IDP Humanitarian Needs Analysis

This study selects data analysis approach with Machine Learning Model learning algorithms with Linear Regression, Support Vector Machine. A model is built for each algorithm and is experimented using a IDP Dataset. In order to compare the performance of both algorithms in respect with their capabilities in predicting the Severities we evaluated their results by graphs and performance rate. Prior to experimentations several steps are conducted, including data preparation, data exploration and model building. These steps are important for data to be ready and could be used as input for the models. Data preparation includes data cleansing that manages error found from the data entry, practically impossible values, missing values, outliers, spaces, typos or error against codebook; data transformation that threats aggregating data, extrapolating data, creating dummies and reducing number of variables; combining data that processes the merging or joining datasets. Following are the steps of data analysis process with Machine Learning Model learning algorithms applied for IDP Severity Prediction occurrence and further shown in Figure 1:

1. Data Analysis
2. Modelling
3. Cross Voting
4. Genetics Optimization
5. Prediction & Decision making

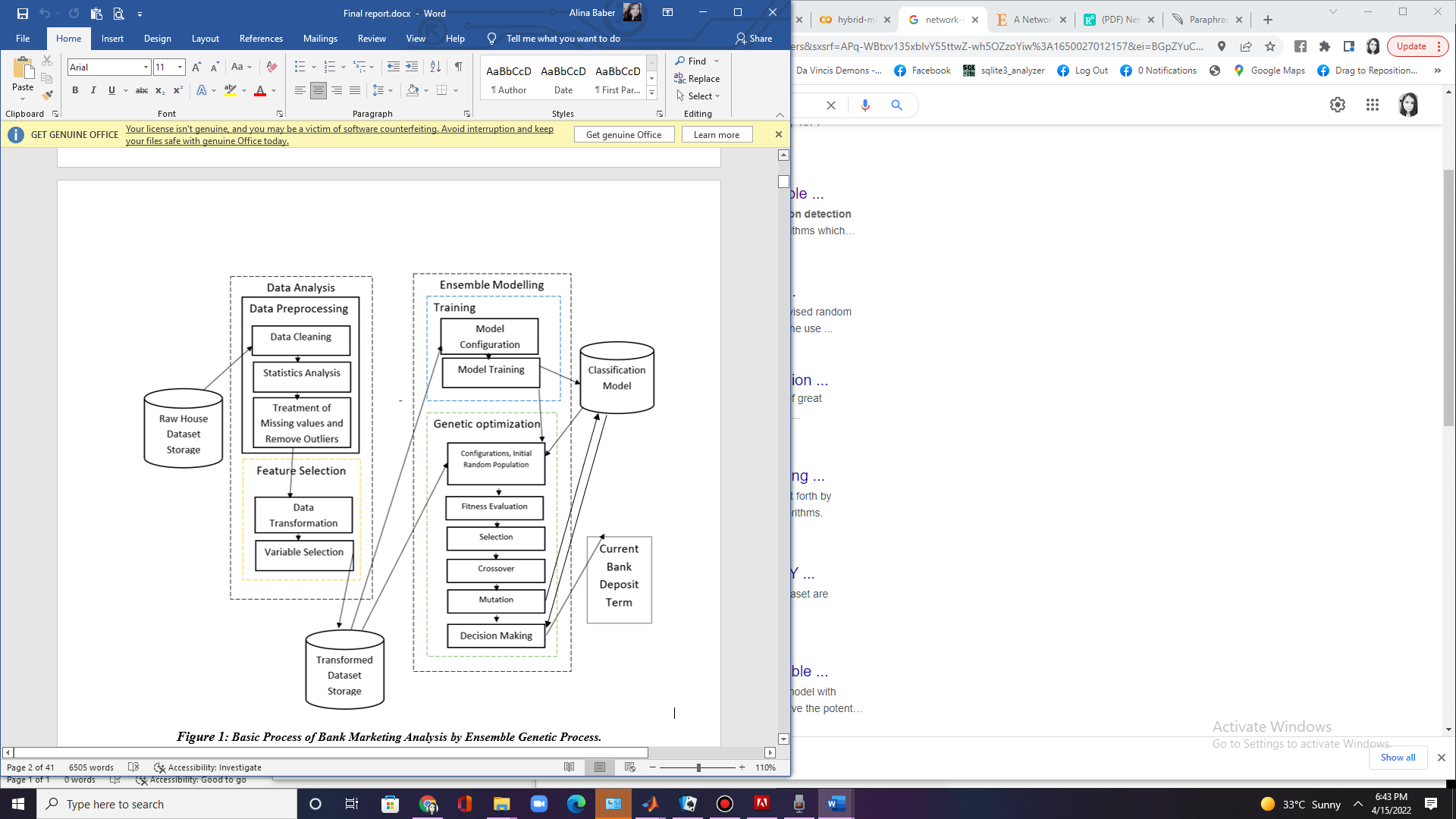
-

Figure Basic Process of Data Analysis with ML.

## Data Outlier Cleaning and Statistical Analysis

After data acquisition Data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate or non-relevant parts of the data and then restoring, remodeling, or removing the dirty or crude data. Data cleaning may be performed as batch processing through scripting or interactively with data wrangling tools.

After cleaning, a dataset should be uniform with other related datasets in the operation. The discrepancies identified or eliminated may have been basically caused by user entry mistakes, by corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores.

## Treatment of Missing Values

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

There are some reasons for the observations missing in a set of data. Such missing values, although aren’t necessarily wrong but are likely to complicate many methods of analysis. Missing data and outliers are potentially associated for contributing to high noise and disturbance leading to unreliable prediction in the model. Exploring and finding the missing values and outliers become such an important step that must be conducted before inputting data to a model, otherwise they can gravely impact the data modelling.

## Data Transformation and Variable Selection

Data Feature transformation is a group of methods that create new features (predictor variables). Feature selection is a subset of feature transformation it is done for Knowledge discovery, Interpretability, to gain some insights and Curse of dimensionality, there are two ways of feature selection Filter type methods select variables regardless of the model. They are based only on general features like the correlation with the variable to predict. Filter methods suppress the least interesting variables. They are mainly used as a pre-process method. Another one Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables.

## Modelling & Prediction

Machine learning classifiers used to produce better predictive performance than a single machine learning classifier. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model. When we \*try to predict the target variable using any machine learning technique, the main causes of difference in actual and predicted values are noise, variance, and bias. ML helps to reduce these factors (except noise, which is irreducible error).

Following are models which we used.

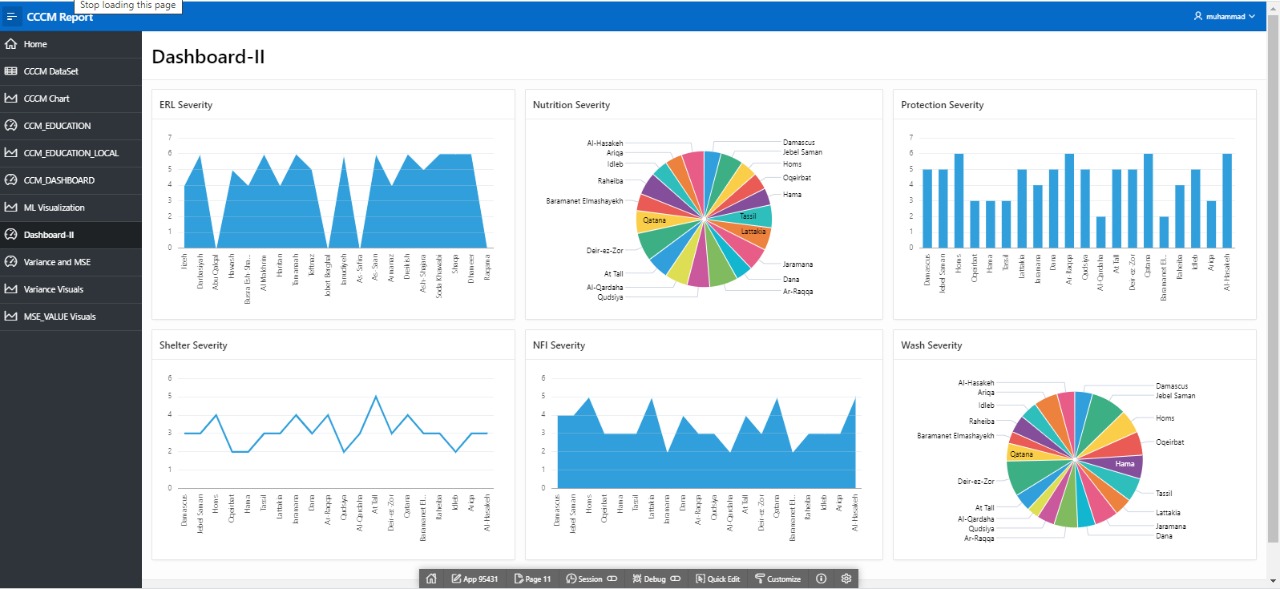
1. [Linear regression model](https://colab.research.google.com/drive/1QGmjgWKsUg7CPohC6cjUaxUT7OI2PF_2?authuser=3#scrollTo=Linear_regression_model)
2. [SVM model](https://colab.research.google.com/drive/1QGmjgWKsUg7CPohC6cjUaxUT7OI2PF_2?authuser=3#scrollTo=SVM_model)
3. [Random forest model](https://colab.research.google.com/drive/1QGmjgWKsUg7CPohC6cjUaxUT7OI2PF_2?authuser=3#scrollTo=Random_forest_model)

# Result

## Statistical Analysis

Following are the results examples of statistical analysis how much severity per city was predicted





## ML Performance Comparison

Following are the Results of ML comparison between Linear Regression , Support Vector Machin and Random Forest: from the comparison we observed that Random Forest have much better accuracy than Linear Regression and Support Vector Machine in the case of IDP severities Analysis.

