Accident Risk Prediction through Machine Learning: A Comprehensive Study

Anushka Jindal
RV College of Engineering
Dept. of Information Science and Engineering
Bengaluru, India
anushkajindal.is20@rvce.edu.in

Tanish Mathur
RV College of Engineering
Dept. of Information Science and Engineering
Bengaluru, India
tanishmathur.is20@rvce.edu.in

Sangya Medhavi Shree Goyal
RV College of Engineering
Dept. of Information Science and Engineering
Bengaluru, India
smedhavisg.is20@rvce.edu.in

Dr. Kavitha S N
RV College of Engineering
Dept. of Information Science and Engineering
Bengaluru, India
kavithasn@rvce.edu.in

***Abstract*—In the realm of road safety, the mitigation of accidents and their potential consequences remains a paramount concern. This research paper embarks on an extensive exploration of accident risk prediction, harnessing the capabilities of Extra Trees Regressor and XGBoost for the critical task of parameter importance analysis. The study is centered around the development of a sophisticated web-based prediction system, encompassing various facets such as system architecture, meticulous data preprocessing, and the intricate intricacies of machine learning model development. The dataset used for the same is from Kaggle comprising of data about accidents in US between 2016-2023 . Through the lens of parameter importance analysis, this research unveils the key determinants that underpin accident risk, casting a spotlight on the profound implications of these insights on prediction accuracy which is currently 85%. The ensuing discourse delves into the nuanced interpretation of results, culminating with a forward-looking perspective that outlines potential pathways for future research endeavors in this consequential domain.**

***Keywords*—*Accident Risk Prediction, Accident Prevention, Road Safety, Data-Driven Analysis, Feature Selection, Risk Assessment, Machine Learning Algorithms, Data Preprocessing, Model Evaluation, Hist Gradient Boosting Regressor, XGBoost, Parameter Importance Analysis, Predictive Systems.*

I. INTRODUCTION

Road accidents remain a persistent challenge with far-reaching implications, necessitating innovative strategies for prevention and mitigation. In response, the field of accident risk prediction has emerged as a crucial avenue for preemptive interventions. This research harnesses the power of machine learning to deepen our understanding of accident risk factors, aiming to revolutionize road safety practices.

Accurate accident risk prediction holds transformative potential. Traditional methods often struggle to capture the intricate interplay of factors contributing to accidents. In contrast, machine learning offers the promise of unraveling complex patterns within voluminous datasets, revealing trends and variables that influence accident occurrence. This paper presents a comprehensive study that advances accident risk prediction through systematic parameter importance analysis, using a dataset spanning US accident records from 2016 to 2023.

Leveraging the robust methodologies of Extra Trees Regressor and XGBoost, this research delves into the relative significance of parameters impacting accident risk. The focal point of our investigation revolves around identifying key drivers of accident risk. By doing so, we enhance predictive accuracy and unlock insights that inform targeted safety measures and policies. The outcomes of this study offer stakeholders actionable intelligence to prioritize safety initiatives and interventions, thus reshaping accident prevention paradigms.

This paper navigates through existing literature, and the specifics of our approach, and culminates in a comprehensive evaluation of our findings. By bridging the gap between traditional approaches and innovative machine-learning techniques, this research contributes to the broader discourse on accident prevention and road safety, showcasing the transformative potential of data-driven insights.

II. RELATED WORK

The insights extracted from the reviewed literature significantly contribute to shaping the foundations of our project [1-5]. Moosavi et al. [1] address the real-time prediction of traffic accidents on sparse data, employing a deep-neural network model that leverages various data attributes such as traffic events, weather data, points of interest, and time. Charandabi et al. [2] introduce a generalized regression neural network optimized with self-

DOI: 10.55041/IJSREM25419

organizing maps, showcasing its accuracy in predicting accident risk based on diverse predictor variables. Their study highlights the significance of attributes like distance from traffic control cameras, day of the week, driver's age, weather, elevation, and vehicle type. Brühwiler et al. [3] emphasize the impact of geographical context on accident risk, evaluating machine learning classifiers' performance while incorporating geographical information like weather, points of interest, and land use. Notably, they find that the inclusion of geographical information, including weather and points of interest, can enhance predictive performance, with land use being the most informative. Kushwaha and Abirami [4] explore multilevel models for accident severity analysis and emphasize the superior performance of the XGBoost algorithm. Lastly, Gutierrez-Osorio et al. [5] propose an ensemble Deep Learning Model, intertwining social media and open data for accident prediction, showcasing its enhanced performance compared to baseline algorithms and other models reported in the literature. These insights collectively guide our approach towards accurate accident risk prediction, emphasizing parameter importance and the evaluation of predictive models, particularly XGBoost, in the context of our project.

III. DATA COLLECTION AND PRE PROCESSING

The dataset used here has been obtained from Kaggle. It is a countrywide car accident dataset, which covers 49 states of USA. The data has been 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 a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.The flow of Exploratory Data Analysis or EDA for the project is as shown in Fig 1.

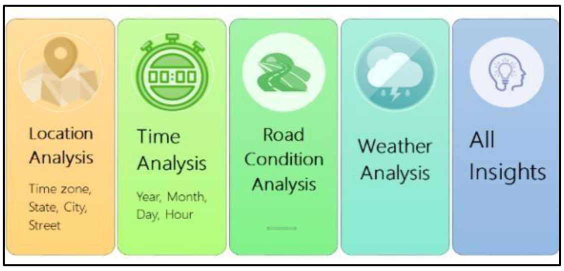
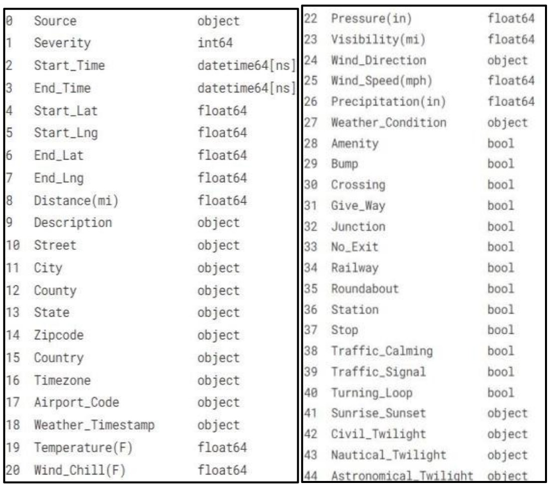


Fig 1. Exploratory Data Analysis

A. Data Collection

The dataset used for this research focuses on accident risk prediction. It contains a total of 44 parameters, encompassing various aspects related to accidents, such as geographical coordinates (start and end), weather conditions (temperature, wind chill, humidity, pressure, precipitation), and more. These parameters are vital for understanding the factors contributing to accidents and subsequently developing an effective prediction model.The dataset includes different data types, such as numerical, categorical, and datetime. Numerical data includes measurements like temperature, humidity, and pressure. Categorical data consists of attributes like city names and various time zone-related columns. The datetime data type involves attributes such as 'Start_Time' and 'End_Time', which provide valuable temporal information. Fig 2. shows the parameters and their data types,



```
0   Source                object
1   Severity              int64
2   Start_Time            datetime64[ns]
3   End_Time              datetime64[ns]
4   Start_Lat             float64
5   Start_Lng             float64
6   End_Lat               float64
7   End_Lng               float64
8   Distance(mi)          float64
9   Description           object
10  Street                object
11  City                  object
12  County                object
13  State                 object
14  Zipcode               object
15  Country               object
16  Timezone              object
17  Airport_Code          object
18  Weather_Timestamp     object
19  Temperature(F)        float64
20  Wind_Chill(F)         float64
```

```
22  Pressure(in)           float64
23  Visibility(mi)         float64
24  Wind_Direction         object
25  Wind_Speed(mph)        float64
26  Precipitation(in)      float64
27  Weather_Condition      object
28  Amenity                bool
29  Bump                   bool
30  Crossing               bool
31  Give_Way               bool
32  Junction               bool
33  No_Exit                bool
34  Railway                bool
35  Roundabout             bool
36  Station                bool
37  Stop                   bool
38  Traffic_Calming        bool
39  Traffic_Signal         bool
40  Turning_Loop           bool
41  Sunrise_Sunset         object
42  Civil_Twilight         object
43  Nautical_Twilight      object
44  Astronomical_Twilight  object
```

Fig 2. Parameters and data types

B. Cleaning and Transforming data

Data preprocessing plays a pivotal role in the quality enhancement and suitability of the dataset for predictive modeling endeavors. The unique and missing values are observed for categorical and numerical columns. Object-based columns are assessed for utility, leading to the removal of the 'Country' column due to its single-value nature and the exclusion of the 'Description' column as it's unsuitable for the regression task. Missing values are managed by imputing them with the most frequent value in each column. This maintains data integrity and completeness. Temporal insights are derived by extracting date-related details from the 'Start_Time' column. This adds depth to the analysis of time-based patterns. Accurate incident duration is determined by utilizing the 'Start_Time' and 'End_Time' columns. This knowledge enriches our understanding of accident timelines. Encoding time-related columns enhances the dataset with structured representations of sun-light conditions and timezones, enabling more insightful analyses. Numerical columns undergo rigorous cleaning, including the removal of uniform columns like 'Turning Loop' and those with substantial missing values, such as 'End_Lat', 'End_Lng', 'Precipitation', and 'Wind_Chill'. Columns weakly correlated with the 'Severity' target are pruned to optimize the dataset for predictive modeling.

C. Exploring the cleaned data

In terms of geographical distribution, California leads in accidents, followed by Florida and Texas. These states might have unique risk factors influencing their accident rates, prompting the need for targeted preventive measures. Cities with the highest accidents include Miami, Houston, Los Angeles, and Charlotte. Accidents are more common on workdays, suggesting a link between traffic volume and accident occurrences. Weather plays a role, as accidents peak on clear days. This is shown in Figure 3.

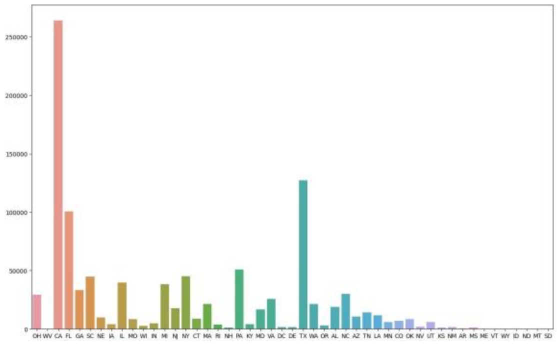


Fig 3. No. of Accidents in each State

Within the context of the cleaned data analysis, Figure 4 presents a graphical representation that illuminates the distribution of accidents across various weather conditions. The x-axis of the graph denotes different weather conditions, including clear, overcast, fair, partly cloudy, and mostly cloudy. On the y-axis, the graph displays the corresponding number of accidents that occurred under each specific weather condition.

This visualization serves as a comprehensive snapshot of how different weather conditions correlate with accident occurrences. By plotting the frequency of accidents against distinct weather categories, stakeholders gain insights into the potential impact of weather on road safety. Patterns and trends become evident as the graph showcases the variations in accident numbers across diverse weather conditions, aiding in the identification of high-risk scenarios and contributing to informed decision-making for accident prevention strategies.

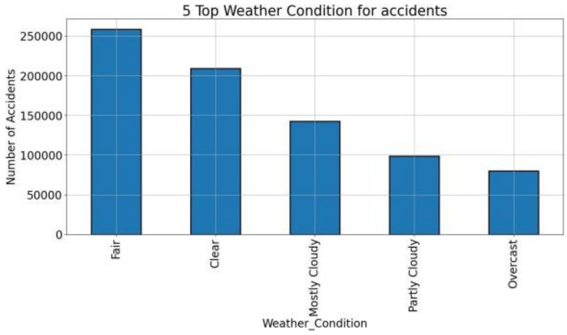


Fig 4. Top Weather Conditions for Accidents

D. Reducing Parameters

Low-correlation columns are removed, enhancing model efficiency by eliminating noise and irrelevant attributes, optimizing predictive capabilities. Excluding 'Descriptions' aligns with task relevance. Retaining essential spatial attributes like city simplifies the dataset while preserving value. Instances with missing values are removed, bolstering data integrity. Instances with practical implications, like zero values for pressure, hour, or distance, are excluded. Undersampling rectifies class imbalance, enhancing the model's handling of diverse accident scenarios and producing more robust predictions. Target encoding with smoothing captures city impact on accident severity while accommodating rarity, enriching the model's understanding. At the end of it the dataset is cleaned and left with just 26 parameters. The heat map generated taking into consideration these parameters is shown in Figure 5, the magnitude of the values in the blocks shows the correlation coefficient between the parameters.

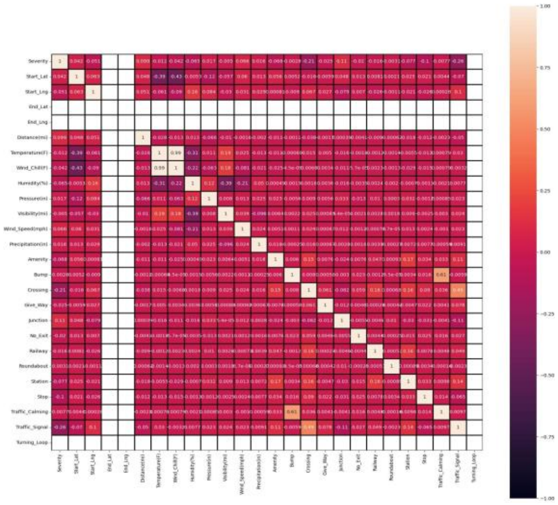


Fig 5. Heatmap of Parameters

IV. MACHINE LEARNING COMPONENT

Under the Machine Learning Components section, we delve into the application of advanced machine learning models to unravel the feature importance in determining the severity of accidents. This analysis sheds light on the pivotal parameters that significantly influence the outcome, contributing to a more nuanced understanding of accident risk prediction. The Hist Gradient Boosting Regressor is a powerful machine learning model that excels in handling regression tasks. It works by constructing an ensemble of decision trees and optimizing their structure using histogram-based techniques. This model efficiently captures complex relationships between input parameters and target variables. In our study, we employed the Hist Gradient Boosting Regressor to assess the importance of different parameters in predicting accident severity. Through iterative optimization, the model revealed the hierarchy of parameter significance, enabling us to discern the critical factors impacting accident outcomes.

XGBoost is another widely-used machine learning algorithm renowned for its performance in predictive modeling. It is an ensemble method that combines multiple decision trees, iteratively refining their construction to enhance predictive accuracy. In our research, XGBoost played a pivotal role in feature importance analysis. By systematically evaluating the contribution of each parameter to the model's predictive power, XGBoost enabled us to identify the most influential factors that drive accident severity.

This analysis informs a comprehensive understanding of accident risk dynamics. Both the Hist Gradient Boosting Regressor and XGBoost were harnessed to perform feature importance analysis. This process involved feeding the models with the collected accident data and corresponding parameters. The models underwent rigorous training, iteratively refining their internal structures to optimize prediction accuracy. As a result, they assigned varying degrees of importance to each parameter, reflecting their contribution to the prediction of accident severity.

The outcome of this analysis was a ranking of parameters based on their impact on accident outcomes. Parameters with higher importance scores were deemed more influential in determining accident severity, while those with lower scores had relatively less impact. This insights-rich ranking provides valuable information for decision-makers, enabling them to prioritize interventions and strategies aimed at reducing accident risk. The feature importance obtained from the Hist Gradient Boosting Regressor is visually depicted in Figure 6, showcasing the relative significance of different parameters in predicting accident severity. Similarly, the feature importance analysis conducted using XGBoost is illustrated in Figure 7, offering a clear graphical representation of the pivotal parameters that drive the model's accuracy..

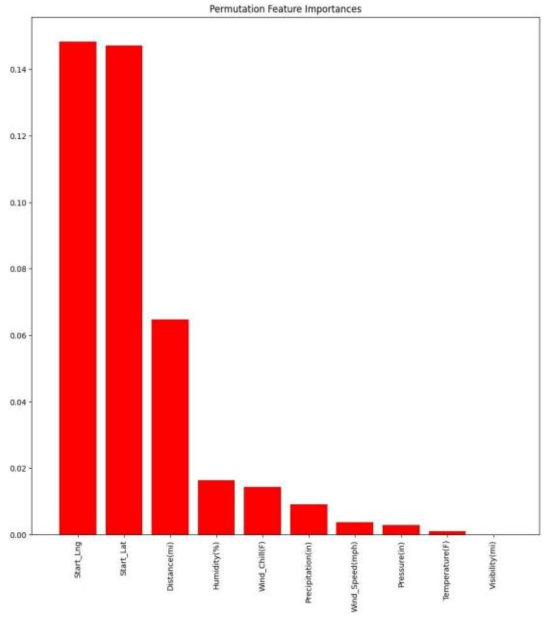


Fig 6. Feature Importance by Hist Gradient Boosting Regressor

In essence, our utilization of Hist Gradient Boosting Regressor and XGBoost for feature importance analysis enhances our understanding of the complex dynamics behind accident severity. These models unravel the intricate relationships between parameters and accident outcomes, empowering stakeholders with knowledge to devise effective preventive measures and enhance road safety.

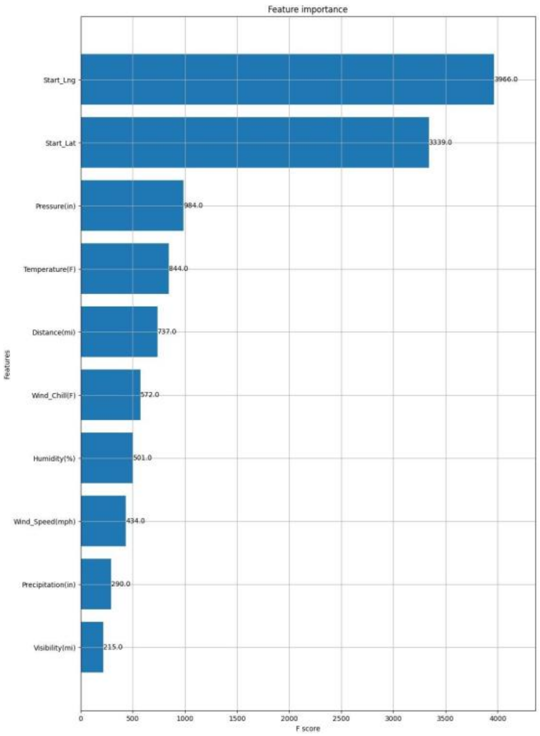


Fig 7. Feature Importance by XGBoost

V. SYSTEM ARCHITECTURE & DESIGN

The proposed research focuses on the design and development of an accident severity prediction system, with the aim of providing actionable insights to users for safer driving practices. The system encompasses several interrelated components, each contributing to the overall effectiveness and accuracy of the prediction process.

The first phase involves meticulous data collection and preprocessing. A comprehensive dataset is obtained from Kaggle, containing diverse attributes such as accident location (latitude and longitude), weather conditions, road features, and temporal factors. Rigorous data cleaning procedures are applied to address missing values, outliers, and inconsistencies, ensuring the dataset's reliability. Through transformation techniques like one-hot encoding, categorical variables are converted into numerical forms suitable for analysis. Notably, feature extraction and correlation analysis are conducted to derive meaningful insights from the data. The latter technique assists in identifying key relationships among different attributes, thereby guiding the subsequent stages. Feature selection is a pivotal step following data preprocessing. Employing methodologies such as Feature Importance Analysis, the most influential attributes are identified. This process aids in reducing model complexity and enhancing computational efficiency, while retaining the predictive power of the system.

Model building and training form the core of the research endeavor. An appropriate regression algorithm is chosen based on its capacity to discern patterns from the data. The dataset is partitioned into training and testing sets, enabling model evaluation. The selected algorithm is trained on the training data, utilizing the top features identified during the feature selection phase. Model performance is rigorously evaluated using pertinent metrics to ascertain its predictive accuracy.

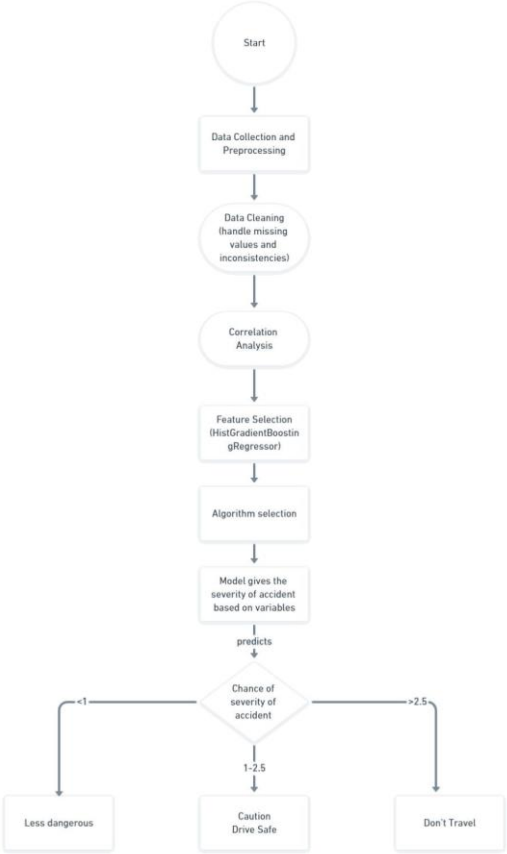

Start
Data Collection and Preprocessing
Data Cleaning (handle missing values and inconsistencies)
Correlation Analysis
Feature Selection (HistGradientBoostingRegressor)
Algorithm selection
Model gives the severity of accident based on variables
predicts
Chance of severity of accident
<1
1-2.5
>2.5
Less dangerous
Caution Drive Safe
Don't Travel

Fig 8. System Architecture

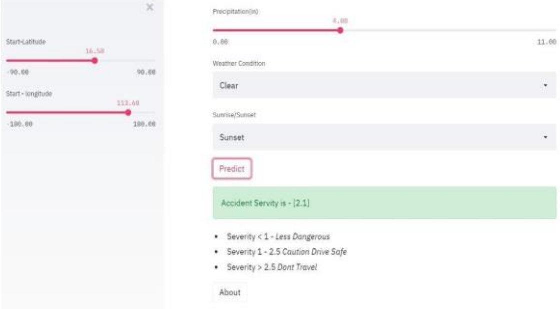
The deployment stage involves the creation of an intuitive user interface. Users input parameters such as starting latitude, longitude, weather conditions, and road features. This input undergoes preprocessing to align with the model's requirements. The trained model then predicts the accident severity based on the preprocessed user data. Predictions are categorized into actionable recommendations, such as "Less Dangerous," "Caution Drive Safe," or "Don't Travel." These outputs are communicated to the user through the interface, providing valuable insights for informed decision-making. The system's performance is continuously monitored, and model updates are executed based on user feedback and performance data. These updates ensure that the system remains adaptive and accurate, aligning with evolving user needs and real-world accident data. Figure 8 shows the sytem architecture in the manner of a flowchart.

VI. TESTING AND IMPLEMENTATION

The accident risk prediction web application's architecture is thoughtfully designed, incorporating vital components to ensure accurate and reliable predictions. The process initiates with users inputting their desired location, which is then translated into precise geographical coordinates. After collecting relevant weather parameters such as humidity in percenatge, wind chill parameter, wind speed in mph, precipitation in inches and spatial data, the application harnesses the power of a sophisticated statistical model that employs advanced machine learning techniques. This model is the core engine that calculates a comprehensive risk score, capturing the intricate interplay between various parameters and prevailing conditions.

What sets the architecture apart is its seamless integration of external data sources, providing users with an intuitive and user-friendly interface. The model's predictive mechanism plays a pivotal role in delivering precise risk assessments, empowering users to make well-informed decisions about their travel plans. This entire process harmoniously combines data acquisition, efficient processing, and complex modeling, resulting in the delivery of a dependable and robust accident risk prediction tool.

At the culmination of this process emerges a numerical risk score, a concise representation of potential risks associated with the specified conditions. Scores below 1 convey a sense of safety, indicating a relatively low risk environment for travel. The range of 1 to 2.5 signifies a heightened level of risk, prompting caution in decision-making. Scores surpassing 2.5 are indicative of extreme danger, strongly advising against travel. This intuitive scoring system empowers users with quick and clear insights into risk levels, enabling them to navigate their travel plans with enhanced safety awareness. The architecture's integration of technology, data, and modeling culminates in an invaluable tool for individuals seeking safer journeys. Figure 9 shows how the information about the parameters is given as input and a numerical score is obtained.


Start-Latitude
16.50
-90.00 90.00
Start - longitude
113.60
-180.00 180.00
Precipitation(in)
4.00
0.00 11.00
Weather Condition
Clear
Sunrise/Sunset
Sunset
Accident Servity is - [2.1]

- Severity < 1 - *Less Dangerous*
- Severity 1 - 2.5 *Caution Drive Safe*
- Severity > 2.5 *Dont Travel*

Fig 9. Web Application

The effectiveness of the model was rigorously assessed through validation using two prominent machine learning algorithms: the 'decision tree classifier' and the 'random forest regressor'. These well-established models are adept at addressing specific tasks within the realm of predictive analytics. The 'decision tree classifier' operates by recursively partitioning the dataset into subsets based on attribute values, ultimately creating a tree-like structure of decisions. This approach effectively categorizes instances into distinct classes, making it particularly suited for classification tasks such as determining risk levels in the accident prediction context. On the other hand, the 'random forest regressor' capitalizes on the concept of ensemble learning. It constructs a multitude of decision trees and aggregates their outputs to generate more robust and accurate predictions. By leveraging the wisdom of multiple trees, this model excels in regression tasks, where the goal is to predict continuous numerical values.

In the testing phase, the accident risk prediction web application demonstrated an impressive accuracy rate of 85% when employing these machine learning models. This substantial level of accuracy underscores the system's reliability and its adept utilization of the 'decision tree classifier' and 'random forest regressor'. The models' synergy and the careful integration of their outcomes contribute to the application's ability to deliver dependable and informative accident risk predictions.

By harnessing the strengths of these models, the application enhances the accuracy and credibility of its risk assessments, enabling users to make more informed and secure travel choices. The successful integration of these machine learning techniques further bolsters the application's position as a valuable tool for promoting road safety and reducing accident risks.

VII. CONCLUSION & FUTURE SCOPE

The research has resulted in a novel approach aimed at mitigating the critical issue of road accidents and their devastating consequences. Through the integration of location and weather APIs, a dynamic and real-time prediction mechanism has been established. The comprehensive evaluation of factors such as weather conditions, visibility, temperature, and time by the machine learning model facilitates the assessment of accident susceptibility and severity in specific regions, enabling timely alerts to users prompting them to exercise caution when traversing areas prone to accidents.

Looking forward, the research lays the foundation for further exploration and progress. The valuable insights gained from analyzing key parameters using tools like Hist Gradient Boosting Regressor and XGBoost provide crucial guidance about the factors that influence accident likelihood. This knowledge forms the basis for potentially refining and optimizing the model to make predictions even more accurate. It can improved further by incorporating advanced methods and broader datasets. By adding real-time traffic information, understanding social behaviors, and considering road conditions, model's ability to predict accidents can be enhanced comprehensively.

VIII. REFERENCES

[1] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights," November 2019.

[2] N. Kaffash Charandabi, A. Gholami, and A. Abdollahzadeh Bina, "Road accident risk prediction using generalized regression network optimized with self-organising map" February 2022

[3] L. Brühwiler, C. Fu, H. Huang, L. Longhi, and R. Weibel, "Predicting individuals' car accident risk by trajectory, driving events, and geographical context," April 2022.

[4] M. Kushwaha and M. S. Abirami, "Comparative Analysis on the Prediction of Road Accident Severity Using Machine Learning Algorithms," February 2022.

[5] C. Gutierrez-Osorio, F. A. González, and C. A. Pedraza, "Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data," June 2022.

[6] A. Azhar, S. Rubab, M. M. Khan, Y. A. Bangash, M. D. Alshehri, F. Illahi, and A. K. Bashir, "Detection and prediction of traffic accidents using deep learning techniques," January 2022.

[7] J. Bao, P. Liu, and S. V. Ukkusuri, "A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data," January 2019.

[8] B. Dimitrijevic, S. D. Khales, R. Asadi, and J. Lee, "Short-Term Segment-Level Crash Risk Prediction Using Advanced Data Modeling with Proactive and Reactive Crash Data," Appl. Sci., vol. 12, no. 2, p. 856, 2022.

[9] R. Zhen, Z. Shi, J. Liu, and Z. Shao, "A novel arena-based regional collision risk assessment method of multi-ship encounter situation in complex waters," Ocean Eng., vol. 246, p. 110531, Feb. 2022.

[10] R. Zhen, Z. Shi, Z. Shao, and J. Liu, "A novel regional collision risk assessment method considering aggregation density under multi-ship encounter situations," Published online by Cambridge University Press, Nov. 19, 2021.

[11] H. Feng, M. Grifoll, Z. Yang, and P. Zheng, "Collision risk assessment for ships' routeing waters: An information entropy approach with Automatic Identification System (AIS) data," Ocean Coast. Manag., vol. 224, p. 106184, Jun. 2022.

[12] G. Hou, K. Xu, and J. Lian, "A review on recent risk assessment methodologies of offshore wind turbine foundations," Ocean Eng., vol. 264, p. 112469, Nov. 15, 2022.

[13] J. Guo and C. Luo, "Risk assessment of hazardous materials transportation: A review of research progress in the last thirty years," J. Traffic Transp. Eng. (Engl. Ed.), vol. 9, no. 4, pp. 571-590, Aug. 2022.

[14] Z. He, C. Chen, and W. Weng, "Multi- hazard risk assessment in process industries: State-of-the-Art," J. Loss Prev. Process Ind., vol. 76, p. 104672, May 2022.

[15] Y. Halabi, H. Xu, D. Long, Y. Chen, Z. Yu, F. Alhaek, and W. Alhaddad, "Causal factors and

risk assessment of fall accidents in the U.S. construction industry: A comprehensive data analysis (2000–2020)," Safety Sci., vol. 146, p. 105537, Feb. 2022.

[16] X. Li, J. Wang, and G. Chen, "A machine learning methodology for probabilistic risk assessment of process operations: A case of subsea gas pipeline leak accidents," Process Saf. Environ. Prot., vol. 165, pp. 959-968, Sep. 2022.

[17] X. Li, J. Wang, R. Abbassi, and G. Chen, "A risk assessment framework considering uncertainty for corrosion-induced natural gas pipeline accidents," J. Loss Prev. Process Ind., vol. 75, p. 104718, Feb. 2022.

[18] J. Zhu, Y. Ma, and Y. Lou, "Multi-vehicle interaction safety of connected automated vehicles in merging area: A real-time risk assessment approach," Accid. Anal. Prev., vol. 166, p. 106546, Mar. 2022.

[19] G. Zhu, Z. Xie, H. Xu, N. Wang, L. Zhang, N. Mao, and J. Cheng, "Oil Spill Environmental Risk Assessment and Mapping in Coastal China Using Automatic Identification System (AIS) Data," Sustainability, vol. 14, no. 10, p. 5837, 2022. DOI: 10.3390/su14105837.