|  |
| --- |
| TECHNICAL REPORT TEMPLATE |

|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
|  |  |

 DSCI - 6007

Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

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

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 5](#_Toc96341556)

[Discussion 6](#_Toc96341557)

[Conclusion 6](#_Toc96341558)

[Contributions/References 7](#_Toc96341559)

|  |
| --- |
| Project Name |

|  |  |  |
| --- | --- | --- |
| Executive Summary This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2021, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 2.8 million accident records in this dataset. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Mary Prashanthi Vongala**  **Sravya Kaitha**  **Alekya Basike**  **Krathik Reddy** | **Questions?**  Contact :  [mvong1@unh.newhaven.edu](mailto:mvong1@unh.newhaven.edu)  [skait1@unh.newhaven.edu](mailto:skait1@unh.newhaven.edu)  [abasi3@unh.newhaven.edu](mailto:abasi3@unh.newhaven.edu)  [kthum1@unh.newhaven.edu](mailto:kthum1@unh.newhaven.edu) |  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| **US CAR ACCIDENTS ANALYSIS** |  |
| Highlights of Project In this project,  We are going to plot the graph between severity and city and state to know the which cities has the highest and lowest severity. We are going to plot the graph between severity and weather, temperature, and humidity to know the how the severity level depends on weather conditions. We are going to plot the graph between severity and wind speed and wind direction to know the how the severity level depends on wind conditions.  We will predict the real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents. Submitted on: |

## Abstract

US-Accidents can be used for numerous applications such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence. The most recent release of the dataset can also be useful to study the impact of COVID-19 on traffic behavior and accidents. We are going to use the CRISM-DM methodology for this project.

Review of available research

The entire process from considering the respective data of inputs and getting the desires values of output by using the K\_NN model is completed by following the CRISP-DM methodology. The cross industry standard process for data mining is a process with six phases that serves as the base for the data science process. The six phases are:

• Business understanding.

• Data understanding.

• Data preparation.

• Modeling.

• Evaluation.

• Deployment.

BUSINESS UNDERSTANDING:

The main aim of this phase of the methodology is the requirement or the need of the project. The first thing to concentrate on is to understand what a customer really wanted to accomplish from a business point of view. In other words, the main objective of the project is to be understood or determined in this phase of the project. Based on the resources available and the situation dependent on the project requirements, risks, outcomes, and success are to be predicted and understood. In addition to all these, a simulated output of the deliverable is also to be determined in a data mining perspective. And what tools and technologies are required for the successful completion of the project are also to be studied and known. This step should be concentrated well because this is the concrete base of the entire project. For our project we decided that tools like programming language python, statistical technique like linear regression, visualization techniques are important to read the data (from the prominent resources like kaggle, Google etc), analyze the data, interpret, predict, visualize, deploy the data and result in the desired output.

DATA UNDERSTANDING:

Understanding the data is the fundamental step for the next process and describing the data to find the interesting patterns. The data gathered from different sources by using various techniques is to be analyzed with data analyzing tools like Pandas, NumPy etc. Analyzing/cleaning is to read the data which is gathered. During this process the unwanted data is cleaned and made precise for the data visualization. Based on this analysis the data study is performed on the gathered data and relevant output or outcomes is/are determined. The data collected is to be understood to the core of its value. So, the data is to be dig deeper and deeper to understand, visualize and to raise a question out of it if possible. In this process the dummy values or the null values or missing values are removed from the collected data by using various techniques. In this process we also checked whether there is any correlation between the independent variables.

Deleting the Missing value:

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.

Deleting the entire row:

If a row has many missing values, then you can choose to drop the entire row.

Replacing with previous value – Forward fill:

In some cases, giving the values with the previous value instead of mean, mode or median is more appropriate. This is called forward fill. It is mostly used in time series data.

Replacing with next value – Backward fill:

In backward fill, the missing value is imputed using the next value.

Pair plots:

A pairs plot is a matrix of scatterplots that lets you understand the pairwise relationship between different variables in a dataset as discussed above. The data is collected from various resources like kaggle, google. First, the data required for the project need to be selected from various sources like Kaggle, data sets or any other platform. The data is gathered and collected into CSV file

DATA PREPARATION:

This phase concentrates on preparing the data based on requirements of the project.

MODELING:

The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data. The analyzed data based on the need may be required to be visualized and represented with visualization. We normally use libraries like sklearn, Matplotlib, Bokeh for the data visualization. Visualization is one kind of technique that converts raw data into an easily understandable format. This may be the representation of the data in the form of bar graphs, pie charts, and scatter plots etc. This process helps in understanding the data and representing the data in a simper format that can be used to draw conclusions quite easily.

EVALUATION:

The most important thing you can do to properly evaluate your model is to not train the model on the entire dataset. A typical train/test split would be to use 70% of the data for training and 30% of the data for testing. As we have discussed previously, it's important to use new data when evaluating our model to prevent the likelihood of overfitting to the training set. However, sometimes it's useful to evaluate our model as we're building it to find that best parameters of a model - but we can't use the test set for this evaluation or else we'll end up selecting the parameters that perform best on the test data but maybe not the parameters that generalize best. To evaluate the model while still building and tuning the model, we create a third subset of the data known as the validation set. A typical train/test/validation split would be to use 60% of the data for training, 20% of the data for validation, and 20% of the data for testing. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

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

DEPLOYMENT:

API and web application are deployed using flask server. API contains the core logic for prediction and web application provides user interface to interact with the API. Both the applications are deployed on the server they interact with each other on a secure protocol. Web application sends necessary data for the prediction of insurance to API using json format. Organizations may use machine learning models for a range of reasons. Examples include streamlining monotonous administrative tasks, fine-tuning marketing campaigns, driving system efficiency, or completing the initial stages of research and development. A popular use is the categorization and segmentation of raw data into defined groups. Once the model is trained and performing to a given accuracy on training data, it is ready to be prepared for deployment.

## Methodology

**DATA ANALYSIS:**

Firstly, we will collect the data from Kaggle and clean the data. We will remove the null values and repeated values in the data.

Tools used: - Numpy, pandas

**DATA MODELLING:**

We are going to use K-NN algorithm to decide the car accident’s location. The k-nearest neighbors’ algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

**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 are going to deploy the project using github and flask server.

Flask server:

Flask server is defined as server software that can run HTTP requests on the public world wide web, private LAN, and private WANs and comprises of one or many computers bundled together and dedicatedly working for running the software application on the worldwide web. Flask has a built-in server, but options of using others are present for the convenience of the developers. The server is capable of handling requests from HTTP on one or more configured websites. In short, the working of a server is to receive the incoming HTTP request and send the processed HTTP request back to the client.

**DATA VISUALIZATION**:

We are going to use matplotlib, seaborn and PowerBi for visualizing the data.

1. We are going to plot the graph between severity and city and state to know which cities have the highest and lowest severity.
2. We are going to plot the graph between severity and weather, temperature, and humidity to know how the severity level depends on weather conditions.
3. We are going to plot the graph between severity and wind speed and wind direction to know how the severity level depends on wind conditions.
4. We are going to plot the graph between severity and distance and city to know how the severity level depends on distance from city.

## Results Section

We are going to find real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence. We are going to plot the graph between severity and city and state to know which cities has the highest and lowest severity. We are going to plot the graph between severity and weather, temperature, and humidity to know how the severity level depends on weather conditions. We are going to plot the graph between severity and wind speed and wind direction to know the how the severity level depends on wind conditions. 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.

## 

## Conclusion

From this project we can conclude/analyze the car accidents by city and weather conditions and distance, temperature, pressure, wind conditions. We will predict car accidents and major hotspot locations. The main reasons behind the accidents we will show on visualizations. From the dataset we will study the impact of COVID-19 on traffic behavior and accidents.

## Contributions/References

<https://www.kaggle.com/datasets/sobhanmoosavi/usaccidents?select=US_Accidents_Dec21_updated.csv>

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761