

CMP4011 Big Data and Cloud Computing

Project Report

Team 10

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Sec** | **B.N** | **Code** |
| Ahmed Osama Helmy | 1 | 5 | 9213061 |
| Omar Mahmoud | 1 | 29 | 9210758 |
| Abdallah Ahmed | 1 | 25 | 9210652 |
| Aliaa Gheis | 1 | 27 | 9210694 |

Contents

[Problem Statement: 3](#_Toc196517379)

[Dataset 3](#_Toc196517380)

[Project Pipeline: 4](#_Toc196517381)

[Data Ingestion 4](#_Toc196517382)

[Data Cleaning 4](#_Toc196517383)

## 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
* Main features of accident are (Severity of Accident, Time, Location, Weather, Road Characteristics)

## Project Pipeline:

### Data Ingestion

This stage involves loading the data and specifying the schema of the loaded data which is as follows:

|  |  |  |
| --- | --- | --- |
| Category | Fields | Data Types |
| Incident Info | ID, Source, Severity, Start\_Time, End\_Time, Description | String, Integer, Timestamp |
| Location | Start/End\_Lat/Lng, Distance(mi), Street, City, County, State, Zipcode, Country, Timezone, Airport\_Code | Double, String |
| Weather | Weather\_Timestamp, Temperature(F), Wind\_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind\_Direction, Wind\_Speed(mph), Precipitation(in), Weather\_Condition | Timestamp, Double, String |
| Road Features | Amenity, Bump, Crossing, [...] (14 boolean flags) | Boolean |
| Time of Day | Sunrise\_Sunset, Civil/Nautical/Astronomical\_Twilight | String |

### Data Cleaning

#### Handling Missing Values and Nulls