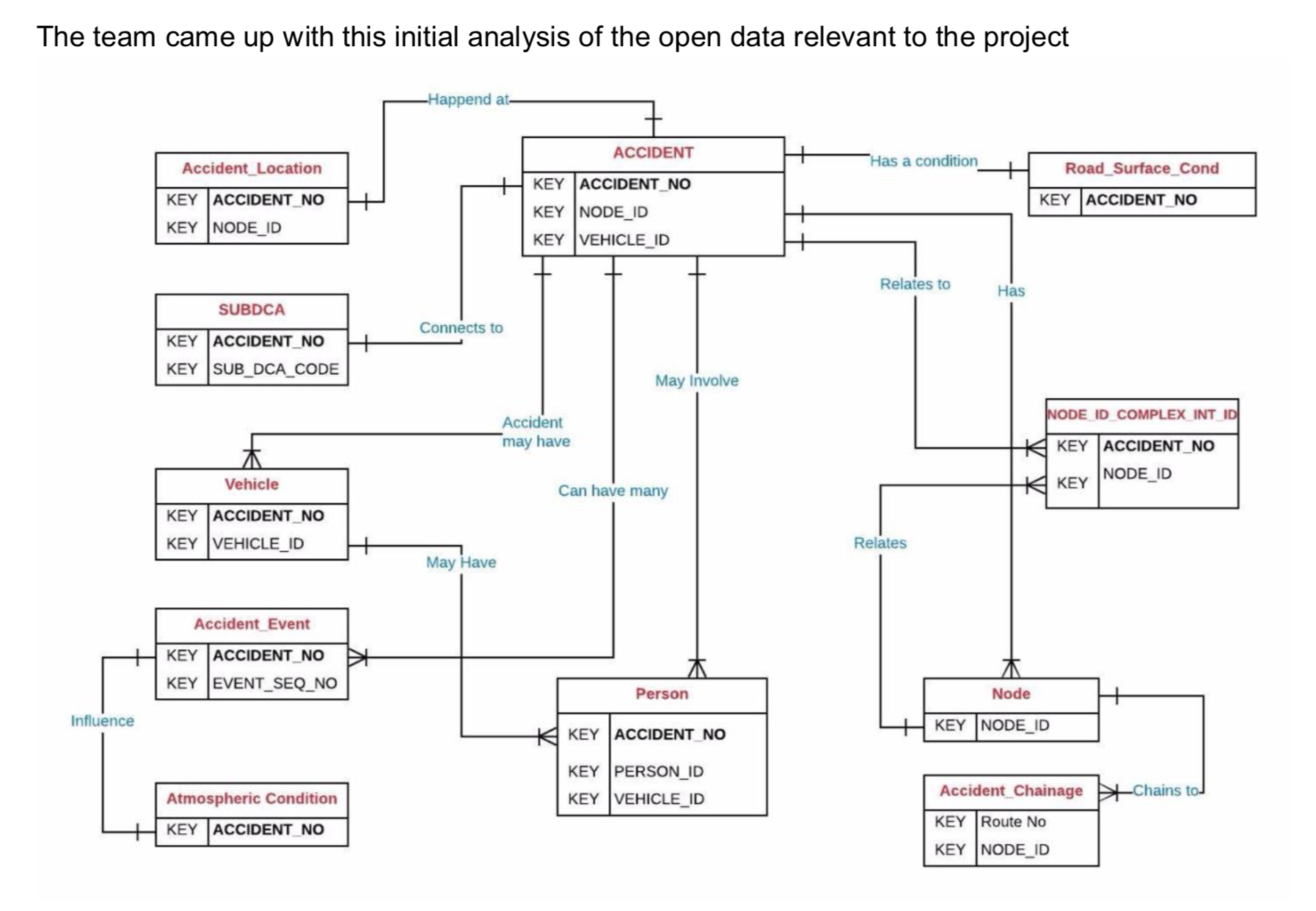
**Product Document - Harsha Jarugu**

## 1.6 Data Source Details

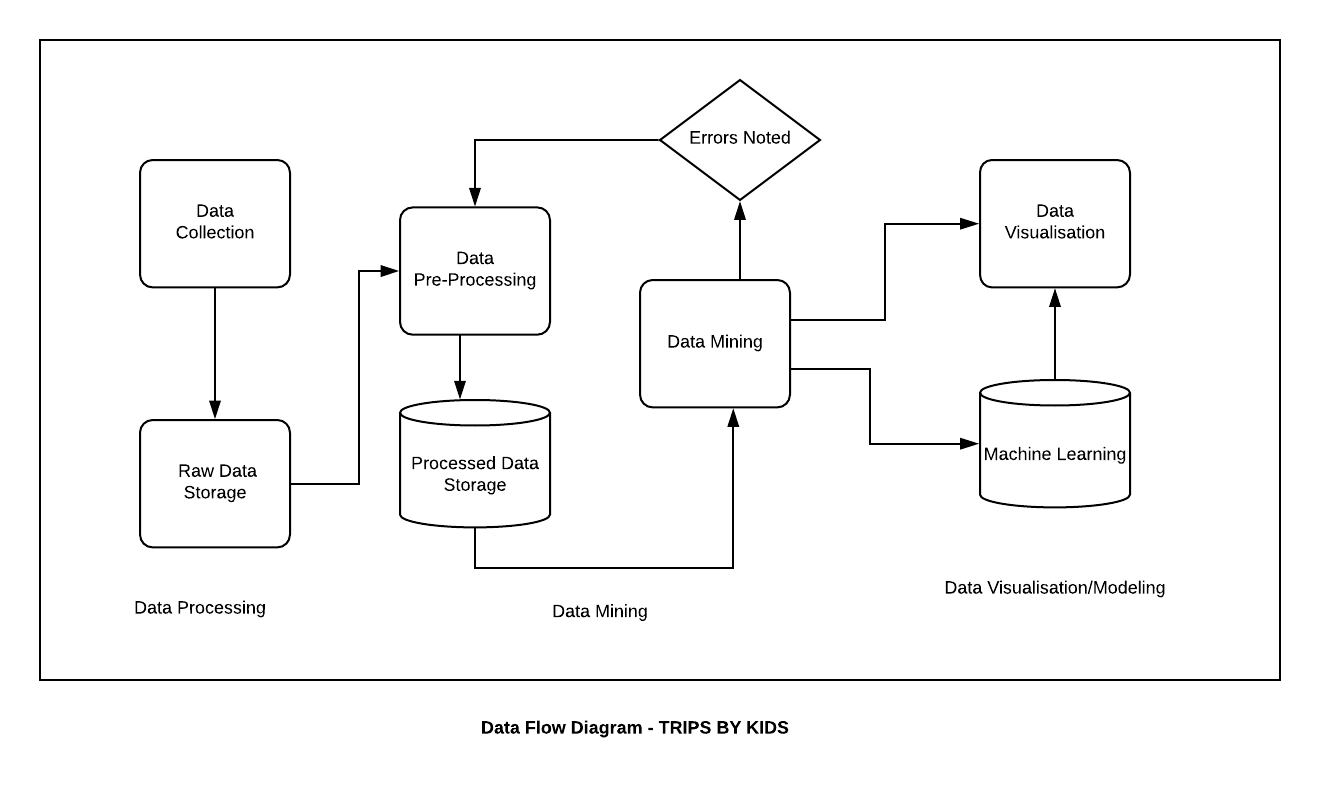
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Source No** | **Data Source Title** | **Format** | **Frequency of Source Updates** | **\*Frequency of Iteration System Updates** | **Granularity** | **Copyright Details** |
| 1 | Crash Stats - Data Extract | CSV | Last Updated 10/5/2018 | Every Iteration | Injury  level | Licensed to the State of Victoria, Public Transport Victoria 2018 |
| Data Resource Link: <https://www.data.vic.gov.au/data/dataset/crash-stats-data-extract/resource/392b88c0-f010-491f-ac92-531c293de2e9> | | | | | | |
| 2 | Gardens and Reserves | CSV | Not Specified | Every Iteration | Address Level | Licensed to the Department of Planning and Community Development |
| Data Resource Link: https://data.gov.au/dataset/gardens-reserves-gfyl | | | | | | |
| 3 | Recreation Groups | CSV | Not Specified | Every Iteration | Address Level | Licensed to the Department of Planning and Community Development |
| Data Resource Link: https://data.gov.au/dataset/recreation-groups-gfyl | | | | | | |
| 4 | Swimming Pool | CSV | Not Specified | Every Iteration | Address Level | Licensed to the Department of Planning and Community Development |
| Data Resource Link: https://data.gov.au/dataset/swimming-pools-gfyl | | | | | | |
| 5 | Ice Skating Centers | CSV | Not Specified | Every Iteration | Address Level | Licensed to the Department of Planning and Community Development |
| Data Resource Link: https://data.gov.au/dataset/ice-skating-centres-gfyl | | | | | | |

Data Modelling

The conceptual and logical model of the data we consider (E-R model is attached at appendix) and try to wrangle the different datasets and merge them with respect to necessity.



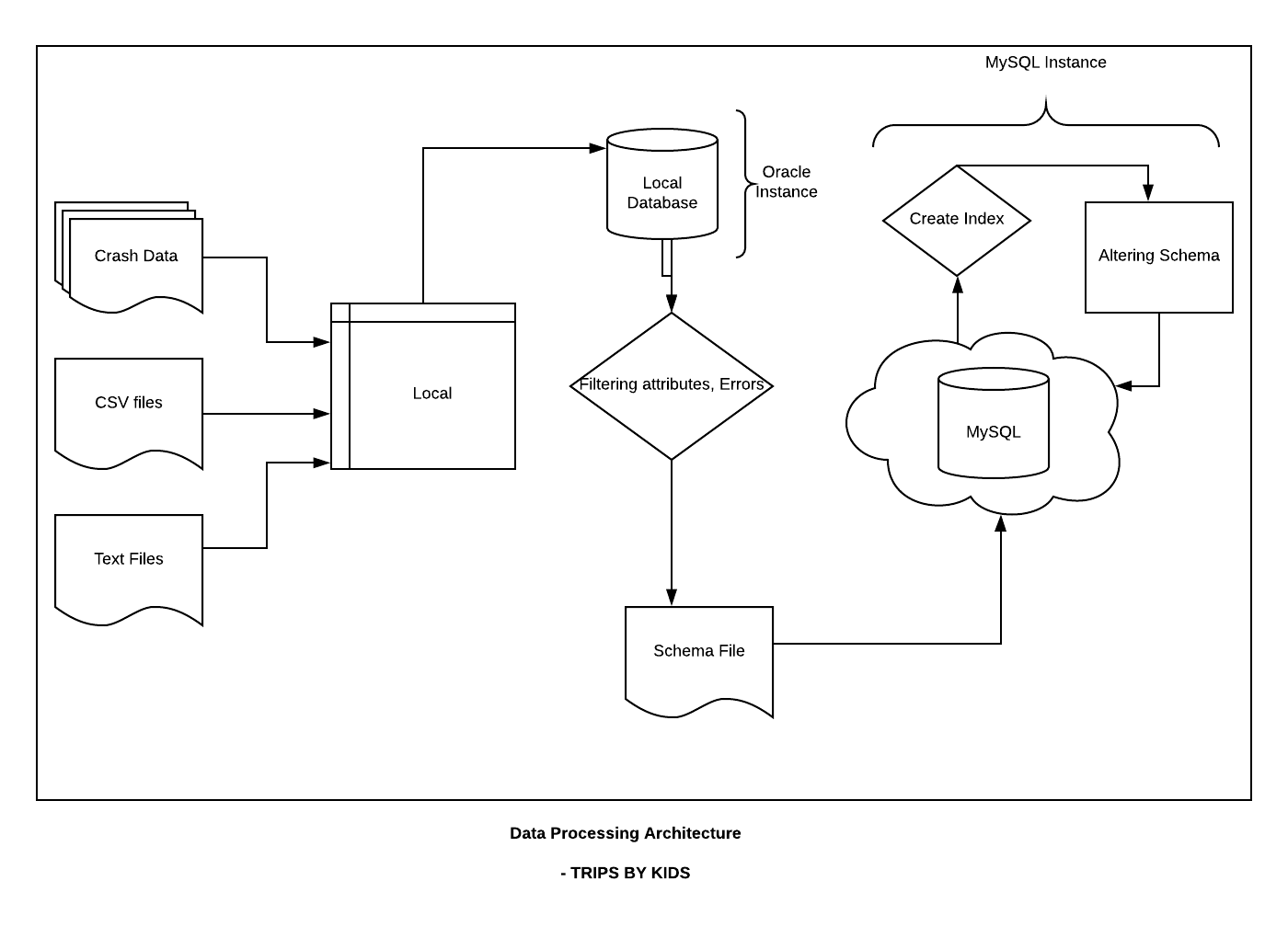
**Data Flow:**



We have followed the above process to collect and normalize the data for our project.

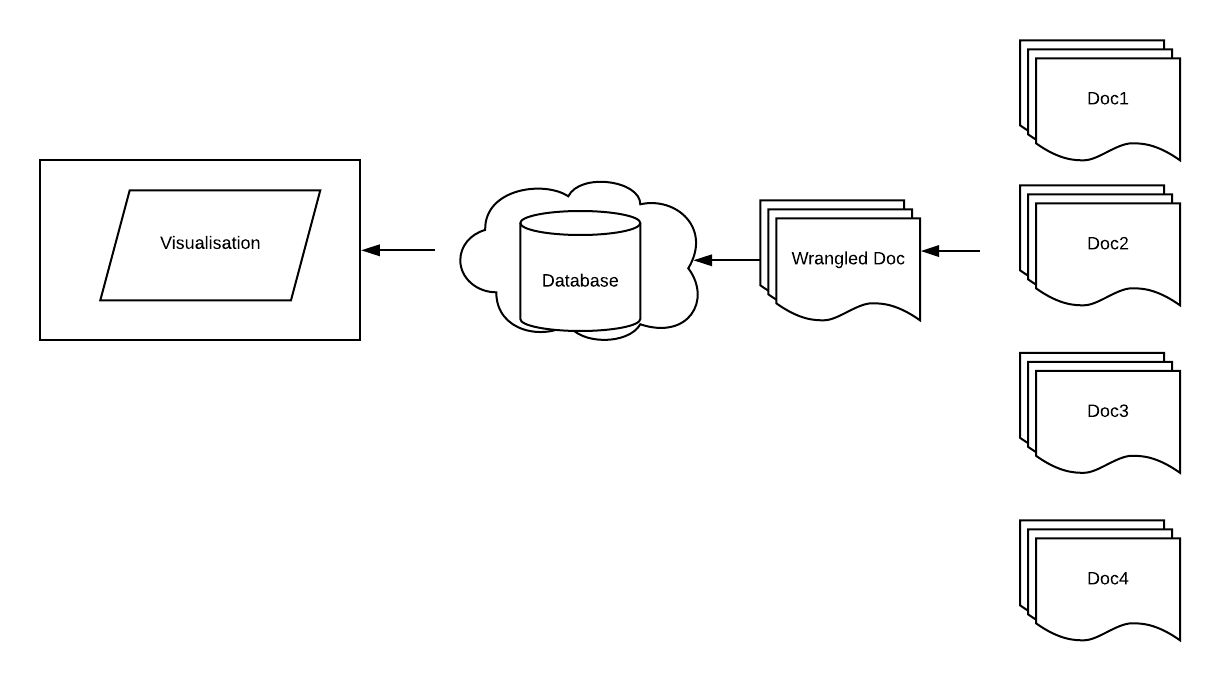
**Data Processing:**

We have followed the following architecture to build our Relational Database. All the steps are indicated with a process flow.



**Data Mining**

* We have taken advantage of T-SQL and PL-SQL concepts and have summarized and grouped the tables as per our requirements and stored them as PROCEDURES which helps us to load data instantly.
* We have found the queries are facing the time complexity. To address this situation, we have created INDEXES on the filter attributes which increased the computing power and decreased the running time.
* All the errors found in this process are cycled back to pre-processing stage and fixed.
* Multiple documents are merged using python and wrangled all lexical, symmetrical errors from the datasets

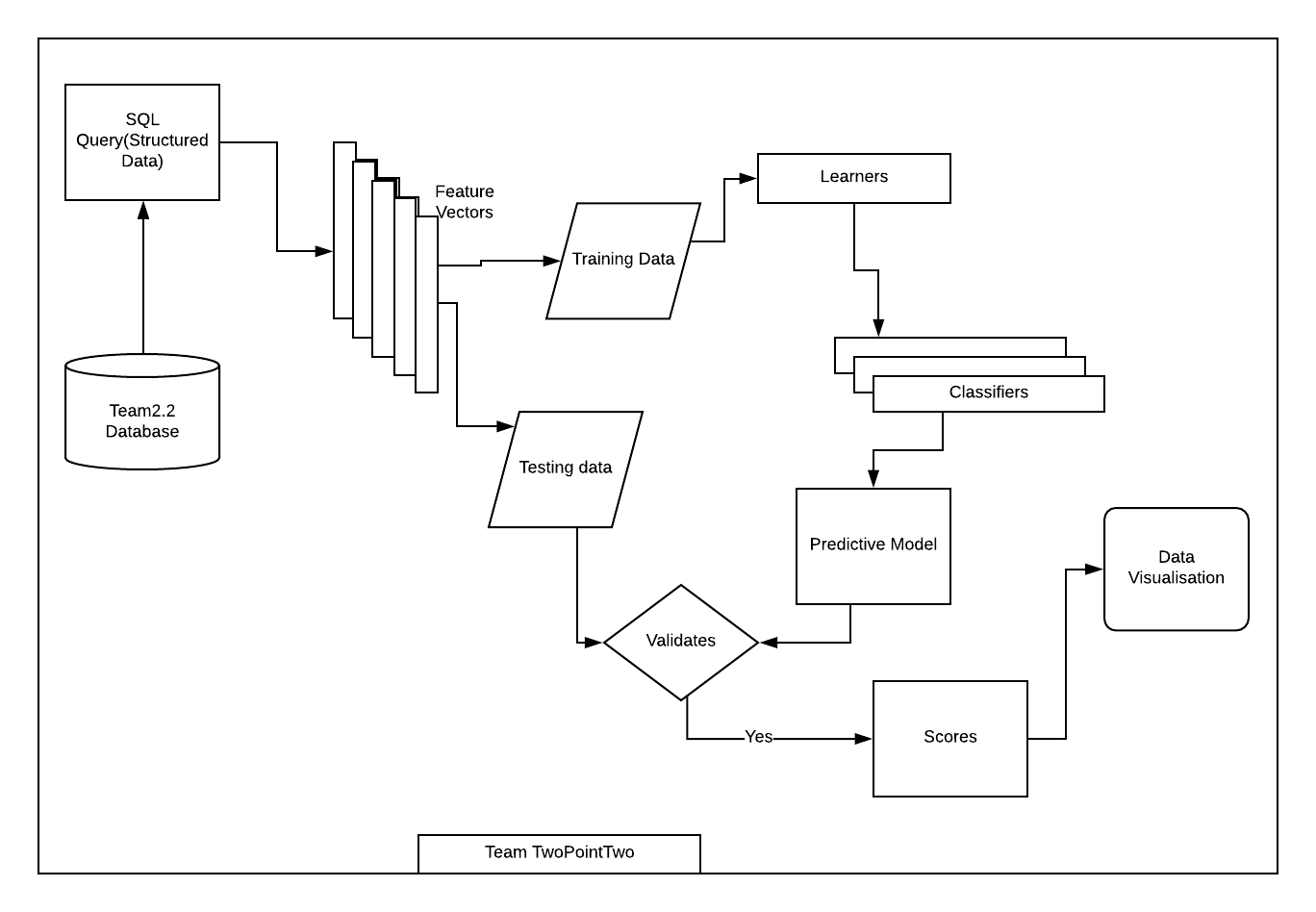


**Machine Learning:**

The Crash data from VicRoads provides us various factors and considering the traffic accidents as one of our major public problem worldwide, for this reason safety we tried to identify the potential factors affecting the severity as consequence of road accidents. In order to identify these factors, Data Mining (DM) techniques such as Decision Trees (DTs), have been used. A dataset of traffic accidents on roads in the Victoria have been processed and analyzed. Decision Tree will allow certain decision rules to be extracted. These rules could be used in future road safety campaigns and would enable parents to implement certain priority actions while travelling with the kids.

**Model Selection:**

We have a big data set, for this reason, we can attempt the models to identify affecting the severity as consequence of road accidents. Different techniques, such as: Regression-type generalized linear models have been used to achieve these objectives. However most of these models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. if these assumptions are violated, the model could lead to erroneous estimations, for example of the likelihood of severity accident. To solve these limitations other method such as, Classification and Regression Trees (CART), have been used in the field of Road Safety. CART is particularly appropriate for studying traffic accident because is non-parametric techniques that do not require a priori probabilistic knowledge about the phenomena under studying and consider conditional interactions among input data



A decision tree (DT) could be defined as a predictive model which can be used to represent both classifiers and regression models (depending on the nature of the variable class). When the value of the target variable is discrete, a regression trees is developed, whereas a regression trees is developed for the continuous target variable. A DT is a simple structure formed by number finite of “nodes” (which represent an attribute variable) connected by “branches” (which represents one of the states of the one variable) and finally, “terminal nodes or leaf’s” which specify the expected value of the variable class or target variable. The principle behind tree growing is to recursively partition the target variable to maximize “purity” in the child node. Severity level is taken as Class Variable and used to identify the factors that are contributing to the Road accidents. Variables that are contributing are observed and recorded and visualized

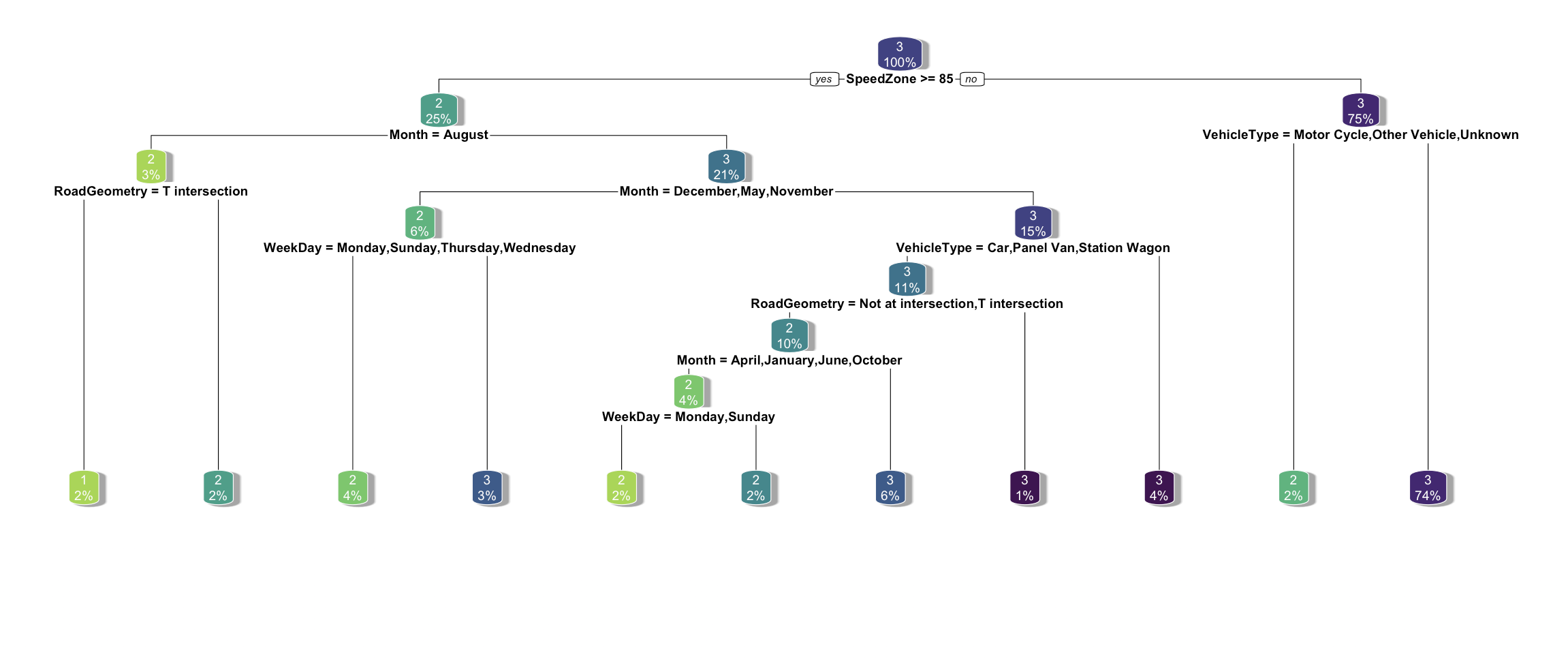
**Variable Importance:**

* The CART modelling process has an important phase in which the variables that are of key importance in the prediction of the dependent variable are identified. Later they are plotted as bar chart as having the greatest influence on accident severity

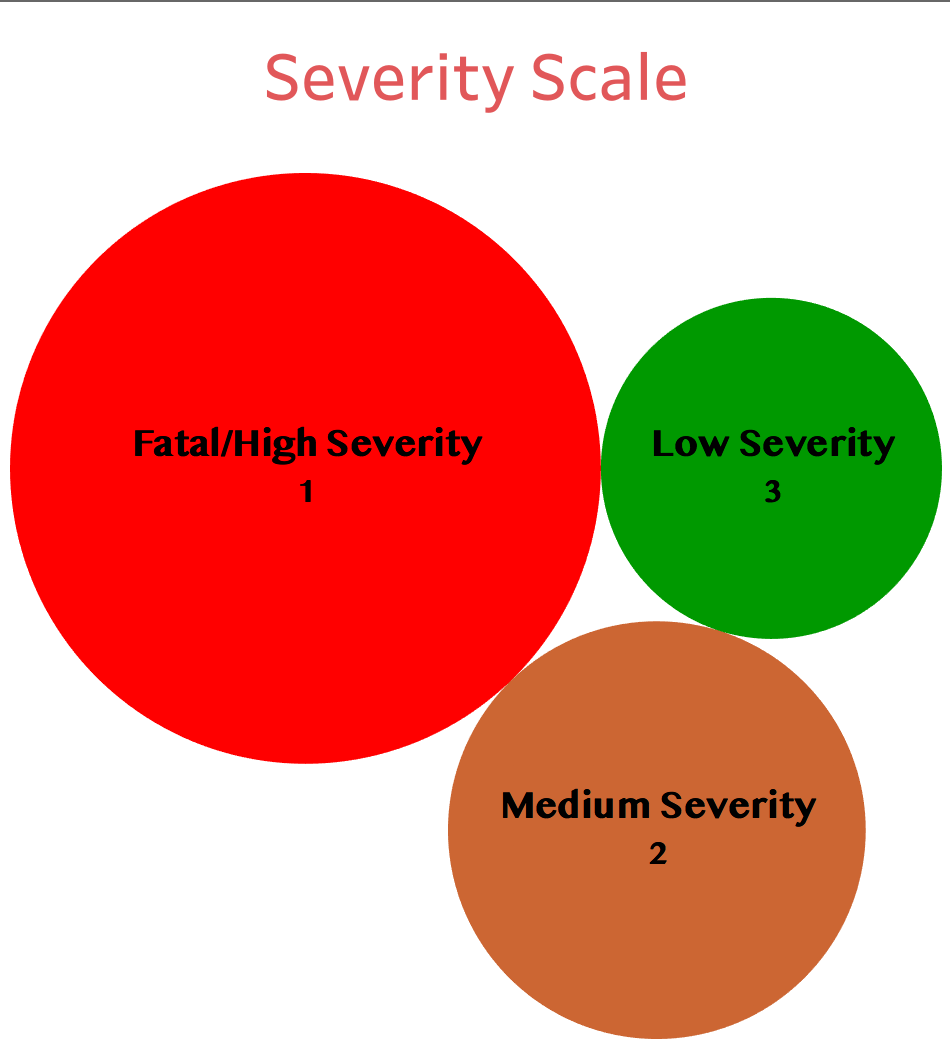


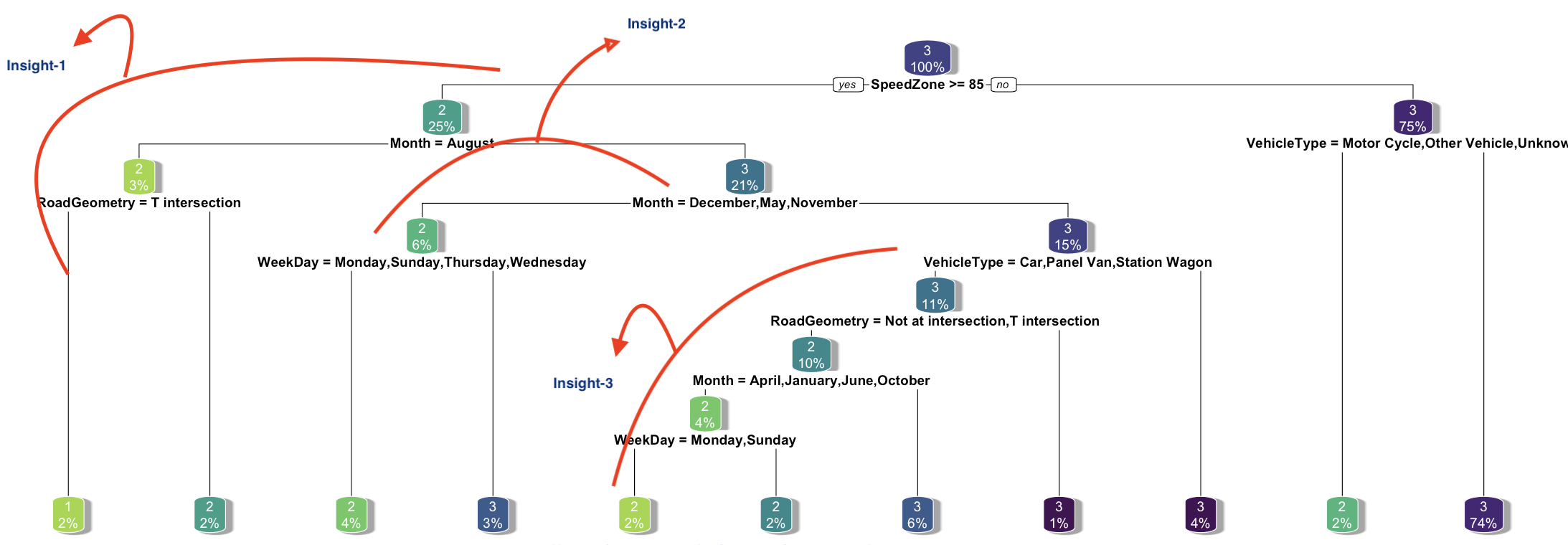
**Decision Tree:**

* Top seven variables are picked dynamically, and they are processed to the random forest model to find the patterns with in these variables They are plotted as Decision Tree.

****

* This is a three-level decision tree generated by using random forest algorithm
* Severity is used as class variable to create the decisions on observing the patterns and segregating them as per the severity level
* Severity level is divided into three modes: 3 being lowest and 1 being highest severity
* Year is used as filter in the dashboard. This facilitates the user to find the patterns for last decade and summaries the patterns of accident.



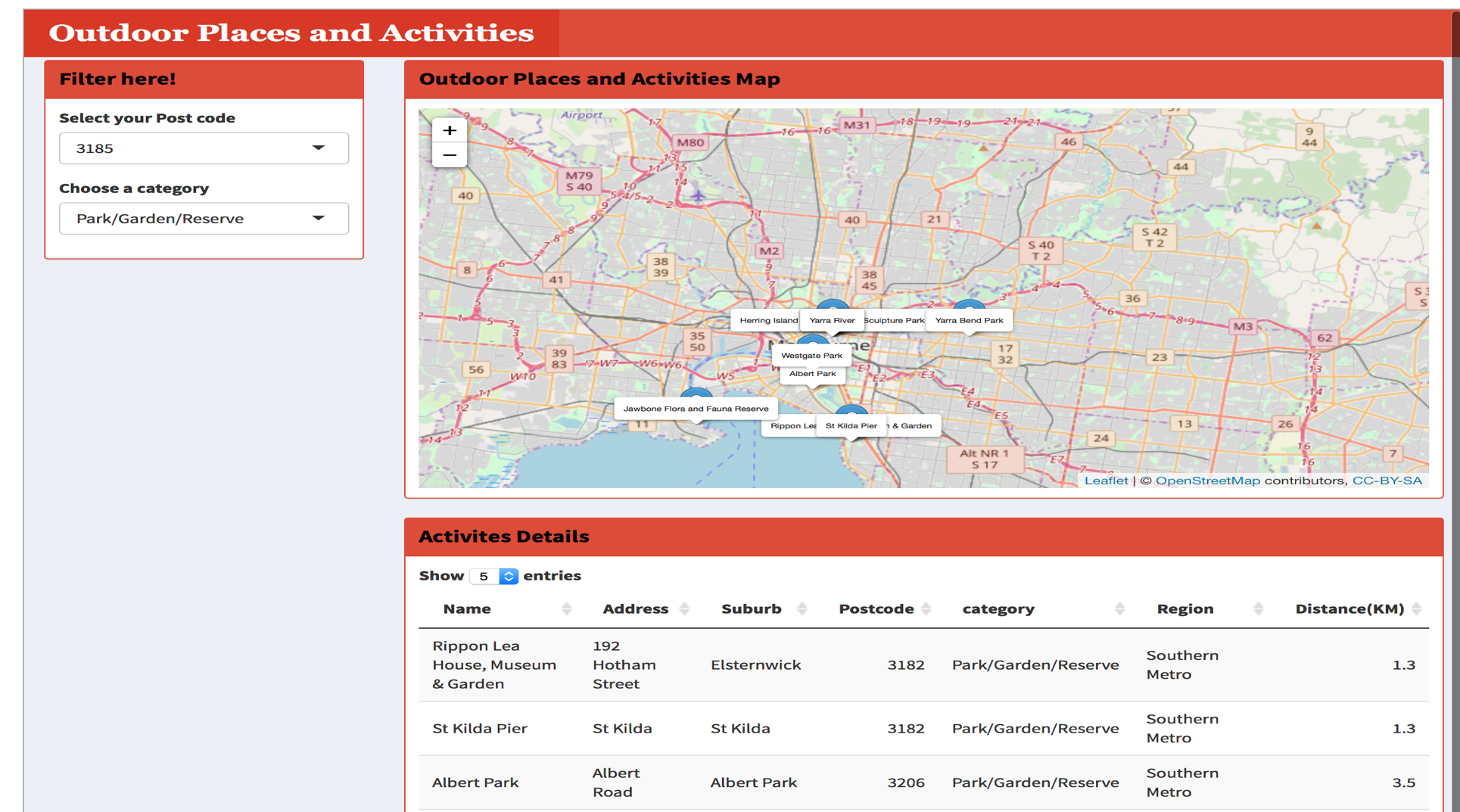
Interpretation

Process to interpret the insight that is generated:

|  |  |  |
| --- | --- | --- |
| **S No** | **Level** | **Insight Description** |
| 1 | 1 | During the Month of August, People travelling on high speed zones (>=85) met accidents at T-Junctions. |
| 2 | 2 | The months with medium – high severe accidents are recorded in months of May, November and December potentially long weekends that include special holidays such as Black Friday, Boxing Day that falls on days such Wednesday, Thursday and Monday. |
| 3 | 3 | People using Vehicle type like car, Panel Van are met with accidents on April, January, June and October with medium severe accidents. |

**Data Visualization (Places of Interest)**

* All the Locations in Victoria with details of parks, little athletic clubs are shown. This visualization helps the kids to understand the places next to each other. They can be filtered based on the choice and suburb kid lives.



**Data Analytic Tools:**

1. Python: Data Wrangling
2. R: Machine Learning, Data Visualization
3. Shiny Dashboard: Interface for ML model and Leaflet visualization