AccidentProphet:

Predicting Road Incidents

Progress report

V 1.0

November 2023

Machine learning course, AIT

Group 11

Mr. Nitesh Ghimire

Mr. Gholamreza Izadi

Mr. Tairo Kageyama

Mr. Rojan Manandhar

Mr. Panithi Sirisatjapipat

Mr. Bakyt Tursaliev

introduction

According to the Centers for Disease Control and Prevention (CDC), each year, 1.35 million people are killed on roadways around the world. It is estimated that fatal and nonfatal crash injuries will cost the world economy approximately 1.8 trillion dollars (in 2010 USD) from 2015–2030.5 That's equivalent to a yearly tax of 0.12% on global GDP.

Traffic accidents on our roadways have significant implications for public safety, traffic management, and the overall well-being of communities. The ability to respond quickly and effectively to accidents, as well as to predict their severity, is paramount. This proposal introduces a machine learning application to improve traffic accident analysis. Our product is driven by the belief that data-driven decision-making, coupled with the power of machine learning, can significantly enhance the way we understand and manage traffic accidents. "AccidentProphet" harnesses the potential of machine learning and data analytics to unlock valuable patterns and trends within accident records.

Related work

Over the past few decades, accident analysis and prediction has been a heavily researched topic. There are three categories of research in this field: analysis of environmental stimuli on accidents, accident frequency prediction, and accident risk prediction. The first category investigates the impact of environmental factors such as weather, traffic flow, and road network properties on the likelihood and severity of accidents. This category provides valuable insights but may not be immediately useful for real-time prediction and planning. The second category predicts the expected number of accidents for a specific road segment or geographical region but relies on information that may not be available in real-time. The third category is similar to the second but defined as a binary classification task that better suits real-time applications. Following are some research papers related to accident analysis.

1. Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. ArXiv. https://doi.org/10.1145/3347146.3359078.
2. Liang, L.Y., Ma'Soem, D.M., & Hua, L.T. (2005). TRAFFIC ACCIDENT APPLICATION USING GEOGRAPHIC INFORMATION SYSTEM. Journal of the Eastern Asia Society for Transportation Studies, 6, 3574-3589.

3. Dhamaniya, A., Sonu, M., Krishnanunni, M., Praveen, P., & Jaijin, A. (2016). Development of Web Based Road Accident Data Management System in GIS Environment: a Case Study. Journal of the Indian Society of Remote Sensing, 44, 789-796.
4. Santos D, Saias J, Quaresma P, Nogueira VB. Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. Computers. 2021; 10(12):157. https://doi.org/10.3390/computers10120157
5. Shakil Ahmed, Md Akbar Hossain, Sayan Kumar Ray, Md Mafijul Islam Bhuiyan, Saifur Rahman Sabuj, A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance, Transportation Research Interdisciplinary Perspectives, Volume 19, 2023, 100814, ISSN 2590-1982, https://doi.org/10.1016/j.trip.2023.

Problem

In various regions, including Thailand, road accidents have become an alarming and persistent concern, significantly jeopardizing public safety, healthcare resources, and traffic management. This challenge is exacerbated by the absence of accessible tools and predictive models that can facilitate data-driven decision-making and accident prevention. Key issues include the lack of proximity-awareness systems to position healthcare facilities near accident hotspots strategically and to improve road adequacy, the absence of personalized driver guidance to reduce accident risks, and the lack of severity classification models to apprise the public about the likelihood of accidents based on contributing factors like weather conditions and nearby points of interest.

The central problem is the dire need for comprehensive and user-friendly solutions to harness the wealth of accident data. These solutions should empower policymakers, healthcare providers, and the general population with critical insights to enhance safety, resource allocation, and preventive measures. The goal is to create a platform that leverages data analytics and machine learning to predict accident trends, identify determinants of accident severity, and provide valuable recommendations. Addressing these problems will ultimately reduce the frequency and severity of road accidents and improve our roadways' overall safety.

Solution

Rationale:

The rationale behind our proposed solution lies in the urgent need to leverage historical accident data to prevent and mitigate road accidents. We can unlock this data's hidden patterns and valuable insights by harnessing machine learning and data analytics. Predicting accidents and understanding their severity determinants can equip policymakers, healthcare providers, and the general public with the knowledge to make informed decisions and reduce accidents' impact on public safety and healthcare resources.

Value Proposition:

- **Accurate Accident Prediction:** Our product will accurately predict when and where accidents will likely occur.
- **Severity Classification:** Users can access a severity classification model that offers insights into the potential impact of accidents.
- **Proximity-Awareness:** Our platform will enable policymakers to position healthcare facilities near accident-prone areas strategically.
- **Personalized Driver Guidance:** Drivers will receive customized guidance and warnings to reduce accident risks.
- **Resource Optimization:** The product will help allocate resources efficiently by directing resources to accident hotspots.

Target Audience:

- **Governments and Policy Makers:** For data-driven decision-making and resource allocation.
- **Healthcare Providers:** For proximity awareness and resource planning.
- **Drivers and Commuters:** For personalized accident prevention guidance.
- **Data Scientists and Analysts:** For insights into accident data and trends.

Requirement Elicitation:

- **High Accuracy:** The predictive model should achieve a high level of accuracy in accident prediction.
- **Real-Time Insights:** The system must provide real-time accident risk assessment and guidance to drivers.
- **Severity Classification:** It should classify accidents based on severity accurately.
- **Resource Allocation:** The solution should help optimize healthcare and traffic management resources effectively.
- **User Engagement:** High user engagement and usability of the platform.

Feasibility:

- **Technical Feasibility:** The project's feasibility depends on the availability of reliable accident data and the technical infrastructure for machine learning and data analytics. Technical requirements are achievable within the timeframe.
- **Economic Feasibility:** A cost-benefit analysis indicates that the economic benefits of reduced accidents outweigh the project's costs. The project is economically viable.
- **Operational Feasibility:** Integration with existing operations, including traffic management and healthcare resources, is feasible.
- **Legal and Regulatory Feasibility:** The project aligns with data privacy and ethical guidelines, mitigating legal risks.
- **Scheduling Feasibility:** The project can be completed within 2 months, ensuring real-time accident prediction.
- **Market Feasibility:** There is a market for the "AccidentProphet" platform, and user adoption is expected.
- **Resource Feasibility:** The availability of resources, including data scientists and developers, is sufficient.
- **Scalability Feasibility:** The platform is designed to scale as the user base grows.

To sum up, the "AccidentProphet" project demonstrates strong feasibility across technical, economic, operational, legal, scheduling, market, resource, and scalability aspects. The project's potential to address a pressing global concern, along with the

market demand, makes it a viable and valuable solution for enhancing road safety and resource allocation.

Risks:

- **Data Quality and Availability:** The project relies heavily on accident data. If the quality or availability of this data is insufficient, it could lead to less accurate predictions and analysis.
- **User Adoption:** The application's success depends on users, including policymakers, healthcare providers, and drivers, embracing the technology. If they resist using the platform or do not trust its recommendations, the impact may be limited.
- **Model Interpretability:** The machine learning models used for accident prediction and severity classification should be interpretable and explainable. If they lack interpretability, it may be challenging to gain the trust of users and stakeholders.
- **Resource Constraints:** The project's feasibility is contingent on the availability of resources, including data scientists, developers, and computational infrastructure. Resource constraints could delay or hinder development.
- **Competing Solutions:** If competing solutions with similar objectives are launched in the same timeframe, it could affect the platform's market reception and user adoption.
- **Regulatory and Ethical Considerations:** Adherence to data privacy and ethical guidelines is crucial. Failure to address these concerns appropriately may lead to regulatory hurdles or public backlash.
- **Technical Challenges:** Developing and deploying machine learning models, ensuring real-time data integration, and creating an intuitive user interface pose technical challenges that may affect the project's timeline and success.
- **Scalability:** As the user base grows, the platform must scale to handle increased data processing and user interactions. Ensuring scalability can be a complex and ongoing process.

Data:

The countrywide car accident dataset that covers 49 states of the USA will be used in this project. The accident data were collected from February 2016 to March 2023 using

multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records. We consider 5 percent of examples as a test set data in ML algorithms. Table 1 shows all attributes in this dataset.

Table 1- US accident dataset (2016 - 2023),

#	Attribute	Description	#	Attribute	Description	#	Attribute	Description
1	ID	This is a unique identifier of the accident record.	17	Zipcode	Shows the zipcode in address field.	33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
2	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	18	Country	Shows the country in address field.	34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
3	Start_Time	Shows start time of the accident in local time zone.	19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	35	Junction	A POI annotation which indicates presence of junction in a nearby location.
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.
5	Start_Lat	Shows latitude in GPS coordinate of the start point.	21	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	37	Railway	A POI annotation which indicates presence of railway in a nearby location.
6	Start_Lng	Shows longitude in GPS coordinate of the start point.	22	Temperature(F)	Shows the temperature (in Fahrenheit).	38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
7	End_Lat	Shows latitude in GPS coordinate of the end point.	23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	39	Station	A POI annotation which indicates presence of station in a nearby location.
8	End_Lng	Shows longitude in GPS coordinate of the end point.	24	Humidity(%)	Shows the humidity (in percentage).	40	Stop	A POI annotation which indicates presence of stop in a nearby location.
9	Distance(mi)	The length of the road extent affected by the accident.	25	Pressure(in)	Shows the air pressure (in inches).	41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
10	Description	Shows natural language description of the accident.	26	Visibility(mi)	Shows visibility (in miles).	42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby loction.
11	Number	Shows the street number in address field.	27	Wind_Direction	Shows wind direction.	43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
12	Street	Shows the street name in address field.	28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
13	Side	Shows the relative side of the street (Right/Left) in address field.	29	Precipitation(in)	Shows precipitation amount in inches, if there is any.	45	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
14	City	Shows the city in address field.	30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	46	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.
15	County	Shows the county in address field.	31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	47	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.
16	State	Shows the state in address field.	32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.			

Some other data sources for Thailand that can be considered during the project implementation are:

- https://cdn.who.int/media/docs/default-source/thailand/roadsafety/overview-en-final-25-7-19.pdf?sfvrsn=f9d7a862_2
- https://www.roadsafetyfacility.org/country/thailand
- https://www.kaggle.com/datasets/thaweewatboy/thailand-road-accident-2019-2022
- https://www.kaggle.com/datasets/thaweewatboy/thailand-fatal-road-accident

Mockups:

visual representations of key screens from the mobile/web application Includes features like the accident prediction dashboard, severity classification, user guidance, and general information about the AccidentProphet. Figures (1 to 3) show the mockups for the application:

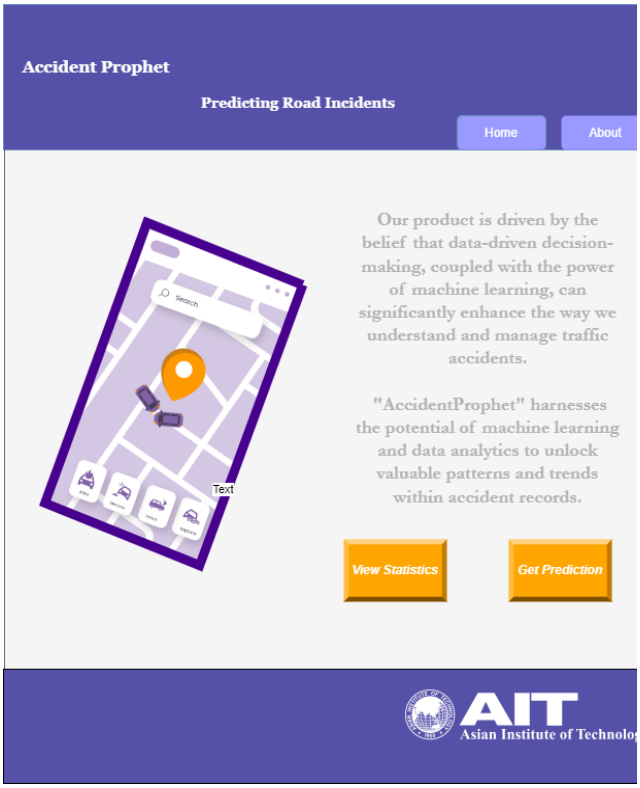

Figure 1- first page of the application including general information.

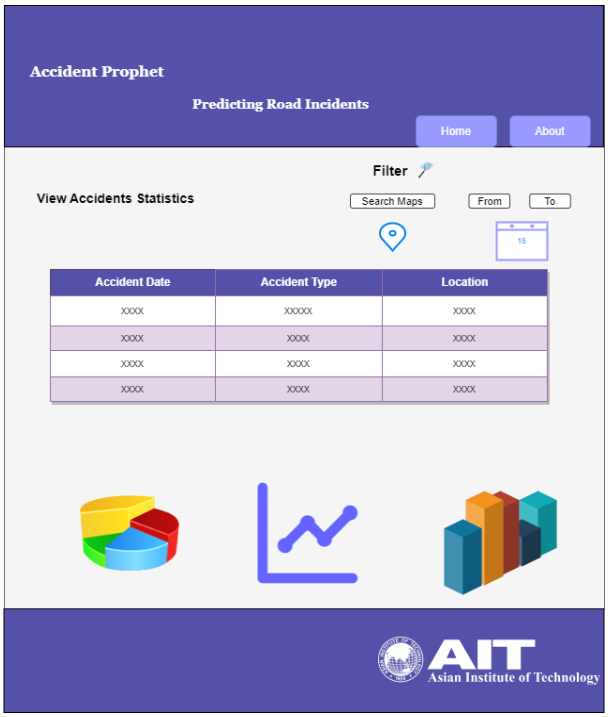

Figure 2- second page of the application including some interesting visuals (dashboard)

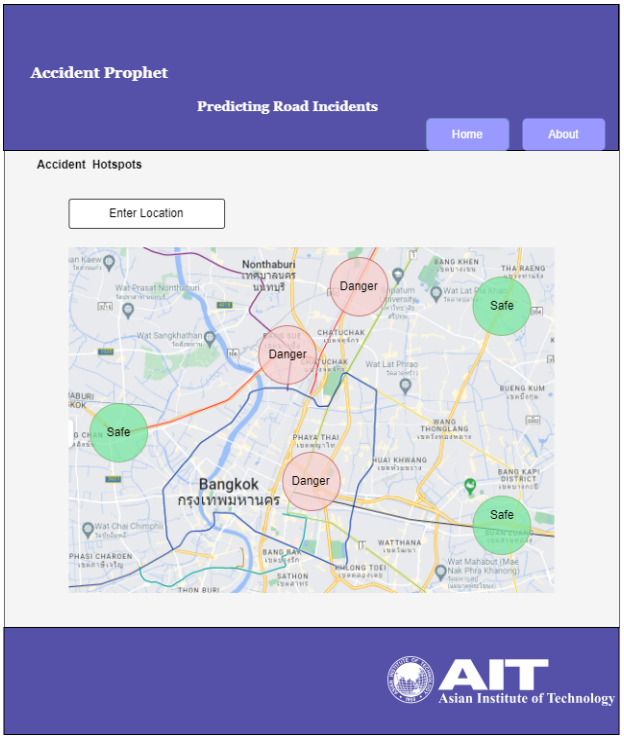

Figure 3- Third page of the application including predictions over the map

Diagrams:

A software architecture diagram develops a detailed software architecture that outlines the system's structure, components, and how they interact. Figure 4 shows the overall structure of the application.

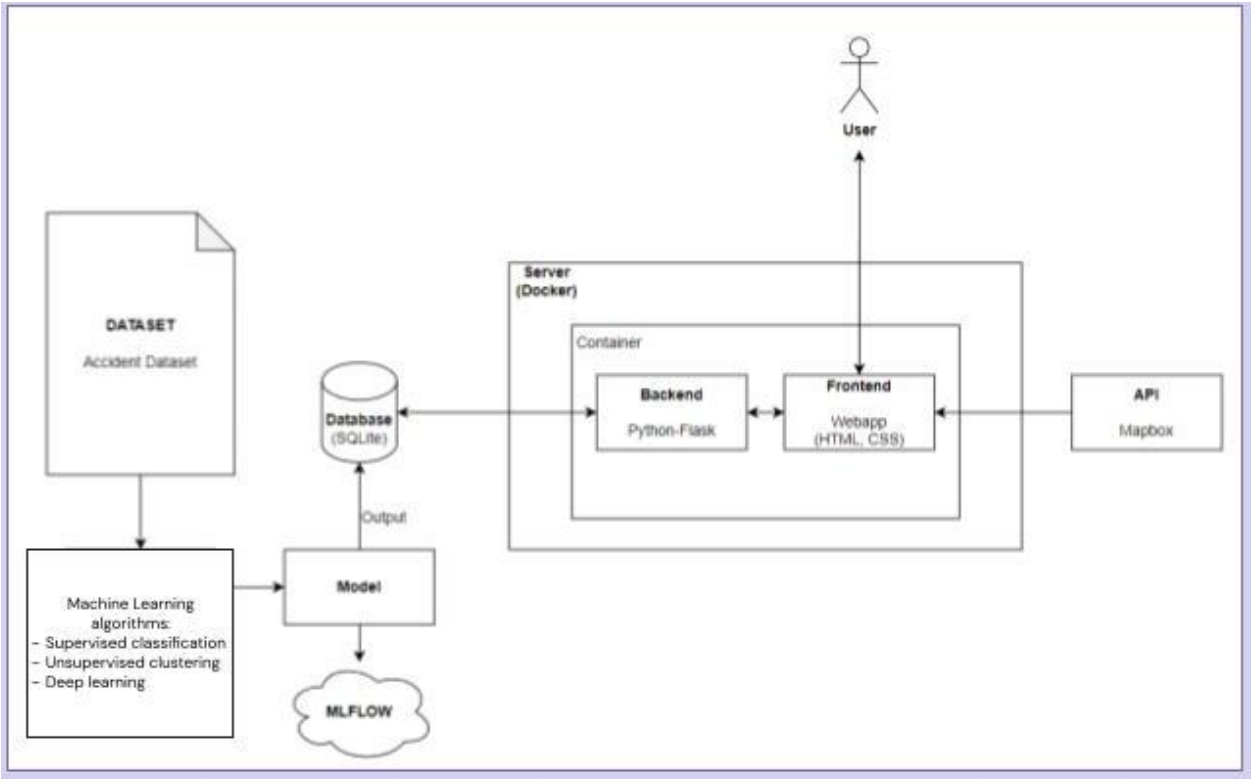

Figure 4- software architecture of the application

Actual Solution:

(Introductory/reminder paragraph what we should do)

The primary function of our application, as previously stated, involves analyzing weather, POI, and time factors to determine severity levels. Additionally, we plan to implement a heatmap layer on the map, as depicted in the mockups (figure 2).

Our team has divided into two subgroups based on the ML big picture guide. The first subgroup, consisting of Reza, Bakyt, and Pun, is working on developing the best classification algorithms. The other subgroup (Tairo and Nitesh) is focused on architecture and the creation of a heatmap.

(Screenshots of your product along with short description how it relates to your requirements)

Our team is currently focused on cleaning data, preprocessing, and conducting feature engineering, as well as software architecture and using map API based on the mockups. Some of the specific tasks related to this include:

- Creating time components
- Creating a new variable: the time between the start and the end of an accident (minute), we are to experiment this as a new label and add a regression problem to our application.
- Create a function to assign grid square IDs
- Recording weather conditions to fewer categories by creating a new column named Weather_Category which categorized similar weather conditions into the same categories.
- One-hot encoding
- EDA
- Train and test splitting
- Try to fit an initial ANN for regression problem (just to get more knowledge about the model)

Figure (5), shows a sample dashboard which we will put on the second page of the application.

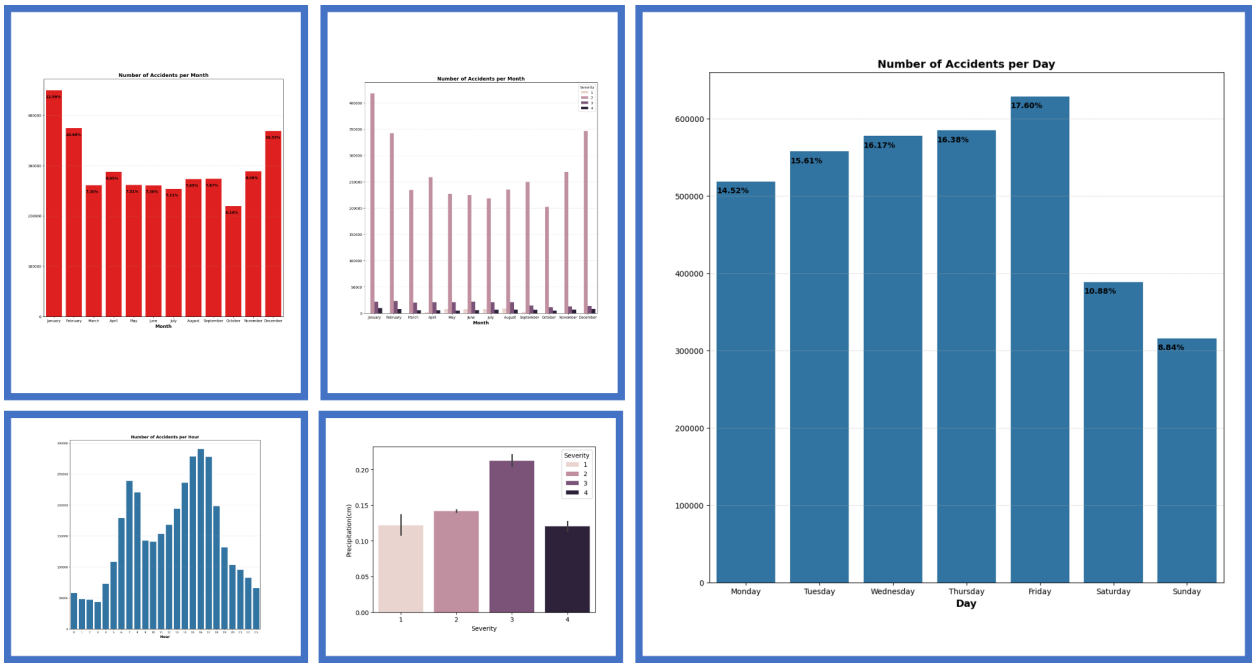

Figure 5. Sample dashboard about accidents

Figure (6) is a sample of using mapbox API in the application.

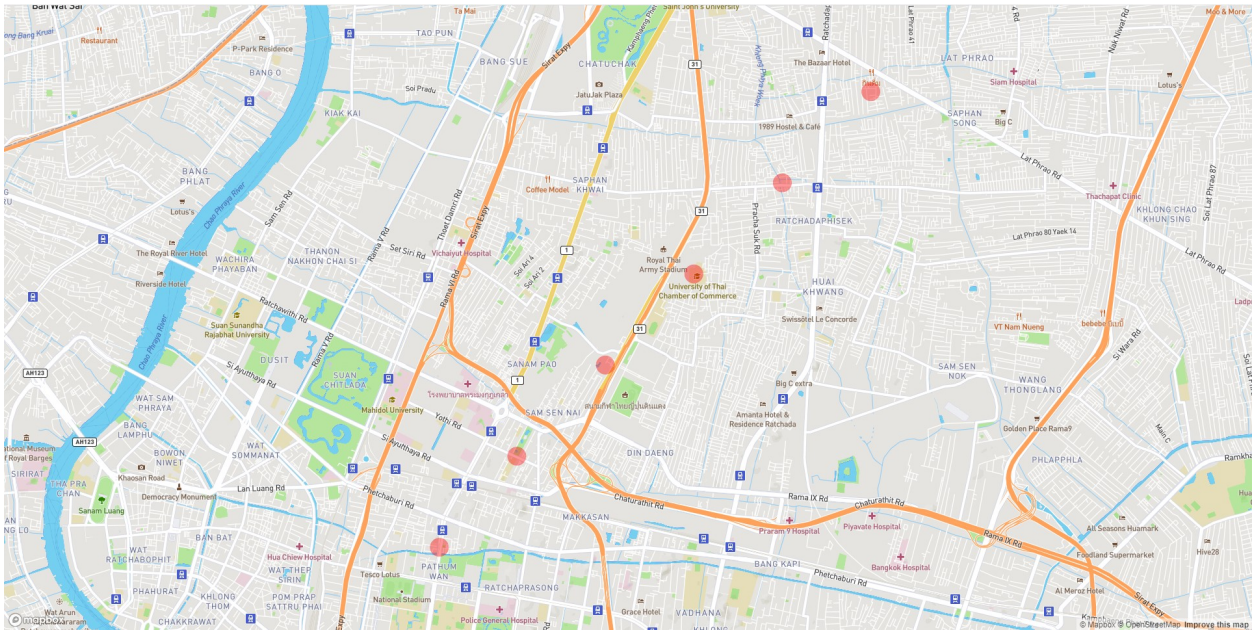

Figure 6. A sample of showing hotspots on the map using mapbox

Figure (7) is the first page of the application repository.

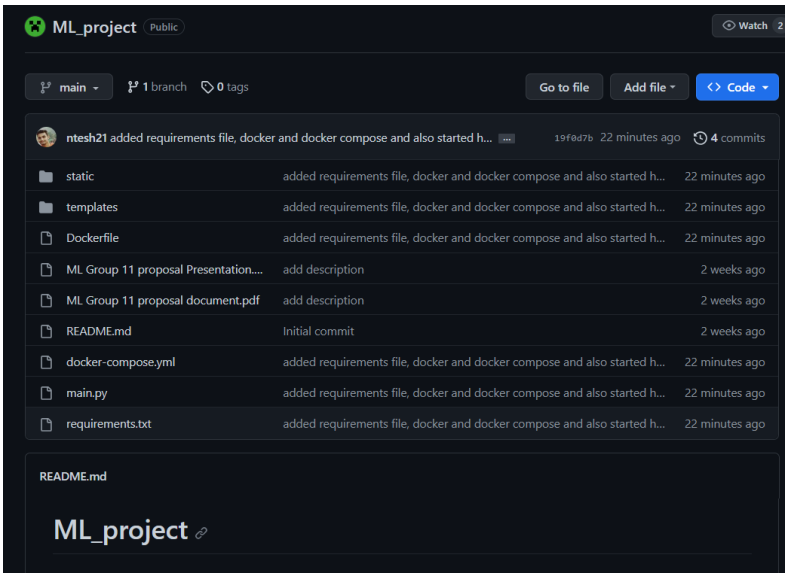

Figure 7. The first page of the application repository.

The learning ANN model codes are displayed in figure (8). However, when it comes to regression problems, ANN algorithms are not very effective, because we have too many features in the data set, using this algorithm can help to get more insight about the data.

```
from sklearn.metrics import r2_score
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.optimizers import Adam
# Build the ANN model
model = Sequential()

# Input layer
model.add(Dense(128, input_shape=(X_train.shape[1],), activation='LeakyReLU'))

# Hidden layers
model.add(Dense(64, activation='LeakyReLU'))
model.add(Dense(48, activation='LeakyReLU'))
model.add(Dense(24, activation='LeakyReLU'))
model.add(Dense(12, activation='LeakyReLU'))

# Output layer
model.add(Dense(1, activation='linear'))  # Linear activation for regression

# Compile the model
model.compile(optimizer=Adam(learning_rate=0.001), loss='mean_squared_error')

# Train the model
model.fit(X_train, Y_train["Distance(km)"], epochs=500, batch_size=100, validation_data=(X_test, Y_test["Distance(km)"]))
yhat = model.predict(X_test)
r2 = r2_score(Y_test["Distance(km)"], yhat)
weights_and_biases = model.get_weights()
# Evaluate the model
loss = model.evaluate(X_test, Y_test["Distance(km)"])
print(f'Mean Squared Error on Test Set: {loss:.2f}')
print(r2)
```

```
31831/31831 [==============================] - 38s 1ms/step - loss: 8.9309 - val_loss: 8.2052
5236/5236 [==============================] - 4s 674us/step
5236/5236 [==============================] - 4s 707us/step - loss: 8.2052
Mean Squared Error on Test Set: 8.21
0.04507700994155206
```

Figure 8. Codes for initial ANN model

Obstacles/risks we face and the plan to mitigate them

- Based on the analysis of our machines, it appears that the dataset is quite sizable, leading to prolonged processing time when implementing different models with varying learning rates, batch sizes, or epochs. To address this issue, we had to narrow down the data selection to include only the data from 2020 onwards, which also allows us to examine more recent information.

Deployment details - where we deploy, and how you deploy. This must match to our "Software Architecture Diagram"

- Basically the application will be deployed the same way as assignments, which is creating an image and composing in an ML server with docker. The image contains both the frontend and the backend part. For the database we plan to use SQLite and put it in the ml server. This depends on whether the professor allows us to do so or not. Another choice is using a cloud server such as AWS. However, our database storage must be less than 5GB in this case.

Contribution of the team members - who do what; justify the amount of work

All team members have made contributions to the project as much as possible. As stated earlier, we have divided the team into two subgroups and worked accordingly. However, we have discussed everything in our line group.

Subgroup one: Bakyt, Reza, Pun

Work on load, data wrangling, feature engineering, ML algorithms

Subgroup two: Tairo, Nitesh

Work on architecture, deployment, ML algorithms

Evaluation

Accuracy:

Accuracy is a crucial metric for evaluating the machine learning model's performance. It measures the model's ability to predict accident occurrences and severities correctly. The accuracy of the prediction model should meet the high-level requirement for precise accident prediction. For example, achieving an accuracy of 90% or higher can be a target.

Recall: Recall, also known as true positive rate, is essential for accident prediction. It measures the ability of the model to identify accidents when they occur correctly. High recall is crucial for preventing accidents. For instance, the requirement could

be to achieve a recall rate of at least 85% to ensure that most accidents are detected.

Inference Speed: Speed is vital for real-time accident prediction and driver guidance. It measures how quickly the system processes and analyzes data to provide timely predictions. It should meet the requirement for real-time insights. For example, the system should process and provide predictions within seconds to meet user expectations.

Human Evaluation:

Satisfaction: User satisfaction is a critical aspect of human evaluation. It measures how satisfied users are with the platform's performance and recommendations. To meet the requirement for user engagement, we can conduct user surveys or collect feedback to ensure that users are content with the system's guidance and accuracy in predicting accidents.

Preference: Preference evaluation can help understand whether users prefer the platform over other solutions or their previous methods. It measures the extent to which users favor the platform for accident prediction and prevention. Meeting the requirement for user engagement may involve conducting preference surveys or comparing user adoption rates before and after the platform's implementation.

Actual Evaluation:

Currently, we are unable to provide any evaluation metrics as we are still in the process of working on the models.

References

- Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. *ArXiv*. https://doi.org/10.1145/3347146.3359078
- https://www.cdc.gov/injury/features/global-road-safety/index.html
- https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents