|  |
| --- |
| MAPPING ROAD MISHAPS TECHNICAL REPORT |

|  |
| --- |
| Distributed and Scalable Data Engineering (DSDE) ****DSCI 6007-04**** |

|  |  |
| --- | --- |
| SEMESTER : SPRING 2024 |  |



Contents

[Executive Summary 2](#_Toc96341550)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Technical Report 3](#_Toc96341551)

Abstract…………………………………………………..5

Review of available research……………….…..6

Methodology……………………………………….….7

[Results Section………………………………………..9](#_Toc96341556)

[Conclusion…………………………………………….10](#_Toc96341558)

[References…………………………………………….10](#_Toc96341559)

|  |
| --- |
| Mapping Road Mishaps |

|  |  |  |
| --- | --- | --- |
| Executive Summary: This dataset, which spans 49 US states, includes information on auto accidents nationwide. Multiple APIs that provide streaming traffic incident (or event) data were used to collect accident data from February 2016 to December 2021. The US and state departments of transportation, law enforcement organizations, traffic cameras, and traffic sensors inside the road networks are only a few of the organizations that provide traffic data that is broadcast via these APIs. This dataset contains approximately 2.8 million accident records as of right now. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Name 1: Sowjanya Kamineni**  **Name 2: Ramya Sree Manubramha**  **Name 3: Ganga Vamsik S** | **Questions?**  Contact : [skami9@unh.newhaven.edu](mailto:skami9@unh.newhaven.edu)  [rmanu4@unh.newhaven.edu](mailto:rmanu4@unh.newhaven.edu)  gsani1@unh.newhaven.edu |  |

Submitted on: 04-22-2024

|  |
| --- |
| Mapping Road Mishaps |

Technical Report:

The CRISP-DM technique is followed throughout the entire process, machine learning algorithm is made for relevant input data and applying the KNN model to obtain the desired output values. The foundation of the data science process is the six-phase cross-industry standard process for data mining.

1. BUSINESS UNDERSTANDING:

The project's requirement or need is the is to focus on determining the actual business objectives of a customer. Furthermore, from a data mining perspective which needs to be applied to determine a simulated output of the deliverable. Studying and being aware of the technologies and tools is necessary for the project's effective completion. As the project's solid foundation, this step needs to be completed with great concentration. In order to read the data (from well-known sources like Kaggle, Google, etc.), analyze it, interpret it, make predictions, visualize it, and produce the intended output, we determined that tools like the statistical technique of linear regression, the programming language Python, and visualization techniques are crucial for our project.

2. DATA UNDERSTANDING:

The first stage in any process is to understand the data, and then you may describe the data to identify any intriguing patterns. Data analysis tools such as Pandas, NumPy, and others should be used to examine the data that has been collected from diverse sources using a variety of approaches. Reading the collected data is the process of cleaning and analyzing it. The undesired data is cleansed and made exact for the data visualization throughout this phase. The data study is conducted on the collected data based on this analysis, and pertinent output or outcomes are established. It is important to fully comprehend the acquired data. Therefore, the data must be examined in greater detail in order to be understood, visualized, and, if feasible, to generate a question. Using a variety of methods, the dummy, null, and missing values are eliminated from the gathered data during this procedure. We also investigated whether there was any association between the independent variables during this procedure.

* **Eliminating the Missing Value:**

It's one of the simplest and fastest methods for handling missing values. No null values results in good visualization.

* **Eliminating an entire row:**   
  If a row include numerous missing data, you have the option to remove the entire row. Also we can eliminate entire row if it does not help in required insights.
* **Replacing with previous value –Forward fill:**

In certain circumstances, it makes more sense to provide the values with the previous value rather than the mean, mode, or median. Forward fill is the term for this. In time series data, it is mostly utilized.

* **Replacing with next value – Backward fill:**   
  The next value is used to impute the missing value in backward fill.

3. DATA PREPARATION:

Here, we prepare the data to obtain required insights from data.

4. MODELLING:

Training a machine learning algorithm to predict labels from features, fine-tuning it to meet business requirements, and verifying it on holdout data comprise the modeling process. Visualization and representation of the analyzed data may be necessary, depending on the requirements. For data visualization, we typically use packages such as sklearn, Matplotlib, and Bokeh. One method of transforming unprocessed data into a comprehensible format is visualization. This could be the data displayed using scatter plots, pie charts, bar graphs, and other formats. Through this procedure, the data are better understood and represented in a simpler format from which inferences can be made with ease.

5. EVALUATION:

Not using the full dataset for model training is the most crucial thing you can do to ensure a thorough evaluation of your model. Using 70% of the data for training and 30% for testing would be a common train/test split. To avoid the risk of overfitting to the training set, as we have already explained, it is crucial to use fresh data while assessing our model. Though there are situations when it's helpful to assess our model as we're developing it in order to determine the optimal parameters, we are unable to use the test set for this assessment; otherwise, we may choose the parameters that work best on the test data but not necessarily the parameters that generalize the best. True positives divided by true positives is precision. We generate a third subset of the data, referred to as the validation set, in order to assess the model while it is still being constructed and adjusted. Using 60% of the data for training, 20% for validation, and 20% for testing would be a typical split of the data for train, test, and validation.

The percentage of accurate predictions made using the test data is known as accuracy. By dividing the total number of guesses by the number of accurate predictions, it is simple to calculate.

Accuracy = correct predictions / all predictions accuracy

Precision is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.

Precision = true positives/true positives + false positives.

6. DEPLOYMENT:

Flask server is used for the deployment of web applications and APIs. The web application offers a user interface for interacting with the API, while the API itself contains the essential logic for prediction. The two programs are installed on the same server and communicate with one another using a secure protocol. The web application uses JSON format to transmit the essential data to the API in order to anticipate insurance. Machine learning models can be employed by organizations for various purposes. Streamlining tedious administrative work, optimizing marketing efforts, increasing system performance, or finishing the preliminary phases of research and development are a few examples. The classification and division of unprocessed data into predetermined groups is a widely utilized method. Once the model is trained and performing to a given accuracy on training data, it is ready to be prepared for deployment.

## Abstract:

There can be many applications be made of US-Accidents like: real-time car accident prediction; identification of hotspots for auto accidents; analysis of casualties and extraction of cause-and-effect rules for auto accident prediction; and investigation of the influence of precipitation and other environmental stimuli on the occurrence of auto accidents. Examining the effects of weather on traffic behaviour and accidents can also benefit from using the dataset's most current version. The CRISM-DM approach will be applied in this project.

Introductory Section:

Road accidents are a major global issue that have a profound impact on societal well-being, economic productivity, and public health. Identifying risk indicators, deriving effective preventive actions, and comprehending the fundamental causes of these catastrophes can be achieved through the potential to leverage the extensive data pool around them.

We explore the complex world of auto accident data in this report, using cutting-edge data engineering methods to uncover interesting trends, correlations, and anomalies. With the use of strong analytical frameworks and the synthesis of diverse datasets, we hope to reveal the intricate web of variables affecting the frequency and seriousness of auto accidents.

Every aspect of the transportation ecosystem creates a digital trail that is ready for investigation, from patterns of urban congestion to meteorological conditions, driving habits, and car specifications. We work to extract actionable insights that can guide evidence-based decision making and eventually improve road safety through rigorous data preparation, feature engineering, and machine learning modelling.

Review of available research:

Understanding the elements that lead to car crashes and crafting efficient prevention tactics is crucial for public safety and welfare. Recently, the utilization of data engineering methods has created fresh opportunities for dissecting and tackling this intricate problem. This evaluation scrutinizes the current research on car accident analysis through the lens of data engineering initiatives, highlighting essential approaches, discoveries, and prospects for further investigation.

While current research has made considerable progress in utilizing data engineering for analyzing car accidents, several opportunities for future investigation arise:

* *Real-Time Data Integration:* Incorporating live data streams from IoT devices, connected vehicles, and social media platforms can improve the immediacy and detail of accident analysis.
* *Causal Inference:* Deepening our comprehension of causal links between variables and accidents using causal inference techniques can guide targeted interventions and policy decisions.
* *Multi-modal Analysis*: Examining the interactions among various transportation modes (such as cars, bicycles, pedestrians) in accident analysis can offer a more comprehensive understanding of road safety dynamics.
* *Ethical and Privacy Considerations:* With the increasing prevalence of data collection and analysis, addressing ethical and privacy issues regarding the utilization of sensitive personal data is crucial.

## Methodology:

DATA ANALYSIS:

Initially, we'll gather the data from Kaggle and perform data cleaning. This process involves eliminating any null or duplicate values within the dataset. Tools utilized: NumPy and pandas.

**DATA MODELLING:**

We will employ the K-Nearest Neighbors (K-NN) algorithm to determine the location of car accidents. K-NN, a non-parametric supervised learning classifier, relies on proximity to classify or predict the grouping of a given data point. Although it can tackle regression or classification tasks, it is commonly utilized as a classification algorithm, operating under the premise that similar points tend to cluster together.

**MODEL EVALUATION:**

We will use two or three machine learning models for this dataset. We will finalize the model which gives the best performance and accuracy.

**MODEL DEPLOYMENT:**

We'll deploy the project using GitHub and a Flask server.

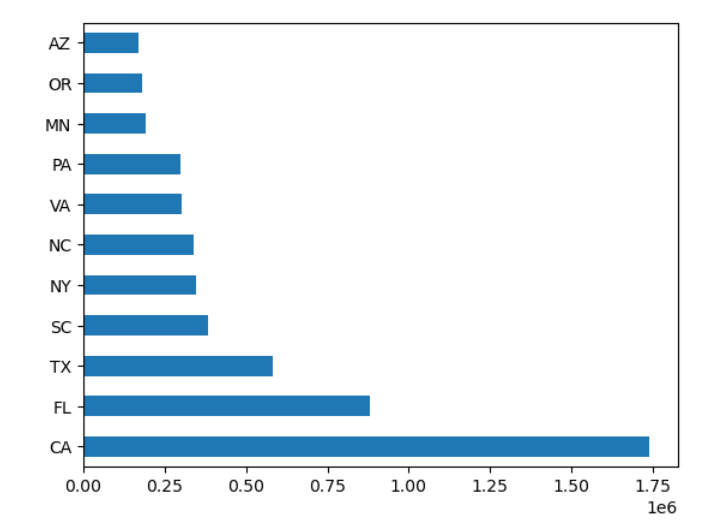
*Flask Server:*

The Flask server is described as server software capable of processing HTTP requests on the public World Wide Web, private LAN, and private WANs. It consists of one or multiple computers grouped together and dedicatedly functioning to run the software application on the global web. While Flask comes with a built-in server, developers have the option to utilize others for their convenience. This server can manage requests from HTTP on one or more configured websites. In essence, the server's role is to receive incoming HTTP requests and send back the processed HTTP responses to the client.

**DATA VISUALIZATION:**

We'll utilize matplotlib, seaborn, and PowerBI for data visualization purposes.

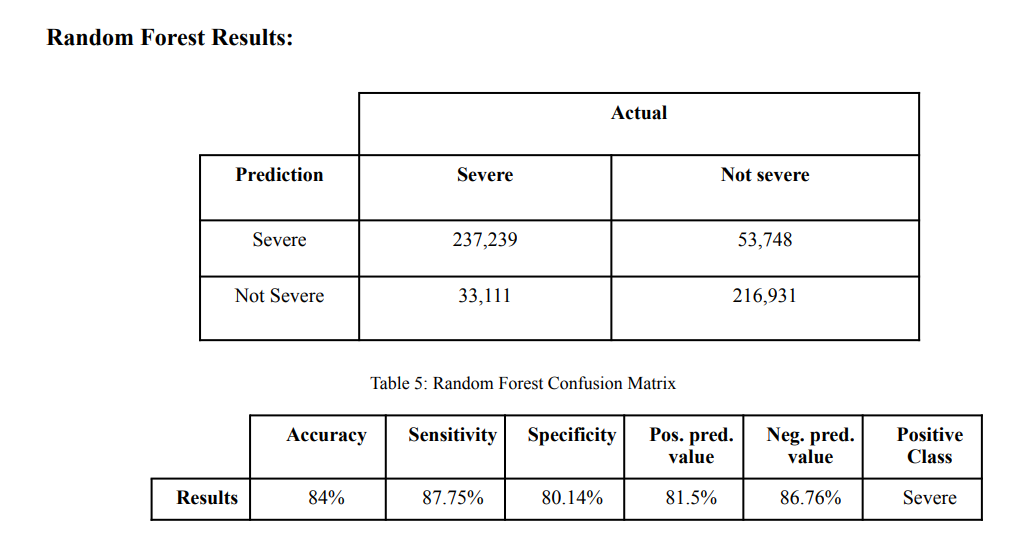
* We'll create graphs plotting severity against city and state to identify cities with the highest and lowest severity levels.
* Graphs will be generated plotting severity against weather, temperature, and humidity to examine the relationship between severity and weather conditions.
* We'll plot graphs correlating severity with wind speed and wind direction to understand how severity levels are affected by wind conditions.
* Graphs will be generated plotting severity against distance and city to analyze the impact of distance from the city on severity levels.
* A pairs plot, as previously mentioned, is a matrix of scatterplots that enables you to comprehend the pairwise relationship between several variables in a dataset. Google and Kaggle are only two of the sites from which the data is gathered. First, the data needed for the project must be chosen from a variety of platforms, data sets, and other sources, such as Kaggle. The information is obtained and stored in a CSV file.



*Bar graph representation of state wise accidents count*

## Results Section:

In order to anticipate auto accidents in real time, we will investigate the following: identifying hotspots for auto accidents, analyzing casualties and deriving cause-and-effect formulas, and researching the influence of precipitation and other environmental factors on event frequency. In order to identify the cities with the highest and lowest severity, we will construct a graph of severity against city and state. In order to understand how the severity level is influenced by weather, we will draw a graph of severity against temperature, humidity, and weather. To determine how severity level varies on wind conditions, we will display the severity vs wind direction and speed on a graph. We are going to plot the graph between severity and distance and city to know the how the severity level depends on distance from city.



Git:

[DSCI-6007-04-Team-03/Mapping-Car-Mishaps: A Data Engineering Study of Car Accident Trends (github.com)](https://github.com/DSCI-6007-04-Team-03/Mapping-Car-Mishaps)

## Conclusion:

With the help of this project, we may draw conclusions about/analyze auto accidents according to city, weather, distance, temperature, pressure, and wind conditions. We'll forecast the areas of significant hotspots and auto accidents. The primary causes of the mishaps that we shall display on the visualizations.

## References:

* <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data>
* [Matplotlib cheatsheets — Visualization with Python](https://matplotlib.org/cheatsheets/)
* [ml-tooling/best-of-ml-python: 🏆 A ranked list of awesome machine learning Python libraries. Updated weekly. (github.com)](https://github.com/ml-tooling/best-of-ml-python)
* [Unintentional injury in the U.S. - Statistics & Facts | Statista](https://www.statista.com/topics/3991/unintentional-injury-in-the-us/#topicOverview)