Time	14.91	0.0
Length	4.03	0.0
AirportFrom_DEN	37.04	0.0
AirportFrom_DFW	48.37	0.0
AirportFrom_LAX	65.57	0.0
AirportFrom_ORD	24.22	0.0
Airline_DL	112.05	0.0
Airline_MQ	65.06	0.0
Airline_OO	122.7	0.0
Airline_WN	23.58	0.0
DayOfWeek_2	0.53	1.0
DayOfWeek_3	0.67	1.0
DayOfWeek_4	0.58	1.0
DayOfWeek_5	0.73	1.0
DayOfWeek_6	1.25	0.01
DayOfWeek_7	0.97	0.62

Stepwise Regression Model

VIRGINIA TECH

```
eliminated features: ['DayofWeek_3', 'DayofWeek_4', 'DayofWeek_5', 'DayofWeek_2', 'DayOfWeek_7', 'Flight', 'DayOfWeek_6']
```

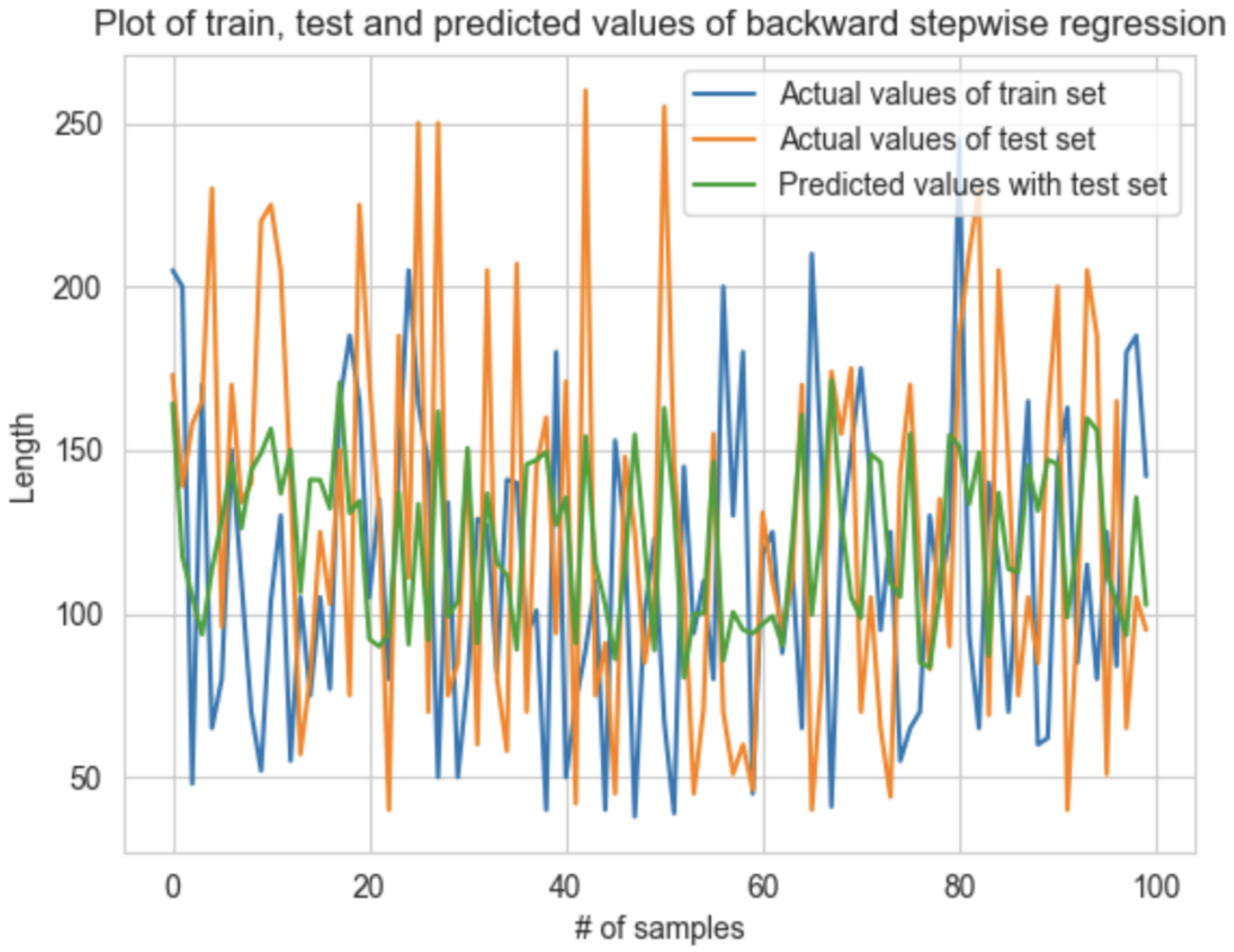
OLS Regression Results

Dep. Variable:	y	R-squared:	0.225
Model:	OLS	Adj. R-squared:	0.225
Method:	Least Squares	F-statistic:	830.3
Date:	Mon, 02 Dec 2024	Prob (F-statistic):	0.00
Time:	11:50:43	Log-Likelihood:	-36914.
No. Observations:	28580	AIC:	7.385e+04
Df Residuals:	28569	BIC:	7.394e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0995	0.028	3.571	0.000	0.045	0.154
Time	-0.0858	0.005	-16.243	0.000	-0.096	-0.075
Delay	0.1626	0.011	15.129	0.000	0.142	0.184
AirportFrom_DEN	0.8005	0.029	27.412	0.000	0.743	0.858
AirportFrom_DFW	0.3528	0.028	12.648	0.000	0.298	0.407
AirportFrom_LAX	0.4303	0.028	15.572	0.000	0.376	0.484
AirportFrom_ORD	0.5101	0.028	18.397	0.000	0.456	0.564
Airline_DL	-0.0987	0.027	-3.681	0.000	-0.151	-0.046
Airline_MQ	-1.0608	0.016	-67.745	0.000	-1.092	-1.030
Airline_OO	-1.2768	0.021	-61.633	0.000	-1.317	-1.236
Airline_WN	-0.7423	0.026	-28.697	0.000	-0.793	-0.692

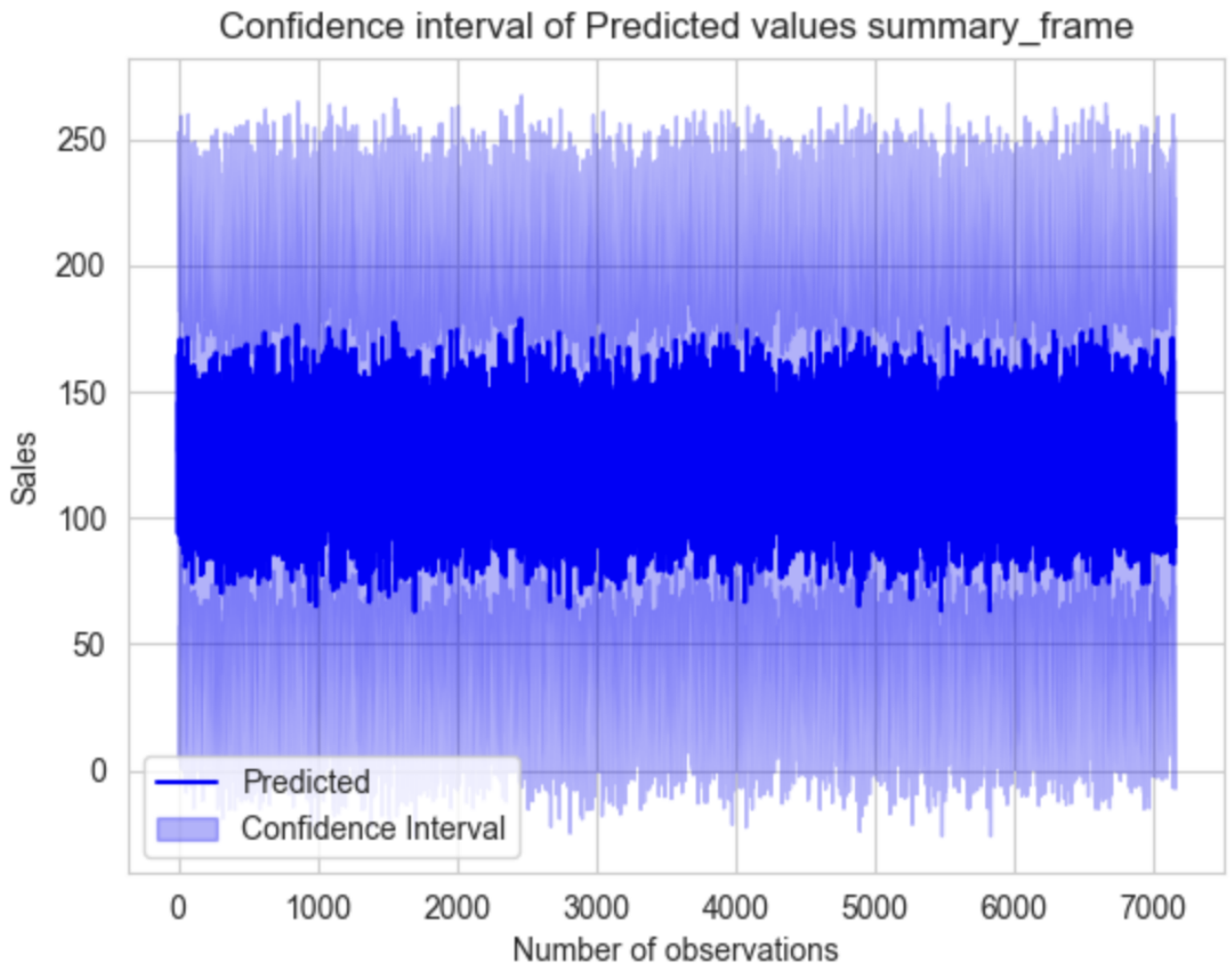
Omnibus:	1539.012	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1798.650
Skew:	0.611	Prob(JB):	0.00
Kurtosis:	3.133	Cond. No.	14.6

Plot with train, test and predicted values



Confidence Intervals

VIRGINIA TECH



Comparing Linear regression and backward stepwise regression

Model	R-squared	Adj. R-squared	AIC	BIC	Mean Squared Error
Linear Regression	0.23	0.22	218035.43	218175.86	2054.45
Backward Stepwise Regression	0.23	0.22	73849.35	73940.22	2032.57

Classification Techniques:

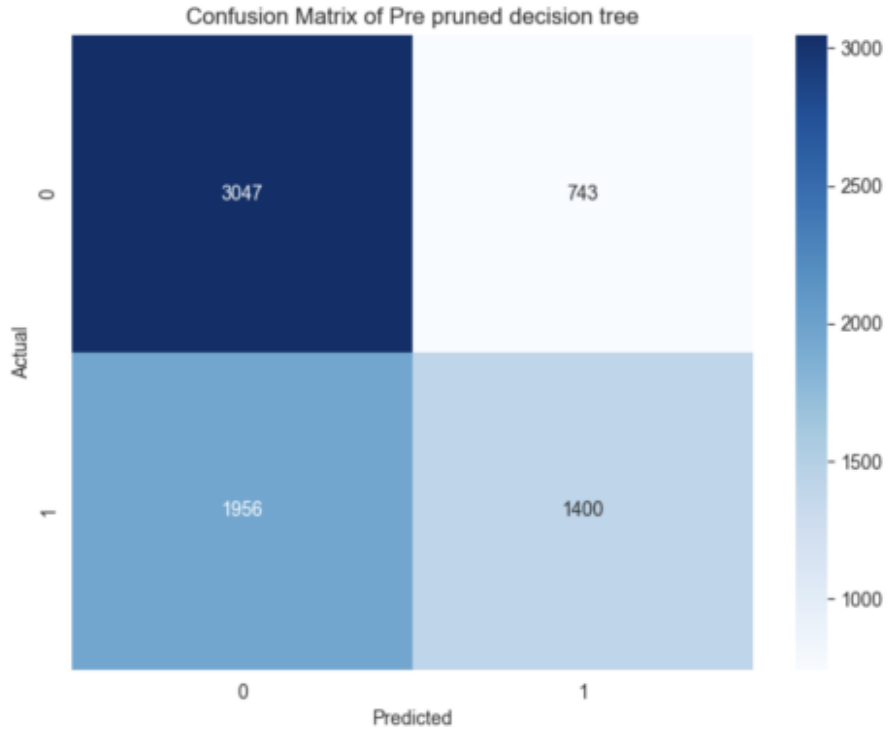
- Pre-pruned Decision Tree
- Post-pruned Decision Tree
- Logistic Regression
- K- Nearest Neighbors
- SVM
- Naïve Bayes
- Multi-layered perceptron
- Comparing all classifier models

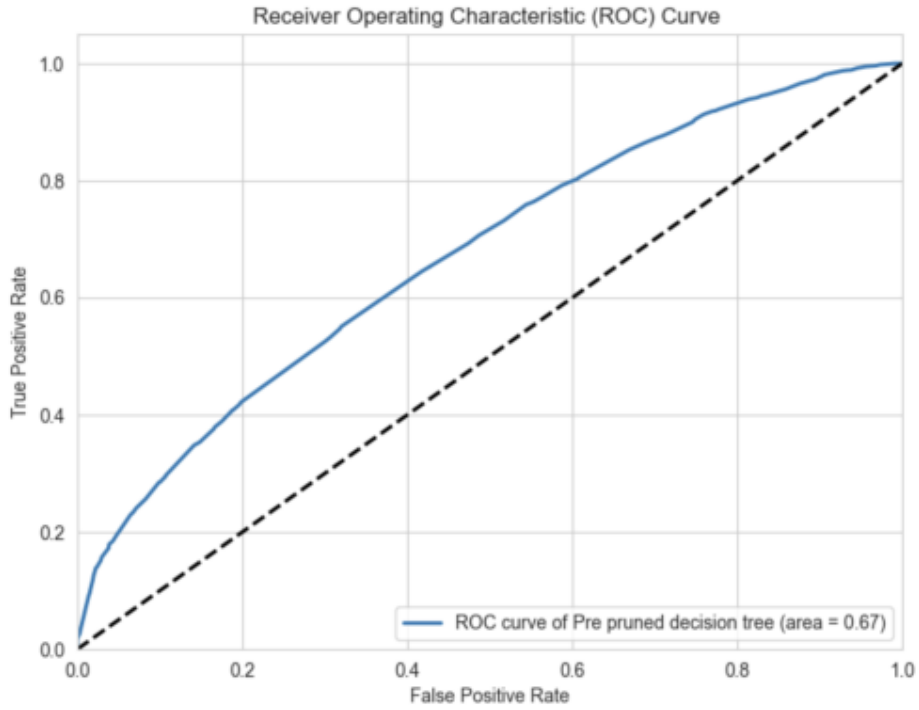
Pre-pruned Decision Tree

VIRGINIA TECH

```
Best Hyperparameters: {'criterion': 'entropy', 'max_depth': 9, 'max_features': 7, 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}
```

Classifier	Accuracy	Confusion Matrix	Precision	Sensitivity or Recall	Specificity	F-score	AUC
Pre pruned Decision Tree Classifier	0.62	[[3047 743] [1956 1400]]	0.65	0.42	0.8	0.51	0.67



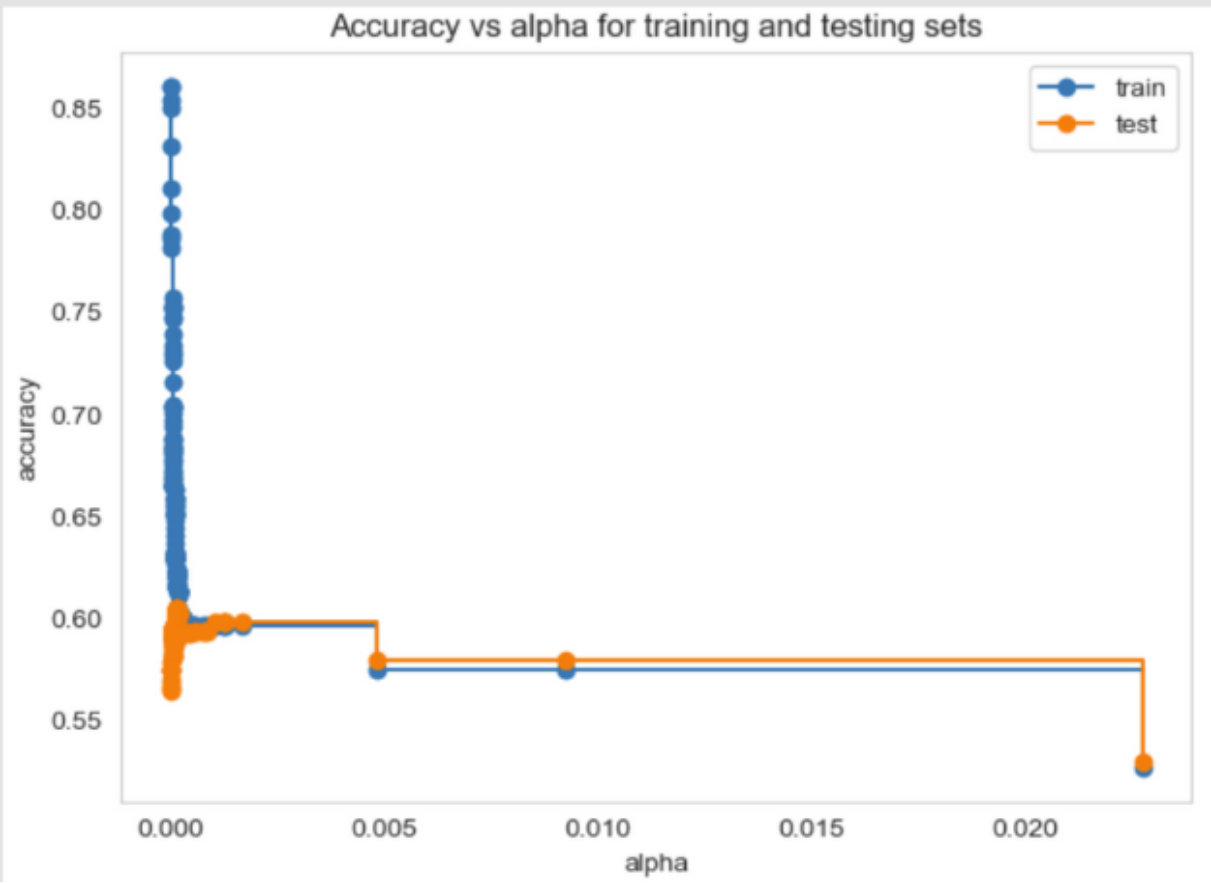


Post-pruned Decision Tree

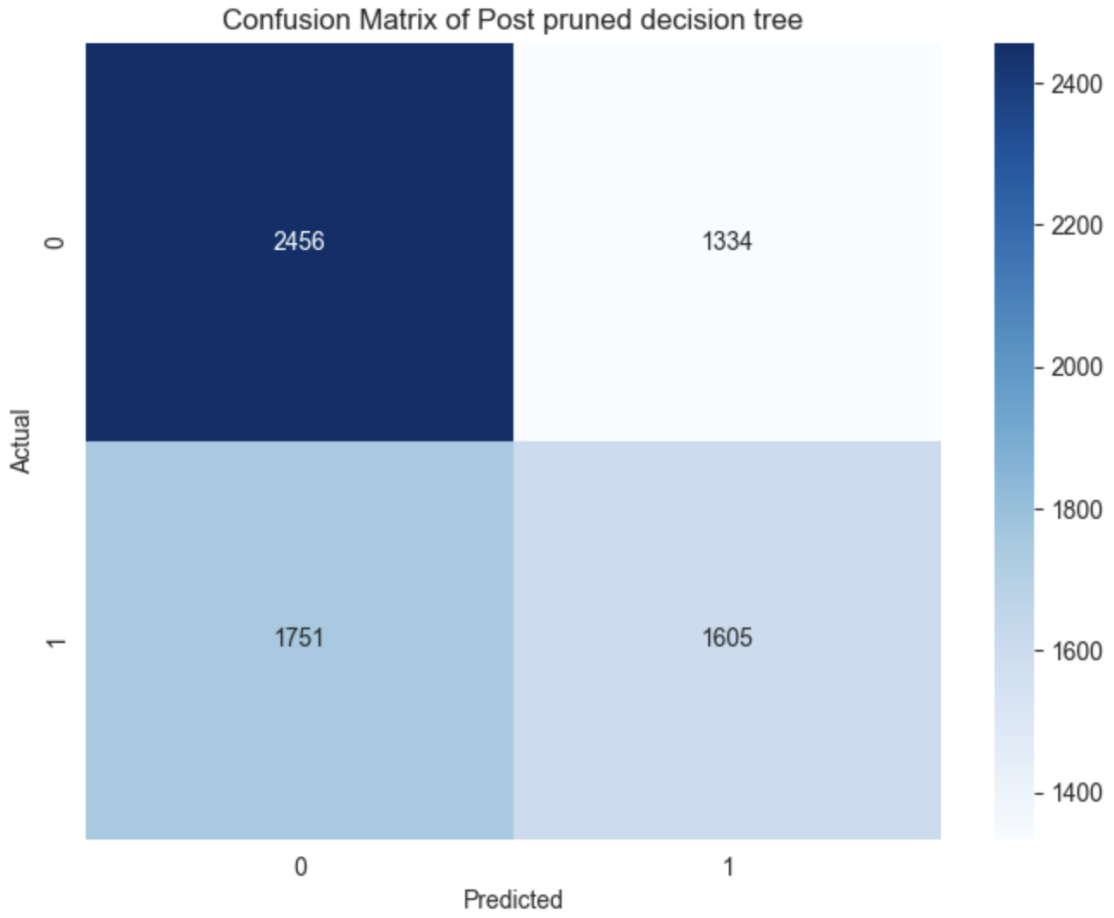
VIRGINIA TECH

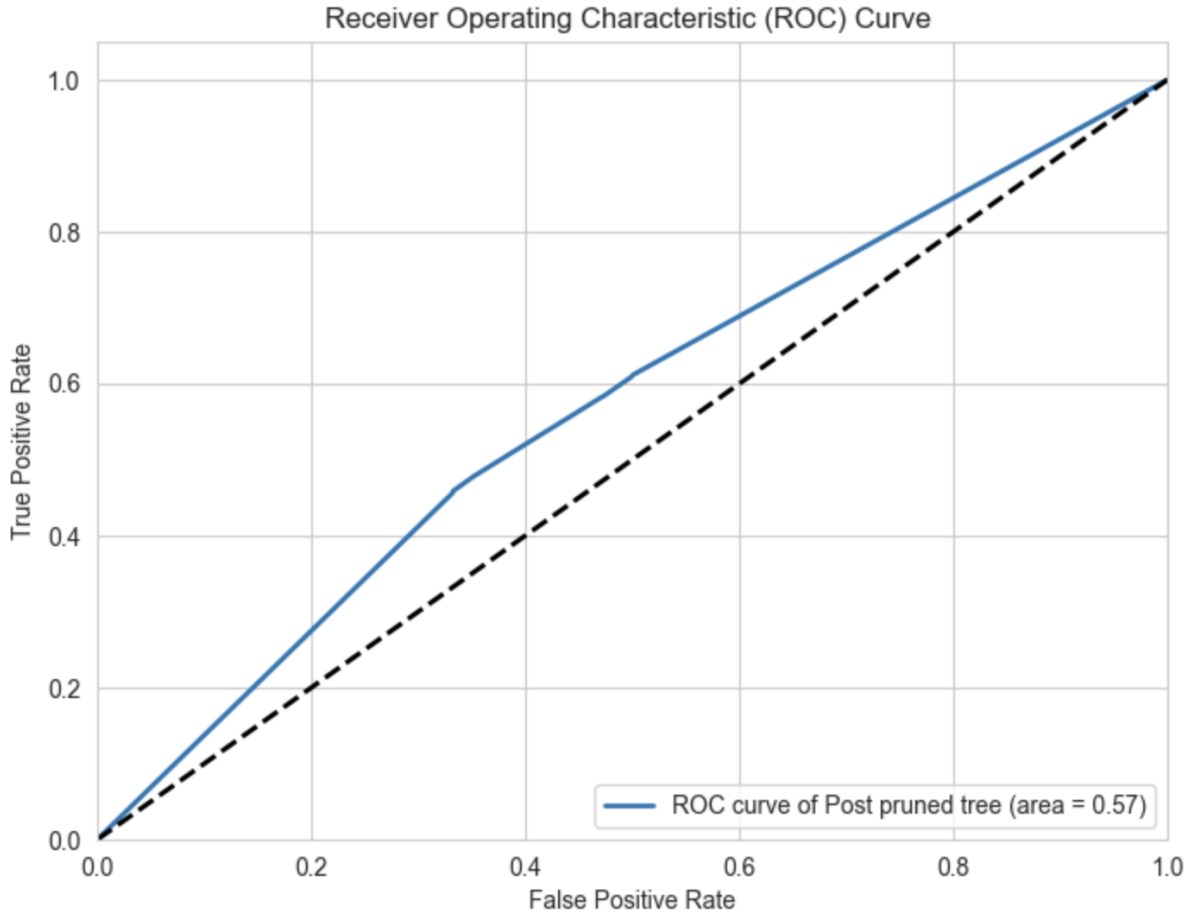
```
Best alpha value of post pruned decision tree: 0.00014000098283027543
```

Classifier	Accuracy	Confusion Matrix	Precision	Sensitivity or Recall	Specificity	F-score	AUC
Post pruned Decision Tree Classifier	0.57	[[2456 1334] [1751 1605]]	0.55	0.48	0.8	0.51	0.57



Post-pruned Decision Tree

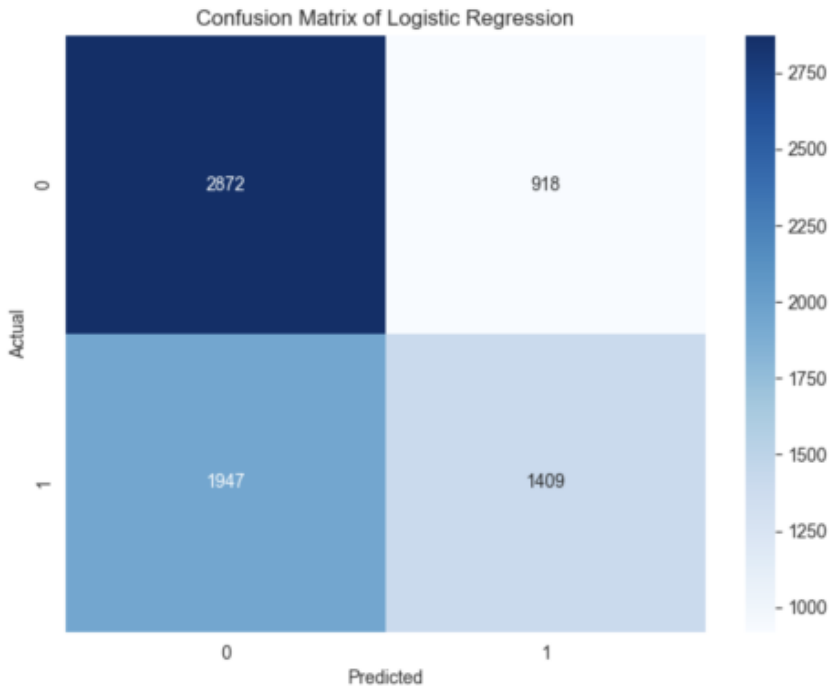


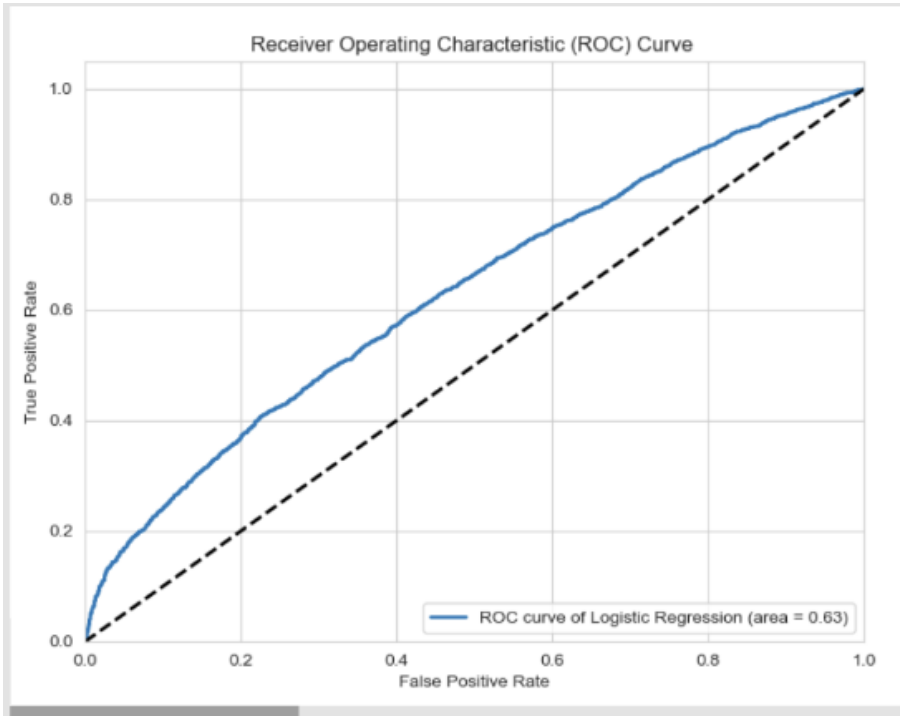


Logistic Regression

```
Best Hyperparameters: {'C': 0.1, 'penalty': 'l2', 'solver': 'saga'}
```

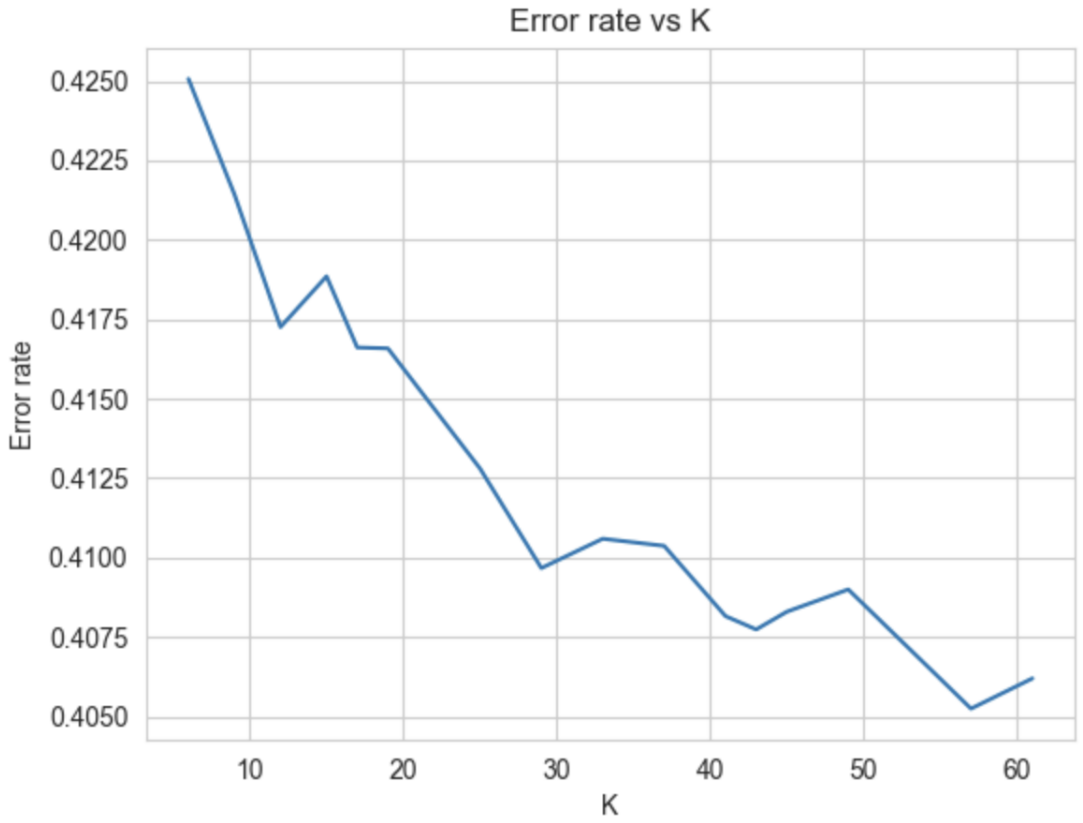
Classifier	Accuracy	Confusion Matrix	Precision	Sensitivity or Recall	Specificity	F-score	AUC
Logistic Regression	0.59	[[2872 918] [1947 1409]]	0.61	0.42	0.76	0.5	0.63



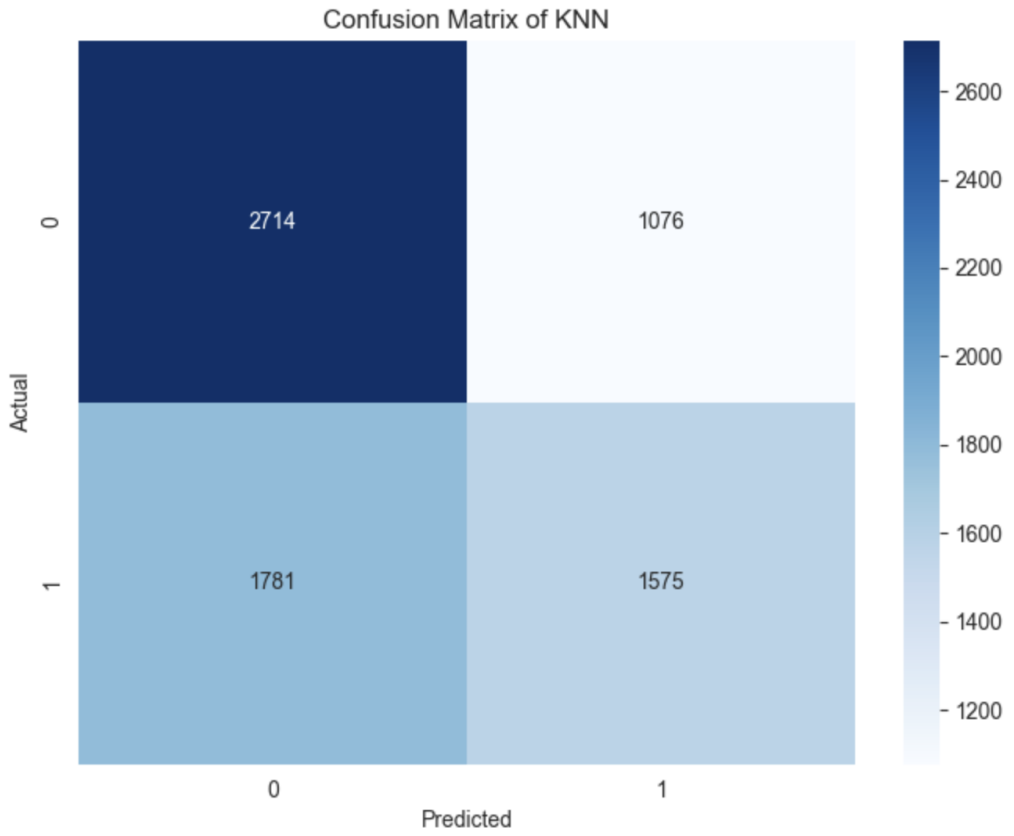


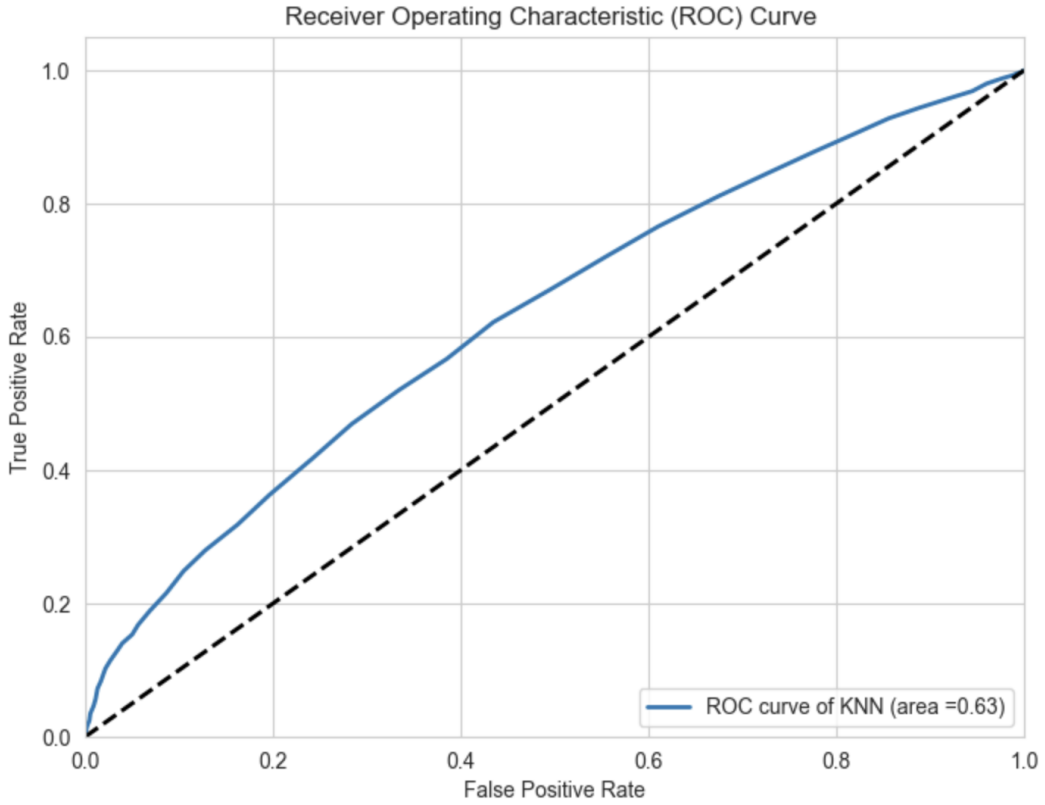
K- Nearest Neighbors

```
Best Hyperparameters: {'n_neighbors': 57}
```



K- Nearest Neighbors





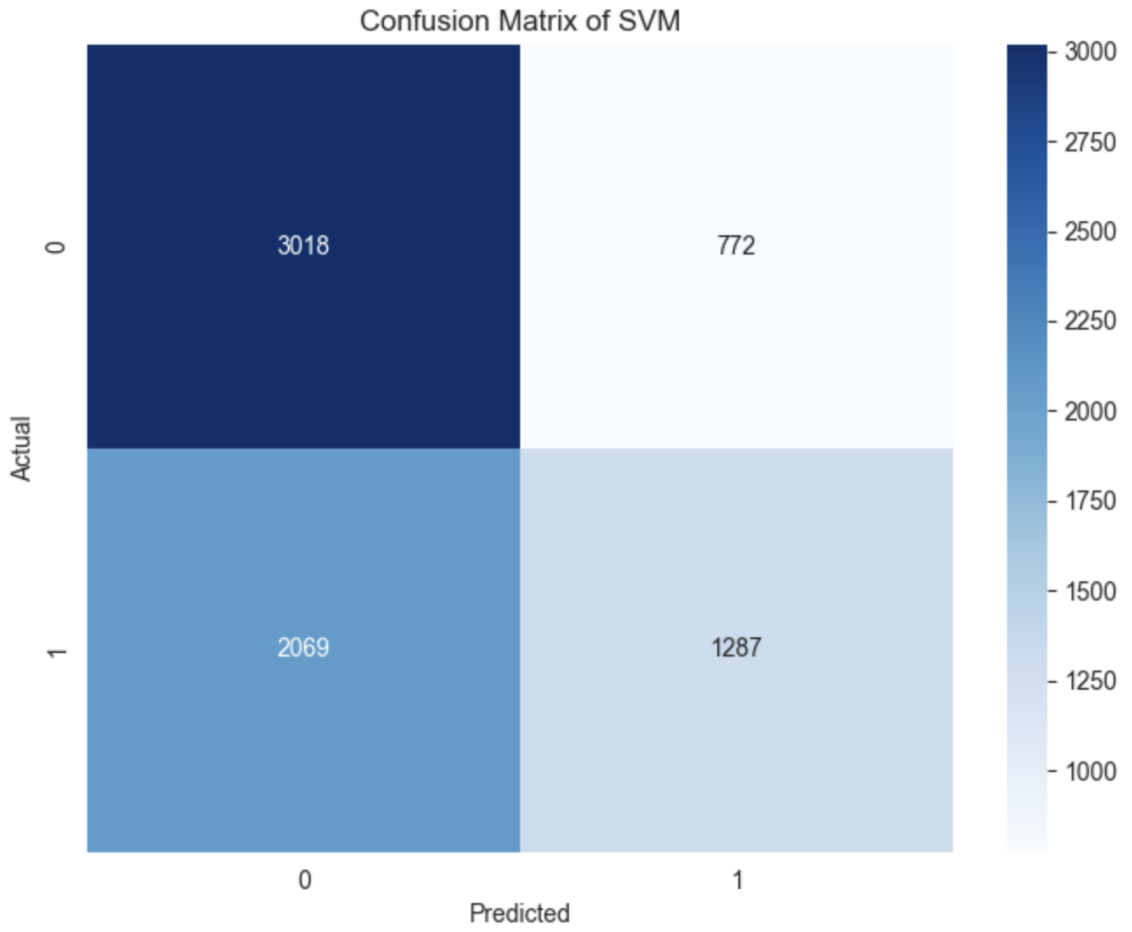
SVM

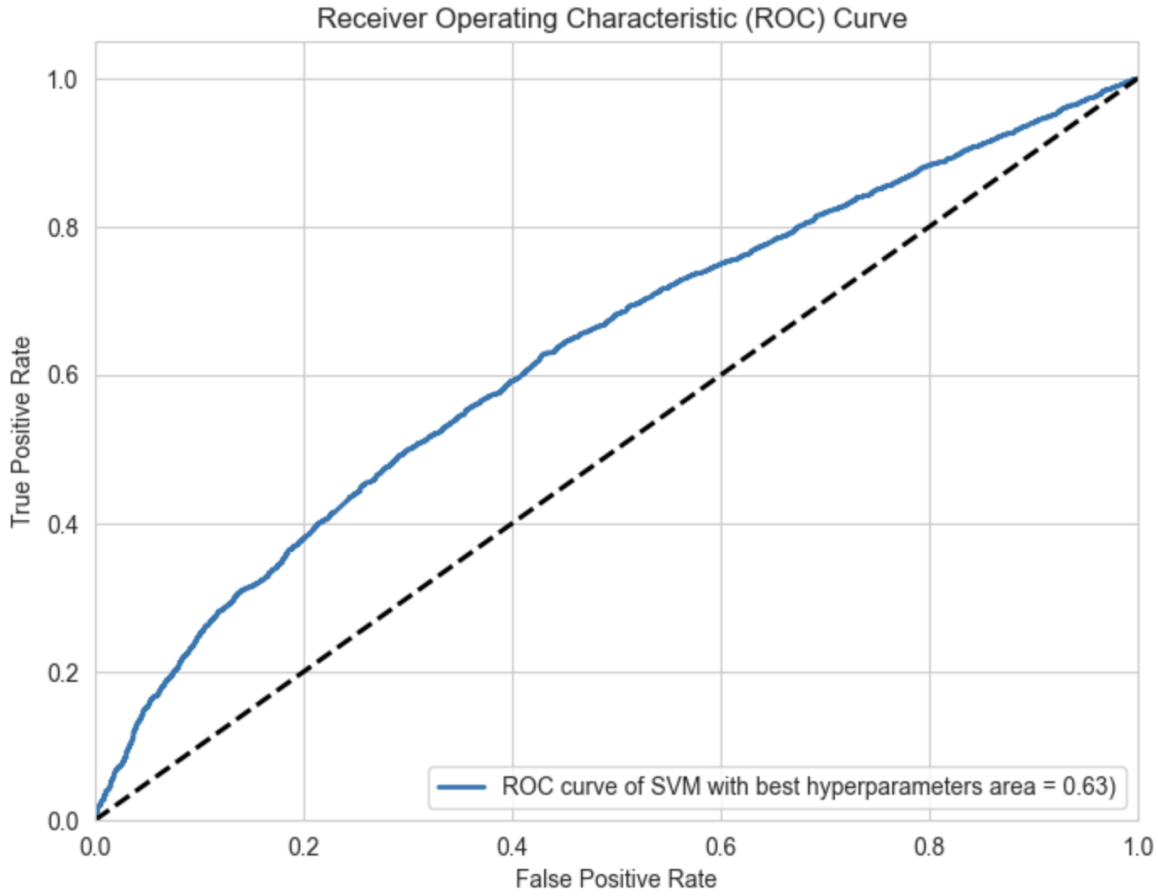
VIRGINIA TECH

```
Best Hyperparameters: {'kernel': 'rbf'}
```

Classifier	Accuracy	Confusion Matrix	Precision	Sensitivity or Recall	Specificity	F-score	AUC
SVM	0.6	[[3018 772] [2069 1287]]	0.63	0.38	0.8	0.48	0.63

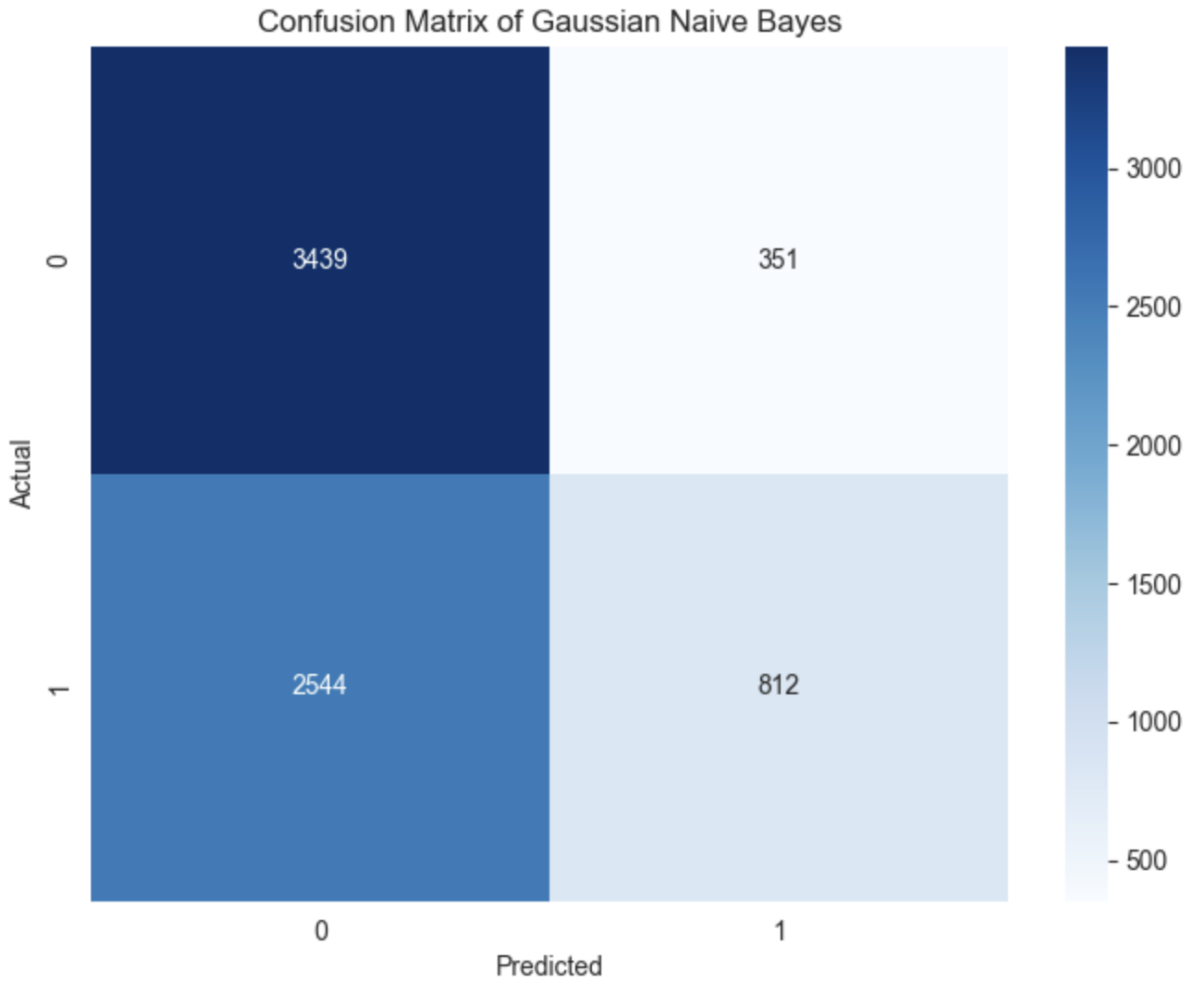
SVM

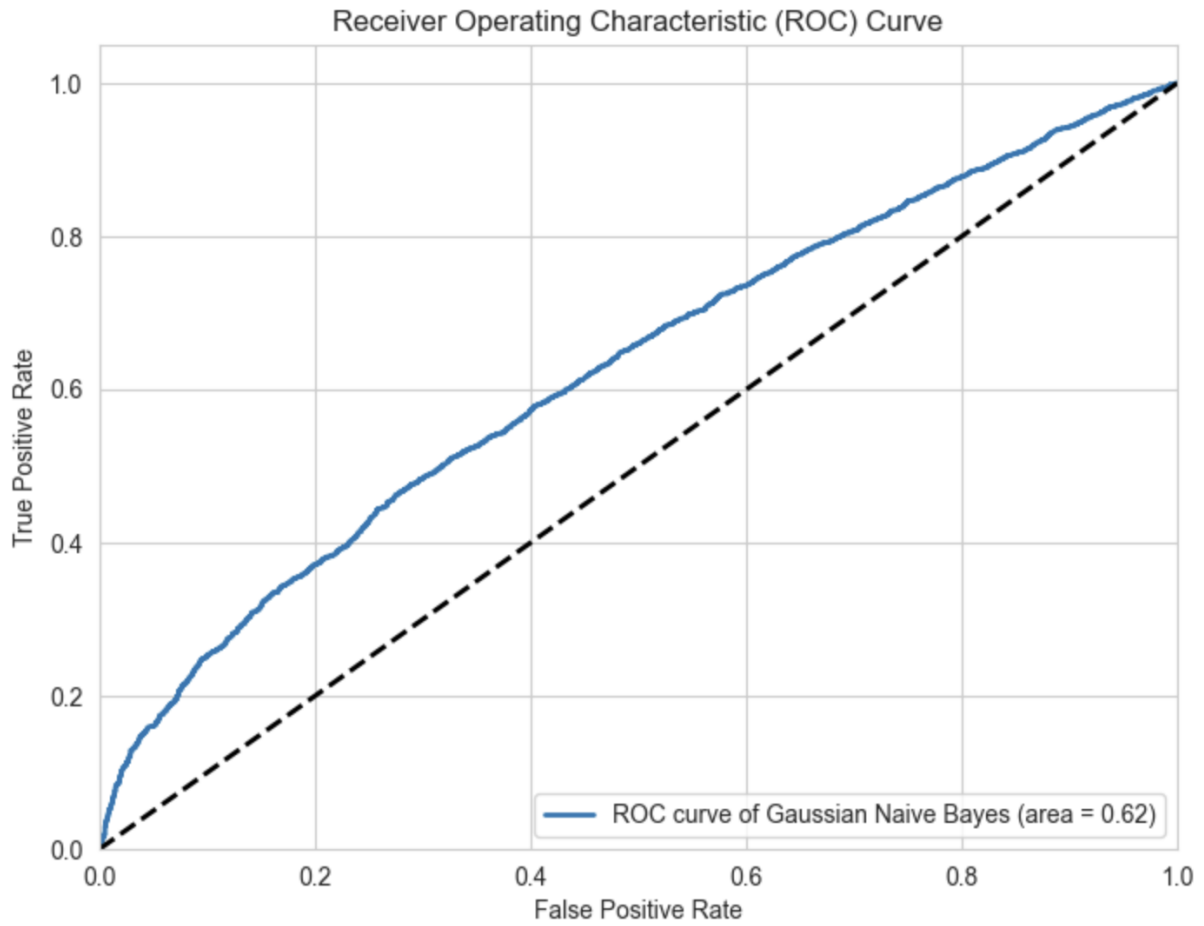

Confusion Matrix of SVM
3018
772
2069
1287
Actual
0
1
Predicted
0
1
3000
2750
2500
2250
2000
1750
1500
1250
1000


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of SVM with best hyperparameters area = 0.63)

Naïve Bayes

Classifier	Accuracy	Confusion Matrix	Precision	Sensitivity or Recall	Specificity	F-score	AUC
Gaussian Naive Bayes	0.58	[[3439 351] [2544 812]]	0.7	0.24	0.91	0.36	0.62



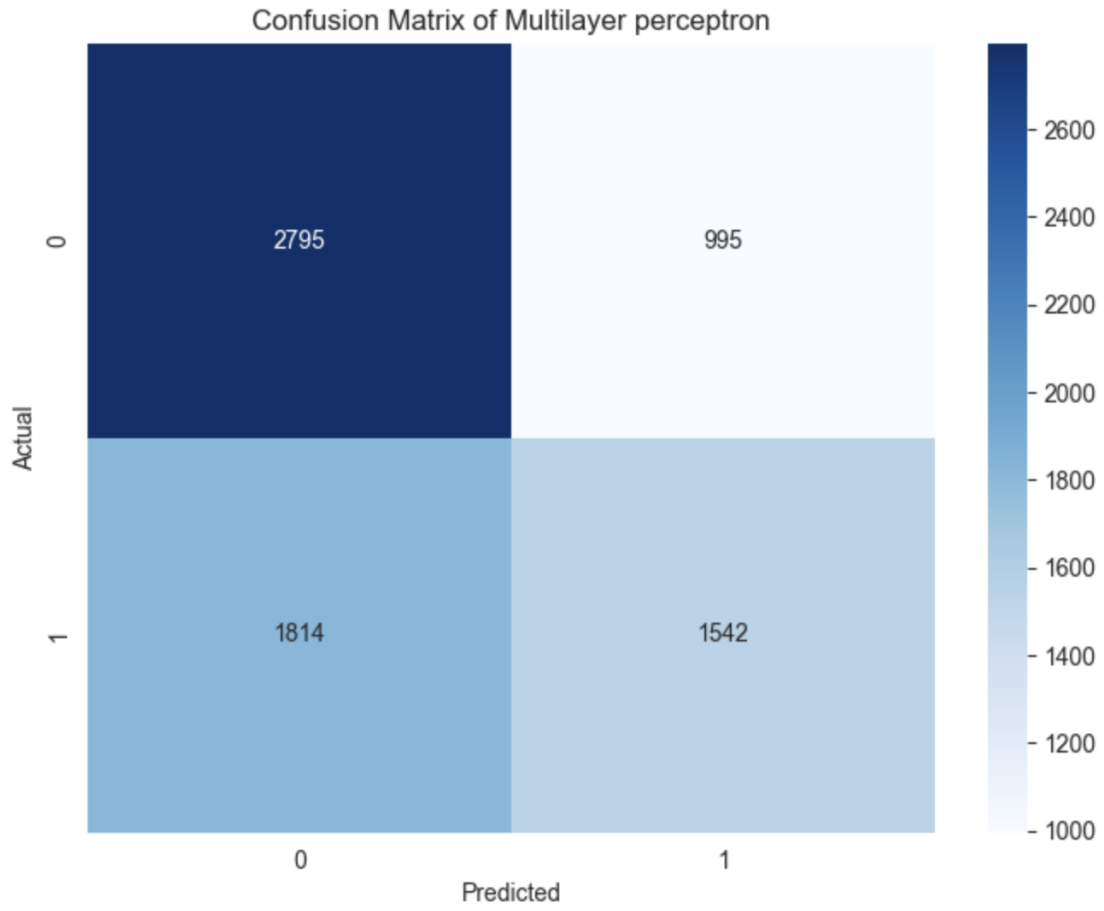


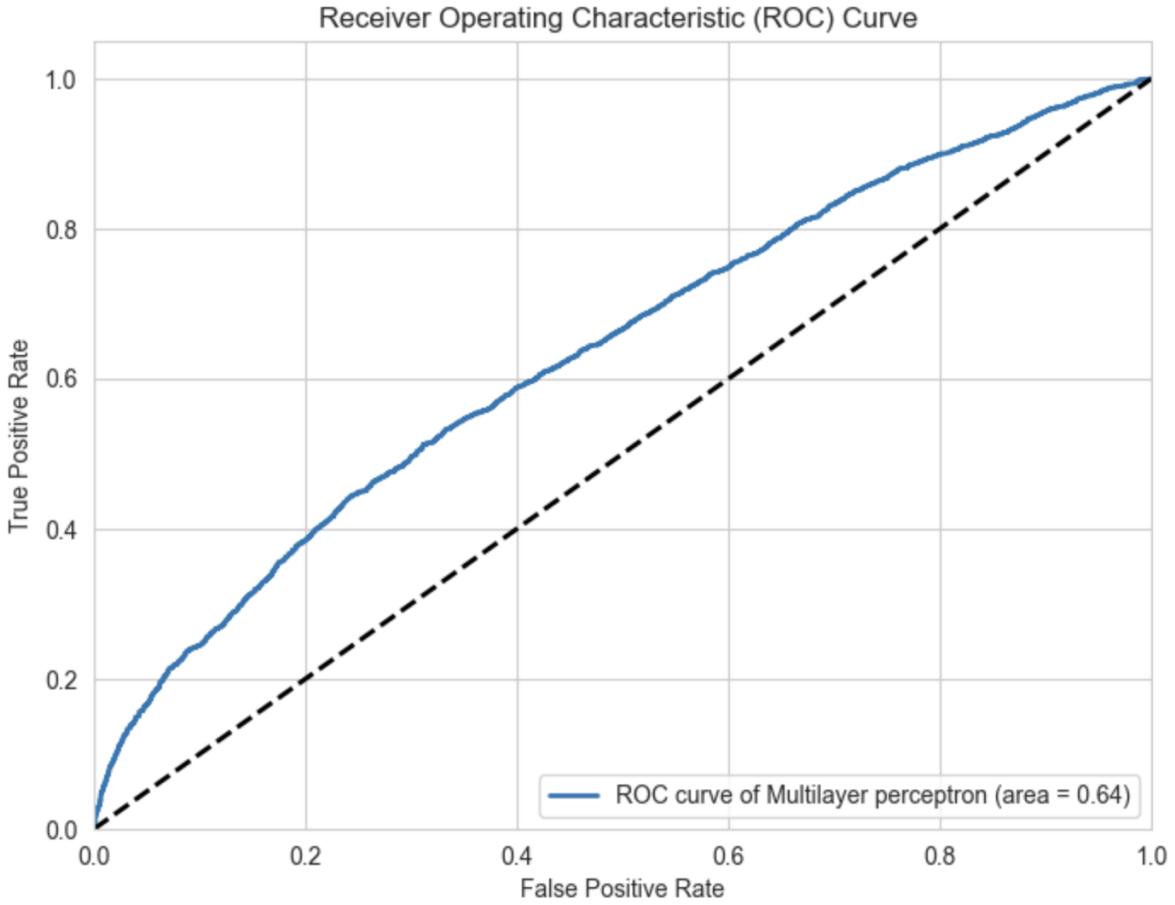
Multi-layered perceptron

```
Best Hyperparameters: {'hidden_layer_sizes': (15, 15)}
```

Classifier	Accuracy	Confusion Matrix	Precision	Sensitivity or Recall	Specificity	F-score	AUC
Multilayer perceptron	0.61	[[2795 995] [1814 1542]]	0.61	0.46	0.74	0.52	0.64

Multi-layered perceptron



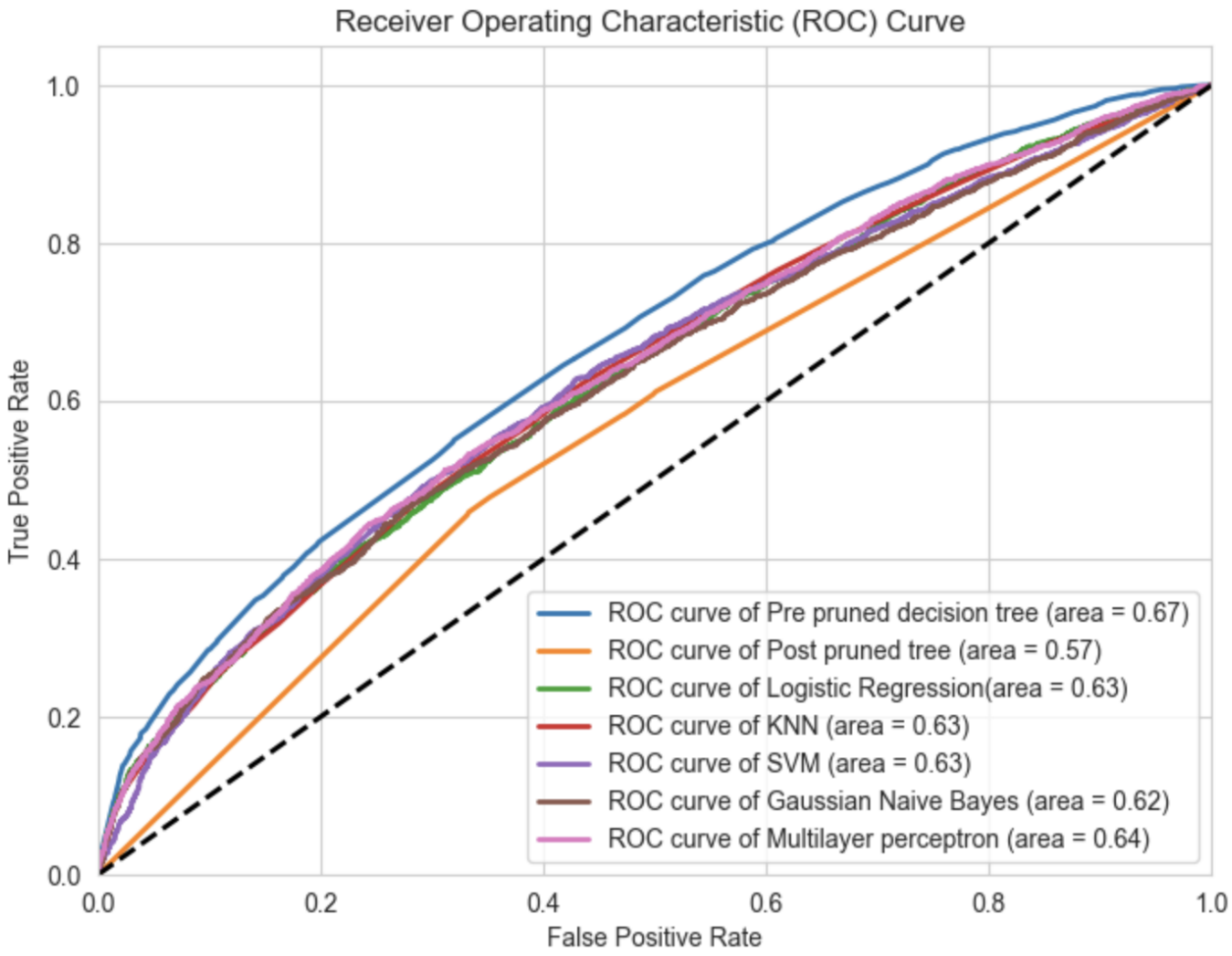


Comparing all classifier models

VIRGINIA TECH

Classifier	Accuracy	Confusion Matrix	Precision	Sensitivity or Recall	Specificity	F-Score	AUC
Pre pruned Decision Tree Classifier	0.62	[[3047 743], [1956 1400]]	0.65	0.42	0.8	0.51	0.67
Post pruned Decision Tree Classifier	0.57	[[2456 1334], [1751 1605]]	0.55	0.48	0.8	0.51	0.57
Logistic Regression	0.6	[[2872 918], [1947 1409]]	0.61	0.42	0.76	0.5	0.63
KNN	0.6	[[2714 1076], [1781 1575]]	0.59	0.47	0.72	0.52	0.63
SVM	0.6	[[3018 772], [2069 1287]]	0.63	0.38	0.8	0.48	0.63
Gaussian Naive Bayes	0.58	[[3439 351], [2544 812]]	0.7	0.24	0.91	0.36	0.62
Multilayer Perceptron	0.61	[[2795 995], [1814 1542]]	0.61	0.46	0.74	0.52	0.64

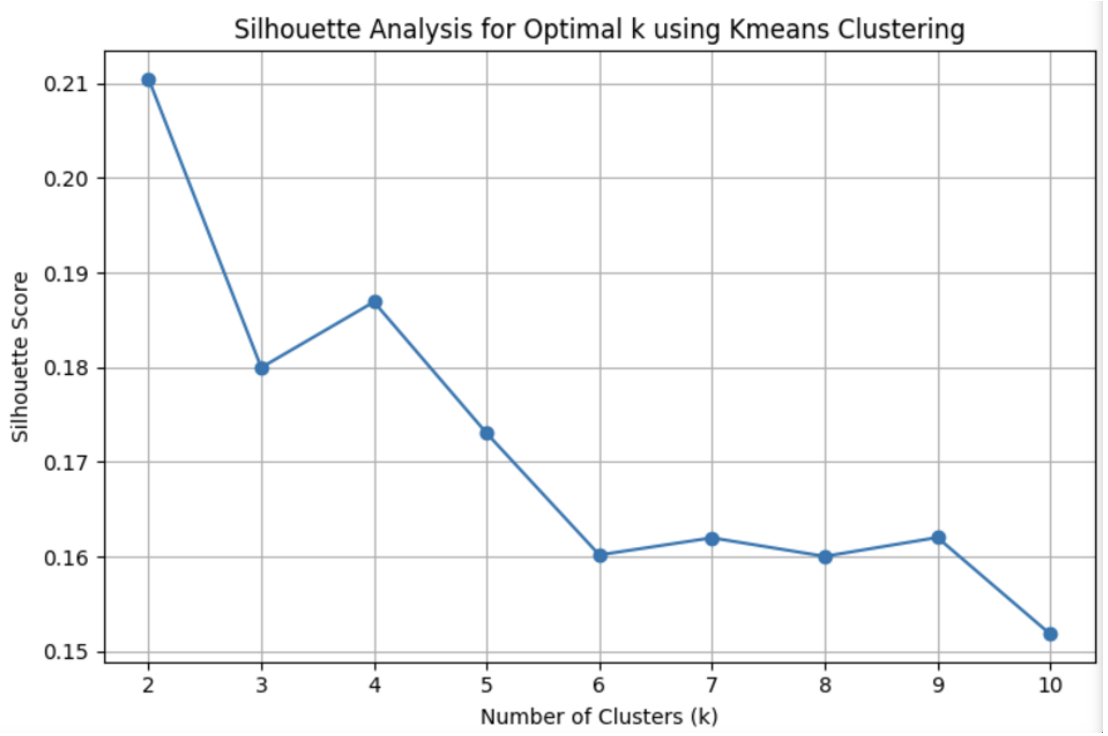
Best Model? – Pre-pruned decision tree

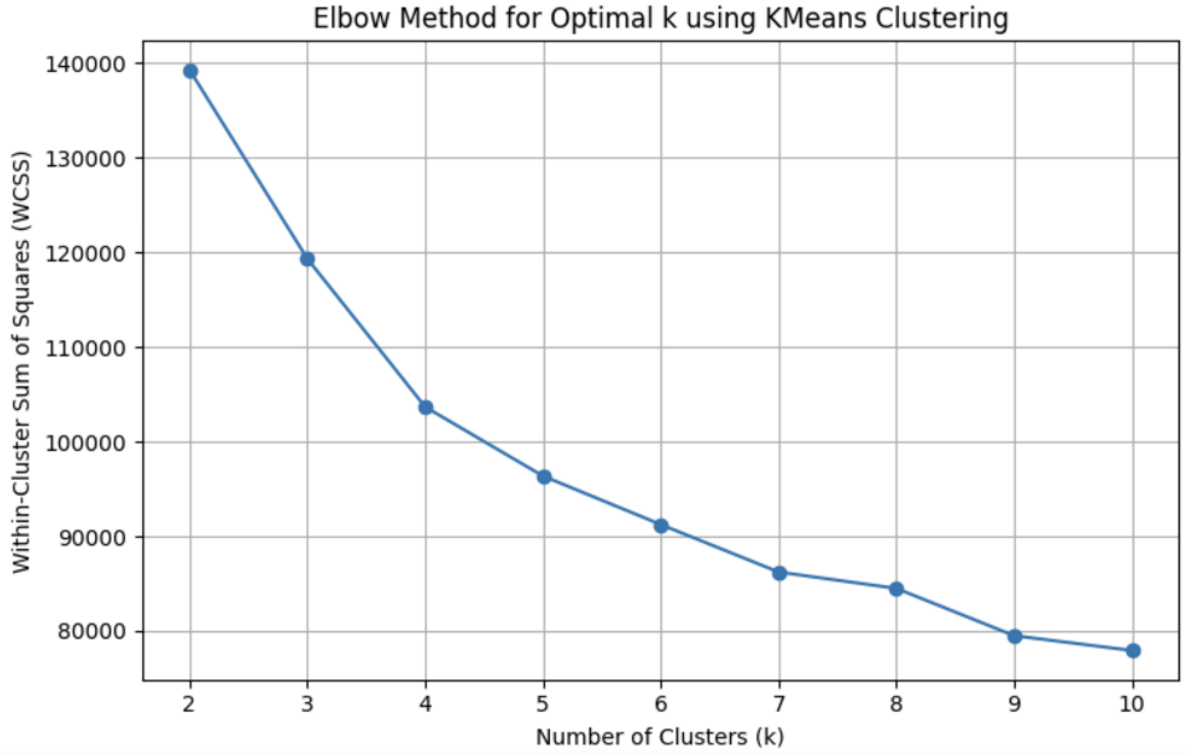


Clustering and Association:

- K-Means Clustering
  - Silhouette analysis
  - *WCSS plot analysis*
- DBSCAN algorithm
- Apriori algorithm

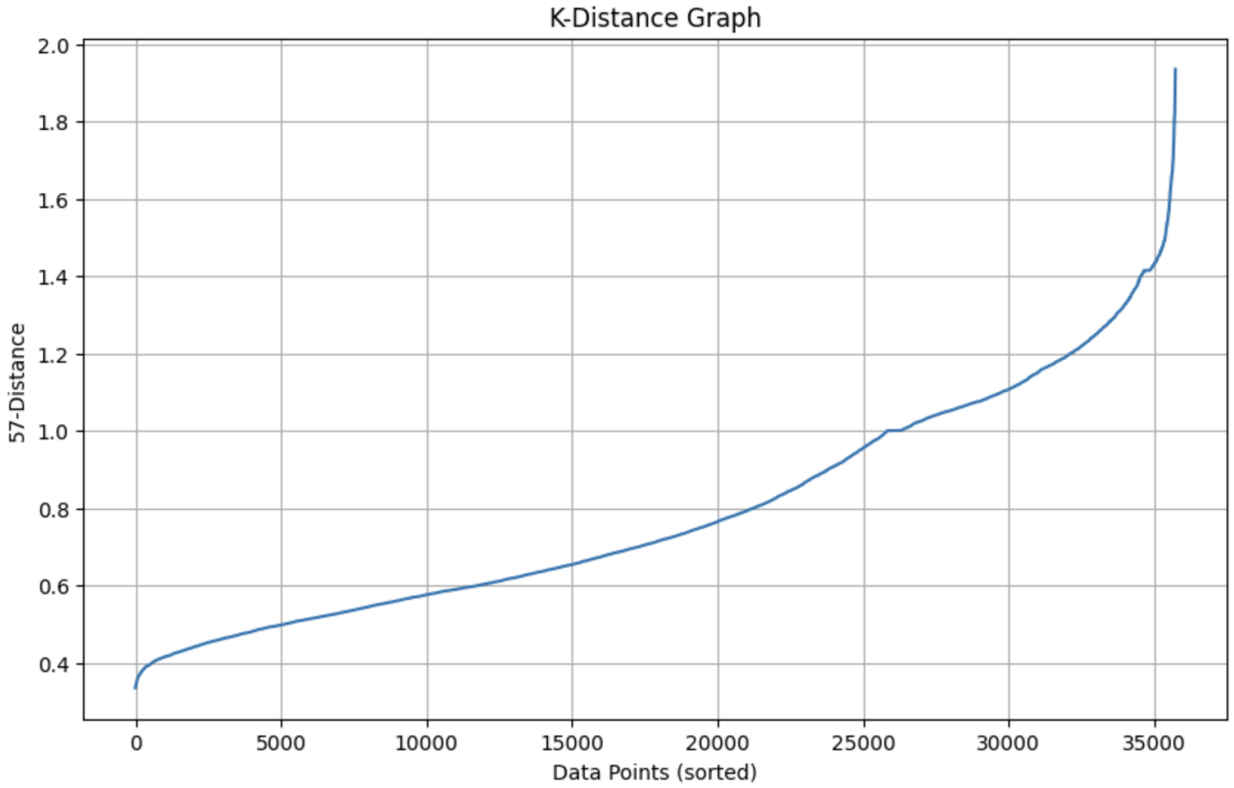
K-Means Clustering





Optimal K- value = 4

DBSCAN algorithm



eps = 1.3

```
Number of clusters: 5
Number of noise points: 17
Silhouette Score: 0.057074034315654 88
```

Apriori Algorithm

VIRGINIA TECH

Processing 4 combinations | Sampling itemset size 4

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
36	(Large Aircraft)	(Short Length)	0.18	0.45	0.12	0.65	1.45	0.04	1.59	0.38
64	(Average Length, StartOfWeek)	(Small Aircraft)	0.18	0.51	0.10	0.56	1.10	0.01	1.12	0.12
9	(Average Length)	(Small Aircraft)	0.43	0.51	0.24	0.56	1.10	0.02	1.12	0.17
3	(Average Length)	(Early Morning)	0.43	0.52	0.23	0.55	1.05	0.01	1.06	0.09
58	(Early Morning, Average Length)	(Small Aircraft)	0.23	0.51	0.13	0.54	1.08	0.01	1.08	0.09

The presence of "Large Aircraft" (18% of occurrences) is associated with "Short Length" (45% of occurrences) in 12% of cases.

The confidence is moderate at 65%, indicating that when there's a large aircraft, there's a 65% chance that the flight will be of short length.

The lift value of 1.45 implies that the occurrence of a large aircraft increases the chance of a short-length flight by 45% compared to what would be expected if the two were independent.

VIRGINIA TECH

THANK YOU