CMPE257

Project Final Report

Group 10

Jinwoo Bae - 007330895
Jeffrey Gu - 013616642
Shyam Kannan - 014132079
Honghao Ma - 010278125
Haroon Razzack - 014311726
Brian Zhang - 018231876

November 22, 2024

1. Idea

1.1 Description

Traffic safety is a significant concern that directly affects the lives of all residents of San Jose and the efficiency of the transportation system. These issues are prominent throughout the globe, with almost 1.2 million[1] people in the world losing their lives in traffic accidents in 2023 alone. For car-centric and urban areas such as San Jose where dangerous road conditions and high traffic congestion is common, the risks of frequent commutes become especially prominent, bringing a concerning level of risk to an activity done by millions every day. Accident attorney Andy Gillin noted that San Jose not only has a high rate of traffic accidents, but car accidents with fatalities have been steadily increasing over the years (Gillin, 2024). Urban areas often face an uneven distribution of traffic accidents, with certain locations being at a significantly higher risk of traffic accidents due to a combination of several dynamic temporal and spatial factors. As mentioned by Gillin, environmental factors such as rain and fog can be difficult to drive in due to low visibility and decreased traction, while human factors such as speeding and distracted driving can cause a driver to be less aware and may be incapable of reacting to a danger in time (Gillin, 2024).

Cities can take proactive steps to reduce accident rates by identifying high-risk areas and analyzing the contributing factors. We aim to predict and prevent accidents by understanding the geographic distribution of traffic risks and the conditions under which these accidents occur. By giving drivers and city planners insights on the factors correlated with traffic risk and forecasting potential hotspot areas, drivers can take the steps necessary to avoid them while city planners can look to make changes to minimize these risks in the future.

We have already seen big efforts in the modern day to reduce fatality numbers, with GPS-based systems that give real-time weather and traffic conditions, as well as initiatives[2] to understand traffic patterns and implement safety measures to circumvent potential hazards. However, these applications focus heavily on historical and static hotspot location data without considering dynamic conditions. For example, a location with a low concentration of accidents overall may experience spikes in accidents under specific conditions. Understanding these rare circumstances can help both drivers and city planners make more informed decisions, reducing the likelihood of accidents and enhancing overall road safety.

[1] https://www.who.int/publications/i/item/9789240086517
[2] https://www.sanjoseca.gov/your-government/departments-offices/transportation/safety/vision-zero

1.2 Goals/Objectives

With road accidents being a prevalent and lethal hazard in San Jose, our project's primary goal is to minimize unnecessary traffic risks and reduce the number of serious accidents in the city. To achieve this, we aim to identify traffic accident hotspots through an analysis of spatial and temporal data and patterns. Additionally, we strive to predict whether a given location qualifies as a hotspot under varying conditions, including temporal and human factors. Our goal is to develop a thorough understanding of the most impactful environmental and temporal conditions associated with traffic risks. By leveraging spatio-temporal analysis, we aim to create a predictive classification model capable of categorizing potential hotspots in San Jose as either high-risk or low-risk, utilizing the features identified as most influential.

To achieve these objectives, we first applied an unsupervised learning model to cluster accident data and uncover spatial patterns across different areas of San Jose. This clustering approach enabled us to identify patterns in traffic accident distribution without relying on predefined labels. Furthermore, we explored twelve supervised learning models to predict if a specific location was in a traffic hotspot. These models included Logistic Regression, Random Forest, Ensemble Voting Classifier, and others. This approach allows us not only to pinpoint accident hotspots through clustering but also to predict traffic hotspot locations in real time under specific conditions.

2. Work developed

For this project, we use the official traffic accident dataset[3] provided by the City of San Jose to identify road crash patterns and create a predictive classification model that uses a combination of both unsupervised and supervised learning. This dataset is currently used as the primary source of data for the Vision Zero Initiative[4] formed by the City of San Jose. As such, we can be confident in the dataset's authenticity and reliability. Figure 1 shows a basic visualization and analysis of the temporal trends of traffic accidents for a typical week. We can observe two peaks on weekdays occurring at 8:00 in the morning and 5:00 in the afternoon while being safer in the evenings. On the other hand, weekends see high accident frequencies all the way until 2:00 AM.

[3] https://www.data.sanjoseca.gov/dataset/crashes-data
[4] https://www.sanjoseca.gov/your-government/departments-offices/transportation/safety/vision-zero

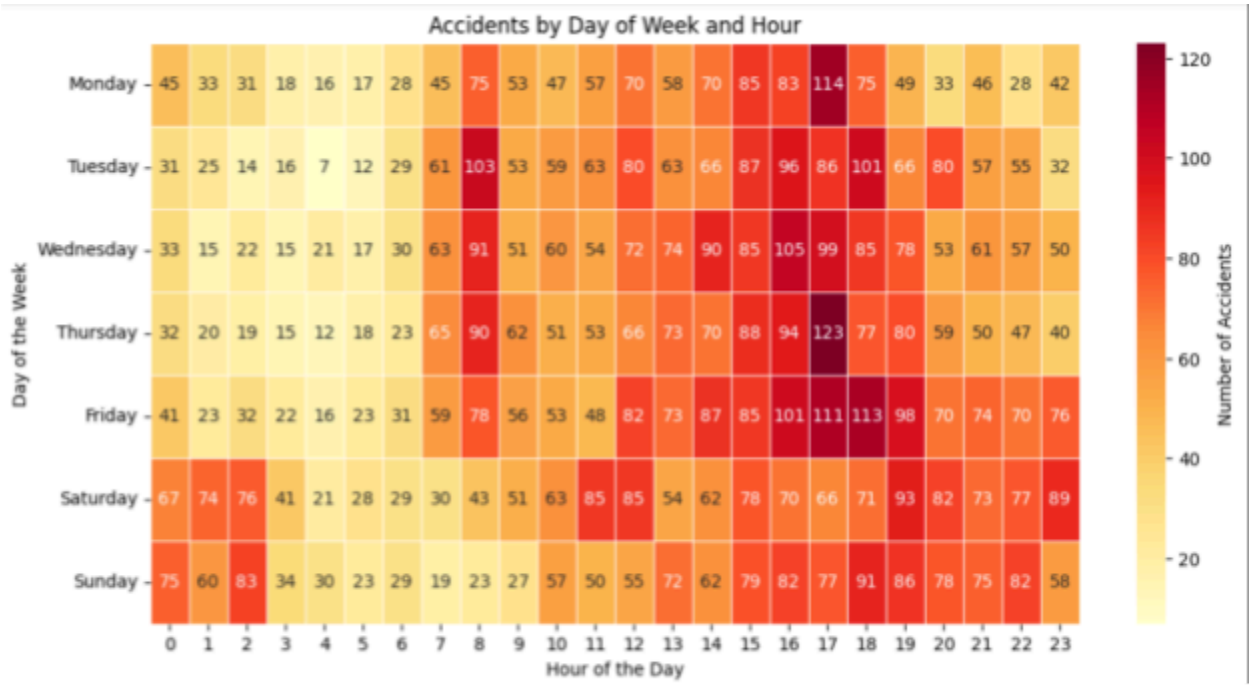

Fig 1. Temporal Trends

These temporal trends provide only a glimpse into the dynamic behavior of different areas in San Jose. A closer examination of the geographical distribution of accidents reveals that the high accident risks on weekends stem from increased traffic congestion in downtown San Jose. Conversely, weekday accidents are more prevalent near freeways and farther from the city center, as commuters travel to work during rush hours. These patterns are further illustrated in Figures 2 and 3.

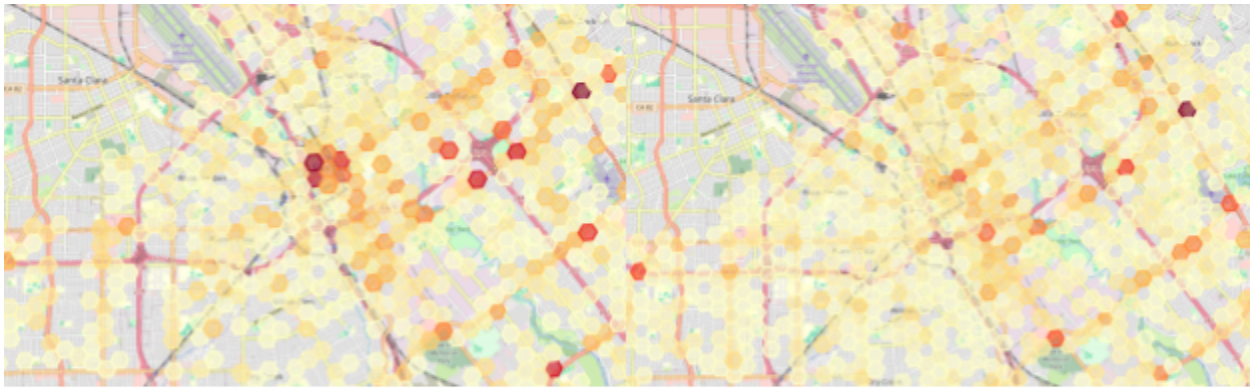

Fig 2. Weekend Accident Frequency

Fig 3. Weekday Accident Frequency

When we further analyze the precise areas across San Jose that are the most accident-prone, we find interesting environmental patterns and trends that persist across San Jose. For example, dense parking areas are shown to be a common

location for high accident frequency in the city. Another common crash location is freeway exits. Examples of these high-risk areas can be observed in Figure 4. In general, we find areas with notable landmarks or points of interest such as gas stations, parking garages, shopping malls, etc. often correlated with higher traffic risk. The work done by Moosavi et al. (2019) recognized the impact of different points of interest and used it to train an accurate predictive model. In the sections below, we also delve deeper into leveraging these spatio-temporal patterns within our dataset, and the numerous factors that correlate with accident risk to create a predictive model for hotspot detection.

Fig 4. High Accident Prone Areas

2.1 Adopted ML algorithms/techniques

For our predictive classification model approach, we incorporate both unsupervised learning and supervised learning techniques. Unsupervised learning enables the machine to learn from the data without supervision to capture spatial patterns around San Jose. As done similarly by Santos et al. (2021), we use DBSCAN, a density-based cluster algorithm, for initial hotspot clustering in order to group our data based on the geographic coordinates of each accident. DBSCAN consists of two parameters, epsilon and minimum points, which correspond to the distance between points required to be in the same cluster and the number of points required to form a cluster, respectively. After several tests and experiments, we found that an epsilon of 150 meters and minimum points value of 20 gave the best results, grouping accidents into dense hotspot clusters while removing sparse accident points that can be classified as noise. A map of San Jose and its clustered hotspots can be seen in Figure 5 below.

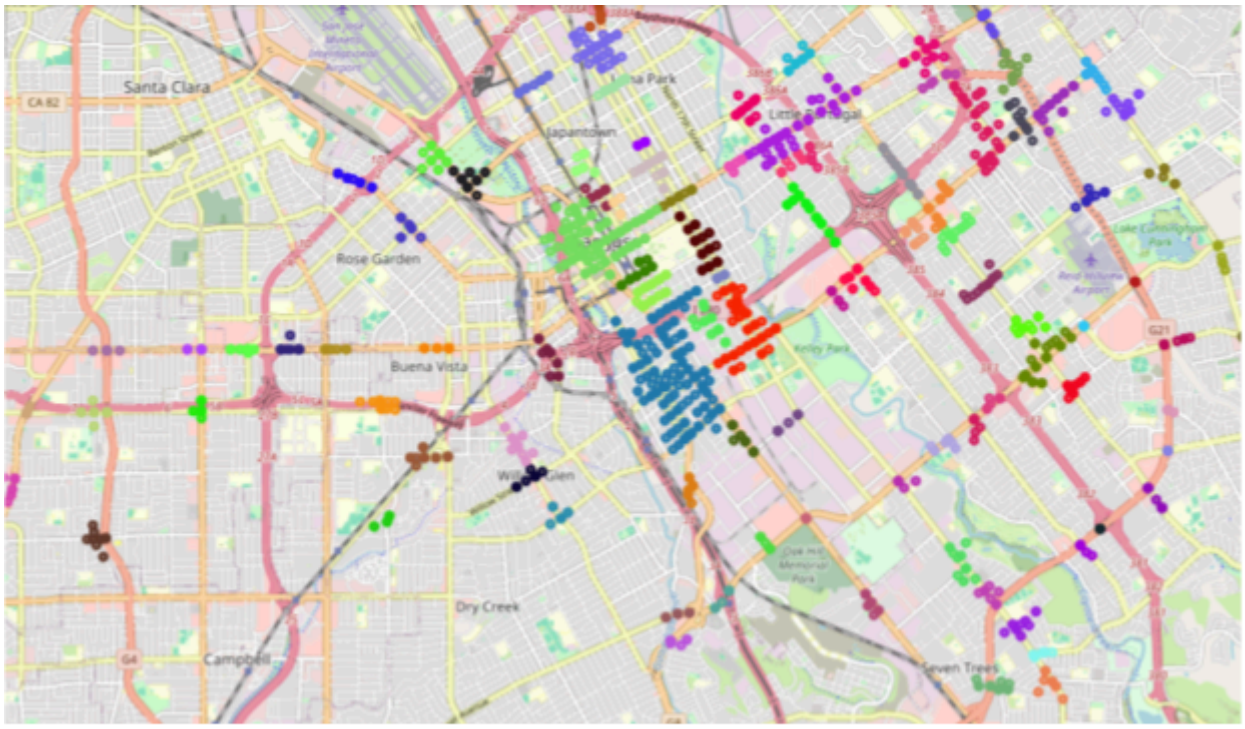

Fig 5. Clustering with DBSCAN

In the case of supervised learning, we wanted to be able to develop a prediction model that can predict if a given location would be in a traffic accident hotspot and determine which supervised learning model would be performant enough to do so. For this project, we explored the following supervised learning models:

2.1.1. Logistic Regression and Naive Bayes

Logistic Regression and Naive Bayes were considered simple models compared to the rest of the supervised models we adopted, and so we wanted to see how well simplistic models like Logistic Regression and Naive Bayes perform for a baseline metric.

2.1.2. K-Nearest Neighbors

As we were dealing with spatial patterns, K-Nearest Neighbors being able to classify data points by their nearest K (5) neighbors may prove to be fruitful for accurate predictions.

2.1.3. Decision Tree and Random Forest

A Decision Tree would be able to show a visualization of the factors that contribute to a traffic hotspot. As an extension, we also explored Random Forest for its ability to aggregate multiple individual decision trees which helps reduce overfitting, as well as its ability to be used for complex relationships since our data would likely not be linear.

2.1.4. Support Vector Machine

As mentioned in 2.1.3, because our data is likely not to be linear, using SVM as another model to capture complex relationships is something that we wanted to explore.

2.1.5. Gradient Boosting, XGBoost, LightGBM, and CatBoost

Since Gradient Boosting[5] is a combination of several weaker learning models, such as Decision Trees, Gradient Boosting would help improve the accuracy of predicting traffic hotspots. We also wanted to explore other Gradient Boosting algorithms, such as XGBoost, LightGBM, and CatBoost, and see how different they are and determine which would be the best Gradient Boosting algorithm to use for our project.

2.1.6. Neural Network

Out of the entire models that we wanted to explore, we believed that Neural Networks were the most complex and powerful model than the rest of the other models, and we were interested in finding out how well this model would perform.

[5] https://www.geeksforgeeks.org/ml-gradient-boosting/

2.1.7. Ensemble Voting Classifier

This supervised learning model combines the prediction of several models[6], and we can leverage this model by using three well performing models to improve the accuracy of predicting traffic accident hotspots.

The classification for a traffic accident hotspot was the latitude and longitude, rounded to the nearest thousandth, exceeding the threshold of accidents which was defined to be more than or equal to five, otherwise it would not be classified as a hotspot. We had many supervised learning models we wanted to experiment with, so in order to judge which model would be more performant than the other, a uniform set of features was used for all of the models to train off of. To do this, we selected critical features from the dataset that were apparent from the onset like latitude and longitude, feature engineered variables based off of the time of the crash to get the month and time of day, and determine which features were important using feature importance with Random Forest to determine the most important features. This was especially useful as we found that a feature that we ignored initially, IntersectionNumber, which identifies the intersection based on an ID, was highly important and would increase the accuracy of the models, 7-8% for Random Forest specifically for instance. We also used K-means clustering, using the elbow method to determine k=9, in order to assign each row a cluster, giving us IntersectionNumber, Latitude, Longitude, Month, Cluster, Lighting, TimeOfDay, and SpeedingFlag as the set of features.

2.2 Performances metrics

Table 1: Supervised Learning Models and metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Notes
Random Forest	**86.23**	**83.97**	**93.17**	**88.32**	**consistent validation**
XGBoost	**91.15**	**88.68**	**96.05**	**92.43**	**Best performing?**
Ensemble Voting Classifier	**88.71**	**85.61**	**95.95**	**90.49**	
KNN	**86.31**	**82.26**	**96.31**	**88.73**	**Interesting.**
Gradient Boosting	**84.9**	**82.38**	**92.85**	**87.3**	**Overfitting**

[6] https://www.geeksforgeeks.org/voting-classifier/

CatBoost	**84.94**	**82.26**	**93.17**	**87.37**	**Decent**
Pruned Decision Tree	65.84	65.18	83.56	73.23	uninterpretable
LightGBM	83.78	84.22	92.22	86.42	overfitting?
SVM	57.35	56.92	97.68	71.93	
Logistic Regression	58.55	59.34	82.29	68.95	
Naive Bayes	57.49	57.29	94.28	71.27	
Neural Network (MLP)	44.2	66.67	0.48	0.95	

2.3 Evaluation

2.3.1 Very Well-Performing Models

The Random Forest model demonstrates strong overall performance in predicting high-risk and low-risk traffic locations. After experimentation, we determined that setting the number of estimators (n_estimators) to 100 was the most optimal configuration for this dataset. The model achieved a respectable accuracy of 86.23%, indicating that it correctly classified most of the data points. Additionally, its high recall value signifies that it is effectively identifying most high-risk locations, a critical metric for this task. However, the precision is slightly lower, which suggests that some low-risk locations are being misclassified as high-risk. The F1-score of 88.3% highlights a solid balance between precision and recall, further underscoring the model's ability to maintain consistency in identifying true high-risk areas while minimizing false positives. Moreover, the ROC AUC score of 0.944 demonstrates excellent discriminatory power, confirming that the model can reliably distinguish between the two classes. Cross-validation results further validate the model's robustness, showing low variance across folds and consistent performance.

Feature importance analysis revealed that geographic features like intersection numbers and latitude/longitude were the most significant predictors, aligning with real-world expectations of accident clustering. Based on the analysis of feature importance, we could experiment with removing features of lower importance to assess whether this improves the model's accuracy and precision. This step could streamline the model and potentially enhance its predictive capability.

XGBoost was one of the top-performing models, achieving an accuracy of 91.15% with a balanced precision (88.68%) and recall (96.50%), leading to an excellent F1-Score of 92.43%. The model consistently performed well across stratified cross-validation, with an average accuracy of 91.45%, displaying its robustness and generalizability. Its ability to handle class imbalances and model complex relationships in the data makes it an ideal candidate for predicting accident-prone areas with precision and reliability.
However, when experimenting with adjustment of features, towards the end of the project, such as the 'DayOfWeek' and 'Hours' we found that XGBoost metrics decreased by 3-5 percentage points while the Random Forest model simultaneously fluctuated its accuracy up to 90%. This instability based on features is worth exploring more, however does not take away from the fact that both models were arguably the 2 top performers.

The Voting Classifier, combining strengths of models like XGBoost and Random Forest, achieved a competitive accuracy of 88.71% with a high F1-Score of 90.49%. It excelled in balancing precision and recall, benefiting from the complementary strengths of its base models. Its cross-validation results (89.02% mean accuracy) indicated reliable performance, though the method is computationally intensive compared to standalone models. Questions remain whether this model can be improved as it uses gradient boosting which is known to be sensitive to parameter settings and dataset splits. Thus, further tuning of the model is required to achieve more clarity and accurate results.

2.3.2 Other Notable Well-Performing Models:
The KNN model achieved 86.31% average accuracy across stratified cross-validation, with high recall (96.31%) and a respectable F1-Score (88.73%). Its strength lies in its simplicity and ability to make predictions based on the local structure of the data, making it a viable option for identifying accident-prone areas. However, KNN is computationally expensive at prediction time, as it requires calculating distances for all data points during inference. It also suffers from sensitivity to feature scaling and noise in the dataset, which can affect performance, especially for high-dimensional data. Despite its strong metrics, these practical limitations make it less favorable compared to models like Random Forest or XGBoost.

CatBoost demonstrated competitive performance, achieving 84.94% accuracy, with strong recall (93.17%) and an F1-Score (87.37%). It excelled in handling categorical data and complex feature interactions, which may explain its balanced performance. Stratified cross-validation results indicated consistency with an

average accuracy of 85.69%, suggesting reliable generalization to unseen data. However, compared to XGBoost and Random Forest, CatBoost's marginally lower precision and feature importance rankings suggest it might not capture geographic patterns as effectively. Nevertheless, its ability to handle missing data and categorical variables efficiently makes it a strong candidate for similar datasets.

Gradient Boosting models achieved 84.90% accuracy, with decent performance across precision, recall, and F1-Score metrics. Cross-validation revealed variability, with an average accuracy of 73.02%, suggesting potential sensitivity to parameter settings or dataset splits. Gradient Boosting leverages sequential learning to refine predictions, which makes it effective in capturing subtle relationships in the data. However, it was outperformed by XGBoost, which offers better optimization and regularization.

LightGBM achieved 83.78% accuracy, with strong recall (92.22%) and an F1-Score (86.42%). This model is known for its speed and scalability, especially on large datasets, making it an attractive option when computational resources are constrained. Cross-validation results showed high consistency, with a mean accuracy of 85.16% and minimal variance. While it slightly lagged behind XGBoost and Random Forest in terms of raw performance, its efficiency and ability to handle large-scale data make it a practical alternative.

2.3.3 Models with Suboptimal Performance:

The Decision Tree performed well in comparison to other models with an accuracy score of 93.78% and F1 score of 94.59%. However, given that the tree was heavily complex with a long depth, the tree was much harder to interpret and thus determining whether the tree was generalizable was difficult even after pruning and reducing the depth of the tree to 5. When reducing the depth, we found that the accuracy and metrics were significantly compromised with an accuracy score of 65.84%.

While most tree-based models performed well, others like Logistic Regression and Naive Bayes struggled due to their inability to capture complex patterns in the dataset. Logistic Regression achieved a modest 58.55% accuracy, primarily limited by its linear assumptions, making it unsuitable for this multi-dimensional and imbalanced dataset. Similarly, Naive Bayes (accuracy: 57.49%) relied on oversimplified probabilistic assumptions that failed to account for nuanced relationships between features.

Models like Support Vector Machines (SVM) and Neural Networks (MLP) also

underperformed. The SVM achieved a near-perfect recall (97.68%) but suffered from poor precision and accuracy (57.35%) due to over-classifying high-risk zones. The MLP model fared the worst, with 44.20% accuracy and negligible recall (0.47%), due to insufficient tuning of hyperparameters or suboptimal architecture for the dataset. These results highlight the importance of balancing complexity and interpretability in model selection.

2.4 Recommendations/Future Work

Future work for this project will concentrate on improving preprocessing stages and the risk clustering model to better comprehend clusters and assess their practical consequences. Additionally, we intend to adjust clustering parameters, investigate clustering methods other than DBSCAN or K-means, and develop a robust hotspot analysis to expose patterns within the data.

Given models like Random Forest, XGBoost and Ensemble methods performed well, we hope to expand upon those approaches with further experimentation of feature engineering and selection. Since XGBoost performed the best, a deployment of this model and an iteration of Random Forest would be well suited.

Additionally, the classification for the supervised learning model should also be adjusted, as if there were a substantial increase in data, the classification would likely consider every accident to be in a traffic hotspot simply due to the sheer number of accidents, signaling a class imbalance. This would be especially true with data located in a city with a smaller land area such as San Francisco with 10-20 years of data.

Finding more datasets that consist of different traffic features and conditions would also be helpful in the future to incorporate more variables that can improve our predictive model. Aspects such as detailed historical weather conditions can be found from weather stations nearby, while websites such as OpenStreetMap can provide information regarding road types and nearby points of interest. By merging these features into our existing dataset, we can capture more traffic risk patterns with a supervised learning model.

In terms of visualization, our future goal is to create geospatial tools, such as interactive dashboards and heatmaps, that can display danger levels and time variations. Since we rely on weekly updates, it is imperative that we integrate real-time data, such as traffic and weather conditions, to improve prediction accuracy. Finally, we intend to develop our algorithms into a scalable, fully

functional risk prediction tool that the public or transportation organizations may use in real time.

3. Project Management

3.1 Final Schedule & Task Distribution (Gantt chart)

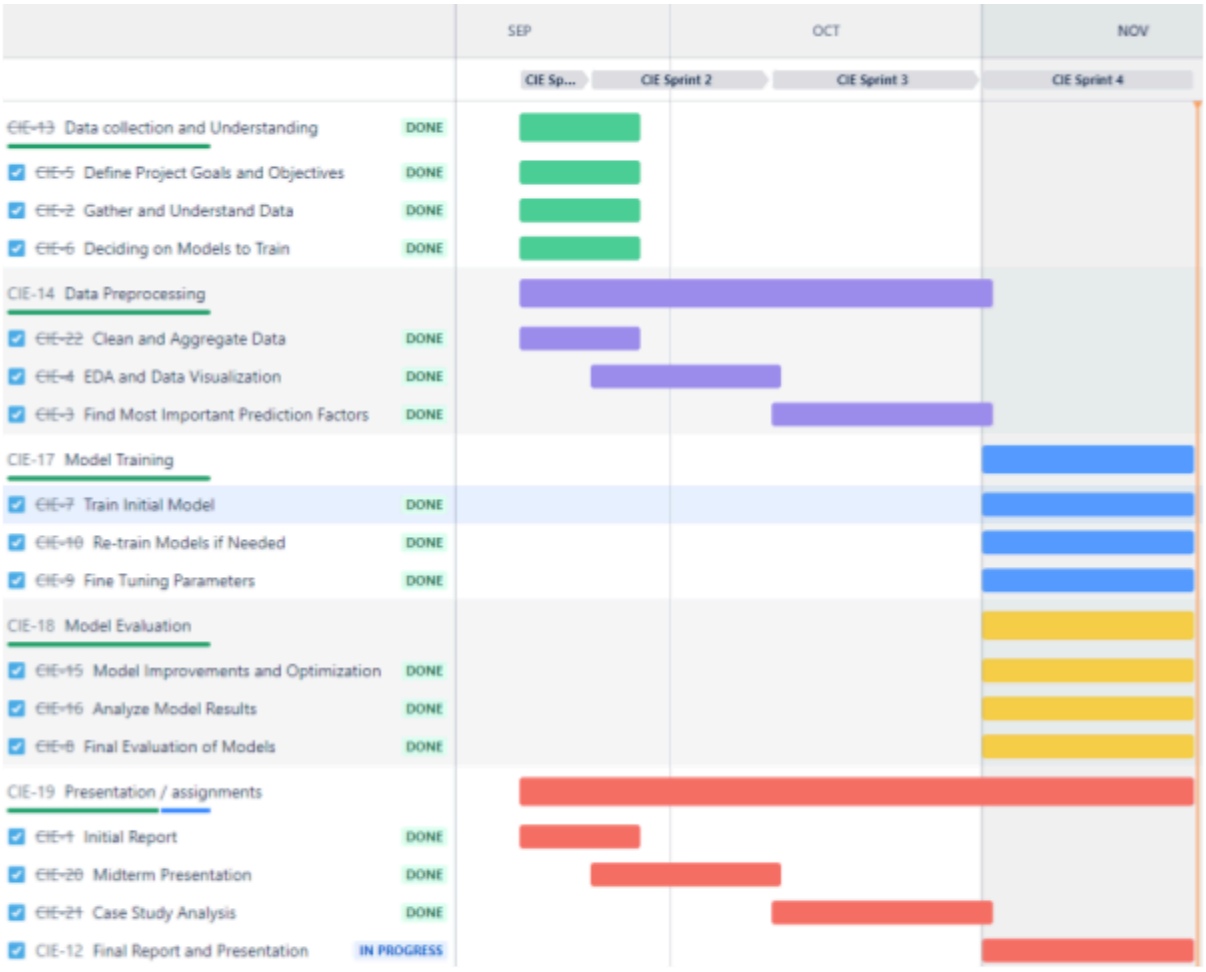

Fig 6. Gantt Chart

3.2 Challenges

Throughout the project, we encountered several challenges that impacted our analysis and model performance.

One of the primary challenges was the broad scope of the problem. Traffic accidents can be influenced by numerous factors, including weather, road conditions, and human behavior. However, the dataset provided limited features, which restricted our ability to explore and incorporate all relevant variables. Some fields contained unclear information or missing values, making certain analyses challenging. Critical fields required for analysis were absent in some datasets, necessitating additional preprocessing steps. We had to perform extensive feature engineering to achieve the desired results.

Another challenge we faced was data imbalance, particularly in weather-related variables. Initially, we hypothesized that weather would play a critical role in predicting traffic accidents. However, since California experiences predominantly sunny weather, the dataset had an overwhelming majority of clear weather entries, with relatively few instances of rainy or foggy conditions. Additionally, our dataset did not include data from the 2024 holiday season, which could have provided insights into seasonal trends. This imbalance made it difficult for the model to learn meaningful patterns related to weather conditions.

Our most significant challenge was overfitting. The dataset contained certain patterns that were highly specific to the training data, such as accident hotspots that may not consistently represent broader trends. This caused the models, particularly the supervised learning algorithms, to perform well on the training data but struggle with unseen cases. The issue of weather data imbalances exacerbated this problem, as the models learned to heavily favor clear weather conditions. To address this, we implemented techniques such as cross-validation, hyperparameter tuning, and careful feature selection to improve the model's generalization ability.

Lastly, we were not able to fully achieve the goals we had envisioned. Our initial objective was to create a comprehensive end-to-end system for real-time traffic accident risk assessment. However, due to time constraints and the complexity of integrating data sources, we focused on developing and evaluating prediction models. While these models provided valuable insights into accident-prone locations and contributing factors, they were not integrated into a complete system capable of functioning in real time.

3.3 Learning Outcomes

Through this project, we focused on analyzing temporal patterns in traffic accidents. This method examined how time-based variables, including the day and month, might influence accident risks and hotspots, providing valuable information for real-time risk prediction. This analysis revealed critical insights into when

accidents were more likely to occur, enabling more precise and targeted risk assessments. For example, certain time periods consistently exhibited higher accident density, which can aid in planning traffic management strategies and raising awareness during peak risk times.

To identify accident hotspots, we applied unsupervised learning methods like DBSCAN. This approach allowed us to group locations with similar traffic accident patterns, effectively highlighting high-risk zones without relying on predefined labels. DBSCAN clustered data points based on density and automatically filtered out noise points, which often represent outliers or less relevant data. The resulting clusters provided well-defined groupings of accident-prone areas while excluding noise points for more accurate insights. As shown in Figure 5, we highlighted the high accident density areas in San Jose, offering actionable information to address areas with frequent accidents.

We also explored the effectiveness of various supervised learning models, including Random Forest, Logistic Regression, and Support Vector Machines. These models were instrumental in predicting accident-prone areas based on features such as location, weather, and time. By comparing their performance, we identified Random Forest, XGBoost, and the Ensemble Voting Classifier as models that performed exceptionally well with our chosen feature set.

Additionally, we evaluated and compared the models using performance indicators such as accuracy, precision, recall, and F1-Score. These metrics provided a clear understanding of each model's strengths and limitations, enabling us to select the most effective strategies for predicting accident risks.

References

Gillin, A. (2024, August 1). *San Jose Car Accident Statistics*. GJEL Accident Attorneys. https://www.gjel.com/blog/san-jose-car-accident-statistics

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

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. *Computers*, *10*(12), 157. https://doi.org/10.3390/computers10120157