Enhancing Road Safety through Crash Data Analysis in San Jose: A Comparative Evaluation of Classification Algorithms

1. Why did you choose this topic? What is your data?

Motivation

Road infrastructure enables people to carry out their day-to-day activities such as commuting to school, work or amusement parks. Assuring road safety is critical to accident prevention. According to San Jose Spotlight news, San Jose tops the list of Silicon Valley cities with the most bicycle crashes, injuries and deaths. The San Jose Department of Transportation's report revealed various factors that commonly contribute to crashes in the city, such as speeding, distracted driving, bad weather conditions, failure to yield, and driving while under the influence of drugs or alcohol. Vehicle crashes can have serious public safety implications, so understanding patterns and trends that may contribute to crashes in the city of San Jose is essential. This can include factors such as weather conditions, time of day, driver behavior, or road infrastructure. Understanding these causes can help inform interventions to reduce the likelihood of crashes and identify high-risk areas or situations where crashes are more likely to occur. With the advancement of data-mining techniques, these patterns can be rapidly extracted to guide policing methods, improve accident-prevention techniques, and increase awareness campaigns.

Goal

This study aims to analyze the post crash data of the city of San Jose to determine risk factors affiliated to road hazard to help establish effective safety protocols to aid accident prevention. The project proposes to implement a multi classification model which best classifies severity of injury coupled with collision type. The risk factors are deduced through the identification of significant features relative to the classification models.

About Data

The data is about crashes of the city of San Jose from 2011 till present including information at both the crash event level and individual crash level.The data is gathered from the San Jose open data gov portal which is updated on a weekly basis. The data at crash event level provides basic information about the crashes, such as the crash date and time, roadway condition, weather as well as the number of injuries(minor,moderate and severe) and fatalities. However, this dataset does not provide detailed information about the vehicles and individuals involved in

the crashes. For this we are taking another dataset that provides more detailed information about the crashes, including information about the individuals like age,sex ,sobriety and about the vehicles involved in the incidents like vehicle type,vehicle direction and violation type. The crash event data and individual crash level data are both available in two time frames at official san jose gov website 2011-2020 and 2021 to present. So all the data was concatenated and later merged using the unique identifier ‘Crash Name’ i.e ‘CR’ which stands for Crash followed by a 10 digit number .The combined dataset consists of 133677 rows and 39 columns which are 'CrashName'(Text), 'Name'(Text), 'Sex'(Text), 'Age'(Numeric), 'Speed'(Text), 'VehicleDamage'(Text), 'PartyCategory'(Text), 'Sobriety'(Text), 'VehicleDirection'(Text), ‘MovementPrecedingCollision'(Text), 'PartyType'(Text), 'OtherAssociatedFactor'(Text), 'VehicleCount'(Numeric), 'ViolationCode'(Text), 'ViolationCodeDescription'(Text), 'CrashFactId'(Numeric), 'MinorInjuries'(Numeric), 'ModerateInjuries'(Numeric), 'SevereInjuries'(Numeric), 'FatalInjuries'(Numeric), 'TcrNumber'(Text), 'CityDamageFlag'(Text), 'ShortFormFlag(Text)', 'Distance'(Numeric), 'CrashDateTime'(TimeStamp), 'PedestrianAction'(Text), 'RoadwaySurface(Text)', 'RoadwayCondition'(Text), 'Lighting'(Text), 'PrimaryCollisionFactor'(Text), 'TrafficControl'(Text), 'Weather'(Text), 'CollisionType'(Text), 'ProximityToIntersection'(Text), 'VehicleInvolvedWith'(Text), 'PedestrianDirectionFrom'(Text), 'PedestrianDirectionTo'(Text), ‘DirectionFromIntersection'(Text) and 'Comment'(Text).

2. Why does this topic need classification methods?

In reference to the systematic literature review by Silva, Andrade, and Ferreira (2020b) three key approaches were identified for classification protocols to enhance road safety. The first model commonly used is the adoption of regression methods using ‘crash frequency’, the second is a classification modeling strategy on ‘crash severity', while the third methodology proposed is a combination of regression and classification models on the two target features. The paper provides evidence that supports the shift from traditional statistical methods to ML models which proved an improvement in safety protocols over the years. The authors analyzed papers from various years for the purpose of their research to identify this improvement. Thus, classification and regression approaches are quite effective for the purpose of this project based on literature review.

The classification methods adopted for this project will lay the groundwork for models which will estimate the chance of traffic accidents based on a variety of variables, including the

type of road, the weather, the behavior of the driver, and the type of vehicle. Based on the accidents' severity, kind, and other features, distinct groups are created using classification algorithms. Through the use of classification approaches and further predictive models that may be used to pinpoint the causes of accidents and create solutions to lessen their impact can be created. They are a crucial tool for deciphering and understanding large and complex data sets, and they can reveal patterns and trends in accident data that human researchers might not immediately notice.We plan to make a multi-classification model which classifies severity of injury along with the collision type. Classifying crashes on collision type and type of injury will help us further to focus on fatal crashes.We can further build models to predict the probability of having a fatal crash based on the data which we classified with our classification model.

3. Literature review

Dias et al. (2023) in their study introduces a solution for anticipating the likelihood of road collisions. The system that was developed involves three stages: gathering and selecting data, preprocessing, and implementing mining algorithms. The data was extracted from the Portuguese National Guard database, and it pertained to crashes that took place from 2019 to 2021. The findings revealed that the greatest frequency of accidents arises between 5:00 pm and 8:00 pm, and that rainfall is the meteorological factor that has the most significant influence on the probability of accidents. Several mining algorithms were tested, including kNN, simple linear regression, Lasso and Ridge, decision tree regression, and a traditional neural network. The models were applied to the initial dataset and to datasets divided by location, such as motorways, national roads or itineraries, and villages. The neural network produced the best overall result with 89% , 87% and 88% accuracy respectively. Furthermore, it was determined that Friday is the day with the highest number of accidents compared to the other days of the week. This information can assist decision-makers in determining the most effective distribution of resources for traffic monitoring.

Hussain et al. (2019) in their paper mentions that accurately analyzing roadway traffic data is essential in identifying variables that have a significant correlation with fatal accidents, in order to provide safe driving recommendations. This paper applies statistical analysis and data mining algorithms to the FARS Fatal Accident dataset as an attempt to tackle this issue. The study investigates the relationship between the fatal rate and other attributes, including collision

manner, weather, surface condition, light condition, and the presence of drunk drivers. The Apriori algorithm is utilized to discover association rules, the Naive Bayes classifier is employed to construct a classification model with 67.95% accuracy, and simple K-means clustering algorithm is used to form clusters with Euclidean distance as the dissimilarity measure, considering two variables: population (in 100,000) and the number of fatal accidents. From the clustering result it is seen that some states/regions have higher fatal rates, while some others lower. Based on the statistics, association rules, classification model, and clusters obtained, various safe driving recommendations can be made.

Mansoor et al. (2020) in their paper outline a machine learning model for estimating the severity of traffic accidents that can improve how quickly and effectively emergency services can respond to incidents. The need for prompt and efficient emergency responses to traffic accidents and the difficulties in anticipating accident severity are covered in the first paragraphs of the study. The severity of accidents can then be predicted by the authors using a two-layer ensemble machine learning model that combines decision tree, gradient boosting, and neural network methods. The authors explore the possible uses using their approach for proactive incident management in their conclusion, as well as the significance of integrating machine learning methods into systems for responding to emergencies. (p. 3-6, 7-9)

Ramya et al. (2021) seek to evaluate traffic incidents and determine the underlying causes using machine learning techniques. To examine the causes causing traffic accidents, the authors suggest a machine learning model that makes use of decision trees, random forests, and logistic regression. The model predicts the risk of an accident occurring by using data from the Indian state of Karnataka, which includes elements like road type, vehicle type, meteorological conditions, and driver behavior. The model's findings, which indicate that driver conduct, the nature of the road, and the weather are the main causes of traffic accidents in Karnataka, are also discussed in the study. (p. 2-4)

Iranmanesh et al. (2022) outline a technique for locating high-risk stretches of rural roads using crash data from the past. The authors suggest an ensemble decision tree-based model that combines decision tree and random forest algorithms to forecast the likelihood of crashes occurring on various rural road segments. Their ensemble decision tree-based model outperformed the other models in terms of accuracy when they compared their model to other

machine learning models. In order to increase safety and efficiency, they also emphasize the significance of employing machine learning techniques in transportation engineering. (p. 6-8)

In their study, Çelik and Sevli (2022) collected data on traffic accidents in Burdur, Turkey, and trained several machine learning models to predict accident severity. The evaluation metrics used was accuracy which classifies the severity of accidents as property damage, injury, or fatality. The authors found that the Random Forest model had the highest accuracy of 84.66%. Additionally, the study identified important factors that contribute to the severity of traffic accidents, including the type of vehicle involved, the driver's age and gender, the time of day, and the weather conditions. They conclude that the findings can be used to develop targeted interventions and policies to prevent accidents and reduce their severity. The study suggests that machine learning techniques can be a useful tool for improving road safety and reducing the social and economic costs of traffic accidents.

Islam, Reza, Gazder, Akter, Arifuzzaman, and Rahman (2021) conducted a study on predicting road crash severity using classifier models and crash hotspots. The study aimed to develop a predictive model which identifies high-risk crash areas and determines the severity of crashes. The researchers collected crash data from Dhaka, Bangladesh, and used machine learning algorithms to develop a predictive model. The results showed that the Random Forest classifier model performed the best in predicting crash severity with an accuracy of 76.14%. The researchers also identified high-risk crash areas using hotspot analysis and found that most crashes occurred in urban areas.They conclude that the findings can be used to improve road safety by identifying high-risk areas and implementing measures to prevent crashes in those areas. The predictive model developed in this study can also be used to assess the severity of crashes and provide timely medical assistance to crash victims. Overall, they highlight the importance of using data-driven approaches to improve road safety and prevent crashes.

Kwon, Rhee, and Yoon (2015) in their paper explore the dependencies that contribute towards risk factors associated with road safety. Classification of severity levels was implemented through Decision Tree, Naive Bayes and Random Forest models with binary logistic regression model as the baseline. Random Forest model outperformed the other swith an accuracy of 93.9%. The paper succeeded in identifying significant risk factors that require interceding through feature selection. The features identified were ‘movement preceding collision’, ‘collision type’,‘state highway’, ‘population’, and ‘at fault’. The research effectively

helped support the adoption of classification methods for the purpose of characterizing risk factors thereby optimizing road safety protocols.

References

Dias, D., Silva, J. S., & Bernardino, A. (2023). The Prediction of Road-Accident Risk through Data Mining: A Case Study from Setubal, Portugal. *Informatics (Basel)*, *10*(1), 17. https://doi.org/10.3390/informatics10010017

Hussain, S., Muhammad, L. J., Ishaq, F. S., Yakubu, A., & Mohammed, I. (2019). Performance Evaluation of Various Data Mining Algorithms on Road Traffic Accident Dataset. *Smart Innovation, Systems and Technologies*, 67–78. https://doi.org/10.1007/978-981-13-1742-2_7

Mansoor, U., Ratrout, N. T., Rahman, S. M., & Assi, K. J. (2020). Crash Severity Prediction Using Two-Layer Ensemble Machine Learning Model for Proactive Emergency Management. *IEEE Access*, *8*, 210750–210762. https://doi.org/10.1109/access.2020.3040165

Ramya, S., Abhijna, Alfiya, S., Kokila, B. V., & Thejaswini, M. R. (2021). Road Accidents Analysis Using Machine Learning. International Journal of Advanced Science and Technology, 30(3), 1570-1578. https://ijrcs.org/wp-content/uploads/IJRCS202107006.pdf

Iranmanesh, M., Seyedabrishami, S., & Moridpour, S. (2022). Identifying high crash risk segments in rural roads using ensemble decision tree-based models. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-24476-z

Çelik, A. & Sevli, O. (2022). Predicting Traffic Accident Severity Using Machine Learning Techniques . Türk Doğa ve Fen Dergisi , 11 (3) , 79-83. DOI: 10.46810/tdfd.1136432

Islam, M. K., Reza, I., Gazder, U., Akter, R., Arifuzzaman, M., & Rahman, M. M. (2021). Predicting Road Crash Severity Using Classifier Models and Crash Hotspots. Safety, 7(3), 39. https://www.mdpi.com/2076-3417/12/22/11354/pdf

Kwon, O., Rhee, W., & Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. *Accident Analysis & Prevention*, *75*, 1–15. https://doi.org/10.1016/j.aap.2014.11.005

Silva, P. S. T., Andrade, M., & Ferreira, S. (2020b). Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering*, *7*(6), 775–790. https://doi.org/10.1016/j.jtte.2020.07.004

As we have seen, age is picked up by most of the feature selection methods where age between 20-40 is seen as the major cause.

because intersections are often high-risk areas for accidents to occur.Accidents that occur closer to intersections tend to have a higher chance of being caused by factors such as failure to yield, running red lights, or making improper turns, which are common causes of intersection-related accidents

head-on collisions are typically more severe than rear-end collisions,and accidents involving pedestrians or cyclists are generally more serious than those involving only vehicles

Lightning and weather are important features .For example, heavy rain or snow can make roads slippery and reduce visibility, increasing the likelihood of accidents. High winds can also make driving more difficult, particularly for larger vehicles like trucks and buses

Sobriety is also considered as an imp feature here, i.e. if the driver had been drinking which could result in an accident.