***Chicago Traffic Crash Analysis and Injury Prediction - Report***

**Project Description:**

In the City of Chicago, traffic crashes pose a significant challenge to public safety, efficient traffic management, and the overall well-being of the city’s residents and visitors. Despite the availability of extensive traffic crash data, there is a lack of effective tools to analyze, understand, and predict these incidents. This gap hinders the ability of city authorities, policymakers, and traffic management teams to make informed decisions, develop proactive safety measures, and implement strategic interventions.

Dataset link: <https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if>

**Data Cleaning:**

Objective: Prepare the dataset for analysis by addressing data quality issues.

Key Steps:

1.Identification and handling of missing values.

2.Removal of duplicate records and irrelevant features.   
3.Transformation and normalization of key variables.   
4.Feature engineering to enhance model input.

**Feature Engineering:**

* categorical variables have been one-hot encoded and assembled into a single feature vector.
* The features column shows the Sparse Vector representation of the encoded features, features = 'POSTED\_SPEED\_LIMIT', 'WEATHER\_CONDITION', 'LIGHTING\_CONDITION', 'TRAFFIC\_CONTROL\_DEVICE', 'DEVICE\_CONDITION','ROADWAY\_SURFACE\_COND','ROAD\_DEFECT','CRASH\_TYPE’
* The label column contains numerical indices representing the target variable.

label = “INJURY\_SEVERITY”

**Model Building:**

Since this is a classification problem, we have used various classification algorithms which are:

* Random Forest Classifier
* Logistic Regression Classifier
* Gradient Boosting Classifier
* Linear SVC classifier
* Decision Tree Classifier

**Target** : ‘INJURY\_SEVERITY’

**Features** : 'POSTED\_SPEED\_LIMIT', 'WEATHER\_CONDITION','LIGHTING\_CONDITION', 'TRAFFIC\_CONTROL\_DEVICE','DEVICE\_CONDITION','ROADWAY\_SURFACE\_COND','ROAD\_DEFECT','CRASH\_TYPE’

**Model Evaluation:**

To evaluate the performance of the predictive models, various evaluation metrics were employed, including accuracy, precision, recall, and F1-score. These metrics provide insights into how well the models are performing in terms of correctly predicting the Injury Severity.

**Results**:

After evaluating the models, the following results were obtained:

- Logistic Regression:

LR Accuracy = 0.8633510102791333

LR Precision = 0.8109876681324307

LR Recall = 0.8633510102791332

LR F1-score = 0.8235537512996971

- Support Vector Classifier:

SVC Accuracy = 0.8875451275112699

SVC Precision = 0.8866783678936732

SVC Recall = 0.8875451275112698

SVC F1 Score = 0.8871031098846179

- Gradient Boosting Classifier:

GBT Accuracy = 0.8895842818393429

GBT Precision = 0.9038458245421899

GBT Recall = 0.8895842818393428

GBT F1 Score = 0.8950201892133081

- Random Forest Classifier:

RF Accuracy = 0.8635177335889758

RF Precision = 0.7456628762226413

RF Recall = 0.8635177335889758

RF F1-score = 0.8002745160750989

- Linear SVC Classifier:

SVC Accuracy = 0.8875451275112699

SVC Precision = 0.8866783678936732

SVC Recall = 0.8875451275112698

SVC F1 Score = 0.8871031098846179

Comparing the models based on these metrics, we can observe the following:

Accuracy: The Gradient Boosting Classifier (GBT) has the highest accuracy (0.8896), indicating it correctly classifies the highest percentage of instances among the models tested.

Precision and Recall: GBT also leads in precision (0.9038), suggesting it has the lowest false positive rate, and has a high recall (0.8896), indicating it correctly identifies a high percentage of actual positives.

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balance between the two. GBT again has the highest F1 score (0.8950), suggesting it has the most balanced performance between precision and recall among the models evaluated.

**Conclusion**:

In this project, we have analyzed traffic crash data and employed machine learning techniques to predict the primary type of traffic incidents. We have extracted valuable insights that can aid traffic safety authorities and law enforcement in devising targeted interventions and policies. Ultimately, this project highlights the capacity of machine learning to assist in the analysis and prediction of traffic crashes injuries severity, contributing to enhanced road safety measures.