response for the dependent variable, i.e. Age of Casualty. The actual data is depicted using the "blue" line & the predicted data is depicted using the "orange" line. As it's evident from the graph, the algorithm is able to predict the expected data to a large extent, as at most of the points in the graph, the lines for the actual data & predicted data overlap.

As shown in Table 1, the performance metric accuracy is calculated for all the machine learning models, applied on the respective attributes. The overall accuracy for Accident Severity ranges between 86.64% & 87.73%. Amongst the ML models executed, SVM gave the highest accuracy of 87.73%, closely followed by RF & KNN, with 86.86% & 87.73% respectively.

The overall accuracy for the attribute Rural or Urban, as shown in table1, ranges between 92.78% & 94.58%. The most robust performance for accuracy is given by Random Forest, i.e. 94.58%. The algorithms KNN & SVM also showed promising results, with an accuracy of 92.78% & 94.18% respectively.

The accuracy for the attribute Sex of Casualty, as shown in table1, ranges between 58.45% & 64.84%. SVM gave the better performance of all the algorithms applied, with an accuracy 64.84%. KNN & Random Forest gave an accuracy of 58.45% & 62.39% respectively.

All the machine learning models gave very low or weak performance for the attribute Sex of Casualty. This may be due to the reason that this attribute has a high covariance & varies due to high degree of dependency on other attributes of the dataset. Due to this, another problem faced while performing the analysis for this particular attribute was that, there was duplication of rows at the time of loading the dataset.

The findings of our study can be used for analysis & prediction of various attributes & scenarios that lead to road accidents & other hazardous situations in the transport industry. The government & concerned authorities can use these results & analysis to better understand the various causes that result in road accidents & enforce strict rules & regulations to prevent such situations from taking place in the future. Also, the hospitals & medical emergency services could be boosted in areas that are most prone to road accidents, so that immediate treatment or help is available.

In our current analysis, we have mainly focused on identifying the reasons that lead to road accidents & identification of important insights/ trends that can be inferred from the data. As part of our future plans for research in this field, we will make an attempt to apply multivariate modelling techniques & other ML methods, such as Artificial NN, Deep Learning etc. which may help us to resolve the stated problem & identify the accident prone area to a greater extent.

References

[1] WHO (2018). "Global status report on road safety 2018 (violence and injury prevention)." Geneva, Switzerland.

[2] Chen. And C. (2017). "Analysis and forecast of traffic accident big data." ITM Web of Conferences EDP Sciences, 12, 04029.

[3] Krishna, S., S, S. K., S, S. K., and Mungara, D. J. (2017). "Traffic management using big data analytic tool." International Journal of Scientific Research In Computer Science Engineering And Information Technology (IJSRCSEIT), 2, 777-781.

[4] Ismael, K.S., and Razzaq (2017). "Traffic accidents analysis on dry and wet road bends surfaces in greater Manchester – UK." Kurdistan Journal of Applied Research, 2(3), 284-291.

[5] Zhang, J., Li, Z., Xu, and C. (2018). "Comparing prediction performance for crash injury severity among various machine learning and statistical methods." IEEE Access, 6, 60079-60087

[6] Zheng, W., & Tropsha, A. (2000). Novel variable selection quantitative structure− property relationship approach based on the k-nearest-neighbor principle. *Journal of chemical information and computer sciences*, *40*(1), 185-194.

[7] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

[8] Joachims, T. (1999). Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, *19*(4).

[9] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, *5*, 30-35.