CMP4011 Big Data and Cloud Computing

Project Report

Team 10

Name	Sec	B.N	Code
Ahmed Osama Helmy	1	5	9213061
Omar Mahmoud	1	4	9202119
Abdallah Ahmed	1	16	9202381
Aliaa Gheis	1	28	9202611

Problem Statement:

Road safety is a critical concern, and understanding accident patterns can help cities improve traffic management and reduce accident rates. This project aims to analyze accident data to identify high-risk locations, contributing factors, and potential mitigation strategies. By leveraging big data processing, we will extract valuable insights for transportation authorities and urban planners.

Dataset

Dataset Name: US Accidents (2016 - 2023)

Link: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Description:

- Contains 7.7 million accident records with 46 columns

Planned Approach or Proposed Solution:

1. **Exploratory Data Analysis (EDA):**

   o Analyze accident severity distribution and trends over time.

   o Identify correlations between weather conditions, traffic signals, and accident severity.

2. **Descriptive Analysis:**

   o Implement K-Means clustering with MapReduce to identify accident hotspots

   o Apriori algorithm for association rule mining to discover patterns in accident contributing factors

3. **Predictive Analysis:**

   o Apply Classification models (Random Forest, SVM) to predict accident severity based on environmental and traffic factors.

   o Implement Regression models (Linear Regression, Gradient Boosting) to estimate accident frequency per location.

4. **Big Data Implementation:**

   o Use Python & Apache Spark.

   o Implement K-Nearest Neighbors (KNN) using MapReduce to classify accident severity based on past accident characteristics.

   o Deploy on Azure (Maybe Azure HDInsight) for real-time accident risk analysis.