BVRIT HYDERABAD

College of Engineering for Women

CRIME CLASSIFICATION

TEAM - 8

TEAM MEMBERS:

20wh1a1238 , P.Anusha
20wh1a1239 , S.Swathi
20wh1a1240 , A.Shivani
20wh1a1241 , N.Saivarshini
20wh1a1242 , U.Sree Lakshmi.K

AGENDA

- Problem statement
- Python Packages used
- Algorithm
- Output
- Comparison table
- Execute the Code

Problem Statement

- Dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, you must predict the category of crime that occurred.

  Dataset Description:

  This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set.

  Data Fields :

- Dates - timestamp of the crime incident
- Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.

Problem Statement

- Descript - detailed description of the crime incident (only in train.csv)
- DayOfWeek - the day of the week
- PdDistrict - name of the Police Department District
- Resolution - how the crime incident was resolved (only in train.csv)
- Address - the approximate street address of the crime incident
- X - Longitude
- Y - Latitude

.

Python Packages used

- pandas
- numpy
- matplotlib.pyplot
- seaborn
- sklearn

Algorithm

- Random Forest
- Decision Tree
- XGBClassifier
- K-Nearest Neighbour

Random Forest

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.
- It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Why use Random Forest?

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Decision Tree

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

  Why use Decision Trees?
- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

XGBClassifier

- XGBoost classifier is a Machine learning algorithm that is applied for structured and tabular data.
- XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.
- XGBoost works with large, complicated datasets. XGBoost is an ensemble modelling technique.

  Why use XGBClassifier
- Performance: XGBClassifier has a strong track record of producing high-quality results in various machine learning tasks.
- Scalability: XGBClassifier is designed for efficient and scalable training of machine learning models, making it suitable for large datasets.

- Customizability: XGBClassifier has a wide range of hyperparameters that can be adjusted to optimize performance, making it highly customizable.
- Handling of Missing Values: XGBClassifier has built-in support for handling missing values, making it easy to work with real-world data that often has missing values.
- Interpretability: Unlike some machine learning algorithms that can be difficult to interpret, XGBClassifier provides feature importances, allowing for a better understanding of which variables are most important in making predictions.

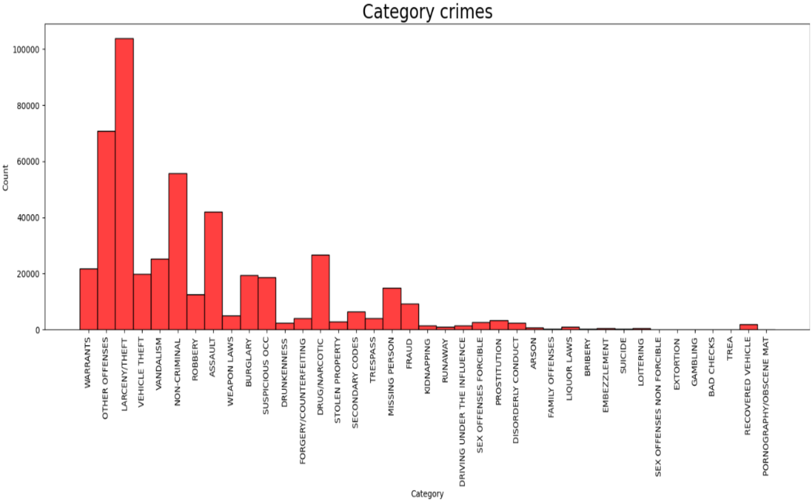
K-Nearest Neighbour

- KNN is a simple, supervised machine learning (ML) algorithm that can be used for classification or regression tasks - and is also frequently used in missing value imputation.
- It is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points.
- By choosing K, the user can select the number of nearby observations to use in the algorithm.

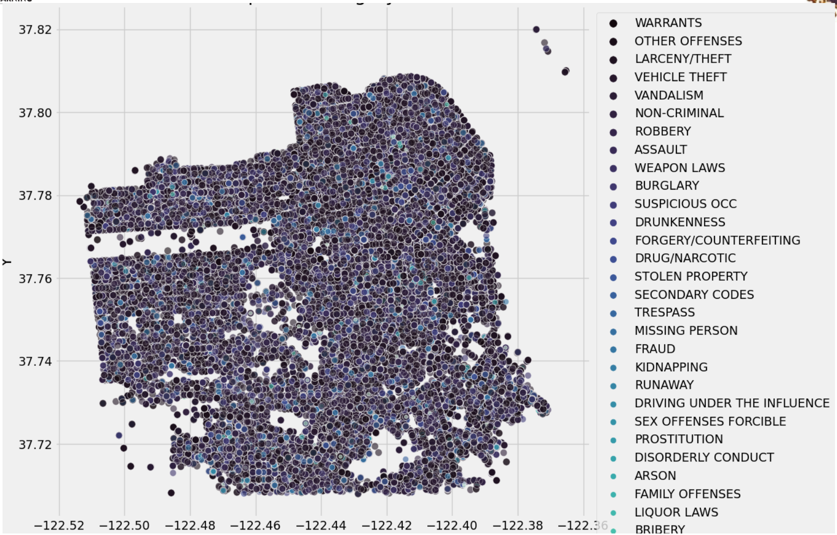
Why use KNN Algorithm

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

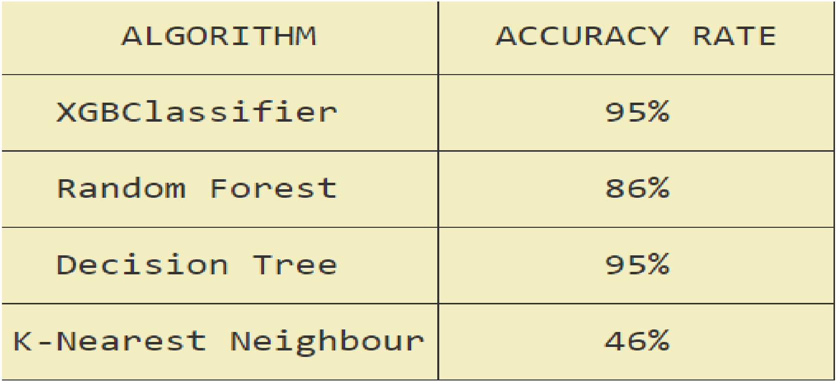
Visualization



Scatter Plot

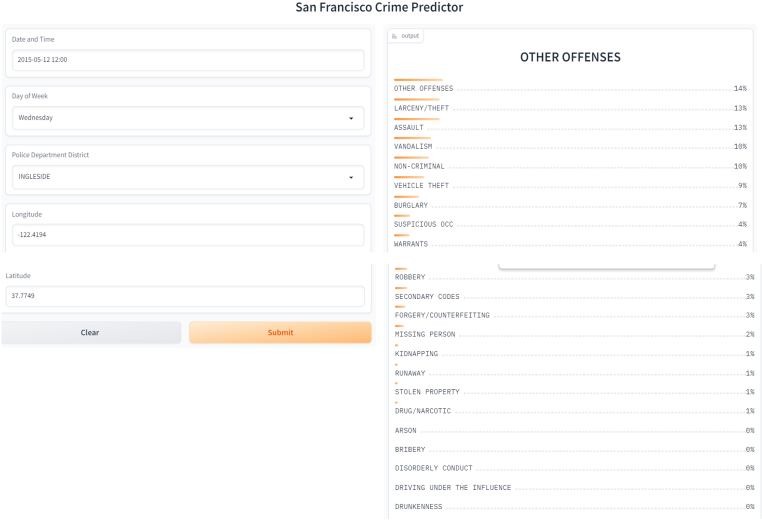


Comparison Table



ALGORITHM	ACCURACY RATE
XGBClassifier	95%
Random Forest	86%
Decision Tree	95%
K-Nearest Neighbour	46%

Output



San Francisco Crime Predictor

Date and Time

2015-05-12 12:00

Day of Week

Wednesday

Police Department District

INGLESIDE

Longitude

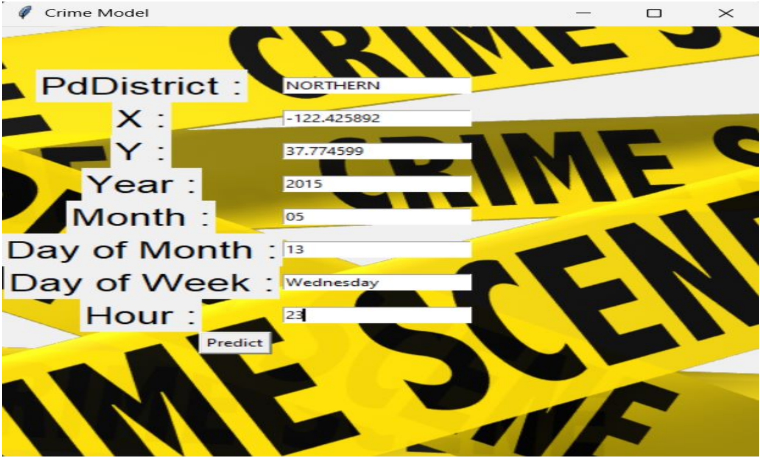
-122.4194

Latitude

37.7749

OTHER OFFENSES

```
OTHER OFFENSES ........................ 14%
LARCENY/THEFT ......................... 13%
ASSAULT ............................... 13%
VANDALISM ............................. 10%
NON-CRIMINAL .......................... 10%
VEHICLE THEFT .......................... 9%
BURGLARY ............................... 7%
SUSPICIOUS OCC ......................... 4%
WARRANTS ............................... 4%
ROBBERY ................................ 3%
SECONDARY CODES ........................ 3%
FORGERY/COUNTERFEITING ................. 3%
MISSING PERSON ......................... 2%
KIDNAPPING ............................. 1%
RUNAWAY ................................ 1%
STOLEN PROPERTY ........................ 1%
DRUG/NARCOTIC .......................... 1%
ARSON .................................. 0%
BRIBERY ................................ 0%
DISORDERLY CONDUCT ..................... 0%
DRIVING UNDER THE INFLUENCE ............ 0%
DRUNKENNESS ............................ 0%
```



PdDistrict : NORTHERN

X : -122.425892

Y : 37.774599

Year : 2015

Month : 05

Day of Month : 13

Day of Week : Wednesday

Hour : 23

THANK YOU