ANALYSIS AND PREDICTION OF ACCIDENT SEVERITY USING MACHINE LEARNING

Submitted in partial fulfilment of the requirements of the

MTech Data Science and Engineering Degree programme

By

SARHAD GAUTAM

2020SC04969

Under the supervision of

HIMANSH GAUTAM

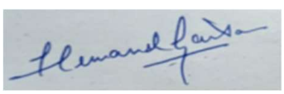
SENIOR DATA ENGINEER

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the Dissertation entitled ANALYSIS AND PREDICTION OF ACCIDENT SEVERITY USING MACHINE LEARNING and submitted by SARHAD GAUTAM IDNo. 2020SC04969 in partial fulfilment of the requirements of DSECLZG628T Dissertation, embodies the work done by him/her under my supervision.

Signature Of the Supervisor,



Name Himansh Gautam

Designation Senior Data Engineer

Place: Hyderabad

Date: 14-03-2023

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

FIRST SEMESTER 2022-23

DSECLZG628T DISSERTATION

Dissertation Title	:	ANALYSIS AND PREDICTION OF ACCIDENT SEVERITY USING MACHINE LEARNING
Name of Supervisor	:	HIMANSH GAUTAM
Name of Student	:	SARHAD GAUTAM
ID No. of Student	:	2020SC04969

Abstract

Everyday either in the newspaper or while chatting with friends we often hear the news of road accidents, maybe we might ourselves be involved with one. Now, one of the simplest and most common conclusions that we can notice that accident severity may vary from life fatality to just a dent in the vehicle of the victims. So, to avoid such things lots of traffic signal or bump in the road or speed limit sign has been installed. However, accidents continue to happen and their severity, well that remains unknown until after the accident has occur. Now, our study will be based on traffic accidents data from UK and their severity is divided into 3 different types. This study aims to predict the two of the highest severity (‘**serious’** and ‘**fatal’**) of the accident based on certain conditions that will be highlighted in the study so that the model can be used to avoid or reduce the damage that maybe caused by the accident to the economy and human life. In this study, we will also provide some brief analytics about the dependency of the accident severity on some of the factors that we are using in for the successful prediction. This study can be used by the logistics or public transportation sector in order to plan the route and journey and take necessary precaution to reduce the accident severity to minimum and hence reduce the overall damage.

Keywords: traffic accidents, fatality, severity, analytics, logistics, public transportation, prediction

SYMBOLS, ABBREVIATIONS AND DEFINITIONS

$	Dollar
USD	United States Dollar
WHO	World Health Organisation
OS&D	Over, Short, or Damaged [5]
Transportation	It is the activity that physically connects the business to its supply chain partners, such as suppliers and customers.[5]
Logistics	The part of supply chain management that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services, and related information from the point of origin to the point of consumption in order to meet customers' requirements.[5]
£	Pound
ML	Machine Learning
NN	Neural Network
Prevention	It can be interpreted in two ways according to UK government i.e., it is the amount of money the Government should spend to likely prevent all road accidents and the loss to society due to the current level of road accidents. [6]
RNN	Recurrent Neural Network
NSE	North-South Expressway
MLP	Multilayer Perceptron
BLR	Bayesian Logistic Regression
ANN	Artificial Neural Network
KNN	K-Nearest Neighbours
SVM	Support Vector Model
SMOTE	Synthetic Minority Oversampling Technique
ENN	Edited Nearest Neighbours
SMOTENN	SMOTE-ENN (Combined over and under sampling technique)
Relu	**Re**ctified **L**inear **U**nit is an activation function defined as the positive part of its argument
SoftMax	It is a normalized exponential function that converts a vector of 'K' real numbers into a probability distribution of 'K' possible outcomes

TABLE OF CONTENTS

INTRODUCTION

Road Accidents are the most common occurring incidents in the world. Its impact is shared not only by the individual victims and its family but also with the industries that supports the overall economy. WHO states that annually 1.4 million people are killed and 20 to 50 million are injured because of road crashes[1]. The white paper cited in [2] suggests that the road accidents bear a cost of USD $518 billion to the global economy. On a high level, the impact of road accidents to the economy is categorized in 4 different but somehow interconnected ways [1]. Industries or businesses are heavily impacted by the road accidents because fatality or serious injuries to the employees leaves a void on the job and separate costs incurred on material/infrastructural damages cuts the revenue of the business. One of the major impacted industries because of these road accidents are transportation and logistics industries, as these industries earn their revenue by reallocating the goods from one place to another in a proper timely manner. These industries become the hotspot of service problems when products delivered to them are OS&D [5], which mostly happens directly (when the truck carrying goods become involved in a road accident) or indirectly (due to road accidents among other vehicles traffic congestion occurs leading to delay in delivery) because of the road accidents. An example of such incident can be viewed in [7] where a mails, letters and parcel transportation company i.e. Royal Mail Group was subjected to a fine of £1.6 million after a yard marshal suffered serious injuries after being hit by a 7.5 tonne truck. Now, this not only lead to bad reputation on this mailing company but also a huge economic loss in the form of fine. Even though incidents like this may be difficult to forecast but if similar incidents based on some geographical features can be predicted even before the accidents happen then it can help the company to regain their reputation and increase their revenue (not only by avoiding fines but also getting business because of good faith).

According to [6] the *prevention value* of fatal, serious, and slight casualties is based on 3 elements i.e., Loss of output due to injury, Ambulance and treatment cost, and cost of life fatalities. Also, according to same report mentioned above for each level of severity, average value for preventing fatal accident is greater than the *prevention value* for fatality. In the report the valuation of accidents occurred in 2012 based on its severity has a huge difference from slight to serious and serious to fatal. Our study is focused on prediction of these severities based on different independent features involved during the accident and without including the conditions of vehicles or drivers (since the dataset many accidents have same accident index and this accident index acts as key to link the accident data with the vehicle data).Through this study we will make sure the accuracy of predicting correct severity should be maximised in order of fatal, serious and then slight so that financial loss incurred during the road accidents can be greatly diminished. The main reason to use ML and NN algorithms to predict the accident severity is because these algorithms are dependent only on input features and doesn't take into account any assumptions or presumptions. Also, the pipeline used for pre-processing data filters out data with multiple missing values making remaining data exact and thus reliable.

In the upcoming sections first, we will be providing the scope of the project, success, and completion criteria of this study, and then will share some more details about the uniqueness of this project. After this we will briefly discuss about some of the related works done in the same field i.e., accidents, from this we will move on dataset description and then finally we will discuss about the models employed for this study and the confusion matrix of the model out that performed the best.

PROJECT SCOPE

The goal of this case study is to minimize the overall *prevention value* without having the major loss/impact on the daily revenue of the businesses. In this study we will use the modern Machine learning algorithms to maximize the overall accuracy in such a way that accuracy percent for accident severity fatal and serious should be maximum. The target party for this study is the transportation and logistics industry as this study will not only provide the way to classify the accident severity based on different circumstances but also will give a high-level valuation of the loss that will be saved once we employ this model. Current solutions that are available in [8], [9] and [10] have focused majorly on either on how the factors/features impact the accident severity or on comparing and improving the overall model performance rather than improving the performance for the high priority accident severity based on high *prevention value*. Moreover, these papers have no interpretation of the results in lieu of logistics sector. This study will not help in predicting whether the accident will occur or not but it will only predict that given the circumstances if an accident occurs what might be the severity of the accident. Since this study is for my learning purpose so I do not have any specific client/stakeholder pertaining to the target sector but any company in logistic sector who has their functioning branch in UK can use this study for their business purpose. Thus, success criteria in this case study are to create the model that can predict all the three kind of accident severity successfully with the best-in-class overall accuracy compared with other papers. However, the completion criteria for this case study are to successfully (i.e., almost 100% accuracy) classify the fatal and serious class of accident severity out of all the expected accident circumstances.

LITERATURE SURVEY

In this section, we will discuss the related works that happened along similar lines in the past and what was its result. In the paper [8], to predict the severity of accident an RNN model was developed this paper used a very small dataset containing the accident records on NSE expressway, Malaysia for a period from 2009 to 2015. This model performance was then competed against MLP and BLR models. All these models achieved overall accuracy of 71.77% (RNN), 65.48% (MLP), and 58.30% (BLR). In the research paper [9], a potential relationship was studied between accident severity and the accident factors. Detailed analysis was made with the help of bar plots visualizing the distributions of accident severity with different factors and based on that a decision tree was constructed to categorize different factors. Based on this analysis when the decision tree is created it resulted in overall accuracy of 84.47%. In the research paper [10], a detailed analysis has been made in order to explore that could have impact on the number of accidents and their associated fatalities, after that ML algorithms like Random Forest, Gaussian, KNN and Decision Tree was employed to classify the severity of the accidents and the respective accuracy achieved in this paper is 82.54, 81.38, 80.39, and 71.08 percent. It has been highlighted in this paper that the overall accuracy is high because the model is able to predict slight accident severity at 94.9 percent accuracy however the accuracy of other two is as low as 8.23% and 0.42% respectively. In the paper [11], conventional ML models' comparison has been made with a proposed hybrid model i.e., BO-RF (Bayesian Optimization – Random Forest). The dataset used in here are from US accidents from Feb 2016 to Mar 2019. Here the F1-score of the BO-RF model (i.e., 0.57) is found to be better than other conventional models like ANN, KNN, SVM, RF with scores as 0.4, 0.28, 0.32, 0.54, 0.57 respectively. One insight that can be procured from this paper is the positive impact of Bayesian optimization on Random Forest is less only as precision of RF was found to be better than BO-RF.

METHODOLOGY

In this case study, I have built a pipeline which will perform some basic data cleaning, remove the missing data, convert categorical data into numeric, and then feed it into training and testing of the model (Splitting data into train and test). The flowchart of this can be seen in *Fig. 1*.

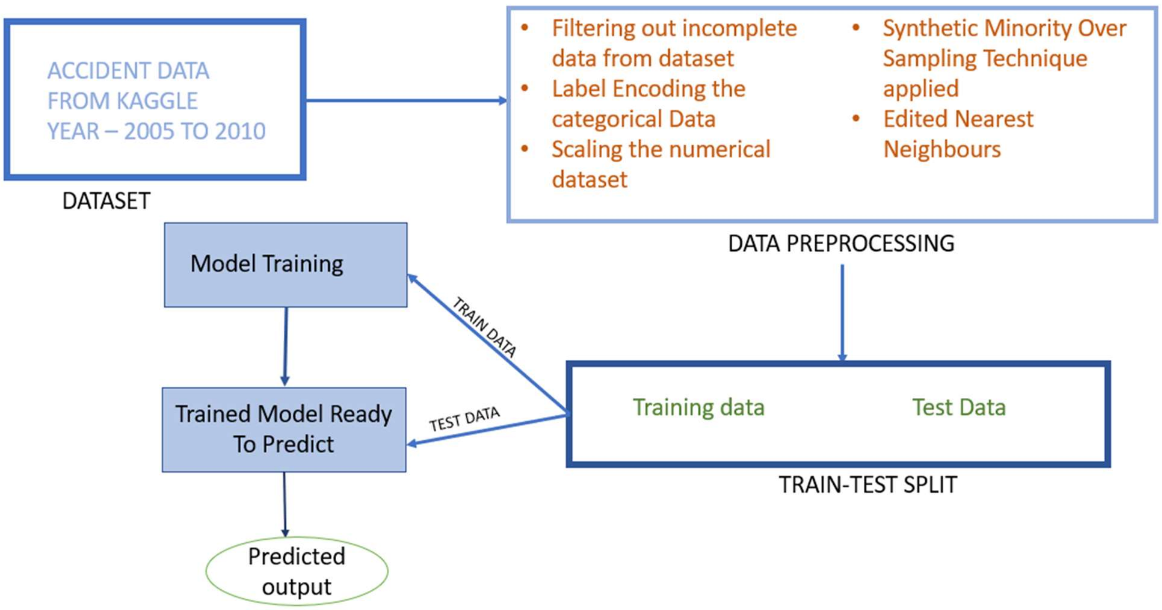


Fig. 1

As mentioned earlier, we had three classes in the target column i.e., "Accident_Severity" which is not properly balanced along to the dataset. In fact, out of 168549 rows of accident data that is to be used to fit and then evaluate model performance only 1367 no of accidents are marked as fatal, and 17828 are marked as serious. So, to make the data balanced we first employed SMOTE for the over sampling of minority classes and then ENN for under-sampling the data in order to remove any outliers. For that, there is a python library in which I used a class called SMOTENN that handles both as their arguments and does it in a sequential manner. Now with this over and under sampling technique combined it becomes more resistant to overfit, and along with this it also help in generating synthetic data for multiple classes with less number of inputs to make it balanced with majority class and removes unnecessary data (outliers kind) from the majority classes. From [9] and [10], it became obvious that Random Forest algorithm has worked best when compared with other conventional Machine algorithm so I used it and along with that I used two other models too to see how well its performing when hybrid ML models and Neural Network are used instead so we bring in XGBoostRandomForest Classifier algorithm and multi-layered Neural Network with "Relu" as input and hidden layer activation function and "SoftMax" for output layer activation function.

In the next page there will be two tables shown in which *Table 1* will be the list of Algorithms along with the parameters used in the case study for classification and the *Table 2* will be the SMOTE and ENN parameters which is passed as an argument in SMOTEEN sampling technique. Please Note: We have not used Logistic Regression with resampling technique because it did quite worse initially before itself.

Algorithms	Parameters
Logistic Regression	random_state=40, multi_class='multinomial', max_iter=100
Random Forest	n_estimators=100, criterion='gini'
XGBoostRandomForest	n_estimators=100, subsample=0.8, colsample_bynode=0.2
Artificial Neural Network	1. Input Layer (Dense): units=8, activation='relu' 2. Hidden Layer (Dense): units=16, activation='relu' 3. Output Layer (Dense): units=3, activation='softmax'

Table 1

SMOTENN Over and Under sampling technique	
Technique	Parameters
SMOTE	sampling_strategy='auto', random_state=0, k_neighbors=5, n_jobs=4
ENN	sampling_strategy='auto', n_neighbors=3, kind_sel='all', n_jobs=4

Table 2

To apply SMOTENN technique for resampling was necessary because when we tried to train the model with the unbalanced data even the best performing model i.e., Random Forest's overall accuracy was quite good but it was highly biased towards the majority class. In the next section, we will discuss about the dataset used for this case study after which we will start the next section coming back to the explanation of above outcome using the results that we got.

DATA DESCRIPTION

The data used for the classification of accident severity has been taken from Kaggle dataset [12] in w hich we have used *accident_data.csv*. The data was very large so while pre-processing whichever row was incomplete, we simply remove that row. But before that we also considered the columns or featu res that have the greatest number of missing values because that means that feature has very less to of fer so we removed most of those columns but retained some of them like *'1st_Road_Class', '2nd_Roa d_Class'*, etc. And the reason was based on common sense that road characteristics play a major role i n the severity of the accidents. Similarly, some other features were also kept in the data while perform ing classification. Here is the list of the features that are used from the original dataset along with the description of it has been used (whether transformed or kept intact).

1. **1st_Road_Class:** This is the road class of 1^{st} road the accident happened on. This is categoric al data with 'A', 'B', 'C', 'Motorway', 'A(M)' as categories. All the categorical data are label encoded in order to make it numeric so that it can be used by different models.
2. **2nd_Road_Class:** This is road class of 2nd road the accident happened on and has values exa ctly same as 1st road class hence handled in the same way.
3. **Carriageway_Hazards:** This is an observation of any hazards in the road at the time of the a ccident e.g., animals or pedestrians in the road. It is also a categorical data and is handled sam e as 1^{st} Road Class.
4. **Date:** The Date when the accident occurred. This feature was first converted to date time obje ct then the date and the month (both real number) is extracted from it.
5. **Day_of_Week:** The day the accident happened. Handled same as 1^{st}_Road_Class.

6. **Did_Police_Officer_Attend_Scene_of_Accident:** This feature has only 3 options: 1 - Yes. 2 - No. 3 - No, the accident was reported by a self-completion form. The transformation is done same as 1st Road Class feature.
7. **Junction_Control:** This tells what controls are in place to control traffic at the junction. Bein g a categorical variable, it has been handled like 1st Road Class.
8. **Junction_Detail:** It gives the info about the type of junction at the location of the accident. H andled like 1st Road Class.
9. **Latitude:** One of the geographical co-ordinates where the accident occurred. It is a numeric d ata so no transformation needed.
10. **Light_Conditions:** The light condition at the time of the accident. This being categorical data has been handled like 1st Road Class.
11. **Time:** The time at which the accident happened. This feature was categorised as morning, aft ernoon, evening and night time but just used numeric values instead.
12. **Longitude:** The other geographical co-ordinate where the accident occurred. It is numeric as well.
13. **Number_of_Casualties:** Number of those killed or injured in the accident.
14. **Number_of_Vehicles:** Number of vehicles involved in the accident.
15. **Pedestrian_Crossing-Human_Control:** This feature tells if there was a human controlled cr ossing present at the scene of the accident.
16. **Pedestrian_Crossing-Physical_Facilities**: This feature tells if there were physical facilities p resent at the scene of the accident.
17. **Road_Surface_Conditions:** The condition of the road when the accident took place. Its categ orical hence encoded like 1st Road class.
18. **Road_Type:** Categorical data with well explanatory name and is handled like 1st road class fe ature.
19. **Special_Conditions_at_Site:** This feature tells if there were any other factors which could ha ve caused the accident ie. oil on the road, faulty traffic lights etc. It's handled like 1st road clas s feature.
20. **Urban_or_Rural_area:** Categorical data with well explanatory name and is handled like 1st r oad class feature.
21. **Weather_Conditions:** It tells what was the weather like at the time of the accident. Its catego rical hence encoded like 1st Road class.
22. **In_Scotland:** Categorical data with well explanatory name and is handled like 1st road class f eature.
23. **Speed_Limit:** It tells the value of the speed limit where the accident took place. It's a numeri cal feature.
24. **Year:** Year of the accident.

After converting the data to numerical I then performed scaling in order to remove domination from a ny feature just because of its high numerical value. In the next and final section, I will provide the res ult of the experiments performed in the case study and provide conclusions based on this.

EXPERIMENTAL RESULTS AND CONCLUSION

As concluded in Methodology section, The reason I need to employ the over and under sampling algo rithm is to remove the bias from majority class. *Table 3* (given below) is the result that proves this ob servation.

Model	Overall Accuracy	Balanced Accuracy	Accident_Severity	Precision	Recall	F1-Score
Random Forest	0.86	0.34	Slight	0.87	0.98	0.92
			Serious	0.21	0.03	0.06
			Fatal	0.06	0	0.01
Logistic Regression	0.87	0.33	Slight	0.87	1	0.93
			Serious	0.22	0	0
			Fatal	0	0	0
XGBoost	0.87	0.335	Slight	0.87	1	0.93
			Serious	0.46	0.01	0.01
			Fatal	0.12	0	0

Table 3

As per the above table all the models were able to identify only accidents with slight severity (the ma jority class) perfectly i.e., 87% however other two classes were almost misidentified. Because of this reason we moved on and used SMOTENN technique and the results obtained are shown below:

Model	Overall Accuracy	Balanced Accuracy	Accident_Severity	Precision	Recall	F1-Score
Random Forest	0.85	0.85	Slight	0.9	0.63	0.74
			Serious	0.72	0.93	0.81
			Fatal	0.99	1	0.99
XGBoost Random Forest	0.57	0.57	Slight	0.64	0.55	0.59
			Serious	0.46	0.37	0.41
			Fatal	0.58	0.78	0.67
Artificial Neural Network	0.59	0.59	Slight	0.67	0.69	0.68
			Serious	0.49	0.38	0.43
			Fatal	0.59	0.7	0.64

Table 4

Comparing the best performing algorithm in both the tables i.e., Random Forest, it was observed that even though overall Accuracy of the model was reduced by 1 percent but there is significant improve ment in the Balanced Accuracy, Precision, Recall and F1-score values after resampling was done.
Comparing it with other research papers it was observed that our model performed slight better than t hat of [9] as their accuracy was 84.47% and our model accuracy is 85.2% which is still better than the accuracy obtained in [10] which is 85.08% although the model trained here is biased towards acciden t_serverity type 3 as its precision is 96.58% and others have precision not even greater than 25 %.
Coming back to the current case study as we have seen our random forest model performed best base d on different scientific scores. But now evaluate the performance of the model based on its usefulne ss in decreasing the *prevention value* of the accidents. To do this let's view the confusion matrix of t he for Random Forest with SMOTENN model below:

Accident data		Predicted Severity			Total Number of accidents / severities
		Slight	Serious	Fatal	
Actual Severity	Slight	18807	10794	125	29726
	Serious	2029	27377	88	29494
	Fatal	90	39	29582	29711

Table 5

Based on [6], in 2012 the cost per accident with severity 'slight' is £23,336 while for 'Serious' is £2 19,043 and for 'fatal' is £1,917,766. To understand the usefulness of this model in real time let's say A logistic company say XYZ has made precautionary plan for accidents for each of the severity uniq uely and trained the driver and made other arrangements involved during transit. Since the order of se verity lies as Fatal (most harmful) to Serious to Slight (least harmful) so whatever arrangements has b een made to tackle fatal severity can still be used in case of Serious or Slight accident severity cases j ust the overall budget/cost will be little more but damage may still be prevented. However, the vice-v ersa will not be true i.e., if the accident that occurred was of fatal nature and the preparation done was for Slight severity than this will not only lead to loss because of the accidents but also the preparation cost incurred will also be lost. Now in *Table 5* the cells coloured green are the safe conditions when t he preparation done is inline with the actual accident that occurred in the given conditions which is e qual to approx. 85% so our model can help the firm be prepared for 85 out of 100 times with the exac t same severity. The cells coloured in light yellow are the situation where the vehicle and driver who are involved in the accident are prepared for worse conditions than which happened thus the only ove rhead expense incurred on preparation will be extra money spent but the cargo and the human casualt y still may be avoided. Now, the cells coloured in red are the harmful ones since in this situation the d river and vehicle are prepared for less severe accidents then what they actually have faced in which c ase the loss will be maximum, but our model has reduced number of accidents that falls under these s ituations to quite less percent to be exact approximately 7 percent of all and only 0.4 percent of fatal a ccidents serious accidents falls under this dangerous situation.
Thus, it can be concluded that my model is economically very useful and reliable for any sector whos e business depends on reallocation of commodity or products from once place to another via road in UK.

ACKNOWLEDGEMENTS

This case study was done under the flagship of Birla Institute Of Technology and Science, Pilani Indi a as part of ongoing Dissertation project.

REFERENCES

[1] 4 WAYS ROAD CRASHES IMPACT THE ECONOMY - https://www.togetherforsaferroads.org/4-ways-road-crashes-impact-the-economy/#:~:text=Road%20crashes%20also%20bear%20a,year%20to%20the%20global%20economy.

[2] Investing in Road Safety A Global Imperative for the private sector: https://www.togetherforsaferroads.org/wp-content/uploads/2020/10/Investing-in-Road-Safety-A-Global-Imperative-for-the-Private-Sector.pdf

[3] Gorea, Rakesh. (2016). Financial impact of road traffic accidents on the society:. International Journal of Ethics, Trauma & Victimology. 2. 10.18099/ijetv.v2i1.11129.

[4] Topolšek, D., Babić, D. & Fiolić, M. The effect of road safety education on the relationship between Driver's errors, violations and accidents: Slovenian case study. Eur. Transp. Res. Rev. 11, 18 (2019). https://doi.org/10.1186/s12544-019-0351-y

[5] The Critical Role of Transportation in Business and the Economy : https://www.informit.com/articles/article.aspx?p=2171313#:~:text=The%20cost%20of%20transportation%20can,cost%20more%20than%20slower%20modes.

[6] Reported Road Casualties in Great Britain: 2012 - Annual Report : https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/244913/rrcgb2012-02.pdf

[7] Royal Mail Fined £1.6 Million For Serious Injuries In Yard Collision : https://www.fisherscogginswaters.co.uk/blog/article/307/royal-mail-fined-16-million-for-serious-injuries-in-yard-collision

[8] Sameen, M.I.; Pradhan, B. Severity Prediction of Traffic Accidents with Recurrent Neural Networks. Appl. Sci. 2017, 7, 476. https://doi.org/10.3390/app7060476

[9] Z. Cheng, B. Liu and J. Huang, "Causal Analysis of Road Safety Accidents in Britain Based on a Univariate Decision Tree Method," 2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI), Zakopane, Poland, 2022, pp. 436-441, doi: 10.1109/ICDACAI57211.2022.00092.

[10] S. Haynes, P. C. Estin, S. Lazarevski, M. Soosay and A. -L. Kor, "Data Analytics: Factors of Traffic Accidents in the UK," 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), Leeds, UK, 2019, pp. 120-126, doi: 10.1109/DESSERT.2019.8770021.

[11] Traffic Accident Severity Prediction Based on Random Forest - https://mdpi-res.com/d_attachment/sustainability/sustainability-14-01729/article_deploy/sustainability-14-01729.pdf?version=1643797753

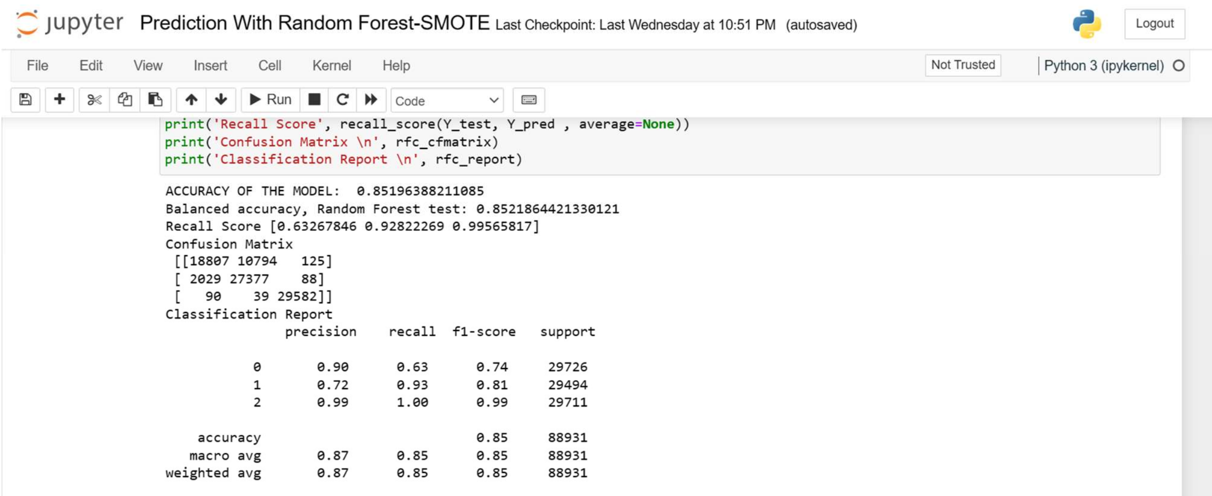
[12] UK Road Traffic Collision Dataset : https://www.kaggle.com/datasets/salmankhaliq22/road-traffic-collision-dataset

APPENDIX-I

HOW TO GET THE CODE FILES AND RUN IT

- First go to github url https://github.com/SarhadGautam/Accident-Severity-Prediction and download the files in the local repository.
- Then open [12] link and download accident_data.csv from Kaggle.
- Unzip the downloaded file in the same folder where all scripts are kept.
- Now, when you execute the script wait for it to complete.

Once executed you can see the results like below picture:



jupyter Prediction With Random Forest-SMOTE Last Checkpoint: Last Wednesday at 10:51 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Help

Not Trusted

Python 3 (ipykernel)

Run Code

```
print('Recall Score', recall_score(Y_test, Y_pred , average=None))
print('Confusion Matrix \n', rfc_cfmatrix)
print('Classification Report \n', rfc_report)
```

```
ACCURACY OF THE MODEL:  0.85196388211085
Balanced accuracy, Random Forest test: 0.8521864421330121
Recall Score [0.63267846 0.92822269 0.99565817]
Confusion Matrix
 [[18807 10794   125]
 [ 2029 27377    88]
 [   90    39 29582]]
Classification Report
               precision    recall  f1-score   support

           0       0.90      0.63      0.74     29726
           1       0.72      0.93      0.81     29494
           2       0.99      1.00      0.99     29711

    accuracy                           0.85     88931
   macro avg       0.87      0.85      0.85     88931
weighted avg       0.87      0.85      0.85     88931
```