Title: **Road Closure Severity Prediction**

Member 1	Rajat Belgundi	rajatrb@vt.edu
Member2	Yash Kulkarni	yashpandharish@vt.edu

Introduction:

Maintenance of roads and construction of new roads or pavements is very important to improve the day-to-day life of the people. These events many times require roads to be closed temporarily or traffic to be diverted onto alternate routes. Due to a variety of reasons, these road closures get prolonged and become an inconvenience for the general public. Every department of transportation in the country always aims to reduce the impact of such road closures to ensure a smooth daily travel routine of the travelers.
The reasons for prolonged road closures could be many. To name a few, environmental reasons - wind, precipitation cold, time of the day, presence of traffic signals or crosswalks, airport or train station in the vicinity and many more. Regardless of the reason, finding ways to speed up the road work, reduce traffic bottlenecks to alleviate the inconvenience caused by the event is the ultimate goal.
We aim to solve this problem by making use of machine learning based classification techniques on the US Road Construction and Closures dataset. Having a machine learning model to classify a particular closure event will help the transport department address the highest and least impacted routes effectively by taking prompt action. During the data analysis we also aim to derive insights from the data to further study the impact of various stimuli on the road closure event. The dataset has been provided by [1] in which the authors also run experiments on spatial data with deep learning.

Project Problem Statement:

Implementing machine learning classification techniques to predict the severity of the road closure event or construction work to assist the department of transportation in either speeding up the current work or offering alternate routes for diverting traffic on busy routes.

Data Set:

US Road Construction and Closures (2016 - 2021)

Available on Kaggle: https://www.kaggle.com/datasets/sobhanmoosavi/us-road-construction-and-closures

About the dataset:

Description

The following is a countrywide dataset of road construction and closure events, which covers 49 states of the US. Construction events in this dataset could be any roadwork, ranging from fixing pavements to substantial projects that could take months to finish. The data is collected from Jan 2016 to Dec 2021, 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. Currently, there are about 6.2 million construction and closure records in this dataset.

Columns: As observed in feature engineering step as we have a variety of features - categorical, datetime, numerical, boolean etc. After categorical encoding of the columns, we have a total of 215 features.

Rows: 6.2 M rows of which we are working with approximately 200,000 instances.

Preprocessing steps:

1. Exploratory Data Analysis
2. Data Cleaning and Preprocessing
3. Feature Engineering
4. Feature Scaling

Methods and Models:

1. Logistic Regression
2. Random Forest (Implemented)
3. Decision Tree
4. KNN Classification (Implemented)
5. Support Vector Machine

References:

[1] Karimi Monsefi, Amin, Sobhan Moosavi, and Rajiv Ramnath. “Will there be a construction? Predicting road constructions based on heterogeneous spatiotemporal data.”, In Proceedings of the 30th ACM SIGSPATIAL 2022.

Implementation:

Data Cleaning and Preprocessing

The dataset consists of data of roadwork closures of various states in the USA. Based on initial analysis of the data, we can see that the state with the highest number of roadwork closures or constructions in the USA is California(CA). Also based on some research we found that the state of California has the second largest road network. This project is aimed to help identify the roadworks of high severity and impact thus we decided to work with data pertaining to the state of California.

The data contains 556,830 instances with 215 columns pertaining to the California state.

To start with and work with a subset of data we have decided to proceed with data from last year i.e., from 2021 to 2022. In the data for this project we have 203,015 instances with 215 columns on which we would build and train two machine learning models namely- Random Forest Classifier & K-Nearest Neighbors Classifier.

We are dealing with features related to weather, geography, points of interest(amenities, station, crossing, traffic signal etc. which means the data types of the features will be different leading to high memory usage.

1. Memory Usage & Optimization:
    a. Handle object dtypes and set them to appropriate data types.
    b. We have many categorical features which are detected initially as dtype objects.
    c. We have two features showing timestamps of start and end time of roadwork which are detected as type objects.
    d. The memory usage is highest for dtype objects which is equivalent to storing a string.
    e. The original dataframe has a memory usage of 8.8GB which needs to be reduced for faster computation.
    f. On converting the dtype of features into their appropriate dtypes we get a memory usage of 2.5GB. (Convert object to category and also object to datetime)

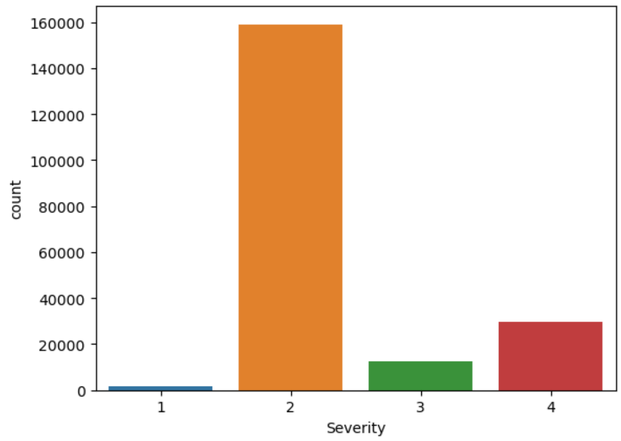
2. Handle Missing Values:
    a. As we have geographic data which has many NaN values, it would be counterproductive to impute data in such columns.
    b. With that point, we have other weather based features which might change with change in latitude and longitude.
    c. We also checked by setting a threshold for allowed non-NA values and found that there are still quite a few missing values.
    d. Thus, we decide to drop the rows containing NaN values.

3. Transforming the target feature:
    a. We have 4 class labels in the severity feature, but we want to be able to identify if the roadwork is severe or not.
    b. Therefore, we can proceed with combining the severity = 3 and 4 into one class and severity = 1 and 2 into another class.
    c. Thus, we have a binary classification problem to solve.

Exploratory Data Analysis (EDA):

1. We start by understanding our target feature which is ‘Severity’ consisting of 4 classes - 1, 2, 3, 4.
    a. Finding: We see that the data is imbalanced with class label 2 having the highest instances.
2. We can further deep dive into understanding which states have the most # of severe roadwork closures/construction.
    a. Finding: The state of California has the highest number of severe roadwork closures/constructions.
3. We now go ahead and deep dive into the California based dataset to see which city has the most # of severe roadwork constructions.
    a. Finding: Los Angeles(LA) has the highest number of severe roadwork closures/constructions.

4. To handle the imbalance in the Severity feature we club classes 3 and 4 into one class with label = 1 (indicating severe) and classes 1 and 2 into one class with label = 0 (indicating not severe)



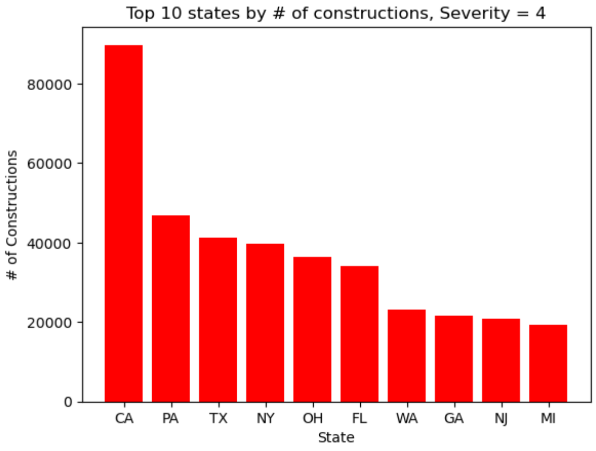


Fig A: Check balance of data

Fig B: Top 10 states by most severe roadwork

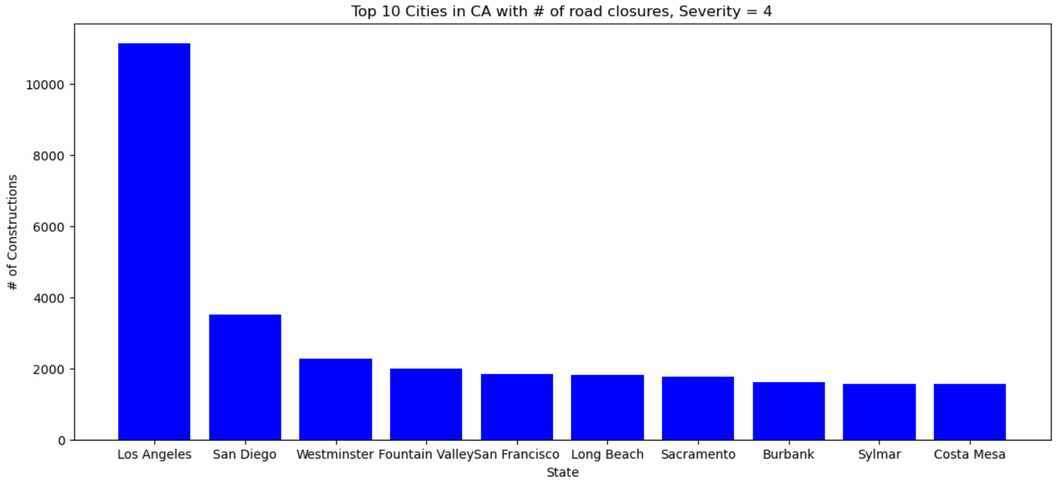


Fig C: Top 10 cities in California by most severe roadwork

Feature Selection:

For features like latitude and longitude we have decided to use latitude only for now, in the final run we will include longitude of the road work as well.

Methodology:

From the given pool of features, we started with Correlation Analysis for feature selection. As this is a classification use case we need to check the correlation of features with the target feature. We then make

a list of features which we want to include and drop the other features. Based on this list of features we now make use of Random Forest Classifier based feature selection to select the features with highest importance. We experiment with various thresholds of feature importance to select the best suited pool of features. We use the threshold based subset of features and apply a Random Forest Classifier with 200 trees/learners to evaluate the results. The metrics used for comparison are Precision, F1-score, and AUC of ROC curve. The metrics recorded are for class label = 1 indicating severe roadwork/closure.
Note: Our work is focused on classifying the most severe roadworks and closures to help draw the attention of state transport department of California to pay attention to.

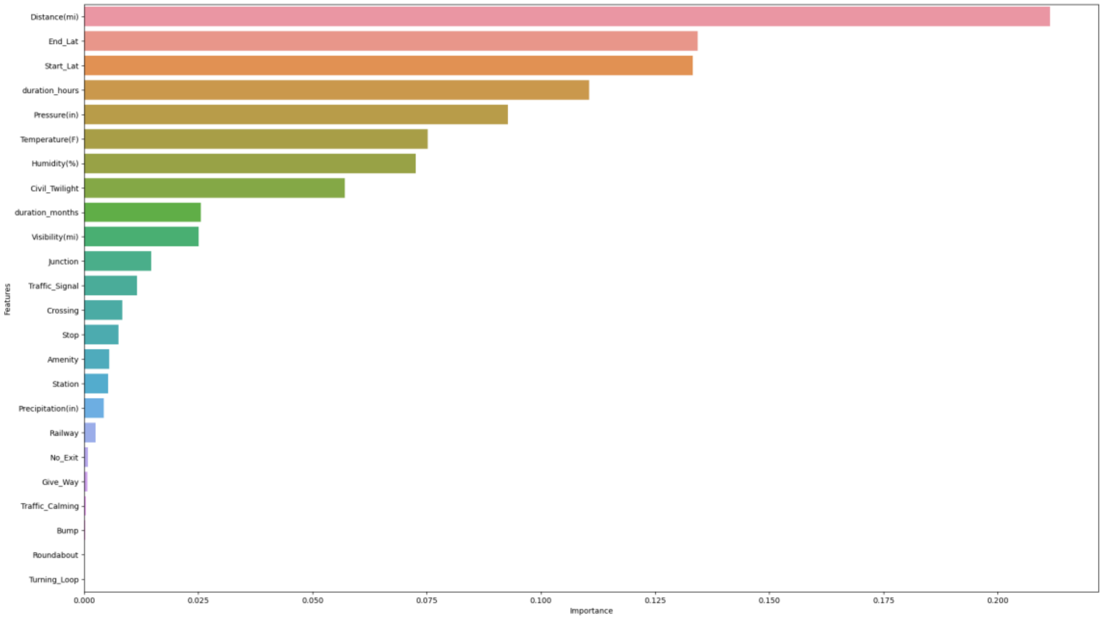
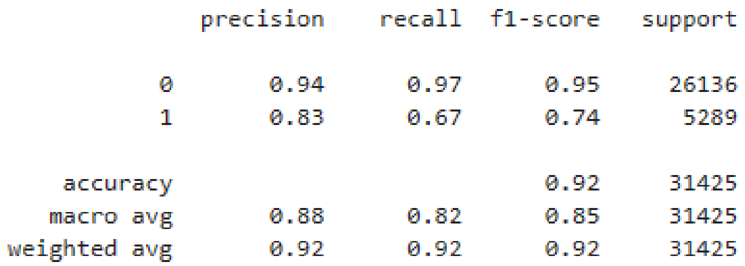


Fig C: Random Forest Based Feature Selection

Model Building and Evaluation (Baseline):

Experiment Data:
As our target feature is imbalance, we make use of ‘stratify’ parameters while splitting the data into train and test sets to ensure that the proportion of class labels (0 & 1) is maintained.
Before arriving at the current subset of features, we evaluated the effect of including and excluding the features like ‘Civil_Twilight’, 'End_Lat', 'Start_Lat', 'duration_hours', 'Pressure(in)'. Inclusion of the mentioned features led to tremendous rise in model performance, the model was able to classify the instances better. To quantify the results, we can take a look at the classification report of inclusion and exclusion of the above mentioned features with Random Forest Classifier.

A) Including these features we get:

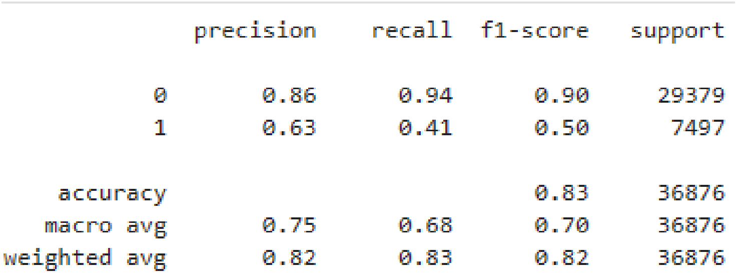


```
              precision    recall  f1-score   support

           0       0.94      0.97      0.95     26136
           1       0.83      0.67      0.74      5289

    accuracy                           0.92     31425
   macro avg       0.88      0.82      0.85     31425
weighted avg       0.92      0.92      0.92     31425
```

B) Excluding the features we get:



```
              precision    recall  f1-score   support

           0       0.86      0.94      0.90     29379
           1       0.63      0.41      0.50      7497

    accuracy                           0.83     36876
   macro avg       0.75      0.68      0.70     36876
weighted avg       0.82      0.83      0.82     36876
```

Thus, we can see that the features [‘Civil_Twilight’, 'End_Lat', 'Start_Lat', 'duration_hours', 'Pressure(in)'] are important in predicting the severity of the roadwork closure.

1. Threshold = 0.001
   Features = ['Distance(mi)', 'End_Lat', 'Start_Lat', 'duration_hours', 'Pressure(in)', 'Temperature(F)', 'Humidity(%)', 'Civil_Twilight', 'duration_months', 'Visibility(mi)', 'Junction', 'Traffic_Signal', 'Crossing', 'Stop', 'Amenity', 'Station', 'Precipitation(in)', 'Railway']
2. Threshold = 0.003
   Features = ['Distance(mi)', 'End_Lat', 'Start_Lat', 'duration_hours', 'Pressure(in)', 'Temperature(F)', 'Humidity(%)', 'Civil_Twilight', 'duration_months', 'Visibility(mi)', 'Junction', 'Traffic_Signal', 'Crossing', 'Stop', 'Amenity', 'Station', 'Precipitation(in)'] (‘Railway’ dropped)
3. Threshold = 0.005
   Features = ['Distance(mi)', 'End_Lat', 'Start_Lat', 'duration_hours', 'Pressure(in)', 'Temperature(F)', 'Humidity(%)', 'Civil_Twilight', 'duration_months', 'Visibility(mi)', 'Junction', 'Traffic_Signal', 'Crossing', 'Stop', 'Amenity', 'Station'] (‘Precipitation(in)’ dropped)
4. Threshold = 0.007
   Features = ['Distance(mi)', 'End_Lat', 'Start_Lat', 'duration_hours', 'Pressure(in)', 'Temperature(F)', 'Humidity(%)', 'Civil_Twilight', 'duration_months', 'Visibility(mi)', 'Junction', 'Traffic_Signal', 'Crossing', 'Stop'] ('Amenity', 'Station' dropped)
5. Threshold = 0.009
   Features = ['Distance(mi)', 'End_Lat', 'Start_Lat', 'duration_hours', 'Pressure(in)', 'Temperature(F)', 'Humidity(%)', 'Civil_Twilight', 'duration_months', 'Visibility(mi)', 'Junction', 'Traffic_Signal'] (‘Crossing, ‘Stop’ dropped)

Classification Models:

1. **Random Forest Classifier:**
    a. A Random Forest Classifier is an ensemble machine learning model that combines multiple decision trees to make predictions.
    b. In a Random Forest, each tree in the ensemble is trained on a random subset of the data and makes an independent prediction.
    c. The final prediction is determined by a majority vote (classification) or averaging (regression) of the predictions from individual trees, which often results in a more robust and accurate model compared to a single decision tree.
    d. Random Forests are widely used for classification and regression tasks in various domains due to their ability to handle complex relationships and reduce overfitting.
2. **KNN Classifier:**
    a. The k-Nearest Neighbors (KNN) classifier is a type of instance-based or lazy learning algorithm used for both classification and regression tasks.
    b. In KNN, the prediction for a new data point is based on the majority class (for classification) or the average of the neighboring data points' values (for regression) among its k nearest neighbors in the feature space.
    c. The choice of k, the number of neighbors, is a key parameter that influences the model's performance and sensitivity to noise in the data.
    d. KNN is a simple and intuitive algorithm but can be computationally expensive for large datasets.
3. **Decision Tree Classifier:**
    a. A Decision Tree classifier is a machine learning model that makes predictions by recursively partitioning the input space into regions and assigning a class label to each region.
    b. The decision-making process involves selecting the best feature to split the data at each node based on criteria such as Gini impurity or information gain.
    c. Decision Trees are interpretable, easy to understand, and widely used for classification tasks, providing a clear representation of the decision-making process in the form of a tree structure.
    d. However, they are prone to overfitting, and techniques like pruning are often employed to address this issue.
4. **Logistic Regression Classifier:**
    a. A logistic regression classifier is a machine learning model that falls under the category of generalized linear models and is widely used for binary classification tasks.
    b. It predicts the probability that a given input belongs to a particular class, and then, based on a specified threshold, assigns the input to one of the two classes.
    c. The logistic function (sigmoid function) is used to model the relationship between the independent variables and the log-odds of the probability of the positive class.
    d. Logistic regression is commonly employed when the outcome variable is binary (0 or 1) and is popular for its simplicity, interpretability, and efficiency in certain scenarios.

Random Forest Classifier Results

Note: All results are obtained when n_estimators are set equal to 200.

Threshold	Precision	Recall	F1 score	AUC of ROC	Accuracy
0.001	0.84	0.67	0.75	0.960	0.92
0.003	0.84	0.67	0.74	0.961	0.92
0.005	0.83	0.67	0.74	0.961	0.92
0.007	0.83	0.67	0.74	0.961	0.92
0.009	0.82	0.66	0.73	0.960	0.92

KNeighborsClassifier Results

Note: All results are when n_neighbors are set equal to 7

Threshold	Precision	Recall	F1 Score	AUC of ROC	Accuracy
0.001	0.69	0.52	0.59	0.885	0.88
0.003	0.69	0.52	0.59	0.885	0.88
0.005	0.69	0.52	0.59	0.885	0.88
0.009	0.69	0.52	0.59	0.885	0.88

Logistic Regression Classifier Results

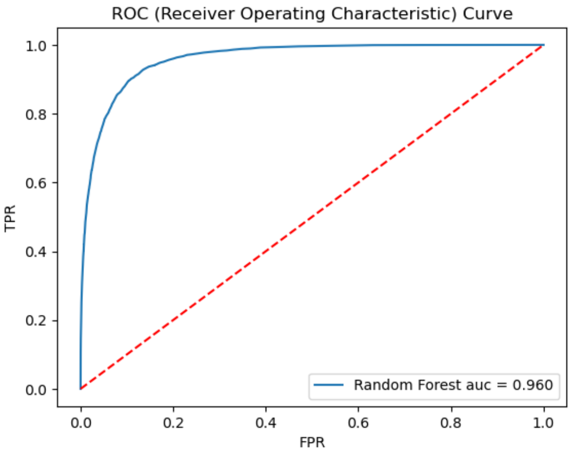
Precision	Recall	F1 Score	AUC of ROC	Accuracy
0.67	0.19	0.30	0.811	0.85

Decision Tree Classifier Results

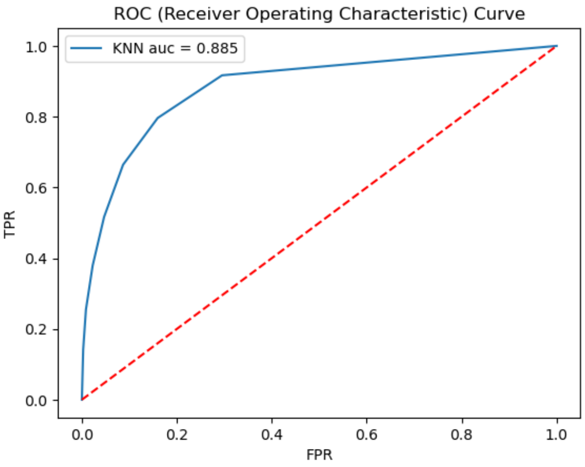
Precision	Recall	F1 Score	AUC of ROC	Accuracy
0.65	0.65	0.65	0.792	0.88

ROC curve:

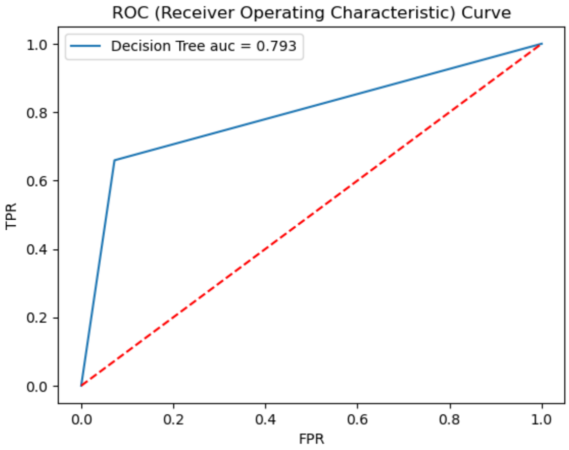
1. Random Forest Classifier



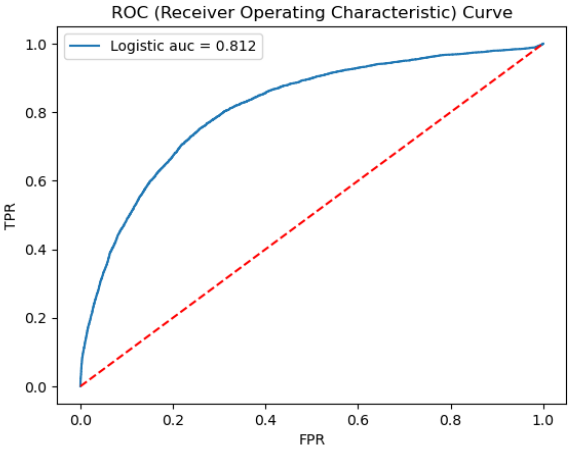
2. K Nearest Neighbors Classifier



3. Decision Tree



4. Logistic Regression



Observations:

1. We see that for our subset of data the subset of features chosen over the experiments seem to show similar results.
2. Our focus is on predicting the severity = 1 class of roadwork closures.
3. The Random Forest algorithm is able to learn better and classify the roadwork closures as it is seen from above results of various metrics like F1, Precision, Recall, AUC of ROC curve.
4. The accuracy of the KNeighborsClassifier is 0.88. The accuracy of the Random Forest Classifier is 0.92. Thus, we can see that the Random Forest performs much better than the KNeighborsClassifier.
5. The accuracy of the Logistic Regression classifier is 0.85. Its precision is 0.67 and Recall is 0.19. The F1 Score is 0.30 and AUC of ROC is 0.811.

6. The accuracy of the Decision Tree classifier is 0.88. Its precision is 0.65 and Recall is 0.65. The F1 Score is 0.65 and AUC of ROC is 0.792.

Conclusion:

In this project we solve a classification problem of machine learning. The problem being solved is extremely relevant and important as it predicts which road construction work or closures have the highest severity. This will assist the government transportation department to focus on speeding up the most severe roadworks. This is helpful in making life easy for the general public. The road maintenance works can be speeded up or decisions can be made to divert the traffic near the roadworks with highest severity. The classification models also take into consideration the various amenities like airports, train stations, hospitals etc in the vicinity of the roadwork or closure.

Based on our observations, we can conclude that random forest classifier performs the best amongst the four classifiers. The random forest classifier performs well on all the evaluation metrics for the classification.

Key Takeaways:

1. Data Preprocessing
2. Exploratory Data Analysis
3. Feature Selection
4. Experimentation for various classifiers
5. Applying classification algorithms to solve the problem
6. Evaluation of classifiers on various metrics

GitHub Link to the Project:

https://github.com/Rajat2312/Term_Project_CS5644

Rajat Belgundi	i) Data Cleaning & Preprocessing:. ii) EDA iii) Feature Selection: Correlation & Random Forest based Feature Selection iv) Implementation of Random Forest,KNN,Logistic,Decision Tree v) Final Report: Content writing
Yash Kulkarni	i) Data Analysis ii) EDA iii) In-depth description of all the features of the dataset in the features_info.txt file. iv) Implementation of Random Forest,KNN,Logistic,Decision Tree. v) Final Report: Content writing