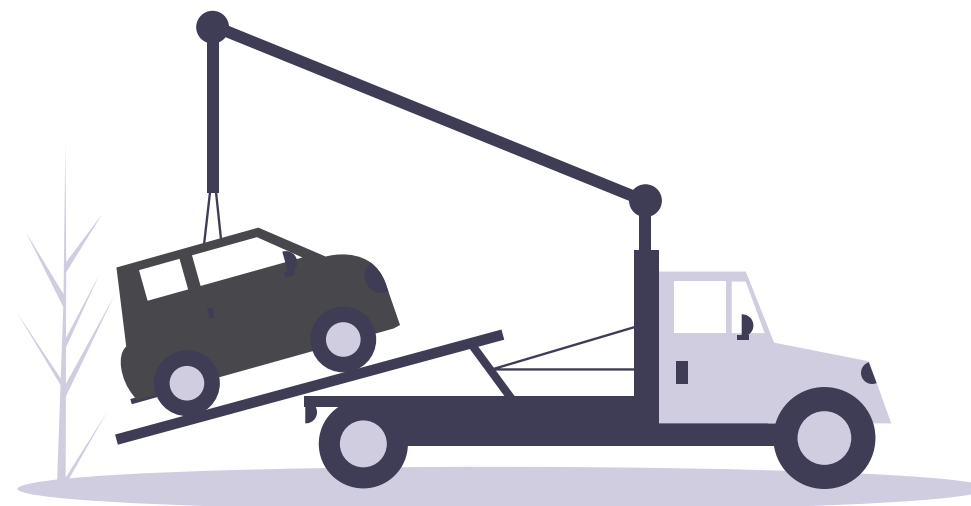


ACCIDENT SEVERITY PREDICTION

MATTEO ORSINI 1795119

FABRIZIO ROSSI 1815023

MINA MAKAR 1804475



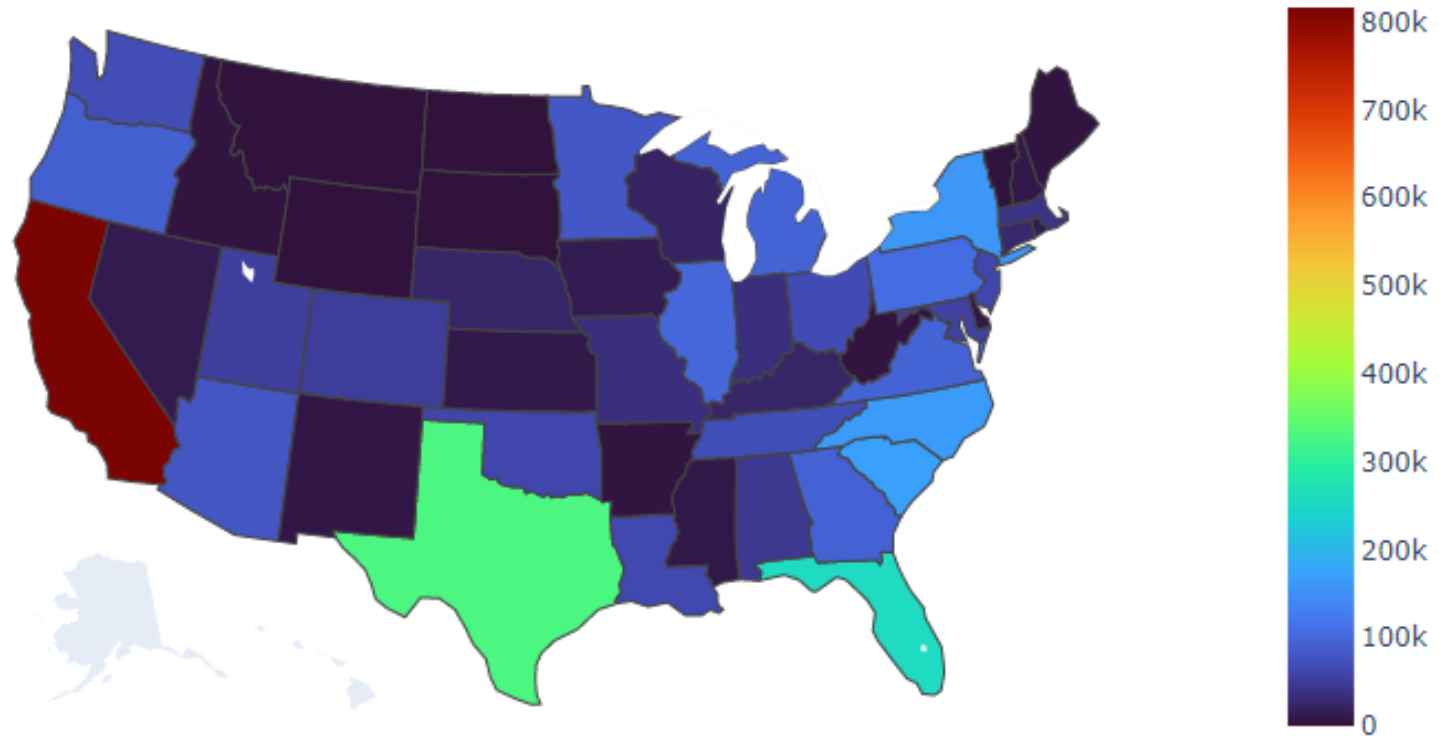
TASK - SEVERITY

- Classify the **impact of an accident** on the traffic (severity)
- The scale of the severity goes from 1 (low impact) to 4 (high impact)
- We used the dataset [«US Accidents»](#) from Kaggle
- Contains about **3.5 million** accidents collected in the US

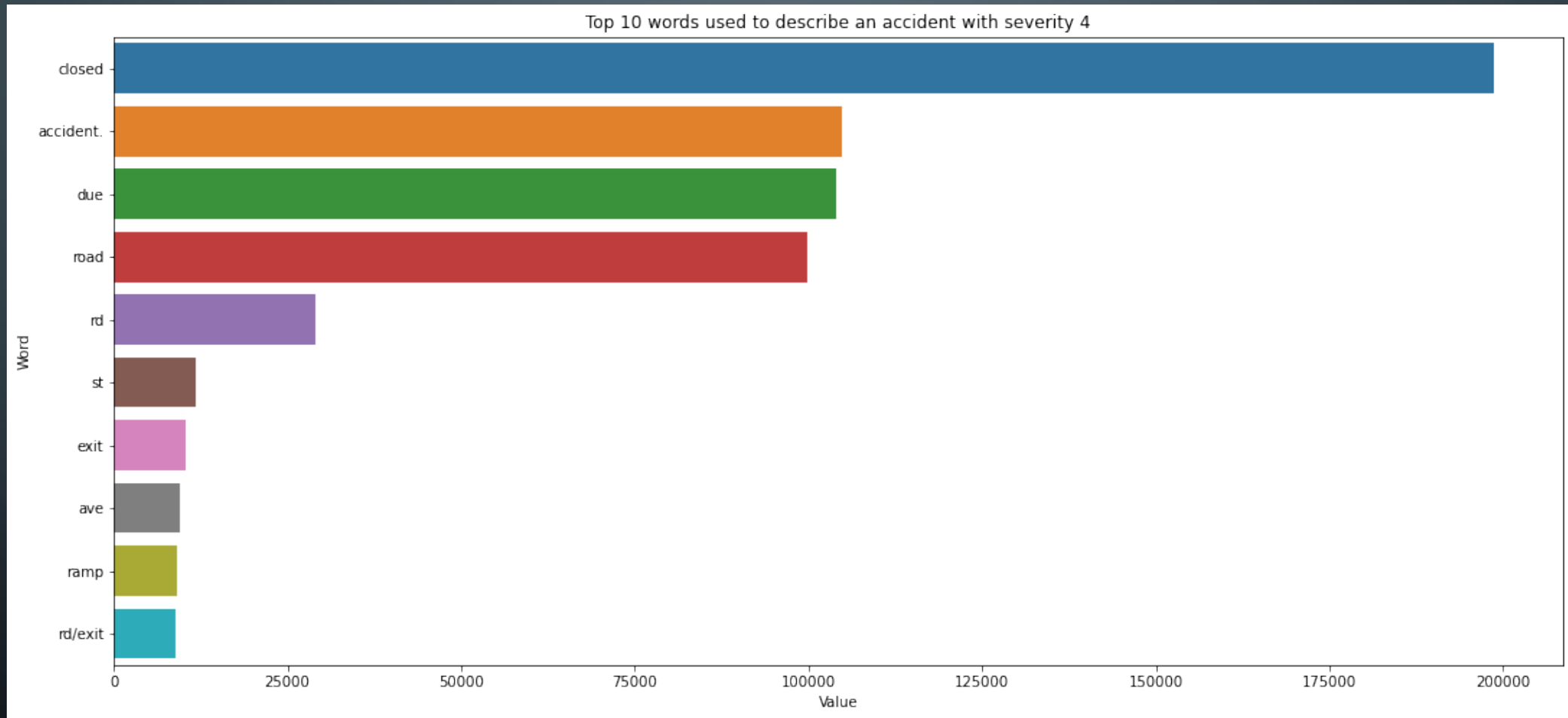


EXPLORATORY DATA ANALYSIS

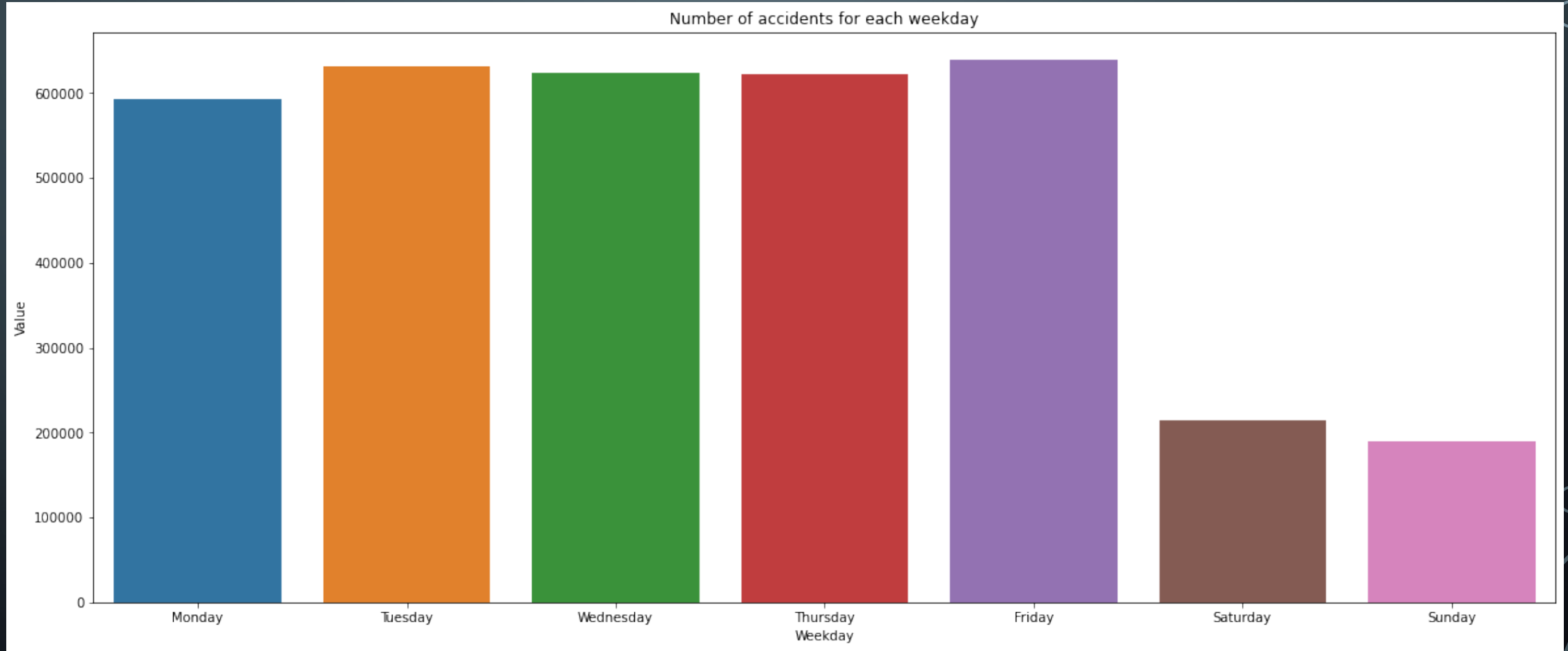
Number of US Accidents for each State



EXPLORATORY DATA ANALYSIS

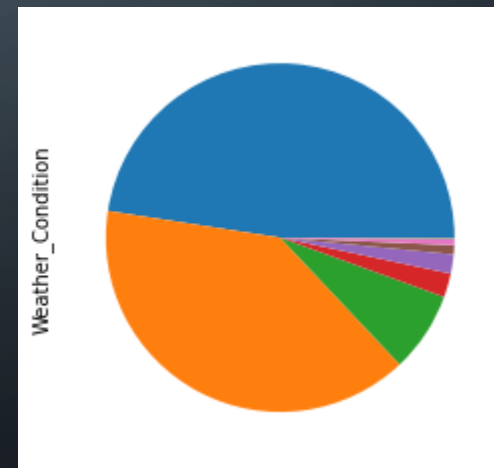
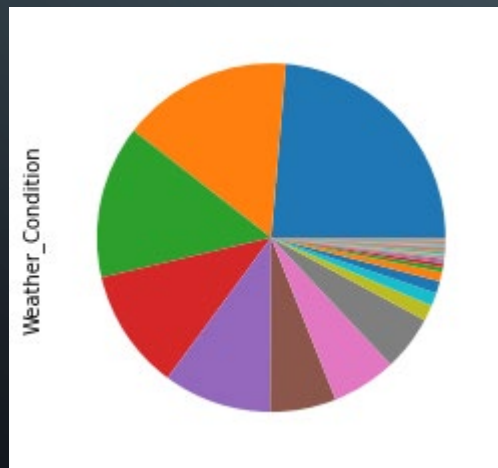


EXPLORATORY DATA ANALYSIS



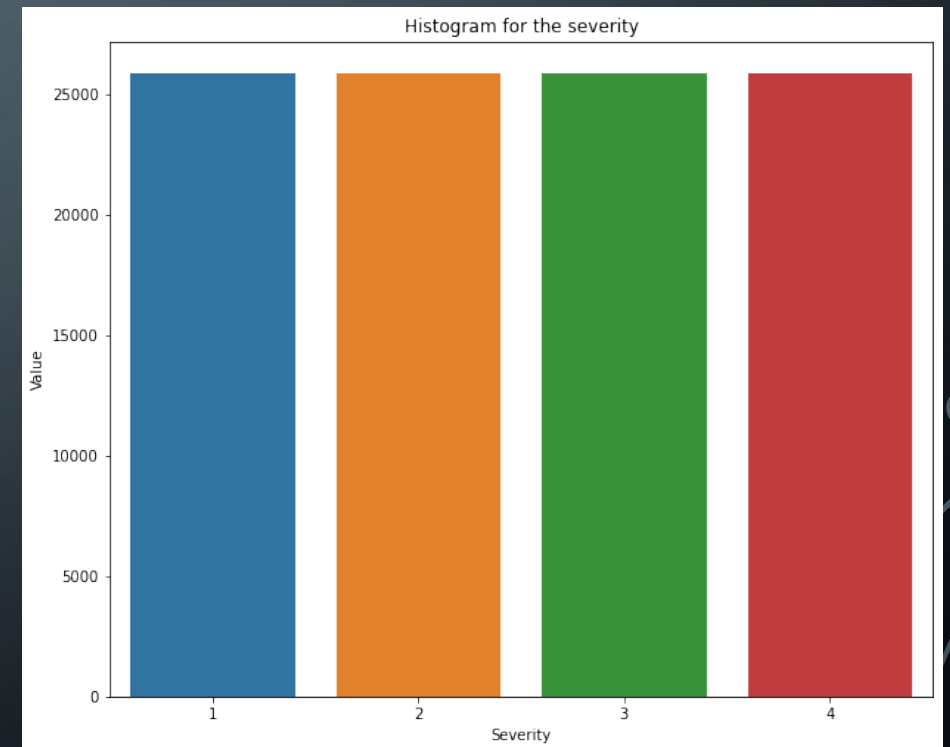
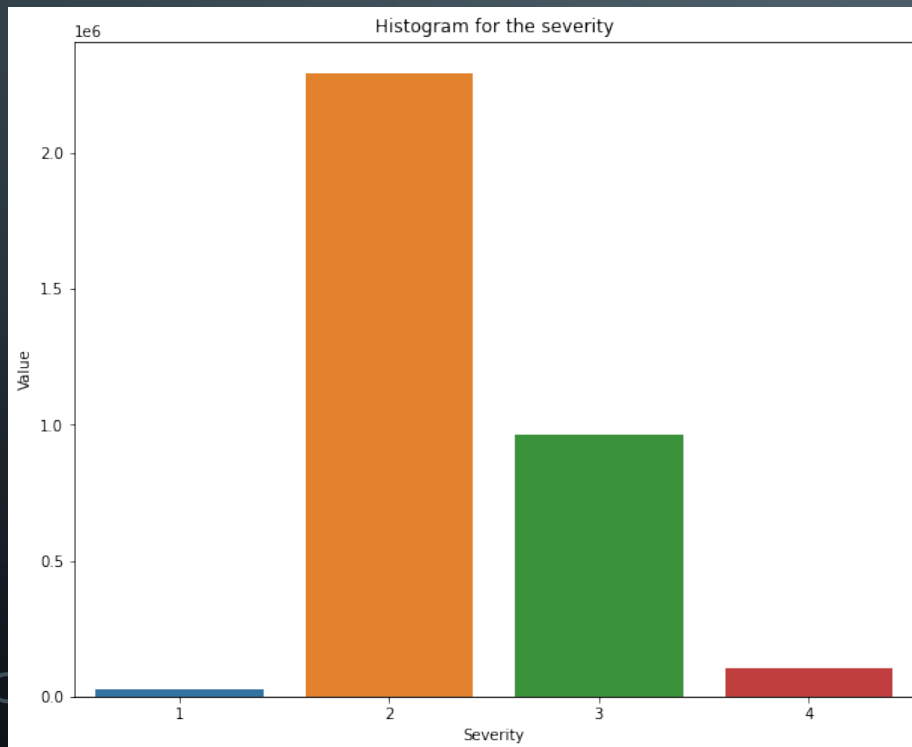
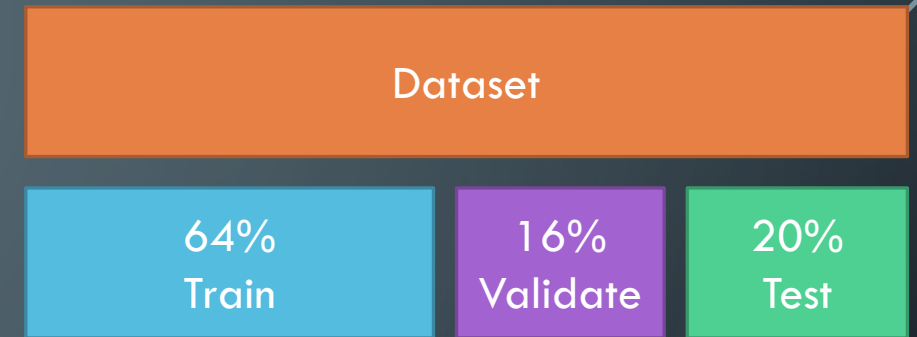
DATA PREPROCESSING

- Reduced number of classes for Weather_Condition (128 \rightarrow 11) and Wind_Direction (24 \rightarrow 10)
- Filled missing values with mean for numerical features
- Removed records with missing values for categorical features
- Scaled and encoded features using the one-hot encoding for the categorical features
- City was encoded using the binary encoder due to the large number of unique values

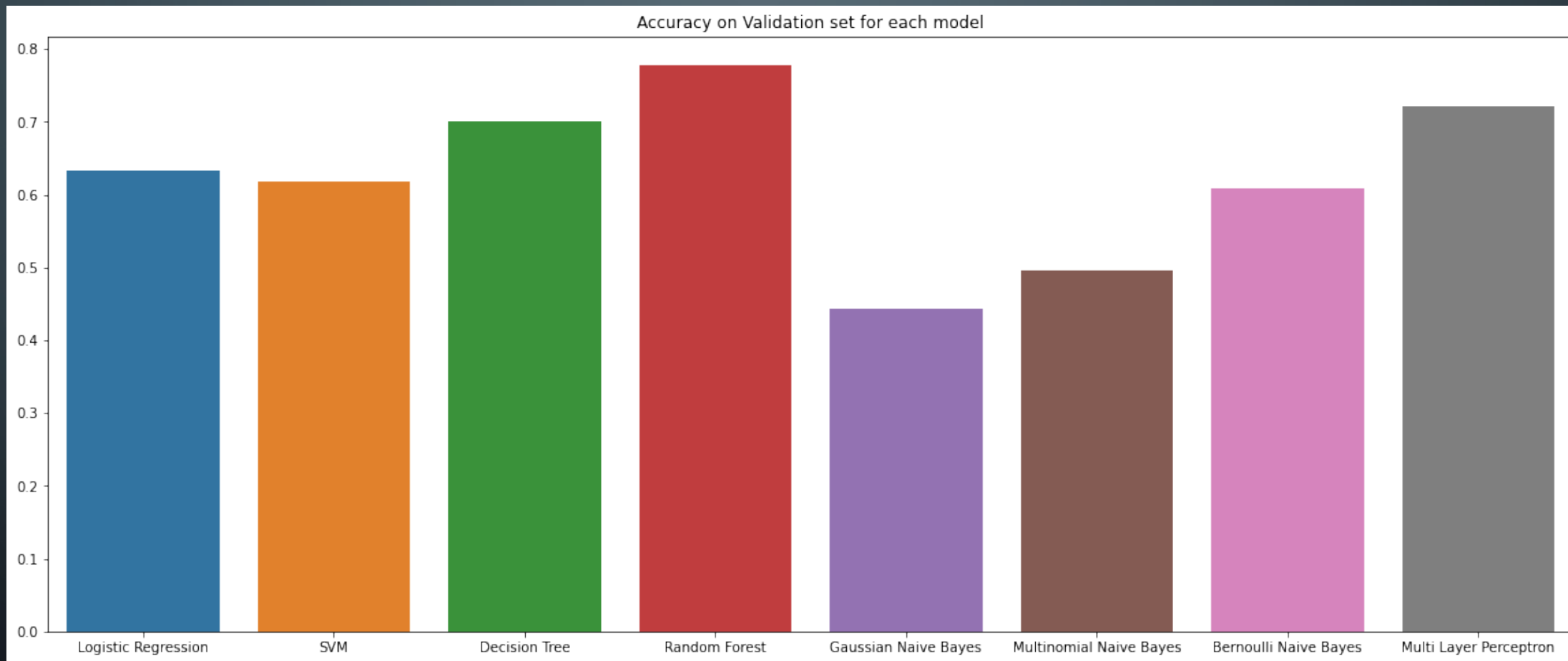


DATA PREPROCESSING

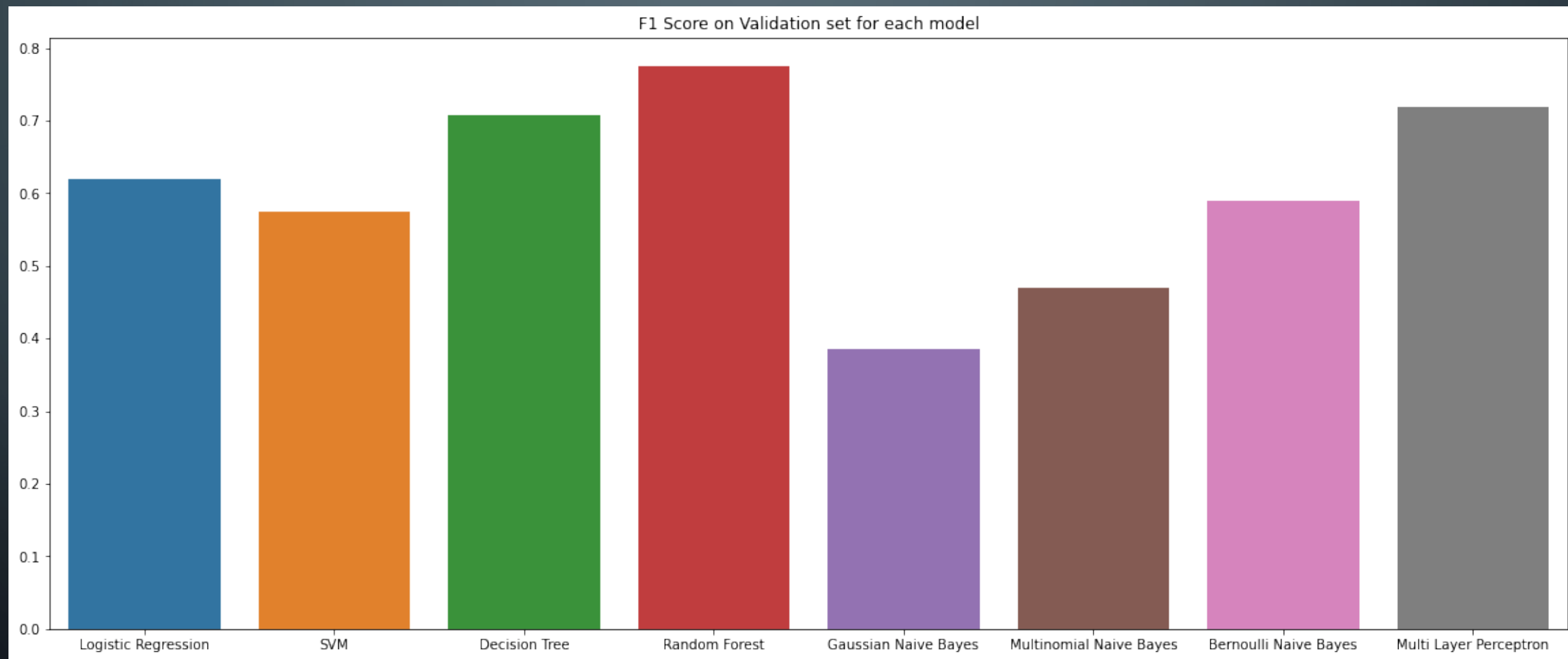
- We handled the unbalanced dataset problem using the undersampling technique
- We splitted the dataset in **train set**, **validation set** and **test set**



RESULTS – ACCURACY



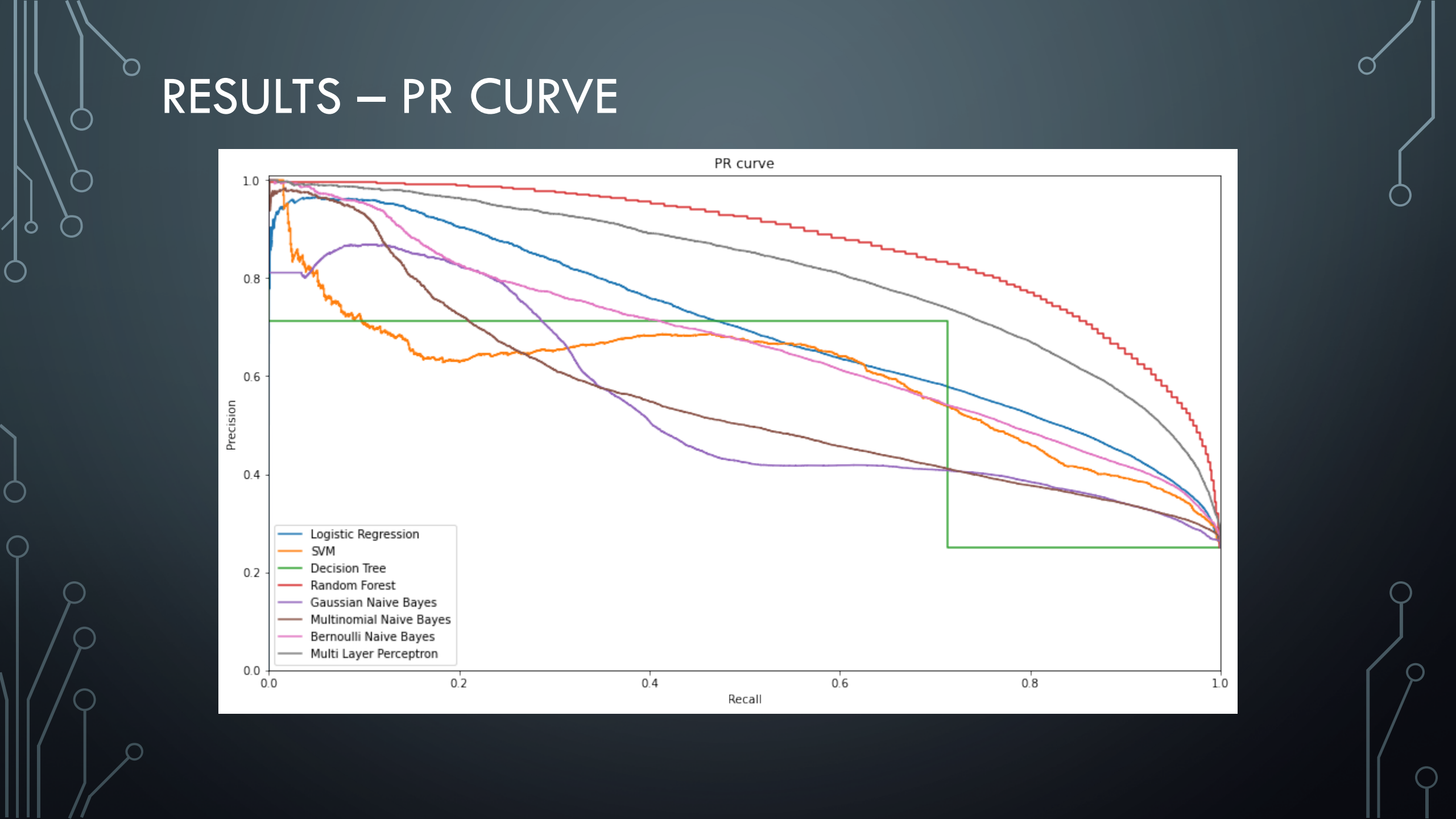
RESULTS – F1 SCORE



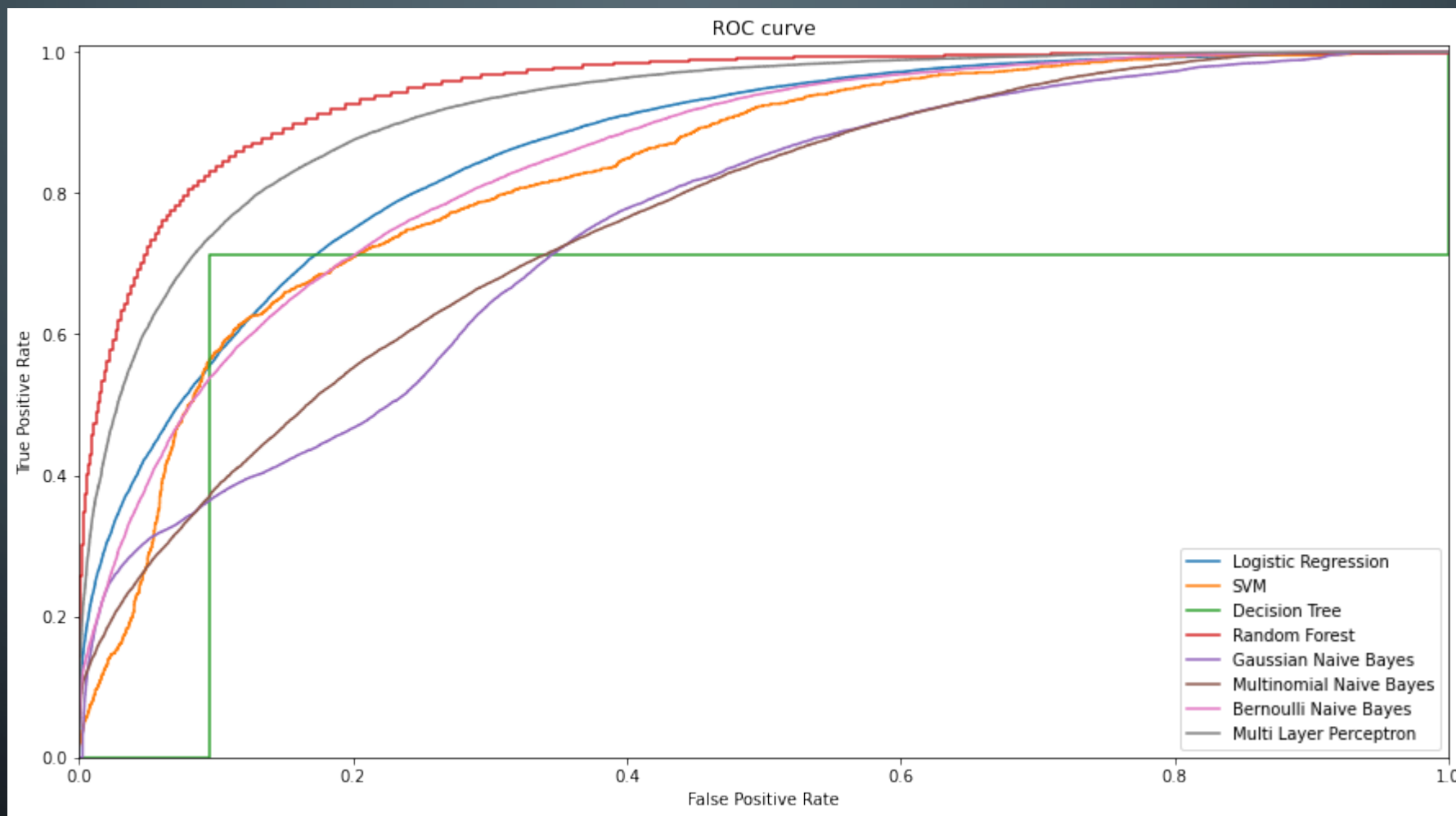
RESULTS – PR CURVE

The PR curve plot displays the performance of eight machine learning models. The y-axis represents Precision (0.0 to 1.0) and the x-axis represents Recall (0.0 to 1.0). A green rectangle highlights the region where Precision is between 0.25 and 0.75 and Recall is between 0.75 and 0.95. The models are ranked by their performance in this region, with Random Forest (red line) showing the highest precision and recall, followed by Multi Layer Perceptron (grey line), Logistic Regression (blue line), Bernoulli Naive Bayes (pink line), SVM (orange line), Multinomial Naive Bayes (brown line), Gaussian Naive Bayes (purple line), and Decision Tree (green line).

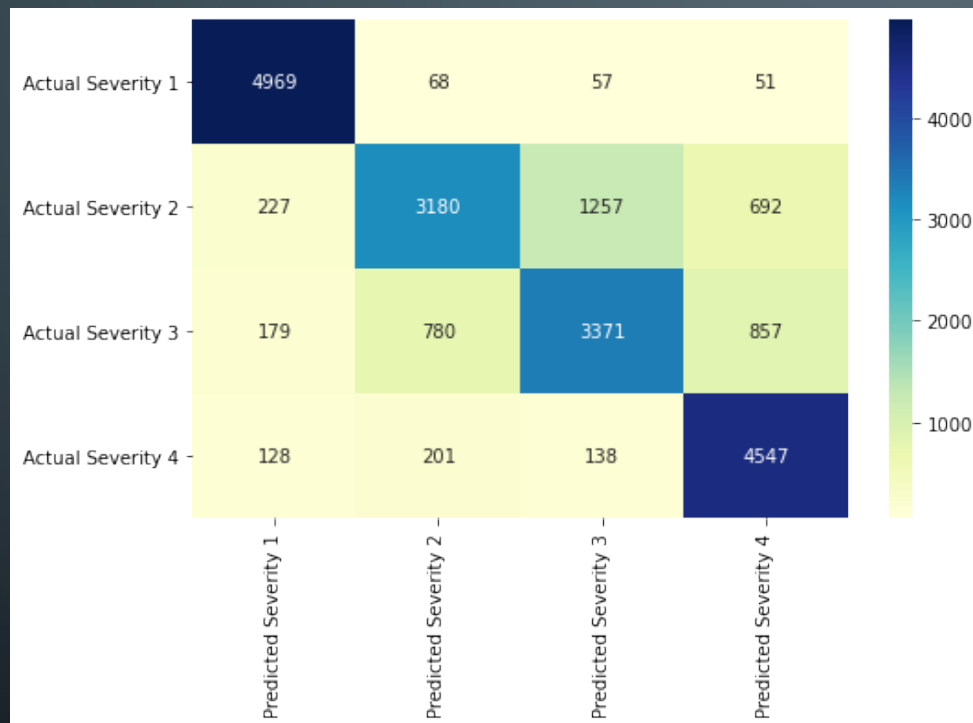
Model	Color	Approx. Precision at Recall 0.95
Random Forest	Red	0.75
Multi Layer Perceptron	Grey	0.65
Logistic Regression	Blue	0.55
Bernoulli Naive Bayes	Pink	0.50
SVM	Orange	0.45
Multinomial Naive Bayes	Brown	0.40
Gaussian Naive Bayes	Purple	0.35
Decision Tree	Green	0.30



RESULTS – ROC CURVE



RESULTS – RANDOM FOREST TEST SET



Class	Precision	Recall	F1-Score
1	0,90	0,97	0,93
2	0,75	0,59	0,66
3	0,70	0,65	0,67
4	0,74	0,91	0,81
Macro avg	0,77	0,78	0,77

RELATED WORKS

