Enhancing Road Safety: A Comprehensive Analysis of Nationwide Traffic Accidents in the United States

Dominic Parosh Yamarthi *Haripriya Raman* *Sushmitha Jakka*

dyamarth *hraman* *sjakka*

Problem Statement:

Road accidents pose a significant threat to public safety, and understanding the underlying patterns and factors is crucial for effective prevention and mitigation strategies. This project aims to analyse a countrywide traffic accident dataset covering 49 states in the United States, comprising data from diverse sources such as transportation departments, law enforcement, and traffic sensors. The primary objective is to uncover insights into the causes, frequency, and impact of accidents, ultimately contributing to the development of informed policies and interventions for enhancing road safety.

To predict future accidents with a certain accuracy or classify the accidents in terms of severity various machine learning algorithms are performed and corresponding accuracies are evaluated. Using these models, the best model can be chosen for prediction or classification. The following machine learning algorithms are performed on the dataset:

1. ***K - Means clustering.***
2. ***K – Nearest Neighbour.***
3. ***Logistic Regression.***
4. ***Linear Regression.***
5. ***Naïve – Bayes or Gaussian.***
6. ***Decision Tree Classifier.***
7. ***Random Forest***
8. ***Gradient Boosting Classifier.***

The above-mentioned algorithms will either cluster or classify the data into various clusters or classes. The classification is done in terms of severity of the accident. To predict the number of accidents that might occur, a regression algorithm should be implemented. It is observed in phase-2 that the random-forest classification algorithm works best, as it yields better accuracy, easy to comprehend, is scalable and is well-suited for a large dataset. Similarly, to predict the number of accidents that might happen, a random forest regressor algorithm is implemented.

Random-Forest Regressor

Random Forest Regressor is a powerful machine learning algorithm that belongs to the ensemble learning family. Ensembles combine the predictions of multiple base models to improve overall performance, and the random forest takes this idea further by constructing a multitude of decision trees during training. In the context of regression tasks, where the goal is to predict continuous values, the Random Forest Regressor excels at capturing complex relationships within the data. Each decision tree in the forest is trained on a random subset of the data and features, introducing diversity, and reducing the risk of overfitting. This randomness and diversity in the model contribute to its robustness and generalization capability.

One key strength of the Random Forest Regressor lies in its ability to handle large datasets with numerous features, making it suitable for a wide range of real-world applications. The algorithm effectively mitigates the risk of overfitting, a common challenge in machine learning, by aggregating the predictions from multiple trees. During the prediction phase, each tree in the forest provides an output, and the final prediction is an average or a weighted average of these outputs. This ensemble approach enhances the model's predictive accuracy and stability, making it particularly valuable for tasks such as predicting housing prices, stock values, or any other continuous variable where capturing intricate patterns is crucial.

Additionally, Random Forest Regressor provides a built-in feature importance measure, which can be insightful for understanding the relative impact of different features on the model's predictions. This transparency is valuable for both interpretability and feature selection in various applications. While Random Forest Regressor exhibits remarkable performance, it is essential to fine-tune hyperparameters to achieve optimal results and avoid potential issues such as overfitting. Overall, the Random Forest Regressor stands out as a versatile and reliable tool for regression tasks in machine learning, balancing complexity, and interpretability to deliver robust predictions.

Steps:

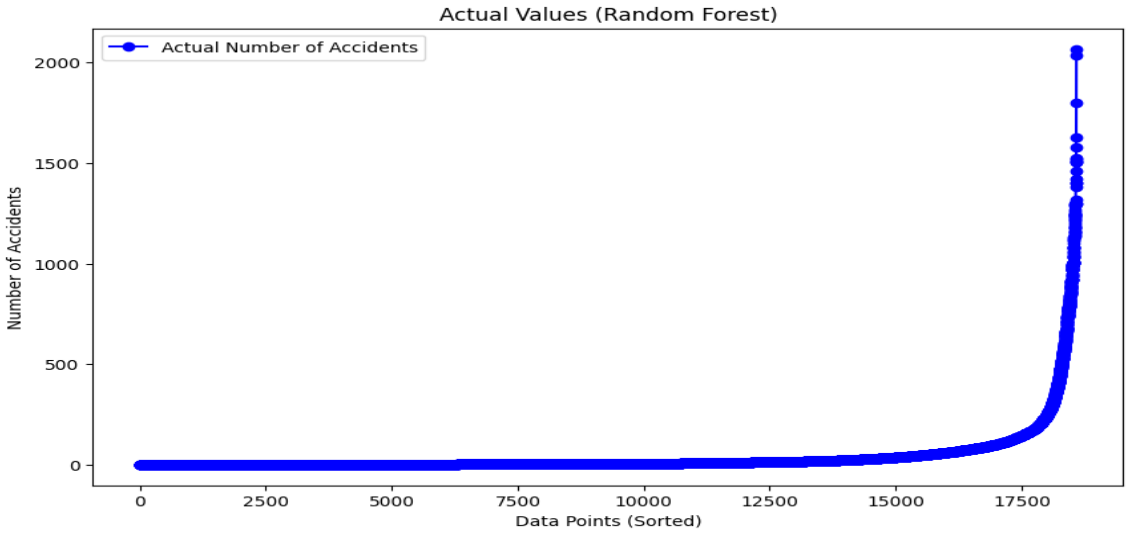
1. The data is pre-processed like phase 1.
2. The data is ordered to get the number of accidents per severity, state, and date.
3. Date column is split into date, month, and year.
4. Number of accidents is taken as dependant variable and severity, state, and date are taken as independent variables.
5. Random-Forest Regression algorithm is performed, and predictions are taken.

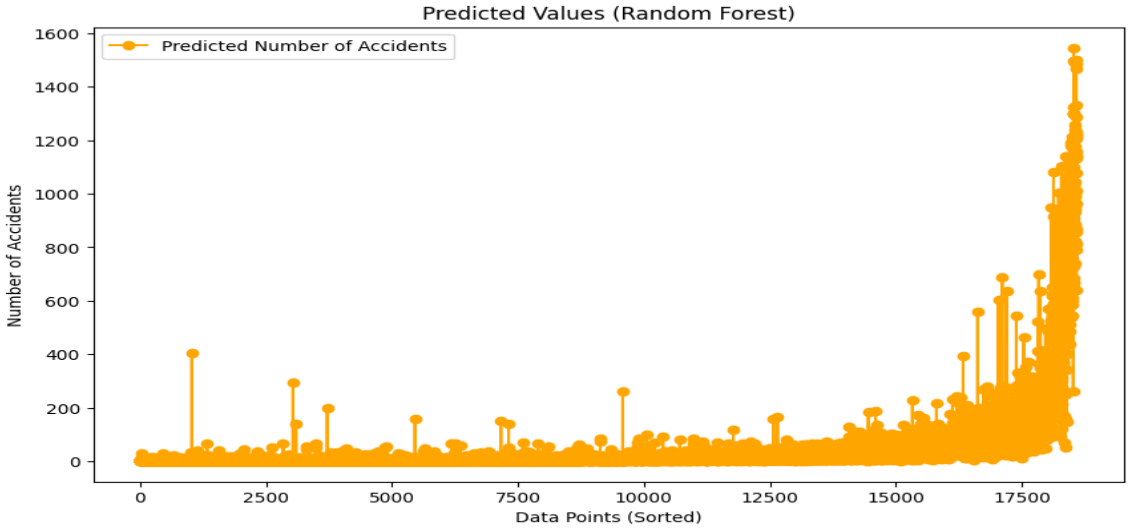
```
# Performing random forest regressor for prediction



# Prepare features and target variable
X = final_df[['Severity', 'State', 'Year', 'Month','Day']]
y = final_df['Number of Accidents']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Build the Random Forest Regressor model
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred_rf = rf_model.predict(X_test)

# Evaluate the model
mse_rf = mean_squared_error(y_test, y_pred_rf)
print(f'Mean Squared Error (Random Forest): {mse_rf}')
```

```
Mean Squared Error (Random Forest): 1736.067680727429
```

```
from sklearn.metrics import r2_score

# Evaluate the model using R-squared
r_squared_rf = r2_score(y_test, y_pred_rf)
print(f'R-squared (Random Forest): {r_squared_rf}')
```

```
R-squared (Random Forest): 0.8828704260519764
```



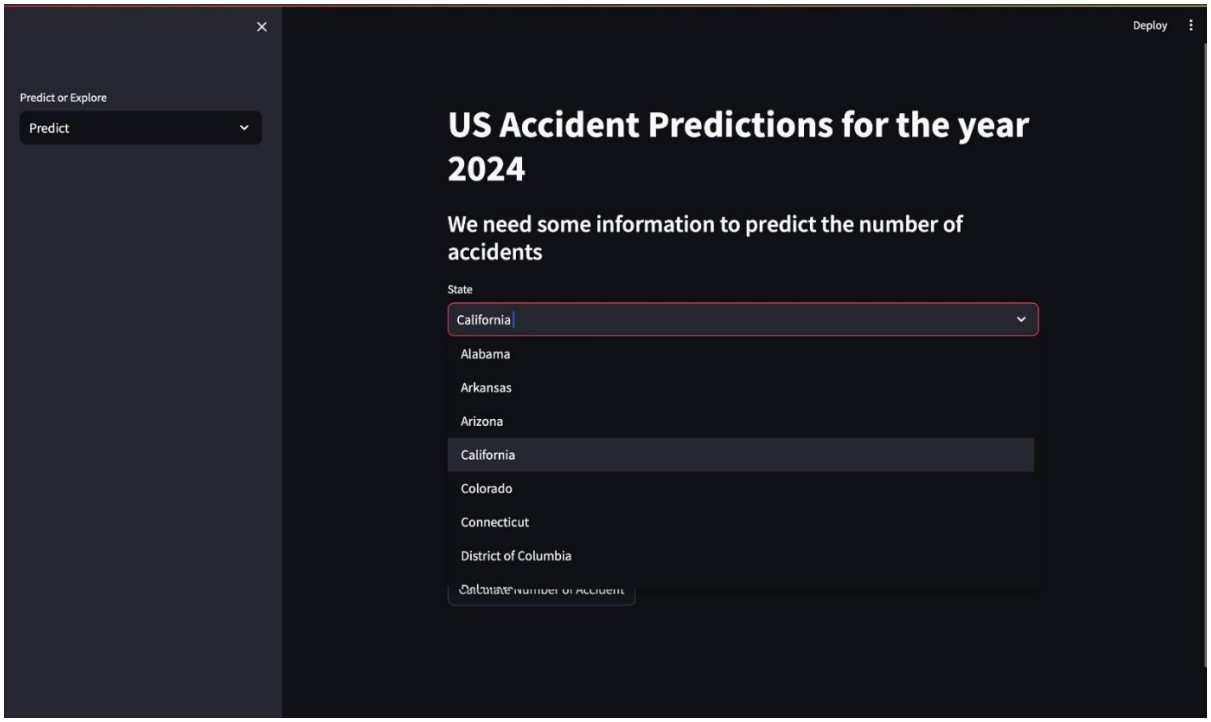



```
    new_instance = pd.DataFrame({'Severity': [2],
                                 'State': [0],
                                 'Year': [2024],
                                 'Month': [10],
                                 'Day':[13]})

    new_instance =new_instance.astype(float)
    print(new_instance)
    # Use the trained Random Forest Regressor model to predict the number of accidents
    predicted_accidents = rf_model.predict(new_instance)

    print(f'Predicted Number of Accidents: {predicted_accidents}')
```

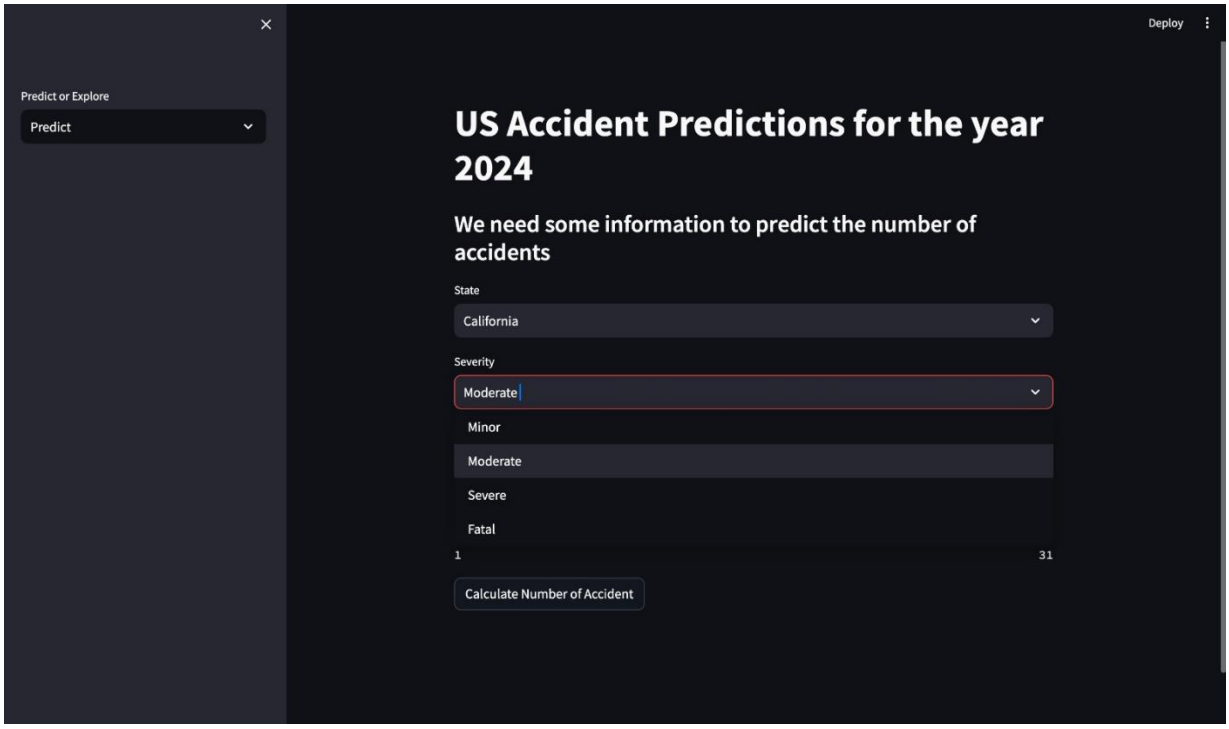
```
   Severity  State    Year  Month   Day
0       2.0    0.0  2024.0   10.0  13.0
Predicted Number of Accidents: [44.18]
```

Working:

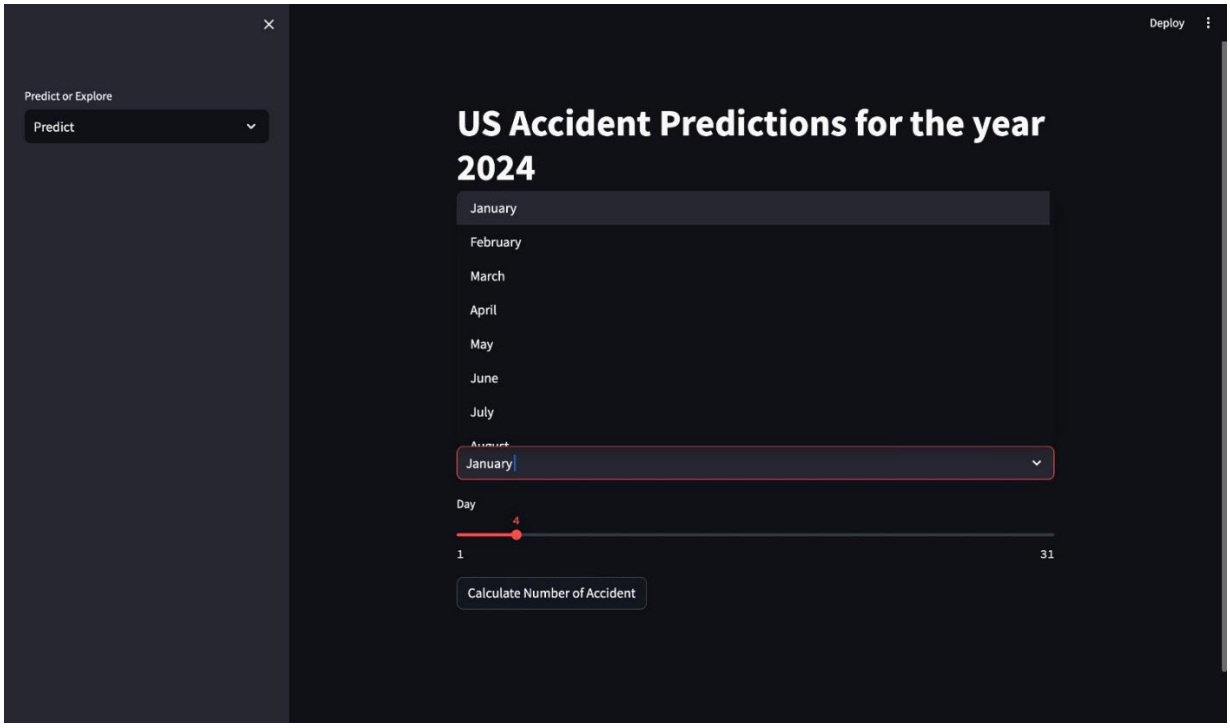
1. A state is selected among a list of 49 states of USA (Alaska is not included)



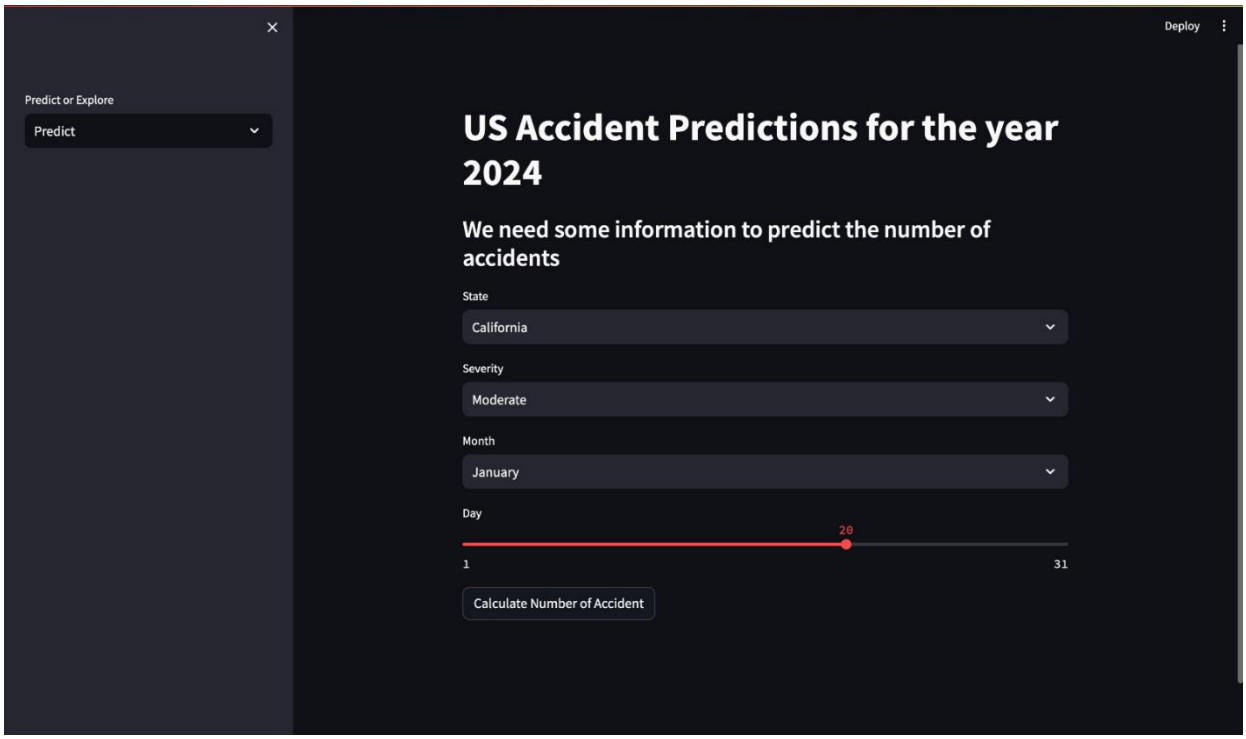
2. Severity of the accident is selected from a list of 4:
    a. Minor: Less damage to property and/or minor injuries
    b. Moderate: High damage to property and serious injuries
    c. Severe: High damage to property and disability inducing injuries possible loss of life
    d. Fatal: Unrecoverable damage to property and adamant loss of life



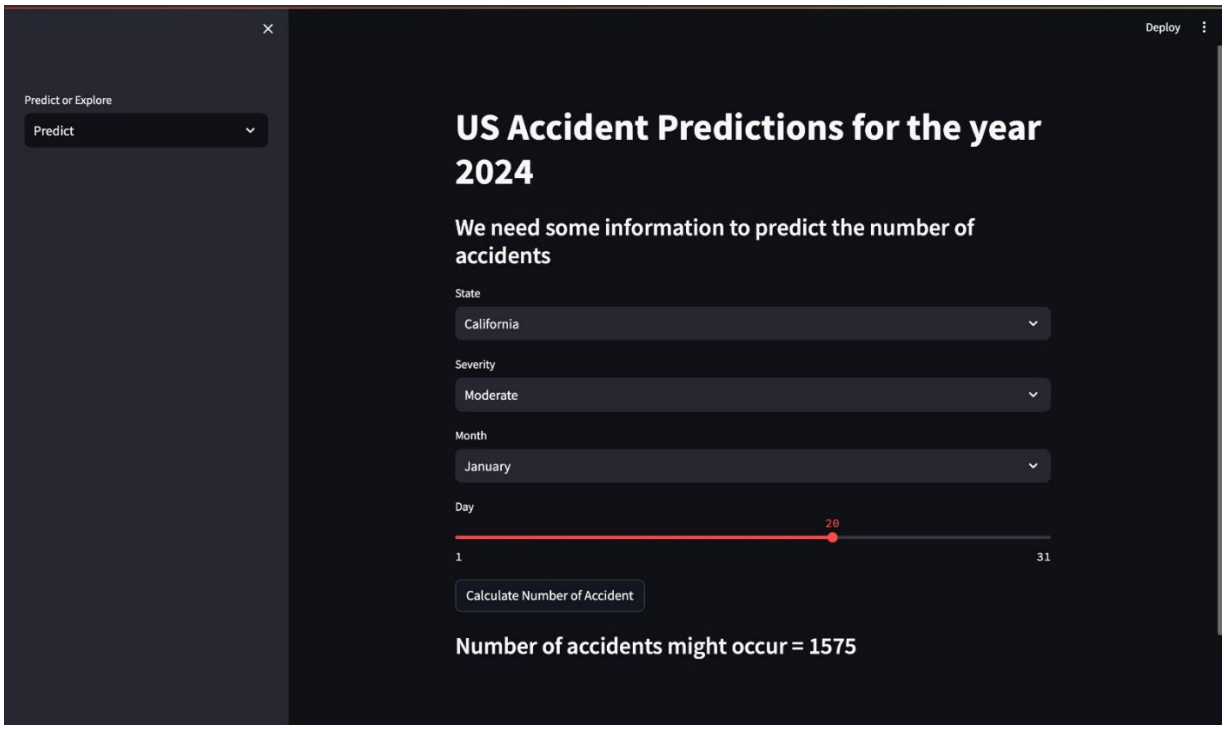
3. A month of the year is selected.



4. A date is selected.



5. Based on the above inputs given, a prediction will be generated for the number of accidents that are most likely to occur on the input date of the input month with the selected severity in the selected state in the year 2024.



Solution to the problem statement:

From this prediction model paramedical team, law enforcement and road safety departments can identify the number of accidents that are likely to occur in the state on any given day and make necessary accommodations for improving road safety.

Possible avenues for improvement:

Due to a lack of hardware components, the model is predicting the number of accidents for a state. With enough resources this model can be extended to predict the number of accidents that are likely to happen per street in a city of a county in any given state for any date, month and year.