ACCIDENT SEVERITY PREDICTION

MATTEO ORSINI 1795119

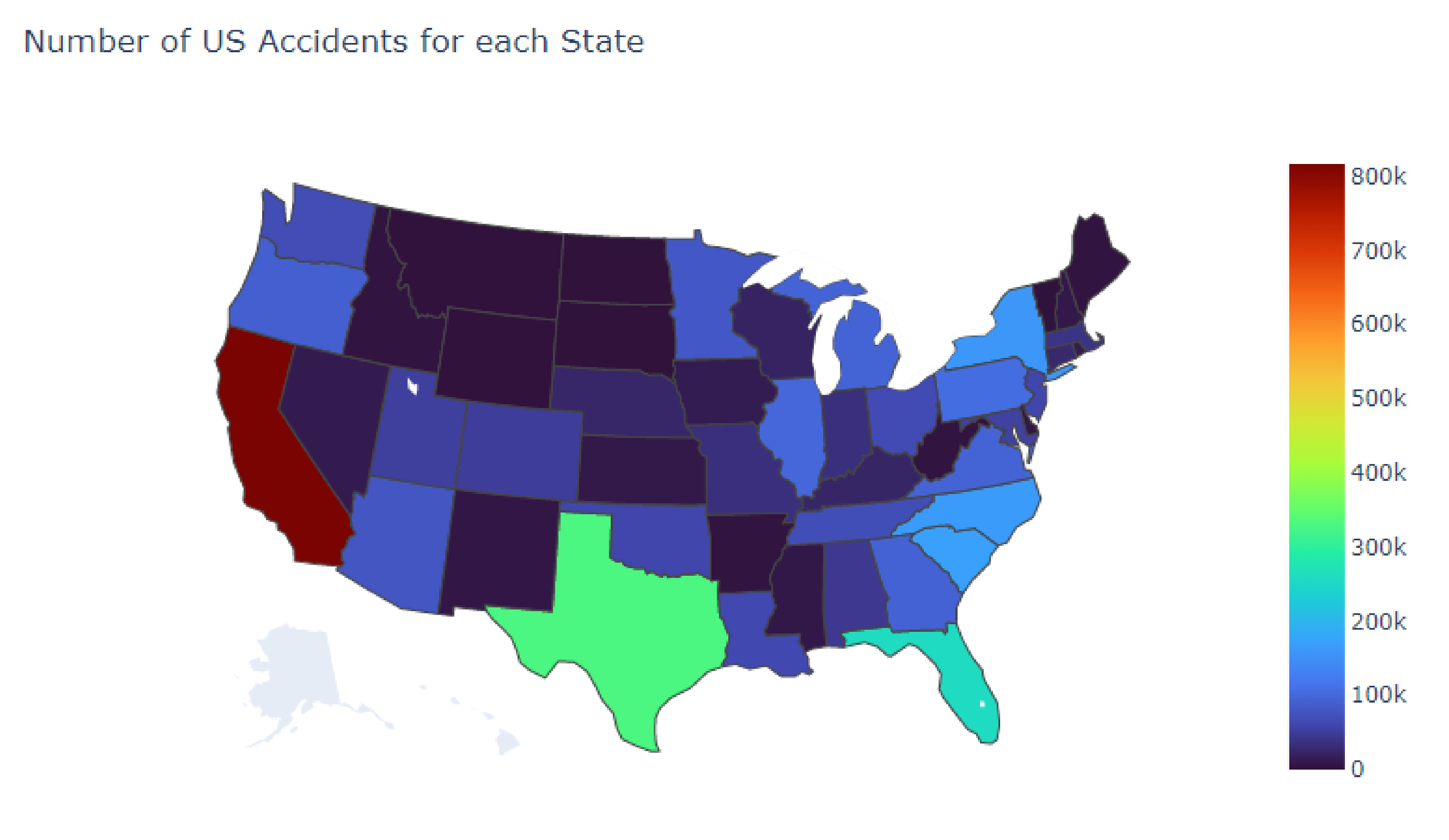
FABRIZIO ROSSI 1815023

MINA MAKAR 1804475

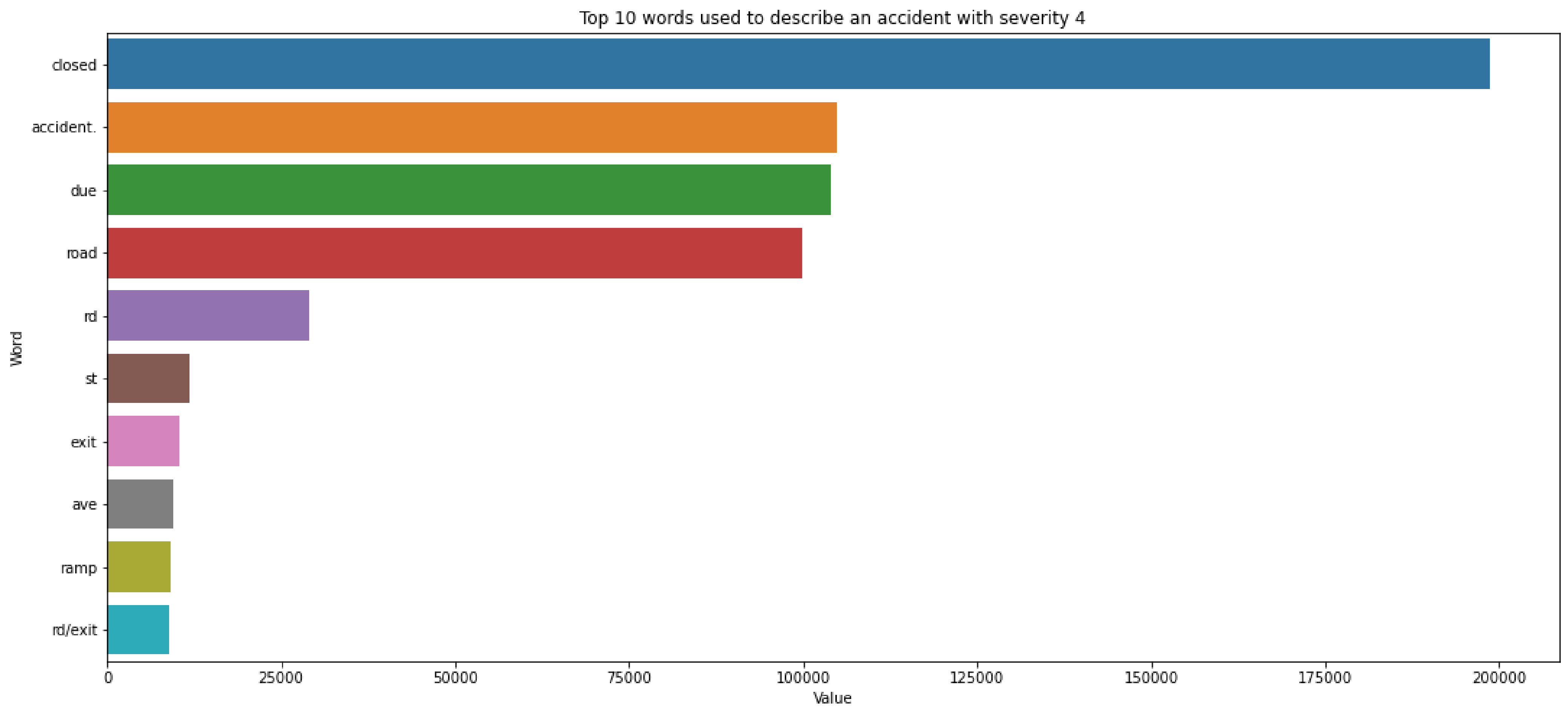
TASK - SEVERITY

- 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

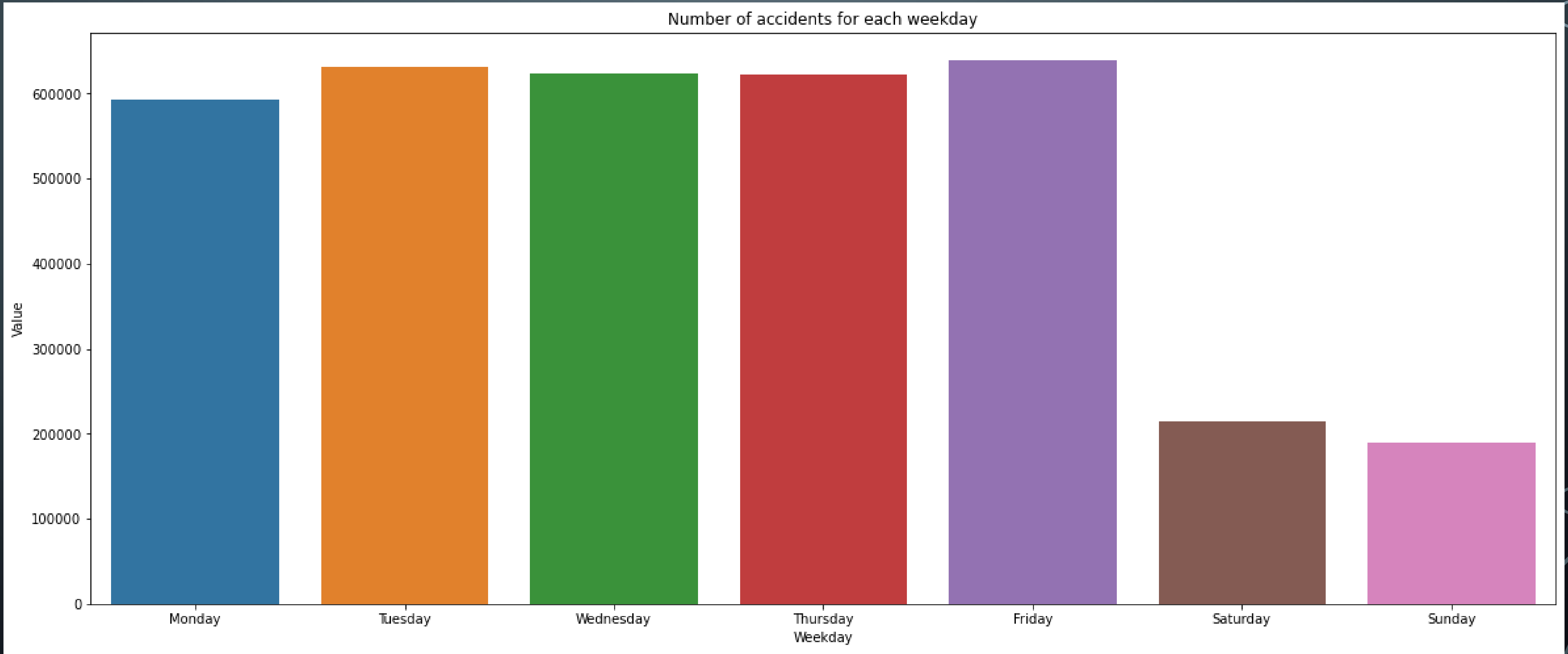
EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

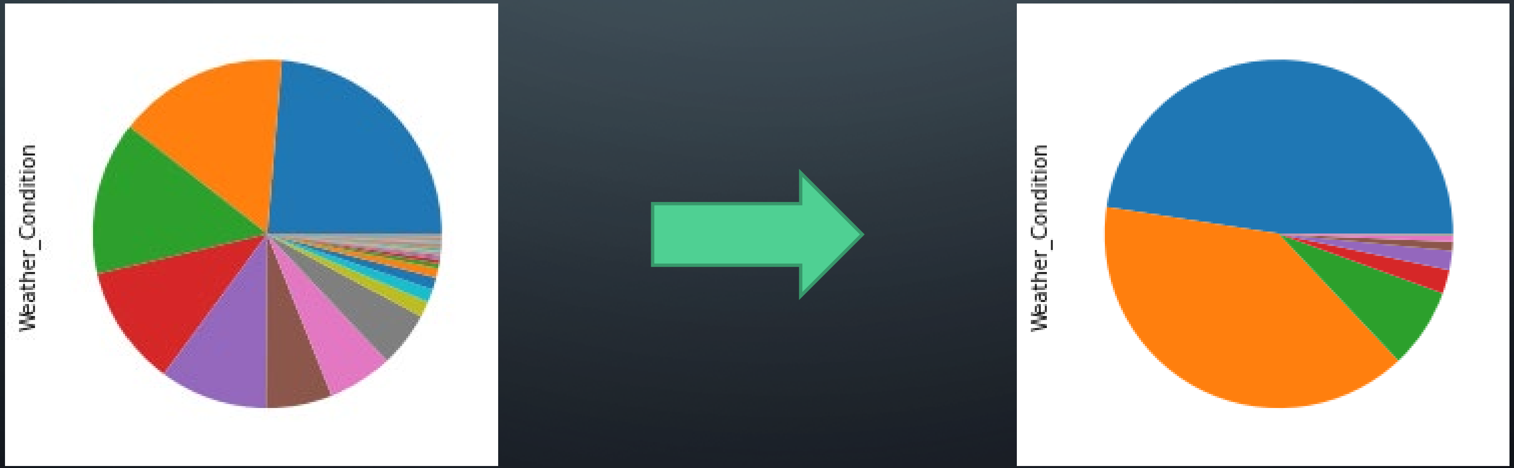


EXPLORATORY DATA ANALYSIS



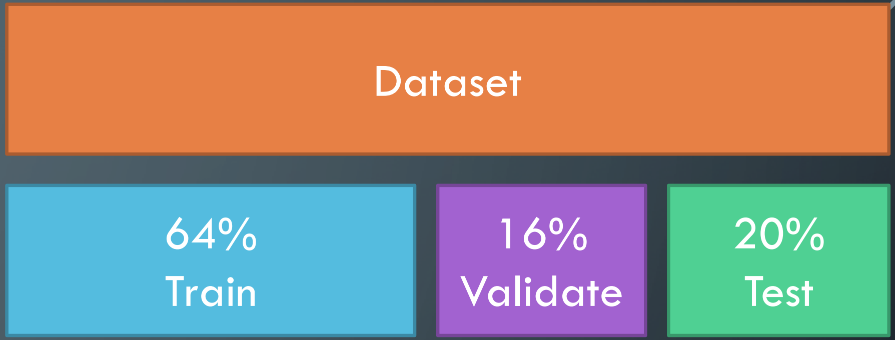
DATA PREPROCESSING

- Reduced number of classes for Weather_Condition (128 → 11) and Wind_Direction (24 → 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**

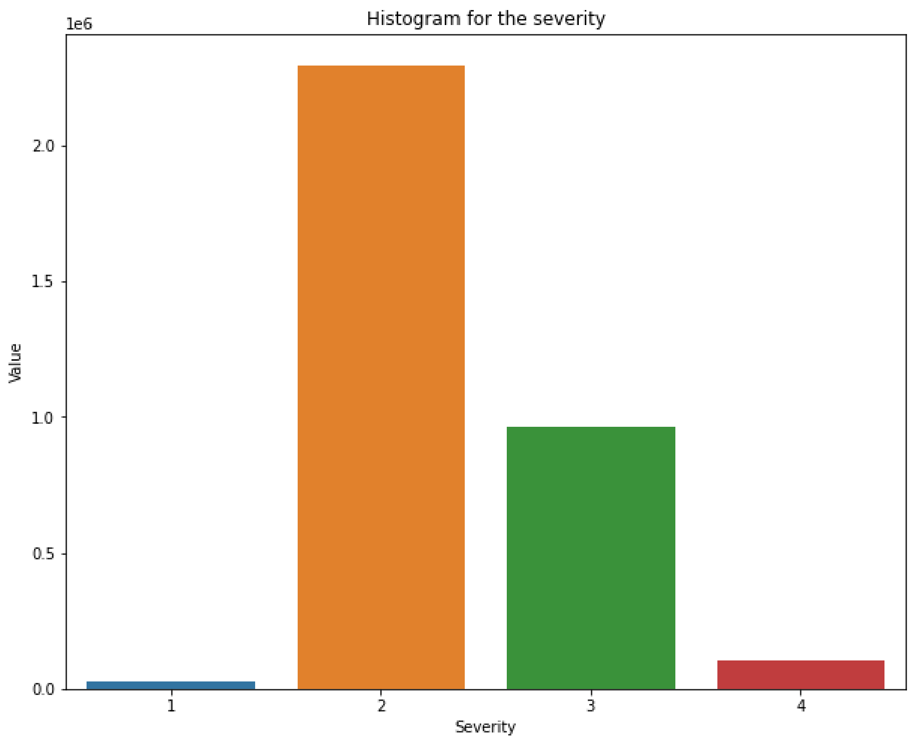


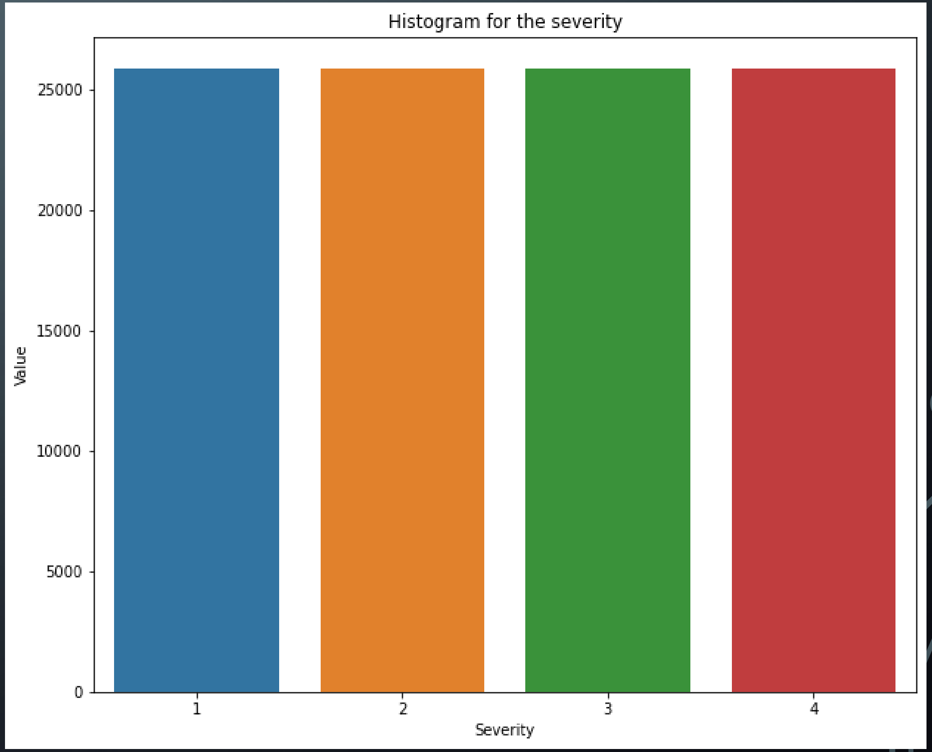
Dataset

64% Train

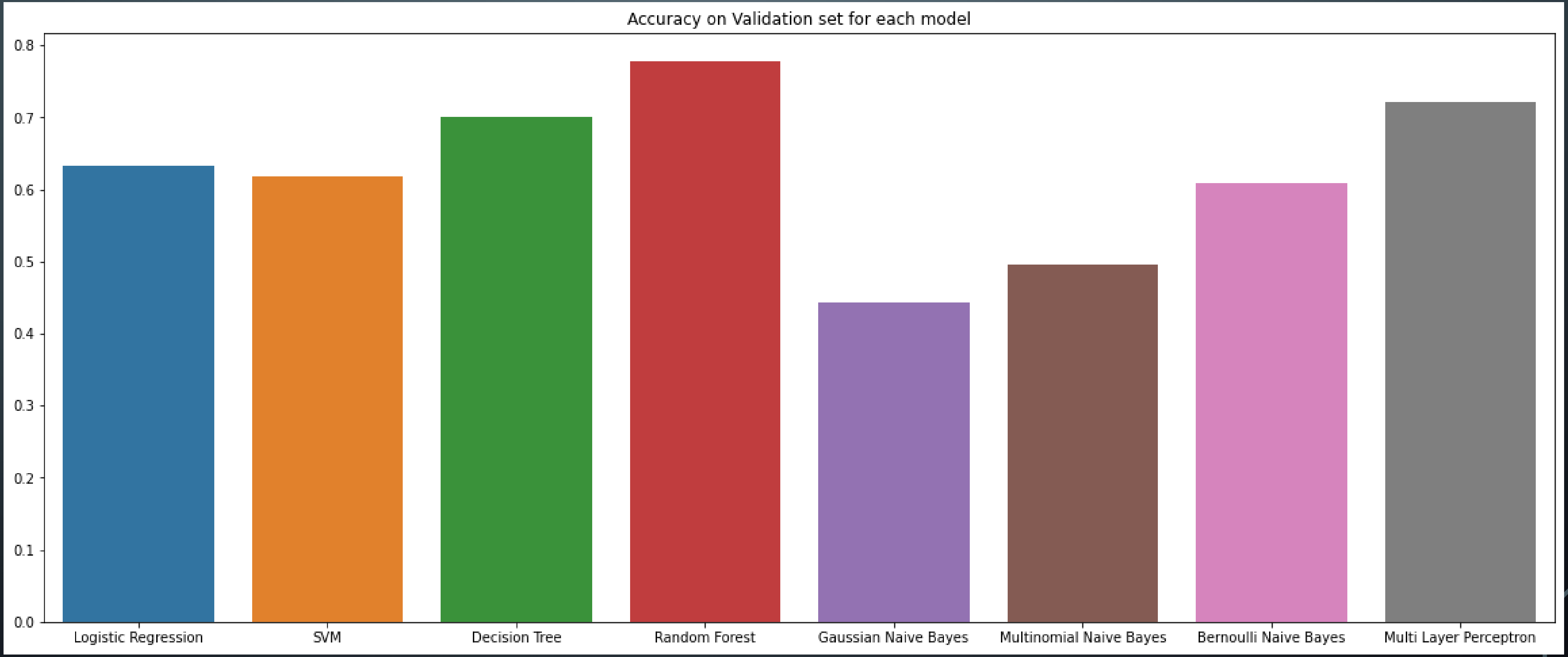
16% Validate

20% Test

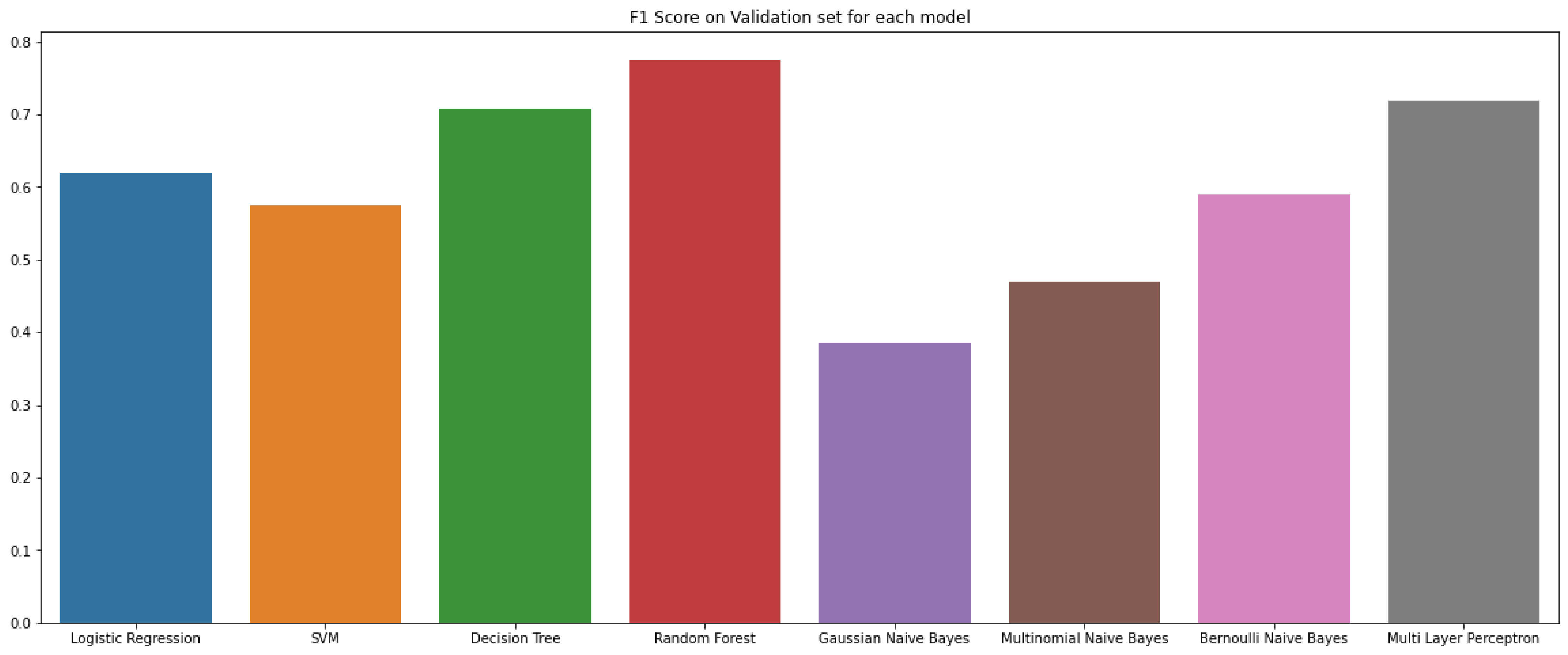




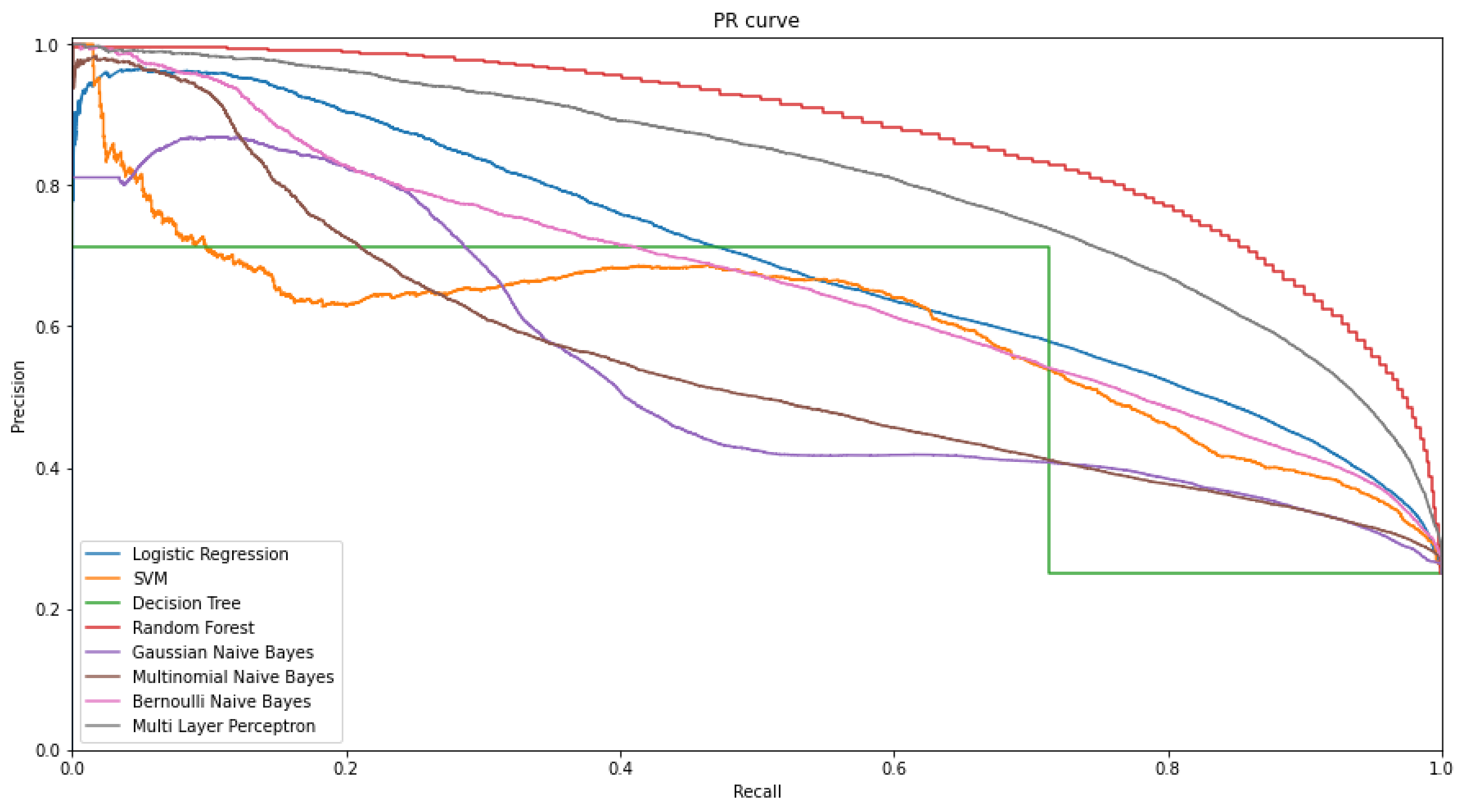
RESULTS – ACCURACY



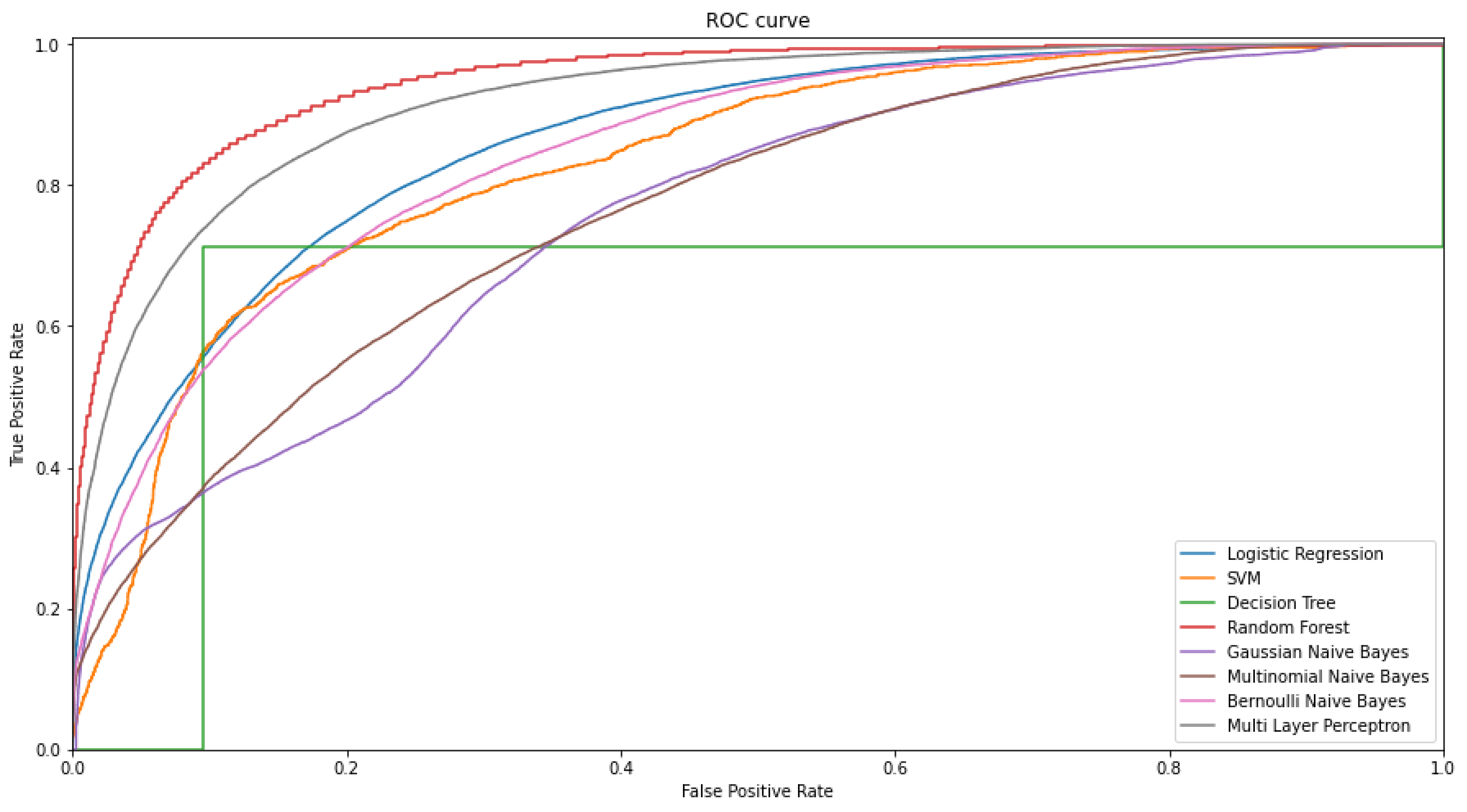
RESULTS – F1 SCORE



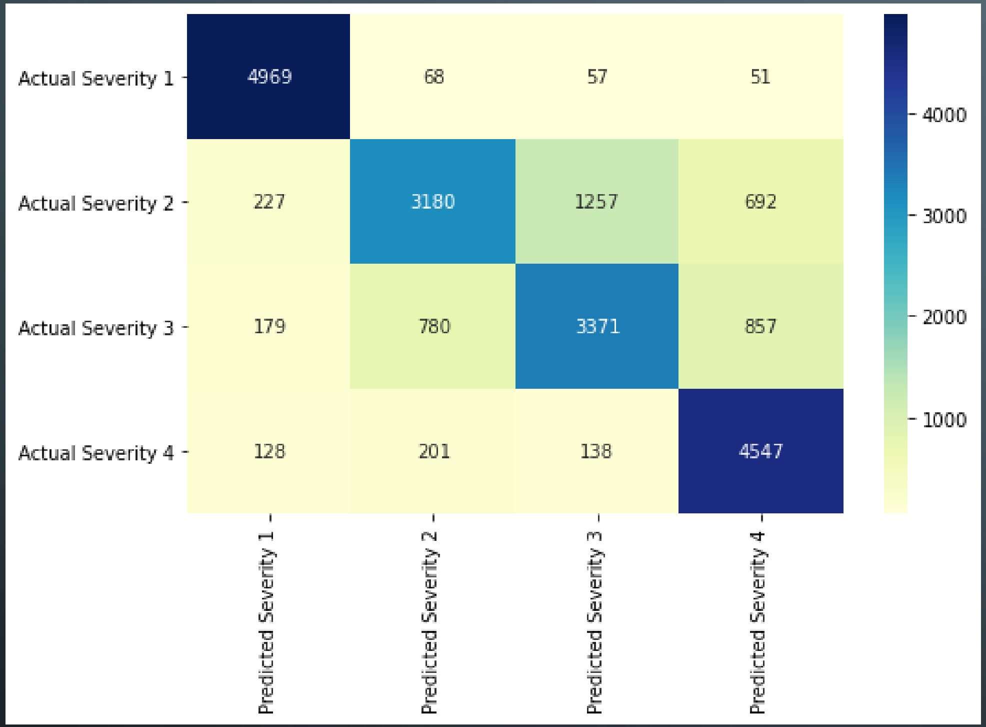
RESULTS – PR CURVE



RESULTS – ROC CURVE

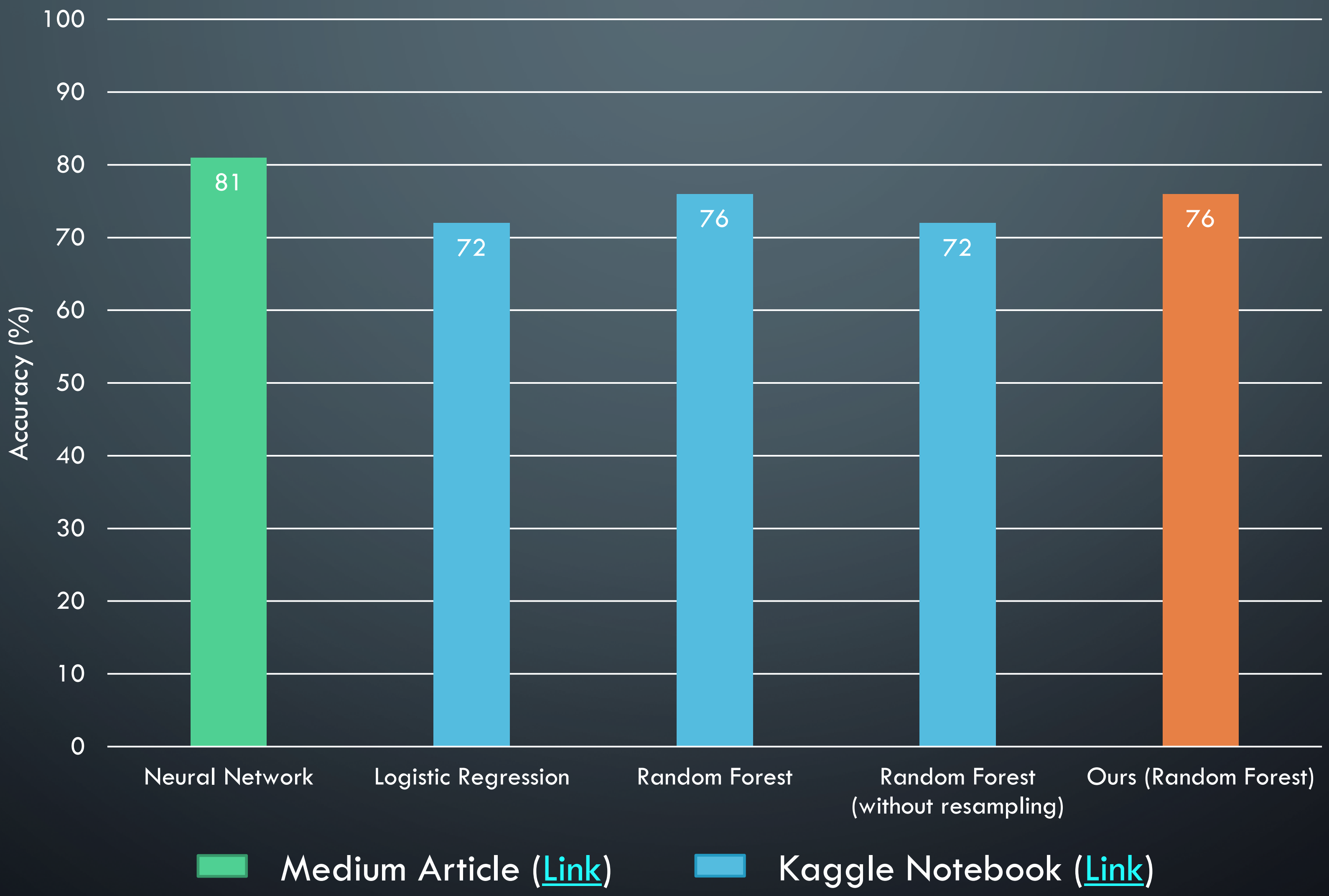


RESULTS – RANDOM FOREST TEST SET



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


Accuracy (%)
Neural Network
Logistic Regression
Random Forest
Random Forest (without resampling)
Ours (Random Forest)
🟩 Medium Article (Link)
🟦 Kaggle Notebook (Link)