Project Proposal

Team 10

|  |  |  |
| --- | --- | --- |
| Name | Sec | B.N |
|  | 1 |  |
|  | 1 |  |
|  | 1 |  |
| Zeinab Moawad Fayez | 1 | 28 |

## Credit score classification.

<https://www.kaggle.com/datasets/parisrohan/credit-score-classification?select=test.csv>

28 Feature

100K+50K

O/P Credit Score [ Categorial Classification]

Since Task is classification most probably we will choose KNN as the algorithm to be implemented using map reduce

## Healthcare Analytics II

<https://datahack.analyticsvidhya.com/contest/janatahack-healthcare-analytics-ii/True/#ProblemStatement\>

#features 16

300K

o/p length stay Categorical Classification 11 Class

Since Task is classification most probably we will choose KNN as the algorithm to be implemented using map reduce

## Vehicle Sales Data Sets

<https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data>

# Features 16

O/P Selling Price [Regression]

550K

1. EDA (exploratory data analysis) Phase:
2. Carry out statistics on the data set like computing mean std …… Plotting Distributions [Data Visualization]
3. Data Cleaning and Handling missing values like using men and std.
4. Checking correlations between features [**Correlation Analysis]**
5. **Reduction of feature space using techniques like PCA**

The algorithms that we think if implementing using Map Reduce Algorithm

1. Clustering Techniques:
   1. Case Classification 🡪 KNN
   2. Case Regression 🡪 KMeans
2. Random Forest for both regression and categorical 🡪 multiple decision trees and combines their predictions through
   1. Case Classification 🡪 voting.
   2. Case Regression 🡪 Aggregate these predictions (e.g., by averaging) to obtain the final regression prediction.
3. **Naive Bayes**: For Categorial Classification Problems
4. **Logistic Regression: For Classification & Regression Problem**

The descriptive analysis methods to be used:

1. Clustering Using Kmeans

The predictive analysis methods: