Project title chosen:

Topic: Spark Streaming with Machine Learning.

Dataset: San Francisco Crime Classification

Problem Statement: Classify Given Data based on categories

Design details

For the implementation of the project the column chosen as determinant variable is description and determined variable is Category variable.

The determinant variable is pre-processed into usable values and is then used in the machine learning models. Then to optimize the models, hyperparameter tuning and data visualization was applied, and the optimized models were obtained.

It was then tested with the models and batch size. The best model is observed to be Logistic Regression.

Then the metrics, graphs, etc. were plotted.

Surface level implementation details about each unit

There are 5 stages that is present in the project. They are:

1. Streaming of Data: The data is initially converted from csv file to json file where the data has been cleaned. The data converted from a json format to a DataFrame format.
2. Pre-Processing of Data: With the Help of Kbest and chi-squared method we were able to obtain the most useful features and dropped those columns which were not needed.
3. Data Visualization
4. Machine Learning: Three models have been tried for classifying the data: Logistic Regression, Random Forest and Naïve Bayes. All models have been hyperparametered tuned and also the relevant graphs have been printed.
5. Clustering: Alternate method in classifying data

Reason behind design decisions

The reason of choosing the description column is because description is found to be able to classify the categories most efficiently. For the Logistic Regression, the max iterations and parameters were optimized for time efficiency improvement. For the Naïve bayes, the smoothing was optimized for time efficiency improvement. For Random Forests, the depth of trees, Number of iterations were optimized for time efficiency improvement.

Takeaway from the project

The project helped us to improve our coding skill. It also helps us to learn implementing machine learning algorithms like naive bayes, logistic regression and random forest. We learnt how to handle large amounts of streaming data using spark streaming.

It also helped us to learn data visualization by doing the plots.