**Road Accident Severity Prediction**

**Name:** Yashwanth Reddy Samala

**Email:** yashwanthreddy.samala@sjsu.eedu

**SJSU ID:** 016014583

**Name: Trivikramm Thopugunta**

**Email:** trivikram.thopugunta@sjsu.eedu

**SJSU ID:** 015955602

**Problem**

***Abstract*— Given the challenges of public safety today, research and analysis of real-time traffic and accident data to predict the risk of accidents is ubiquitous. Accident risk prediction can significantly improve public safety by warning the public. In this project, the probability of accident risk for a given case with selected conditions is predicted.**

***Keywords—prediction, classifiers, preprocessing, encoders, scaling***

This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including multiple APIs that provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road- networks. Currently, there are about **2.8 million** accident records in this dataset.

The economic and social impact of traffic accidents cost U.S. citizens hundreds of billions of dollars every year. And a large part of losses is caused by a small number of serious accidents. Reducing traffic accidents, especially serious accidents, is nevertheless always an important challenge. The proactive approach, one of the two main approaches for dealing with traffic safety problems, focuses on preventing potential unsafe road conditions from occurring in the first place. For the effective implementation of this approach, accident prediction and severity prediction are critical. If we can identify the patterns of how these serious accidents happen and the key factors, we might be able to implement well-informed actions and better allocate financial and human resources**.** Features like weather, traffic volume, road conditions, time of the day of previous accidents are utilized from the dataset. Machine learning algorithms like Logistic Regression, Decision Tree, Neural Networks and random forest classifiers are used and their results are compared to provide the best prediction.

**Dataset:**

<https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

**Objective**

The first objective of this project is to recognize key factors affecting the accident severity. The second one is to develop a model that can accurately predict accident severity. To be specific, for a given accident, without any detailed information about itself, like driver attributes or vehicle type, this model is supposed to be able to predict the likelihood of this accident being a severe one (on a scale of 1 to 4, 1 being the least severe and 4 being the most). The accident could be the one that just happened and still lack of detailed information, or a potential one predicted by other models. Therefore, with the sophisticated real-time traffic accident prediction solution developed by the creators of the same dataset used in this project, this model might be used as a tool to capture factors that cause an accident.

**Approach**

Data cleaning was first performed to detect and handle corrupt or missing records. EDA (Exploratory Data Analysis) and feature engineering were then done over most features.

Finally, Logistic regression, Support Vector Classifier, Decision Tree, Random Forest Classifier, and MLP classifier were used to develop the predictive model.

Out of all the models the one which has best accuracy is chosen and saved to be used in our web-based dash application for predictions

1. ***Data Preprocessing***

Our dataset consists of 2845342 rows and 47 columns. We have Traffic Attributes (12), Address Attributes (9): Weather Attributes (11), POI Attributes (13). We have implemented correlation matrix among the features. After observing the graph, we found that start and end GPS coordinates of the accidents are highly correlated. We can consider just one of them for the machine learning models. wind chill and temperature are also highly correlated to each other. So we can also drop one of them.

Our dataset consists of 2845342 rows and 47 columns. We have Traffic Attributes (12), Address Attributes (9): Weather Attributes (11), POI Attributes (13). We have implemented correlation matrix among the features. After observing the graph, we found that start and end GPS coordinates of the accidents are highly correlated. We can consider just one of them for the machine learning models. wind chill and temperature are also highly correlated to each other. So we can also drop one of them.

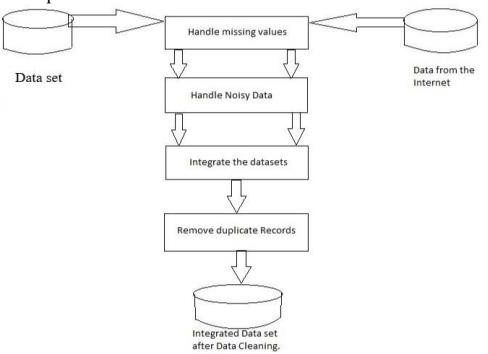


Fig. Data Preprocessing Other features like-

* + ID- since they don’t carry any information.
  + Start\_Time: because it was decomposed by the time features added before (day, month, weekday)
  + End\_Time: Beause we cannot know when traffic going to clear.
  + Description: most description only report the name of the road of the accident, so decided to omit the feature.
  + Number, Street, County, State, Zipcode, Country: there are lots of unique values for street. There are lot of null values for number. And moreover, we just focus on the city where the accident happened
  + Timezone, Airport\_Code, Weather\_Timestamp: because they are irrelevant for our task.
  + Turning\_Loop: since it's always False
  + Sunrise\_Sunset, Nautical\_Twilight, Astronomical\_Twilight: because they are redundant

1. ***Handling missing and erroneous data***

We have removed the duplicate values in the dataset consists of 2646365, 33 columns now. We have cleaned up the erroneous data by dropping the error rows. Null values are replaced with mean in the numerical features and the rows having nulls in categorial data are removed from the dataset.

1. ***Exploration and engineering***

Extracted needed information from a single feature into multiple features which will be helpful for preparing data for better accurate models.

And, we have given a more generalized names to some redundant values of some features. Thus, removing erroneous, redundant data and making encoding easy.

Based on the exploration, the accidents with severity level 2 are much more serious than accidents of other levels. The dataset it unequally distributed between the classes. This may result in bias results when model predicts the data. The model may neglect the low class. Therefore, we need to do resampling. There are 2 techniques. We have used under sampling of the data randomly by reducing all the distribution among the classes to least class number i.e., severity 1.

We have used one hot encoders for the some categorial features and for one particular feature, we have used binary encoder as there are nearly 7000 unique values in that feature. Min max scaler is used into to scale the dataset which will be helpful for training the model better.

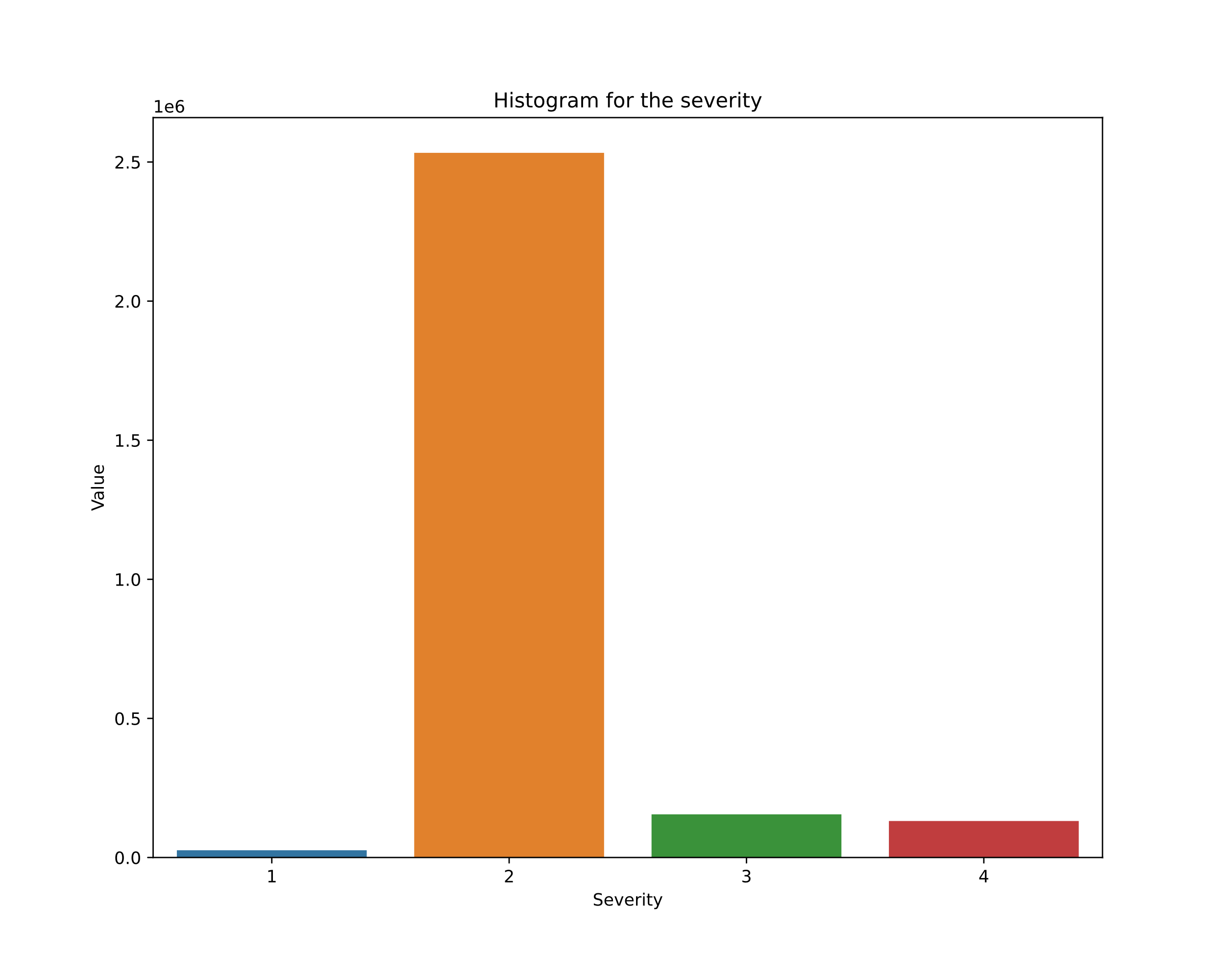


Fig. graph which shows data distribution among four different classes before under sampling.

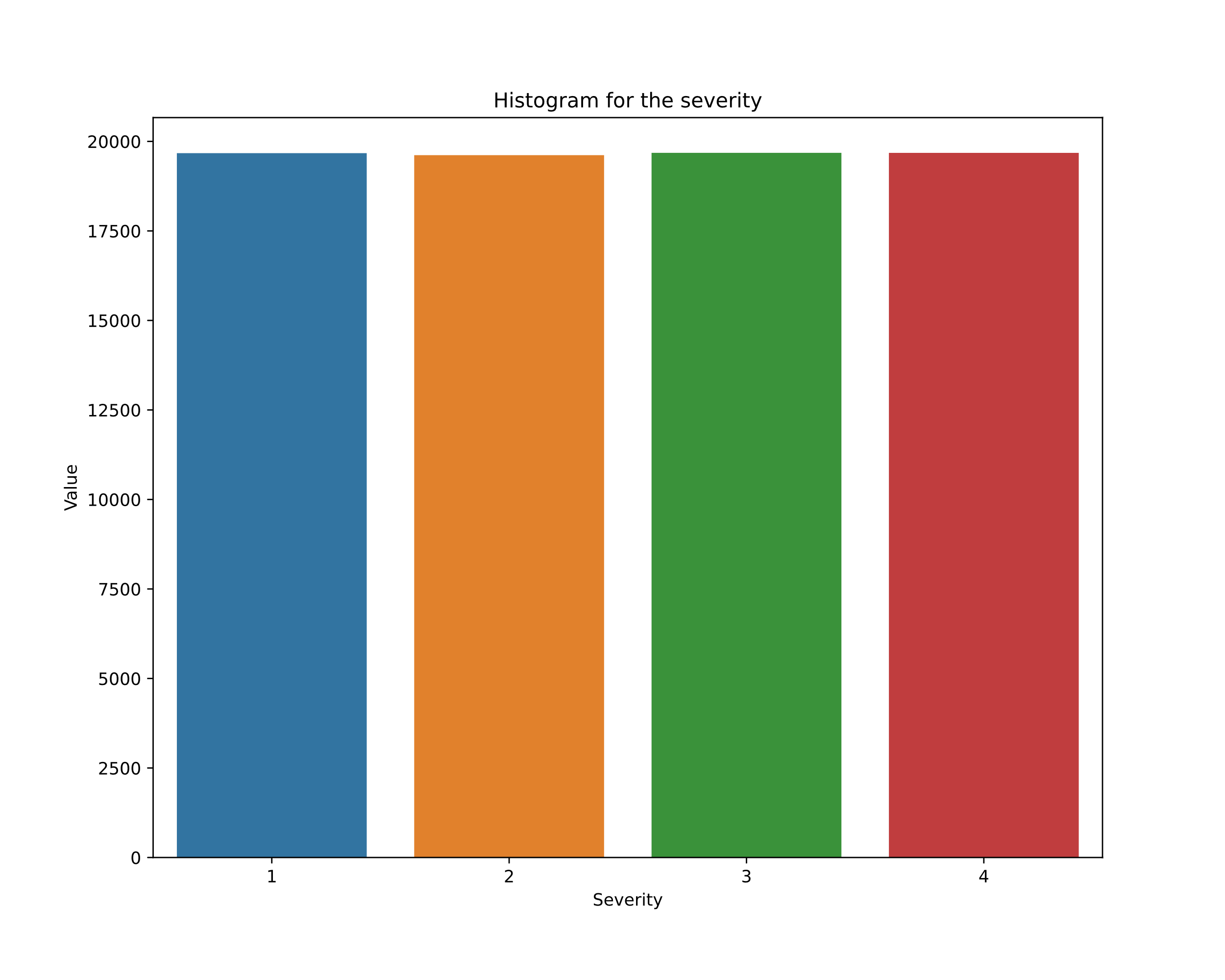


Fig. graph which shows data distribution among four different classes after under sampling.

1. ***Model***

Stored our prediction feature into one data frame and other features into another data frame X and Y. Split each data frame into two parts. Training and testing data. We have divided our data into 80 % Training data and 20% of Testing data. Later we trained our models using that training data. The idea is that to predict the severity. Using the model by giving test data as input and comparing the result with the expected test output, we can check the accuracy of our model for prediction of severity.

1. ***Logistic Regression****: It is used to*understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help us predict the likelihood of an event happening or a choice being made.

In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

***Logit(pi) = 1/(1+ exp(-pi))***

We get the following results after using a logistic regression model

precision recall f1-score support

1 0.70 0.87 0.78 14788

2 0.66 0.66 0.66 14672

3 0.56 0.56 0.56 14792

4 0.61 0.45 0.52 14734

accuracy 0.64 58986

macro avg 0.63 0.64 0.63 58986

weighted avg 0.63 0.64 0.63 58986

precision recall f1-score support

1 0.70 0.88 0.78 4884

2 0.67 0.66 0.67 4944

3 0.56 0.56 0.56 4889

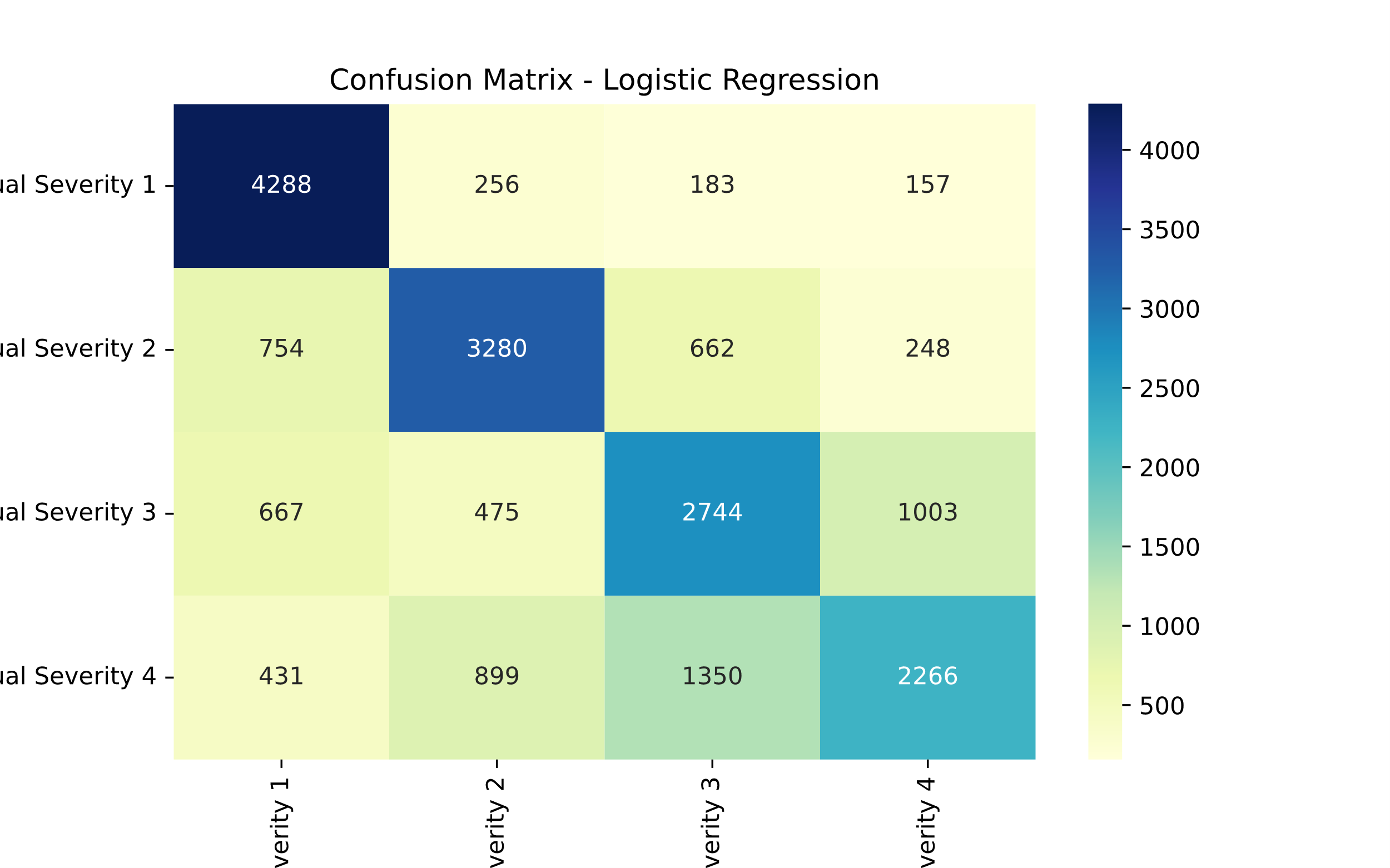
4 0.62 0.46 0.53 4946

accuracy 0.64 19663

macro avg 0.63 0.64 0.63 19663

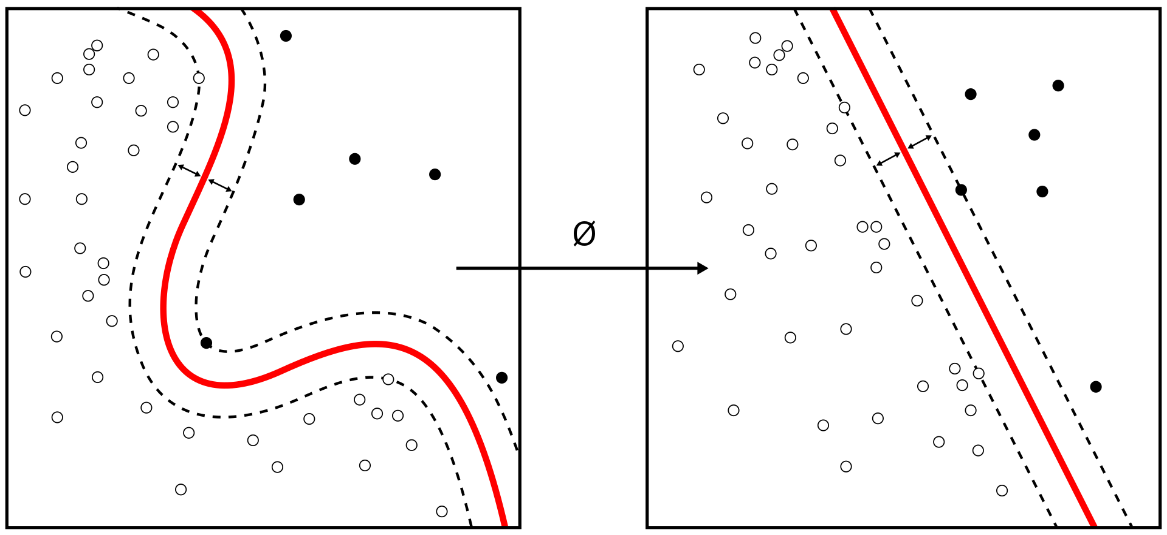
weighted avg 0.63 0.64 0.63 19663

***cunfusion matrix:***



1. ***Support Vector Machine****:* support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

SVM maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.



*We get the following results after using a* Support Vector machine

precision recall f1-score support

1 0.70 0.95 0.81 1998

2 0.71 0.66 0.69 1967

3 0.57 0.61 0.59 1996

4 0.67 0.43 0.52 2039

accuracy 0.66 8000

macro avg 0.66 0.67 0.65 8000

weighted avg 0.66 0.66 0.65 8000

precision recall f1-score support

1 0.71 0.98 0.82 489

2 0.70 0.67 0.68 494

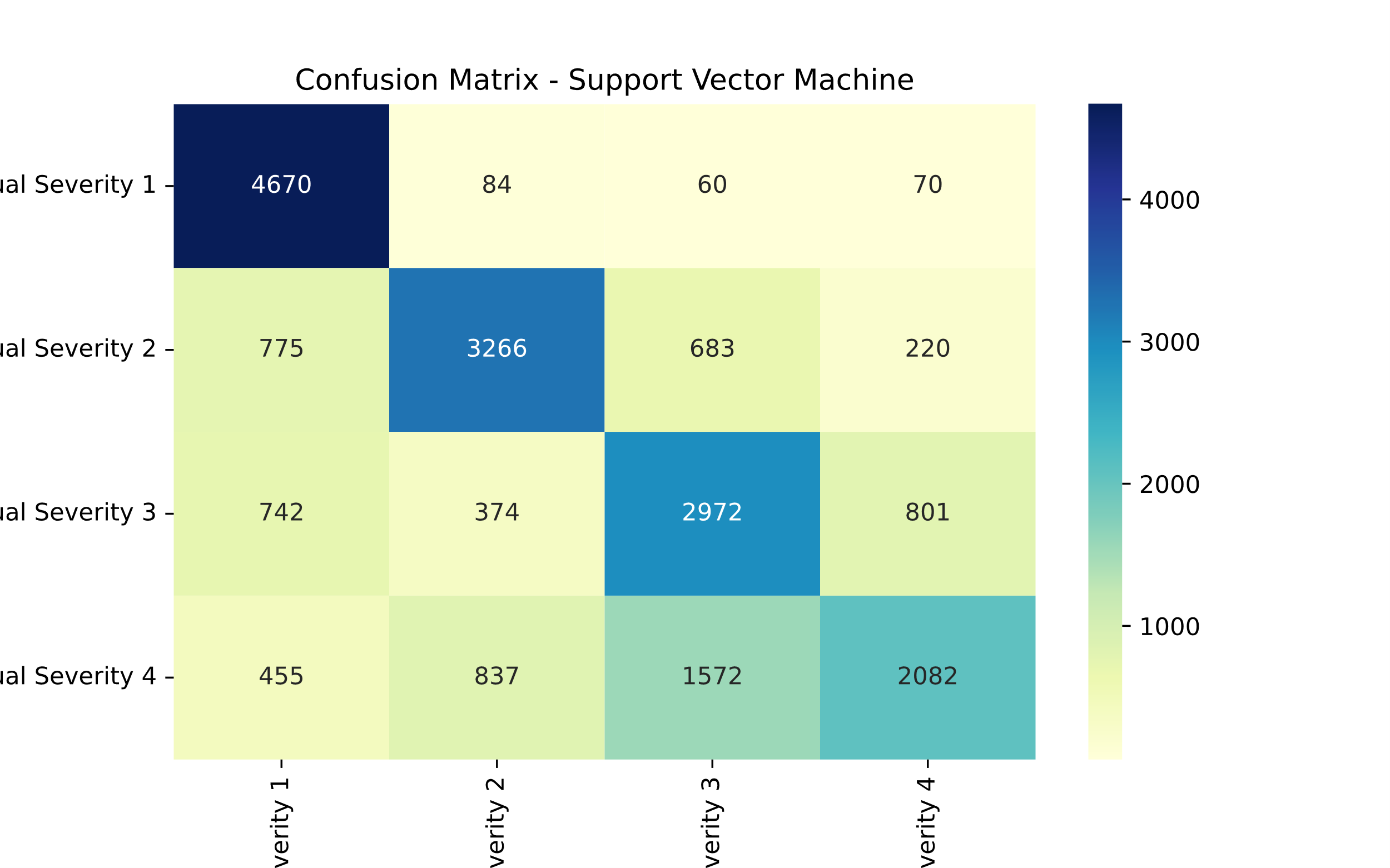
3 0.57 0.59 0.58 516

4 0.65 0.40 0.50 501

accuracy 0.66 2000

macro avg 0.66 0.66 0.65 2000

weighted avg 0.66 0.66 0.64 2000



1. ***Decision tree:***The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

*After using a decision tree we get the following results.*

precision recall f1-score support

1 1.00 1.00 1.00 14788

2 1.00 1.00 1.00 14672

3 1.00 1.00 1.00 14792

4 1.00 1.00 1.00 14734

accuracy 1.00 58986

macro avg 1.00 1.00 1.00 58986

weighted avg 1.00 1.00 1.00 58986

precision recall f1-score support

1 0.90 0.90 0.90 4884

2 0.72 0.71 0.72 4944

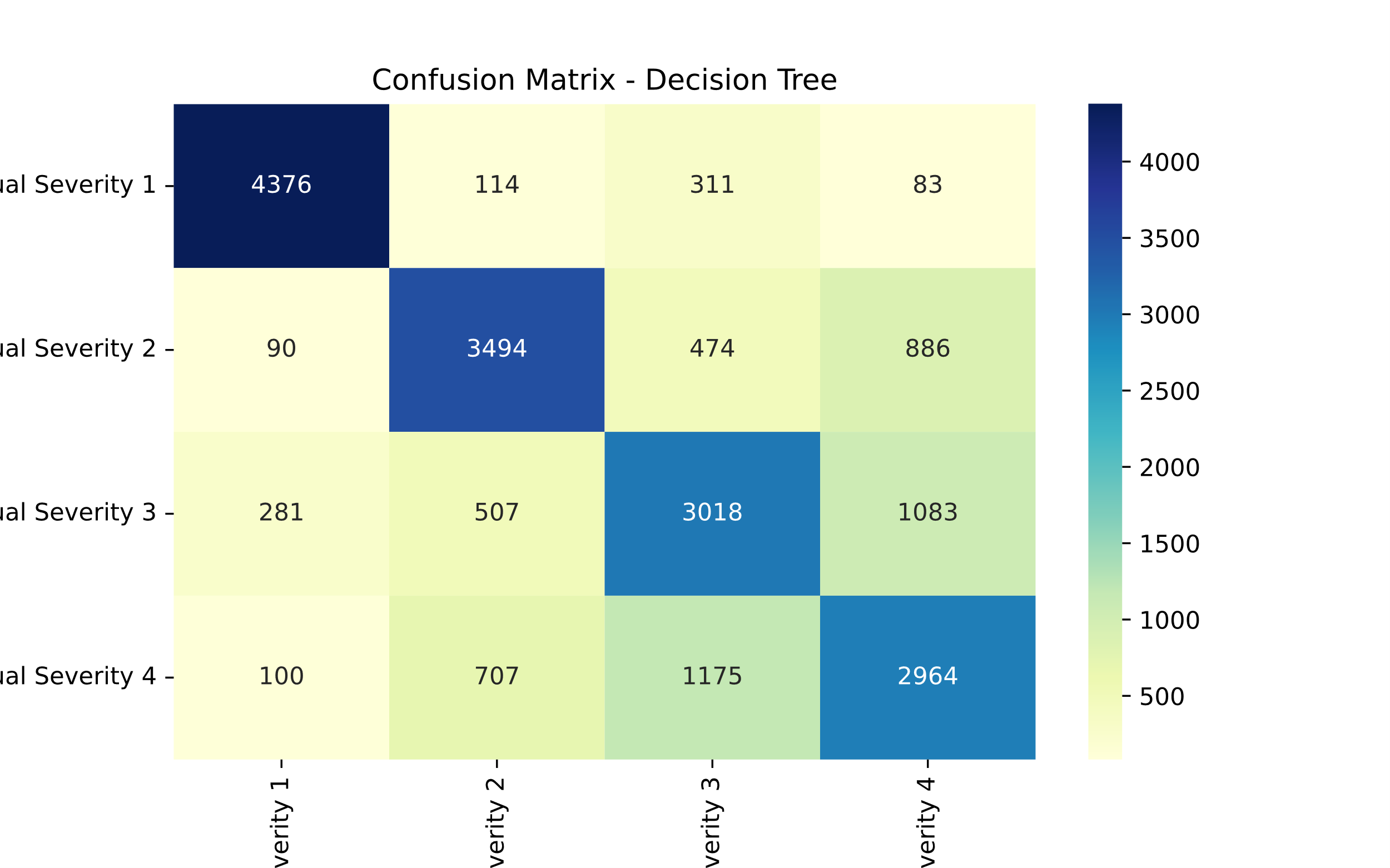
3 0.61 0.62 0.61 4889

4 0.59 0.60 0.60 4946

accuracy 0.70 19663

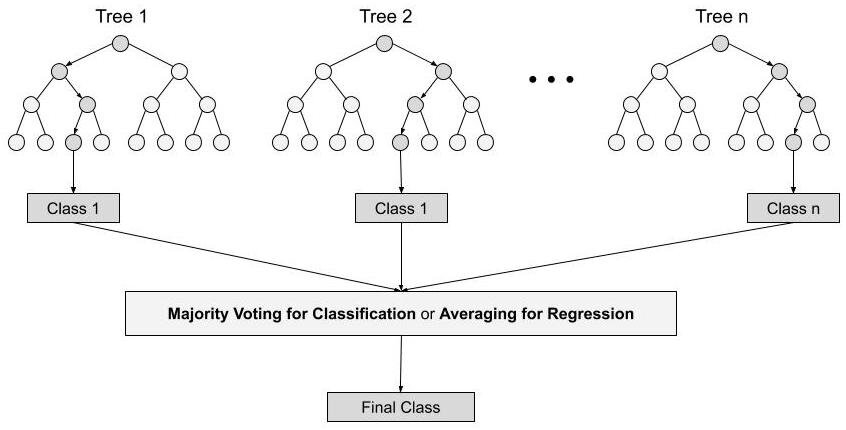
macro avg 0.71 0.70 0.71 19663

weighted avg 0.71 0.70 0.71 19663



1. ***Random Forest****:* It is a *Supervised Machine Learning Algorithm* that is *used widely in Classification and Regression problems*. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing *continuous variables* as in the case of regression and *categorical variables* as in the case of classification. It performs better results for classification problems.



*After using a random forest algorithm we get the following results.*

precision recall f1-score support

1 1.00 1.00 1.00 14788

2 1.00 1.00 1.00 14672

3 1.00 1.00 1.00 14792

4 1.00 1.00 1.00 14734

accuracy 1.00 58986

macro avg 1.00 1.00 1.00 58986

weighted avg 1.00 1.00 1.00 58986

precision recall f1-score support

1 0.89 0.97 0.93 4884

2 0.82 0.80 0.81 4944

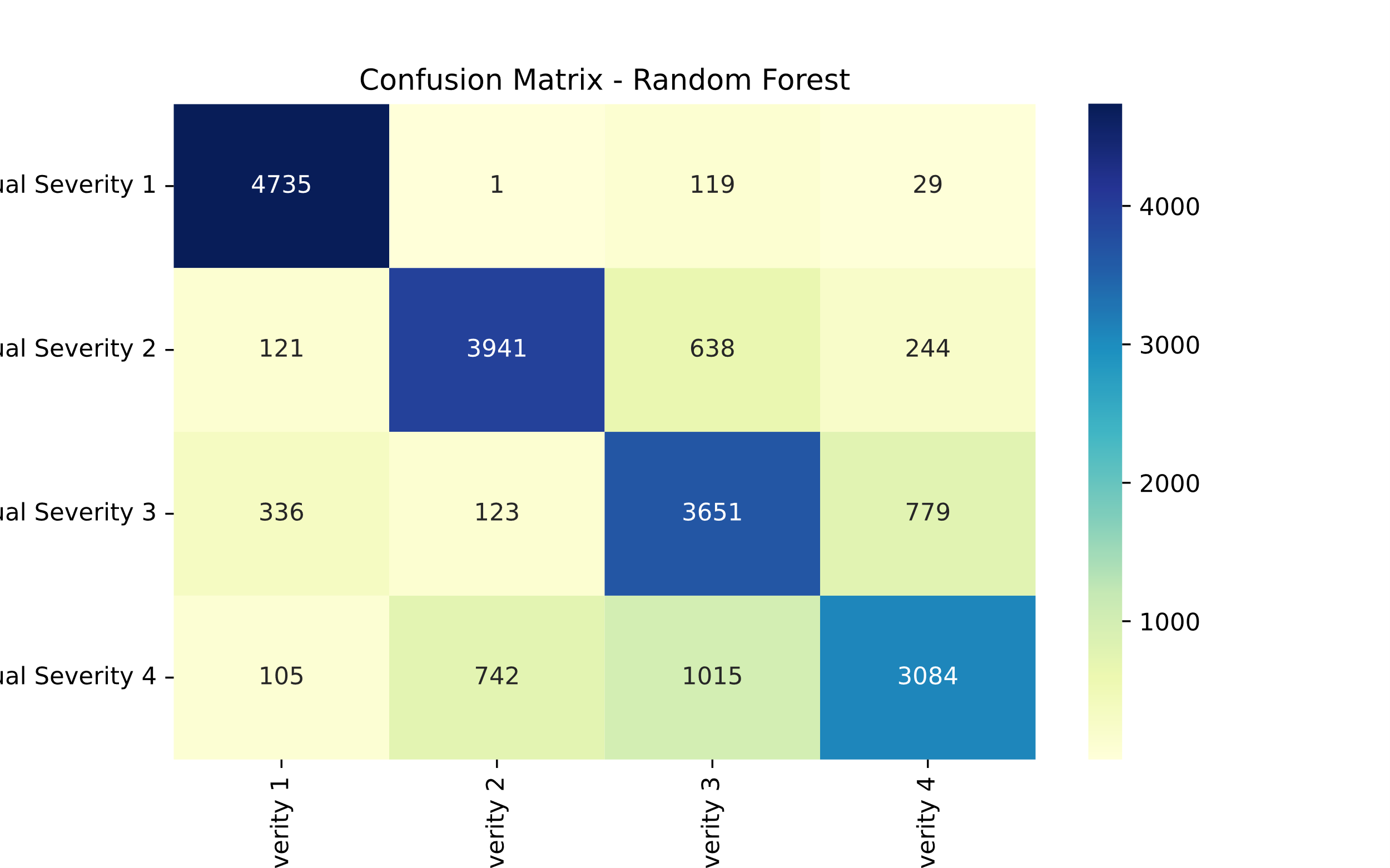
3 0.67 0.75 0.71 4889

4 0.75 0.62 0.68 4946

accuracy 0.78 19663

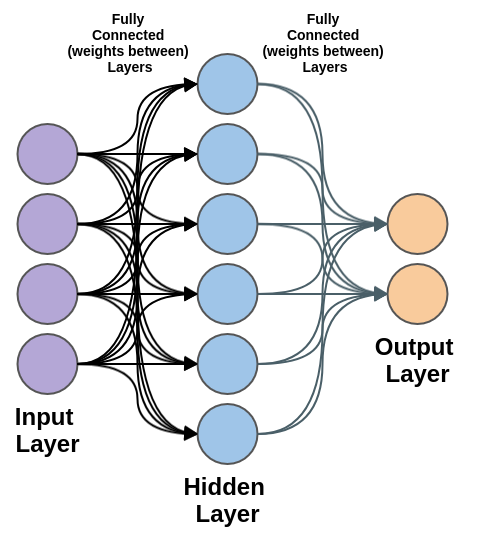
macro avg 0.78 0.78 0.78 19663

weighted avg 0.78 0.78 0.78 19663



1. ***MLP classifier:*** Multilayer Perceptron’s, or MLPs for short, are the classical type of neural network. They are comprised of one or more layers of neurons. Data is fed to the input layer, there may be one or more hidden layers providing levels of abstraction, and predictions are made on the output layer, also called the visible layer.

*After using a random forest algorithm, we get the following results.*



precision recall f1-score support

1 0.88 0.97 0.92 14788

2 0.82 0.77 0.80 14672

3 0.67 0.76 0.71 14792

4 0.74 0.61 0.67 14734

accuracy 0.78 58986

macro avg 0.78 0.78 0.78 58986

weighted avg 0.78 0.78 0.78 58986

precision recall f1-score support

1 0.86 0.95 0.90 4884

2 0.79 0.74 0.77 4944

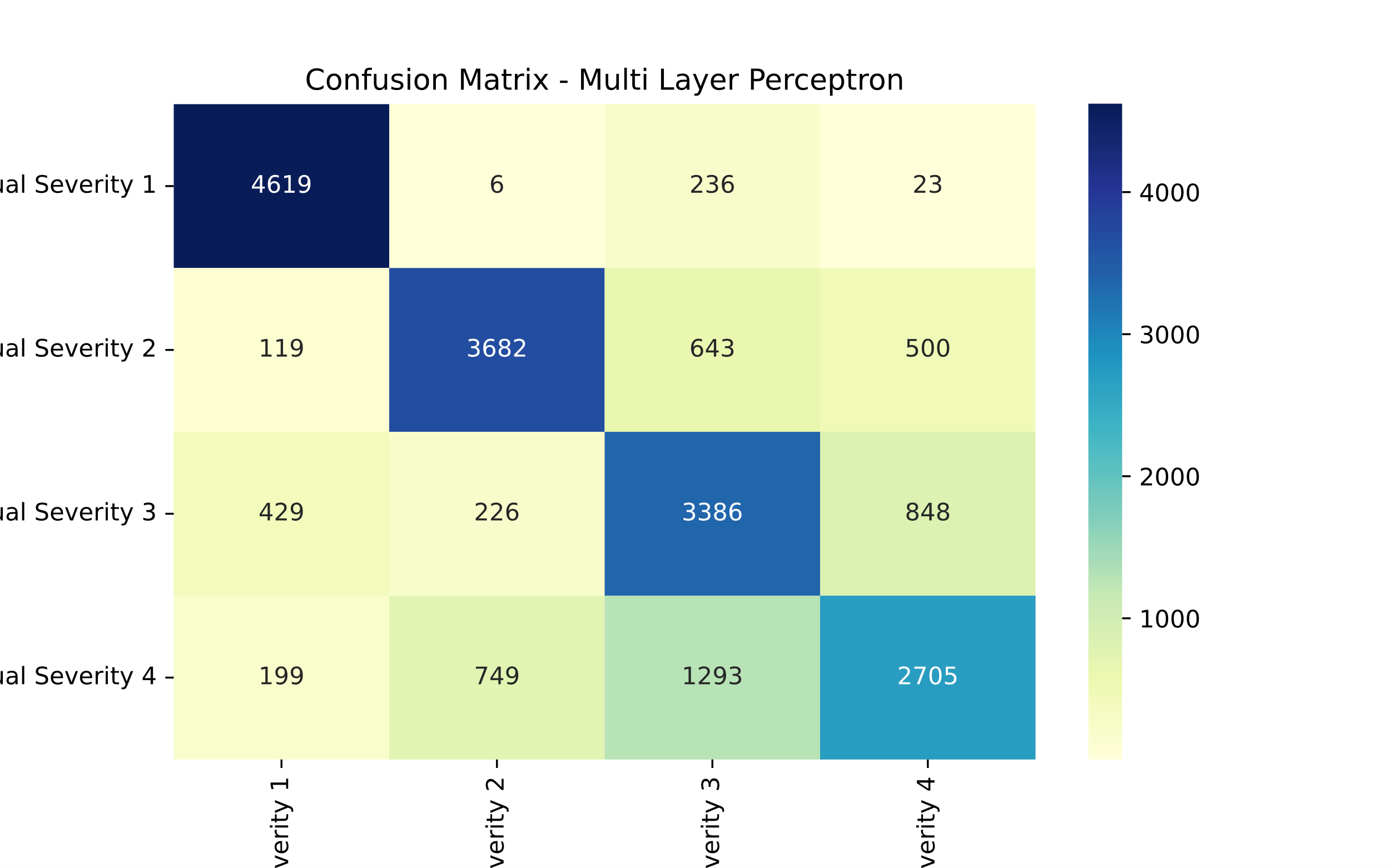
3 0.61 0.69 0.65 4889

4 0.66 0.55 0.60 4946

accuracy 0.73 19663

macro avg 0.73 0.73 0.73 19663

weighted avg 0.73 0.73 0.73 19663



**Comparing all the models**

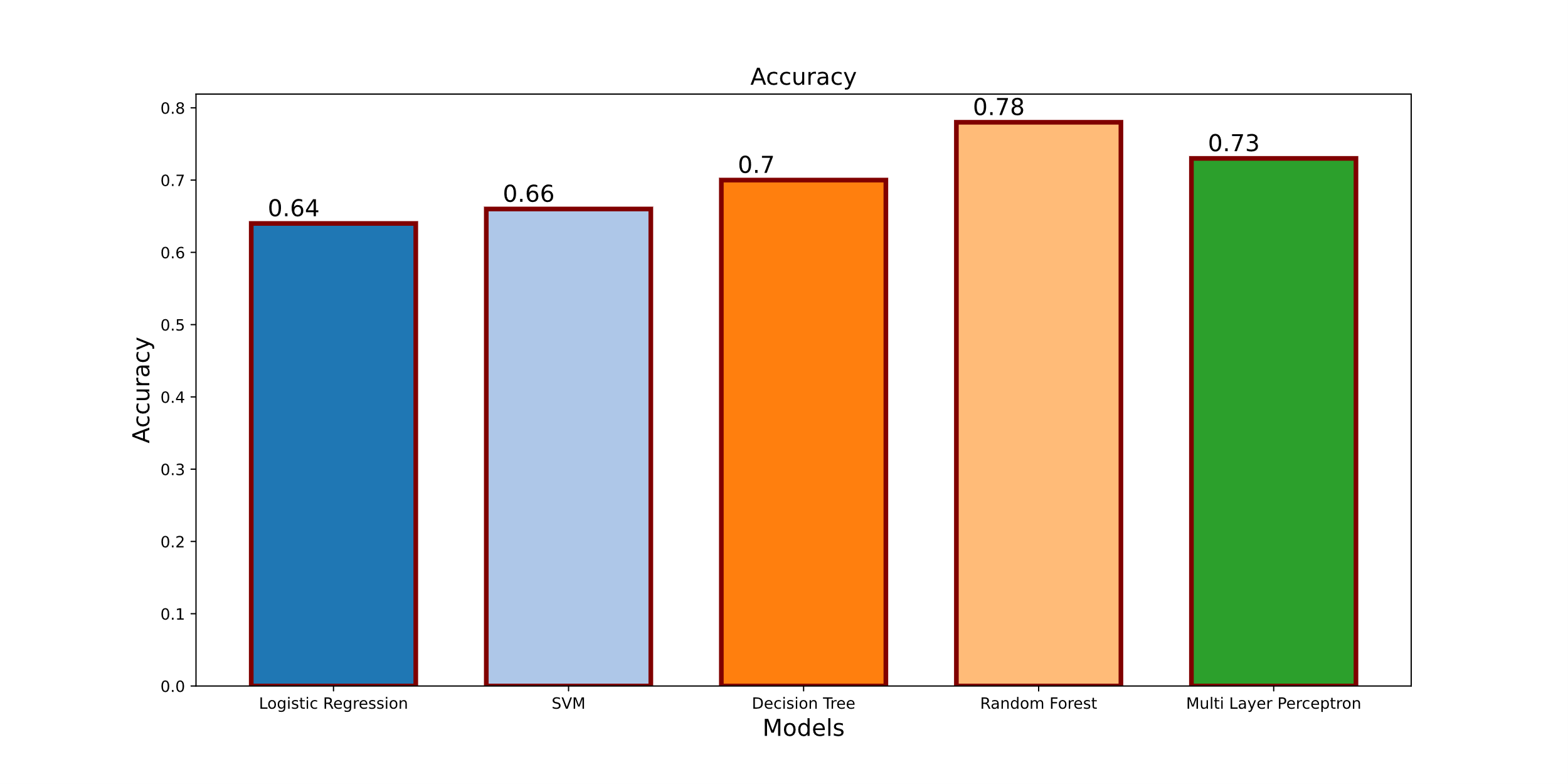


Fig. graph comparing accuracy of all the models used .

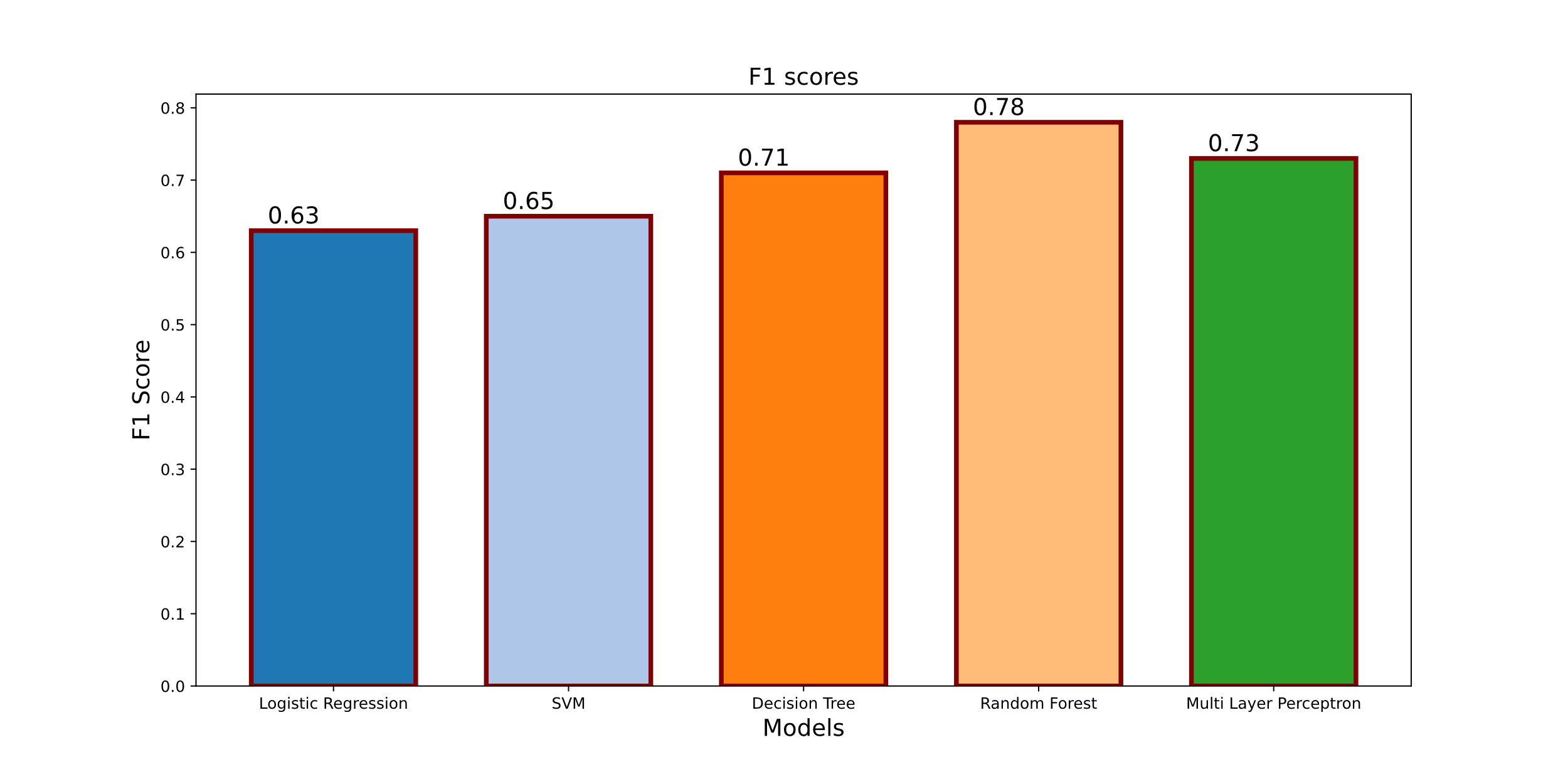


Fig. graph comparing F1 of all the models used .

**Hyper-parameter tuning:**

We obtained the most accurate model as random forest and multilayer perceptron. We can tweak the accuracy of model by hyperparamenter tuning.

**CONCLUSION:**

The causalities in road accidents are terrible to the society as well as to the families. These accidents create a great impact on the impact loss as well. So, it has ended up an essential requirement to control and organize activity with a progressed framework to diminish the number of street accidents in our nation. By taking basic safeguards, based on forecast or notices of a modern framework may avoid activity mischances. In addition, it’s an essential requirement for our nation presently, to handle these circumstance where each day, many individuals were encountering with accidents and this is increasing day by day. The implementation of We and this is increasing day by day. The implementation of machine learning could be a useful and an incredible approach to require an accurate decision with the involvement to oversee the current circumstance and the discoveries of the investigation portion can be suggested to traffic authorities for diminishing the number of mischances. We are able to utilize proposed approaches to actualize machine learning here since of their demonstrated and higher

exactness to anticipate activity mischance seriousness.

machine learning could be a useful and an incredible approach to require an accurate decision with the involvement to oversee the current circumstance and the discoveries of the investigation portion can be suggested to traffic authorities for diminishing the number of mischances are able to utilize proposed approaches to actualize machine learning here since of their demonstrated and higher exactness to anticipate activity mischance seriousness.

**ACKNOWLEDGMENT**

“Road Accident Severity Prediction” incorporates a framework which examines and analysis the accidental data and gives a legitimate and successful visualize report that highlights the parameters based on which factors might have had an impact due to which the accident might have occurred. Thereby this project helps to overcome the accidents by examining which factor might have caused an accident so that there would be diminishment in casualty rate. This project will continuously energize us to do our work superbly and profession. We would like to extend our sincere thanks to our Professor Dr. Mahima Agumbe Suresh who has been a constant support and encouragement throughout the project. Prof. Mahima has been helping us at every stage of development of this project.

**REFERENCES**

1. L. Breiman, H. Freidman, A. Olshen and J. Stone, “Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software", 1984.
2. M. S. Satu, S. Ahamed, F. Hossain, T. Akter and D. M. Farid, "Mining traffic accident data of N5 national highway in Bangladesh employing decision trees," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 722-725.
3. S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations," Journal of Modern Transportation(2016), vol. 24, issue no. 1, pp. 62-72
4. M. M. L. Elahi, R. Yasir, M. A. Syrus, M. S. Q. Z. Nine, I. Hossain and

N. Ahmed, "Computer vision based road traffic accident and anomaly detection in the context of Bangladesh," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, 2014, pp. 1-6

5. https://[www.tandfonline.com/doi/epub/10.1080/08839514.2021.201864](http://www.tandfonline.com/doi/epub/10.1080/08839514.2021.201864) 3?needAccess=true