Predicting Accident Severity with Machine Learning: A Critical Weapon in Emergency Management of Road Traffic Accident

**Abstract:**

**Business Understanding:** Road traffic accidents (RTAs) and the associated injuries are a significant public health challenge in countries worldwide. About 1.25 million people and up to 50 million people are injured yearly, and RTAs represent the 10th leading cause of death globally. RTAs are also the number one cause of mortality of the young between the ages of 5 and 29, especially in developing countries where timely pre-hospital emergency care and subsequent transportation of accident victims to appropriate health facilities are often limited. Timely and appropriate medical care of RTA victims can significantly reduce accident and injury outcomes. RTA victims’ survival can be significantly improved if they are attended to in a timely manner and cared for by appropriate and qualified medical personnel (Sam et al., 2019).

Objective: To predict accident severity using machine learning (ML) and factors associated with RTAs.

**Methods:** A publicly available dataset of collisions or RTAs data from the Seattle Department of Transportation (SDOT) containing attributes or observations relating to the accident in Seattle City in the period 2004 to present. The data was cleaned, prepared for training three ML algorithms. The performance of each of the algorithms was measured and compared to select an optimal classification model.

**Results:** Conclusion: The K-Nearest Neighbor classifier (F1-score = 66%) outperformed both the Decision Tree (F1-score = 63%) and the ensemble Gradient Boosting Tree classifier (F1-score = 65%). However, a performance score of 66% indicates that the model development could be improved.

1. **Introduction**

Road traffic accidents (RTAs) and the associated injuries are a significant public health challenge in countries worldwide. About 1.25 million people and up to 50 million people are injured yearly and RTAs represent the tenth leading cause of death globally (Assi et al., 2020; Gebresenbet & Aliyu, 2019). RTAs are also the number one cause of mortality of the young between the ages of 5 and 29 , particularly in developing countries where timely pre-hospital emergency care and subsequent transportation of accident victims to appropriate health facilities are often limited (Gebresenbet & Aliyu, 2019). The number of RTAs and victims have been increasing globally due to the growing global population and rapid adoption of motorization (Assi et al., 2020). Timely and appropriate medical care of RTA victims can significantly reduce accident and injury outcomes. RTA victims’ survival can be significantly improved if they are attended to in a timely manner and cared for by appropriate and qualified medical personnel (Sam et al., 2019). There is a need for improvements in pre-hospital emergency services to improve the survival rate of RTA victims.

***Business Problem***

Capturing data about RTAs, particularly the factors that affect accident severity levels could be helpful in proactively predicting accident severity. Traditional statistical techniques that have been employed to predict the severity of RTAs have, such as Ordered Probit (OP) and Logistic Regression (LR) have shown some inherent limitations and often lead to inaccurate predictions (Assi et al., 2020). It is expected that techniques based on ML algorithms, which can model the non-linear relationships of the factors associated with RTA severity, could produce more accurate predictions. In fact, some researchers have reported better performance of the ML algorithms compared to the traditional statistical techniques (Assi et al., 2020).

The objective of this project was to predict accident severity using information that can be easily captured at collision or crash sites. Capturing such information could help emergency management teams enabling them to predict accident severity, to dispatch of appropriate emergency equipment, vehicle, and personnel or to provide appropriate and timely guidance to the nearest health facility. Such a capability would empower emergency management teams, trauma centres, and ultimately lead to improvements in road traffic safety, even in remote areas with limited access or emergency services.

1. **Data Understanding**

The data for this study was obtained from a publicly available dataset provided by the SDOT, which is collisions data containing weekly information about accidents in Seattle City for all collision types from 2004 to present. The dataset contains a total of 40 attributes, such as the location of the collision (in latitude, longitude, and a description of the general location of the collision), the collision type, total number of people involved, the number of vehicles involved, the number of injuries, the number of fatalities, the date and time of the accident, whether the a driver involved was speeding, whether a driver involved was under the influence of drugs or alcohol, collision type, weather condition, road condition, the accident severity, and many other factors.

1. **Methodology**

***Data Cleaning and Preparation***

The sample collisions dataset shared as part of the course was missing several key attributes such as the number of injuries and number of fatalities. Therefore, the dataset was downloaded directly from the SDOT web portal. The data was then cleaned, pre-processed in readiness for training the ML algorithms selected. The SDOT codes accident severity into four categories (0=unknown, 1=property damage, 2=injury, 2b=serious injury, and 3=fatality). In the pre-processing, the severity was coded such that all injuries or fatality categories were recoded as injury (1=injury) and everything else was property damage or similar (0=property damage). The severity, which was initially labelled as “SEVERITYCODE” was renamed to simply as “severity”. The incident date-time column was changed to a Pandas datatime. All data rows with missing entries for severity, latitude, or longitude were dropped. The resultant useful dataset included 2132674 rows.

The data was further prepared and transformed for training the ML algorithms. The data types of the various data columns were checked and revised, for data float and integer columns, where necessary. The latitude and longitude, which captured the location information, were transformed into a single attribute using a Haversine formula (StackExchange, n.d.) and named simply as “location”. A review of the computed location data revealed that most of the collision, particularly injuries, occurred within approximately one standard deviation from a hot spot location (or mean point), as seen in Figure 1. A further variable named “locality” was created based on how the location data was distributed in the location histogram bins, to capture information about the general location of the accidents and the proximity to the hot spot. The date-time data was transformed into an “hourofday” variable given the distribution of collisions and, particularly injuries, was spread throughout the day. Further visualization and exploration of the data showed that most of the collisions, particularly those with injuries, involved fewer than ten people. Similarly, collisions involving pedestrians involved fewer than two people. Refer to Figure 1. The focus of analysis for these variables (person and pedestrian count) was restricted to those counts between 1 and 10 or 2 respectively. The same approach was applied to the injuries, serious injuries, fatalities, the number of vehicles involved, and SDOT code.

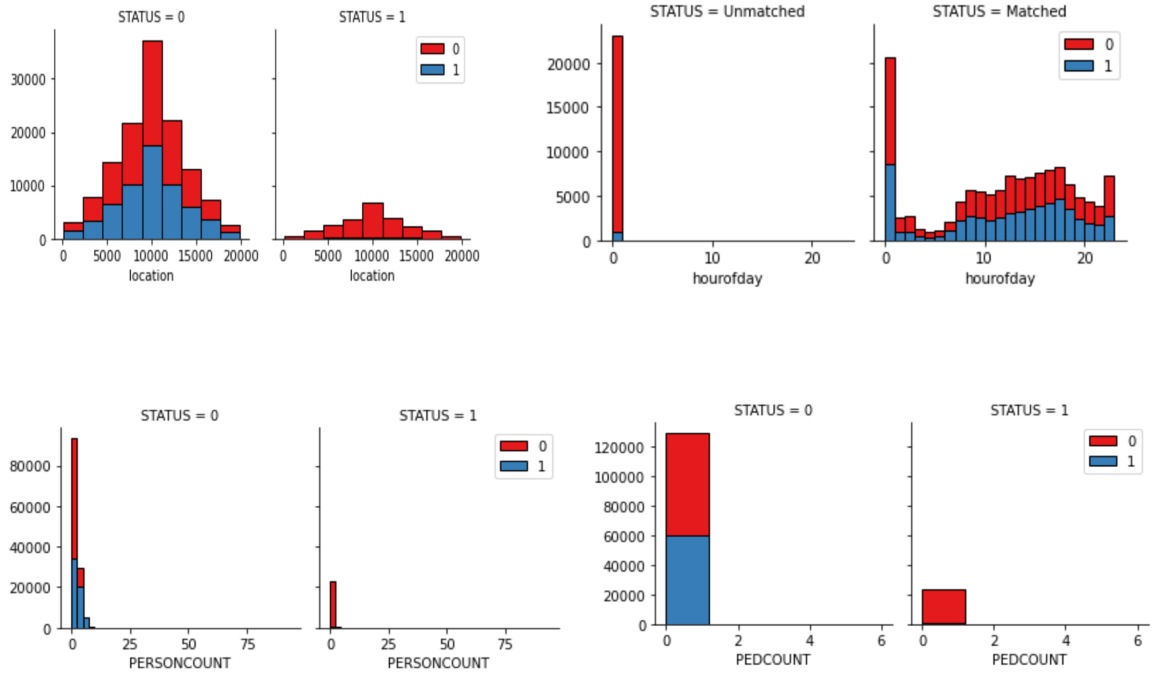


Figure 1: Distribution of collisions by location, hour of the day, person count, and pedestrian count

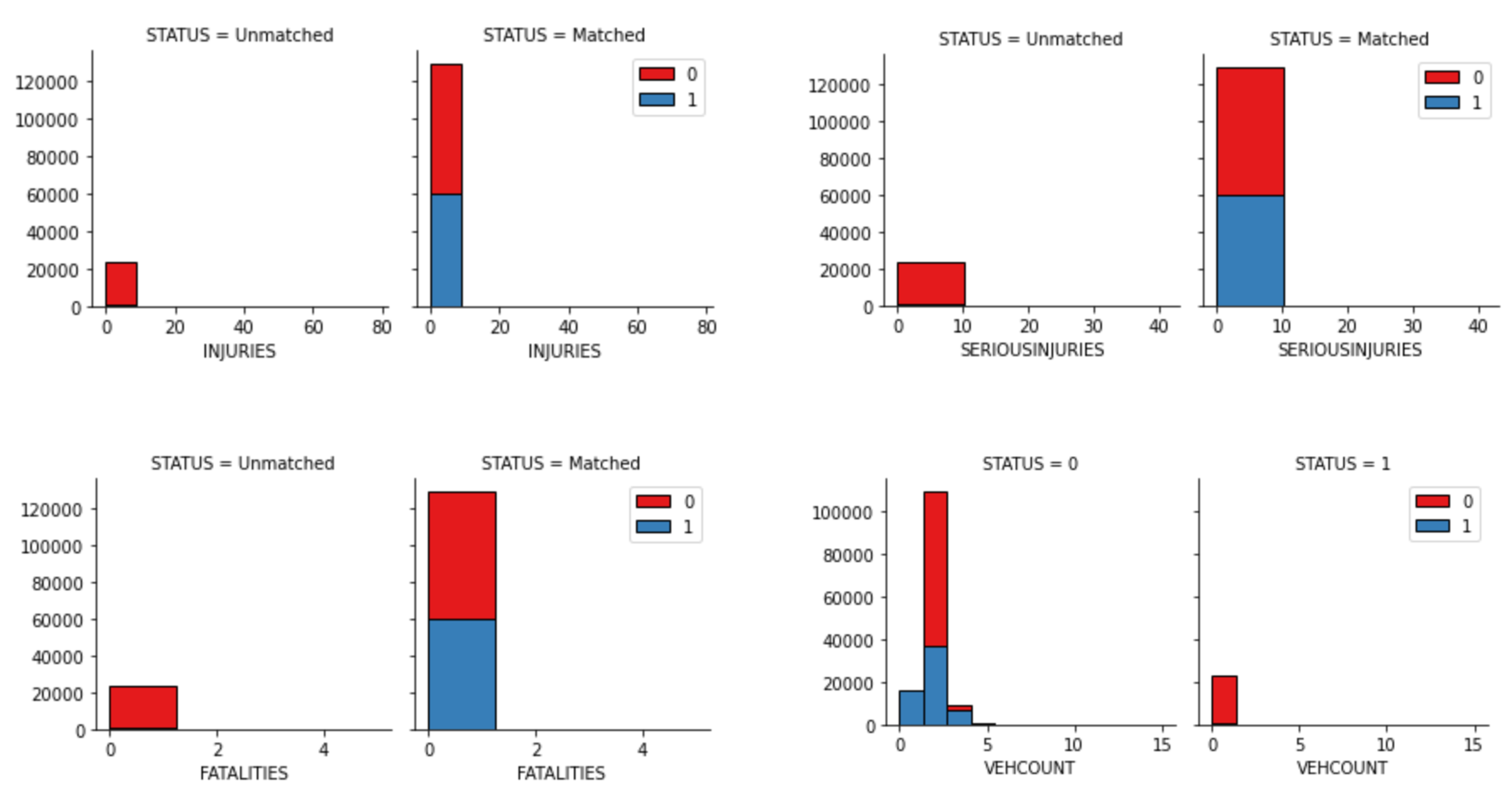


Figure 2: Collisions by number of injuries, serious injuries, fatalities, and number of vehicle involved

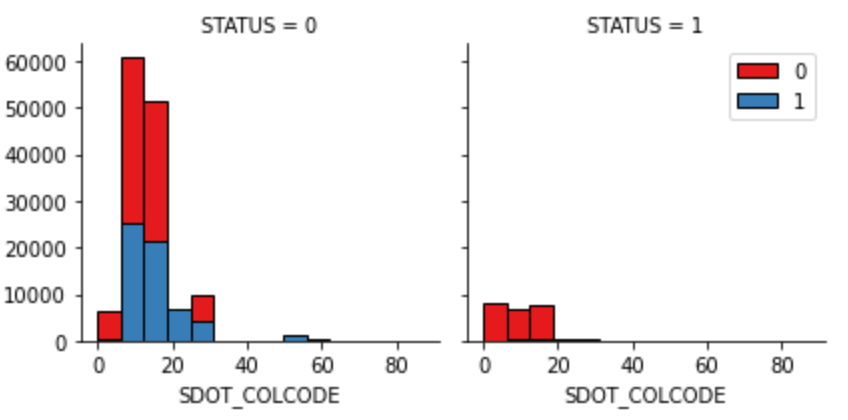


Figure 3: Collisions by SDOT code

The distributions of the collisions for the injuries, serious injuries, fatalities, and vehicle count can be seen in Figure 2 and 3. The data showed that most collisions were associated with codes between 10 and 15 and most injuries matching that range involved vehicles entering at an angle, sideswipes, read ending, or sudden right or left turns.

Pandas one-hot encoding was applied to categorical variables such as collision type, weather conditions, road conditions, and light conditions. The added additional but valuable data attributes to the overall set of collision 59 candidate attributes. The 59 candidate attributes included some redundant and some non-numerical attributes from the original dataset and these were eliminated from further consideration, recognizing that they may be considered in the future if needed to improve the predictability of the ML models. A subset of 48 numerical features was selected from the 59 to be considered for feature selection for the model development and testing.

***Balancing Dataset***

In the data preparation stage, it was evident that the dataset was unbalanced with respect to the two classes (1=injury and 0=property damage). Therefore, both an up-sampling and a down-sampling balancing techniques were evaluated for balancing the dataset (EliteDataScience, 2017; Moosavi et al., 2019; Ramya et al., 2019). Eventually, the down-sampling approach was selected since the original dataset was large, the resultant reduced balanced dataset was still large, and down-sampling retains the benefit that all data points are contained in the original dataset.

***Splitting Dataset***

The dataset was randomly split into two for model development and testing: 1) reserved for phase 1 (model development) training and evaluating the ML algorithms, and 2) reserved for phase 2 (model testing) final testing only (Sanjay, 2020). The first dataset would subsequently be split again in the phase 1 for training and testing to select the optimal ML algorithm.

***Data Normalization***

Data normalization was conducted using the standard scaler in Scikit-Learn. The scale differences between the dataset attributes was not too, nevertheless, the normalization was done ate minimal computational cost.

***Selecting the Best Features***

With a large number of features, it is to be expected that some of the 48 attributes will not be good enough features for subsequent model development as not all features will contribute equally to the predictive power of the algorithm. The importance of features was determined by a Decision Tree algorithm and the top 24 features were selected for further processing and model development (Serengil, 2020).

***Dimensionality Reduction***

Generally, if the number of attributes in a dataset is large using all the attributes or features typically does not result in a much better predictive model performance. In fact, with a large dataset a large number of features may actually result in poorer performance depending on the type of algorithm (Vickery, 2020). In this case, the number of futures was large (24 or more) and it was necessary to reduce the number of features to minimize the computational cost without compromising the predictive accuracy of the models. So, dimensionality reduction by principle component analysis using Scikit-Learn was used to reduce the 24 attributes to two latent variables that were used for model development and testing.

***Model Development***

Three models were developed in phase 1, using the first test and evaluation test dataset. The algorithms developed included K-Nearest Neighbor (KNN), Decision Tree (DT), and Gradient Boosting Decision Tree (GBDT). Each model was developed for an optimal set of parameters using the training data and tested using the test data. The KNN was chosen because it is recognized as one of the simplest, it makes no assumptions about the nature of the data, and topmost classification algorithms commonly used by Data Scientists in a variety of applications, even though it can be computationally intensive (Bronshtein, 2019; Navlani, 2018). The Decision Tree algorithm was chosen because it is generally considered attractive for exploratory analysis, it can handle multi-dimensional data, it uses different feature subsets and decision rules at different stages of classification, and it facilitates an easier interpretation of the results (Du & Sun, 2008; Serengil, 2020; Tyagi, 2019; Vickery, 2020). The Gradient Boosting Decision Tree algorithm was chosen because it is an ensemble ML algorithm and it effectively reduces the impact of class imbalance (Xuan et al., 2019).

***Model Testing***

As indicated above, the model development, evaluation, and testing were conducted in two phases. Phase 1 involved developing the three models and evaluating them using a separate randomly split dataset for training. In phase 2, the performance of each model was then tested using a separate dataset that was set aside after the first random splitting exercise. The dataset was first normalized, then the 24 features were then selected as input to dimensionality reduction, and then the resultant two latent features were used as input to cross validation using scikit-learn. The model performances were compared using F1-Scores, which is the harmonic means of model precision and recall and better for when there is class imbalance, and confusion matrices (Huilgol, 2019).

1. **Presentation of Results**

Out of the 48 features available, 24 were selected for the model development stage based on a decision tree algorithm ordered list of feature importances, as seen in Figure 4. Dimensionality reduction was used to reduce the 24 features to two (2). A scatter plot of the final two features representing the two classes (1=injury and 0=property damage) can be seen in Figure 4, which seems to indicate that a complex decision boundaries would be required to achieve high classification performance.

The default accuracy measure was used to determine the optimal parameters for the three classifier algorithms. The optimal parameters, as seen in Figure 5, were then used to develop the final models and the performance of the three algorithms were compared to select the optimal model was then recommended for the accident severity prediction classification task. The KNN classifier produced the best performance overall (F1-score = 66%) both during model training and evaluation testing as well as during the final cross validation, as seen in Figure 6. The corresponding confusion matrices can be seen in Figure 7.

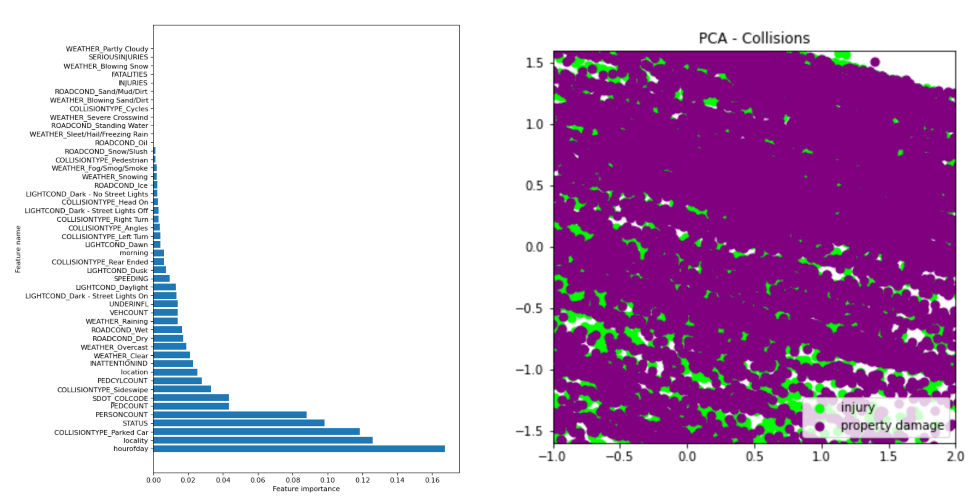


Figure 4: Feature importances in ascending order and the scatter plot for the two classes

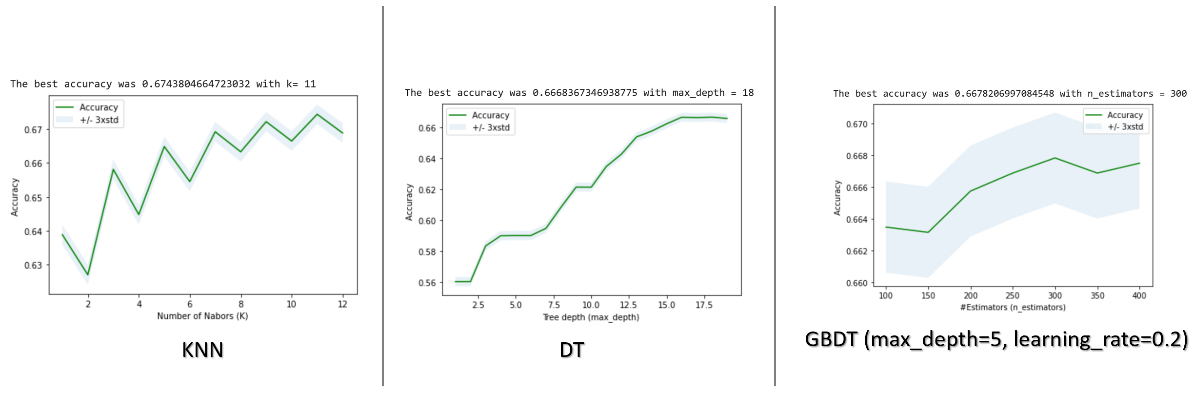


Figure 5: Plots of the optimal parameter for the three algorithms

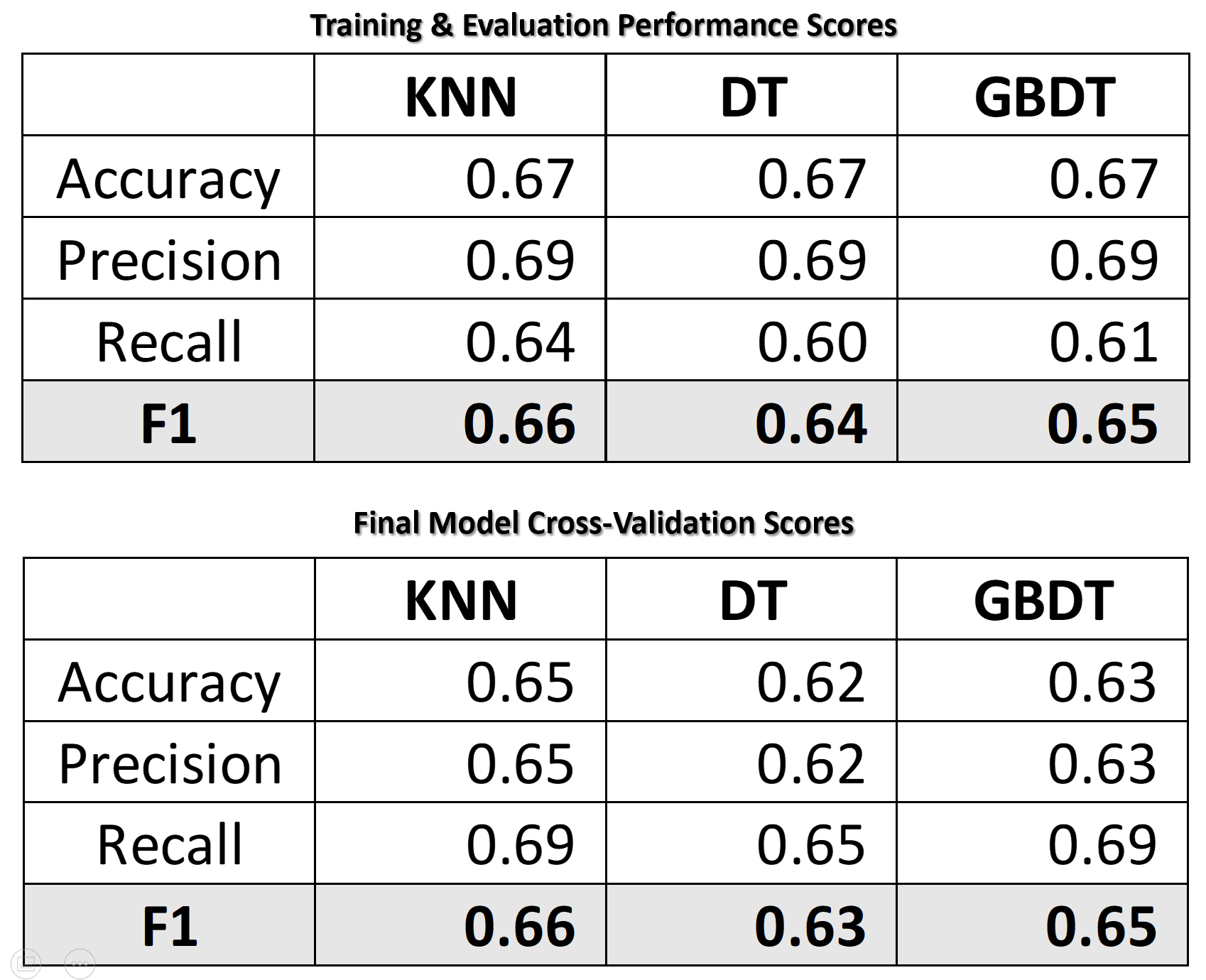


Figure 6: Training & testing performances vs final cross validation scores

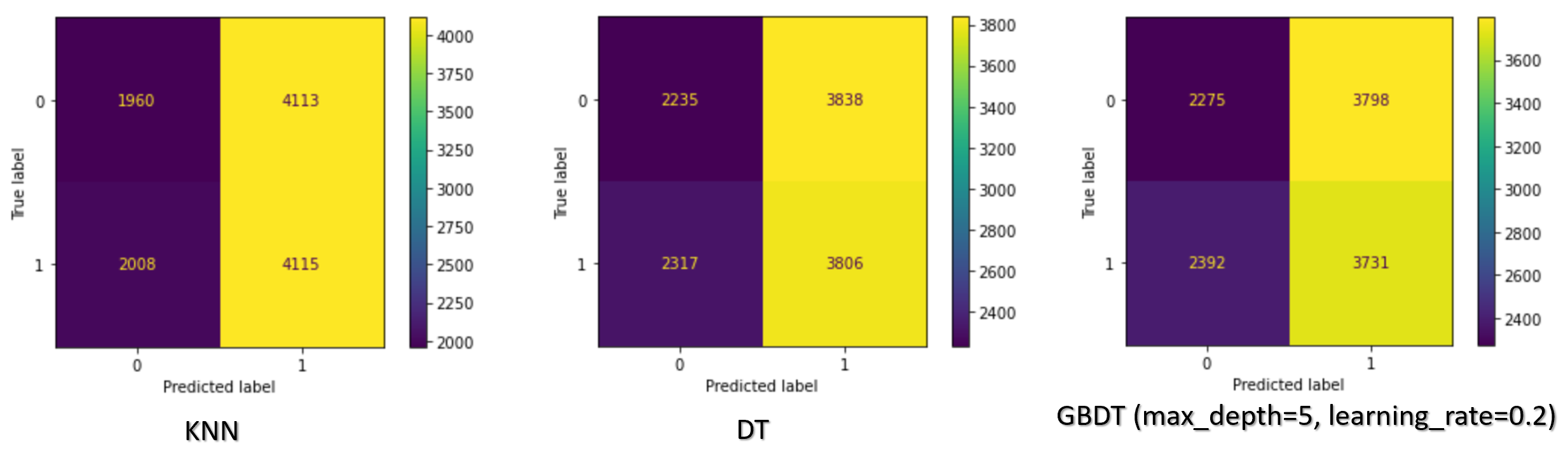


Figure 7: Confusion matrices from cross validation

1. **Discussion**

Overall, the KNN classifier produced the best performance (F1-score=66%) in the final model testing, as seen in Figure 6. The training and testing performance scores were very comparable with the final test scores indicating that there was no overfitting. However, with a performance score of 66% it was concluded that there was room for further refinements in the model development. Nevertheless, the final testing scores revealed improvements in recall (true positive) scores, which is a positive sign as a higher proportion of true positives (severity = injury) would be identified as positive or a better ability of the classifier to identify true positives or a reduction in false negatives. This is also backed up by the results of the confusion matrices.

1. **Conclusion**

Overall, the KNN classifier produced the best performance (F1-score=66%) in the final model testing, as seen in Figure 6. The training and testing performance scores were very comparable with the final test scores indicating that there was no overfitting. However, with a performance score of 66% it was concluded that there was room for further refinements in the model development process, possibly in evaluating different classifier algorithms, a more extensive optimization by using grid search to determine model parameters, or a review of the data preparation and pre-processing steps.

**References**

Assi, K., Rahman, S. M., Mansoor, U., & Ratrout, N. (2020). Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. *International Journal of Environmental Research and Public Health*, *17*(15), 5497. https://doi.org/10.3390/ijerph17155497

Bronshtein, A. (2019, May 6). *A quick introduction to k-nearest neighbors algorithm*. Medium. https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

Du, C.-J., & Sun, D.-W. (2008). Object classification methods. In D.-W. Sun (Ed.), *Computer Vision Technology for Food Quality Evaluation* (pp. 81–107). Academic Press. https://doi.org/10.1016/B978-012373642-0.50007-7

EliteDataScience. (2017, July 5). *How to Handle Imbalanced Classes in Machine Learning*. EliteDataScience. https://elitedatascience.com/imbalanced-classes

Gebresenbet, R. F., & Aliyu, A. D. (2019). Injury severity level and associated factors among road traffic accident victims attending emergency department of Tirunesh Beijing Hospital, Addis Ababa, Ethiopia: A cross sectional hospital-based study. *PloS One*, *14*(9), e0222793. https://doi.org/10.1371/journal.pone.0222793

Huilgol, P. (2019, August 24). *Accuracy vs. F1-Score*. Medium. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 33–42. https://doi.org/10.1145/3347146.3359078

Navlani, A. (2018, August 2). *KNN classification using scikit-learn*. DataCamp Community. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

Ramya, S., Reshma, SK., Manogna, V. D., Saroja, Y. S., & Gandhi, G. S. (2019). Accident severity prediction using data mining methods. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *5*(2), 528–536. https://doi.org/10.32628/CSEIT195293

Sam, E. F., Blay, D. K., Antwi, S., Anaafi, C., & Adoma, J. A. (2019). Pre-hospital and trauma care to road traffic accident victims: Experiences of residents living along accident-prone highways in Ghana. In *Emergency Medicine and Trauma*. IntechOpen. https://doi.org/10.5772/intechopen.86118

Sanjay, M. (2020, August 19). *Why and how to cross validate a model?* Medium. https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

Serengil, S. (2020, April 6). Feature importance in decision trees. *Sefik Ilkin Serengil*. https://sefiks.com/2020/04/06/feature-importance-in-decision-trees/

StackExchange. (n.d.). *Combining Latitude/Longitude position into single feature—Data Science Stack Exchange*. Retrieved September 4, 2020, from https://datascience.stackexchange.com/questions/49553/combining-latitude-longitude-position-into-single-feature

Tyagi, P. (2019, January 7). *Decision Tree*. Medium. https://medium.com/@pytyagi/decision-tree-ac0c9e3b8258

Vickery, R. (2020, April 28). *The art of finding the best features for machine learning*. Medium. https://towardsdatascience.com/the-art-of-finding-the-best-features-for-machine-learning-a9074e2ca60d

Xuan, P., Sun, C., Zhang, T., Ye, Y., Shen, T., & Dong, Y. (2019). Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Frontiers in Genetics*, *10*, 1–11. https://doi.org/10.3389/fgene.2019.00459