Data Mining Project - Road Safety Analysis

Team - Data Miners

Team Members' Names and UIS Emails

Names	UIS Email ID
Revathi Gunasekaran	rguna2@uis.edu
Hari Krishna Goalla	hgoal2@uis.edu
Cury Velpuri	mvelp2@uis.edu

Guided by: Dr. Neetu Singh

Abstract

Data mining has become an integral part of analyzing, predicting, and exploring the data, an entity which is the most powerful in today's world. SAS Enterprise Miner is one of the significant tools used to implement such vital processes on data to derive insightful results. We carried out our analysis on a data set that is based on road accidents pertaining to U.K, obtained from Kaggle, so as to derive meaningful analysis that can help in suggesting crucial measures to mitigate the intensity of the number of accidents, as road safety is one of the most crucial area where lives are at stake every second. We started off with preliminary data analysis using excel, followed by data exploration using SAS Miner.

Tableau has been instrumental for data visualization in our case that led us to view very interesting visualizations, analysis and dependencies. We cleaned the data based on the results of data exploration from SAS, eliminated a few variables that were giving issues like missing data, data types while a few variables were automatically rejected by SAS while importing the data. We performed both predictive and explorative analysis to help find solutions to our research questions that focus on important factors that determine the accident severity (target variable) and creation of homogenous clusters respectively.

We used Classification tree, Regression tree and Neural networks with respect to predictive analysis followed by comparison of all the models to identify the best model out of all based on performance metrics like misclassification rate, overfitting etc. In case of exploratory analysis, we went with ward clustering, centroid clustering and average clustering to compare the results, followed by trying a 90% sample on all the models to test the stability after which we attempted k-means clustering to check for alternative clustering options.

We determined that number_of_vehicles, speed_limit, weather_conditions are significant factors in predicting the accident severity while light_conditions, road_surface_conditions and weather_conditions are crucial variables in terms of clustering. From the results of both these analyses, we can observe that weather_conditions, light_conditions and road_surface_conditions play an important role, which is why actions aligned towards making these conditions better can help reduce the number of accidents or at least reduce the severity of accidents.

Problem Description

The problem we are trying to solve is related to road-safety. In our project, we have determined the most important factors that were used to predict the severity of accidents. The reason we chose road-safety is because it is a serious concern that needs to be looked after as it involves a lot of lives at stake. This problem needs to be addressed because each year, roughly 1.3 million people die in road accidents around the world, with between 20 and 50 million people suffering non-fatal injuries (Kashani et.al, 2014). Pedestrians, cyclists, and motorcyclists, as well as their passengers account for more than half of all road traffic deaths and injuries (Kashani et.al, 2014).

So, by predicting the factors which account for such accidents, in the long-term, we will be able to reduce the accidents as well as the deaths.

Our entire project answers two questions:

1. Identifying the most important factors for predicting the severity of road-accidents (Predictive analysis)

2. Exploring the data to group into homogeneous clusters based on the significant variables (Exploratory analysis)

Most of the accident investigation is done based on different accident scenarios and simulations. As we know, fatalities and injuries have a significant impact on our society. Engineers and researchers in the automobile industry have been attempting to design and build safer vehicles, yet traffic collisions are unavoidable. They have been using real-life data to analyze many elements of traffic accidents in recent years.

On the other hand, steps must be taken to limit the number of accidents. It is critical that the measures be based on scientific and objective studies of the causes of accidents and the severity of injuries (Jayasudha, et.al, 2009). Our project is focused on finding the best model to predict the accident severity based on several variables: number_of_vehicles, speed_limit, weather_conditions, first_road_number and light_conditions and also we have investigated accidents and categorized them based on exploratory analysis using various clustering models like Ward clustering, Centroid and Average clustering.

Technical/Analytical Methods Used In Data Mining Process

Brief Description of Data Set and Data Source:

Data Source - https://www.kaggle.com/datasets/qasimhassan/reducing-the-number-of-high-fatality-accidents?select=accident-data.csv

The data set we have used is a secondary data set we obtained from Kaggle.com while the data set finds it origin from the website of road safety data at data.gov.uk/ published by department for transport of U.K. The mentioned website provides timely data, statistics, summaries, supporting documents collected by the police and transport departments each year classified by vehicles, casualties, accidents, blood alcohol content, digital breath test and many more for public use (Department for Transport, n.d).

Kaggle does a great job in presenting a great range of data sets for everyone and in giving an opportunity to everyone in publishing data sets, collaborating with others in terms of data science. Our data set obtained from Kaggle had 27 variables including all the unique IDs and references ranging from accident_index, accident_year to urban_or_rural_area and light_conditions (Hassan, 2022). This data set is a very interesting case as the variables available give a wide range of perspective, for e.g. - road_type, junction_control, weather_conditions, special_conditions, speed_limit, time, day_of_week, accident_severity etc, all of which present a diverse point of view with respect to each case (Hassan, 2022).

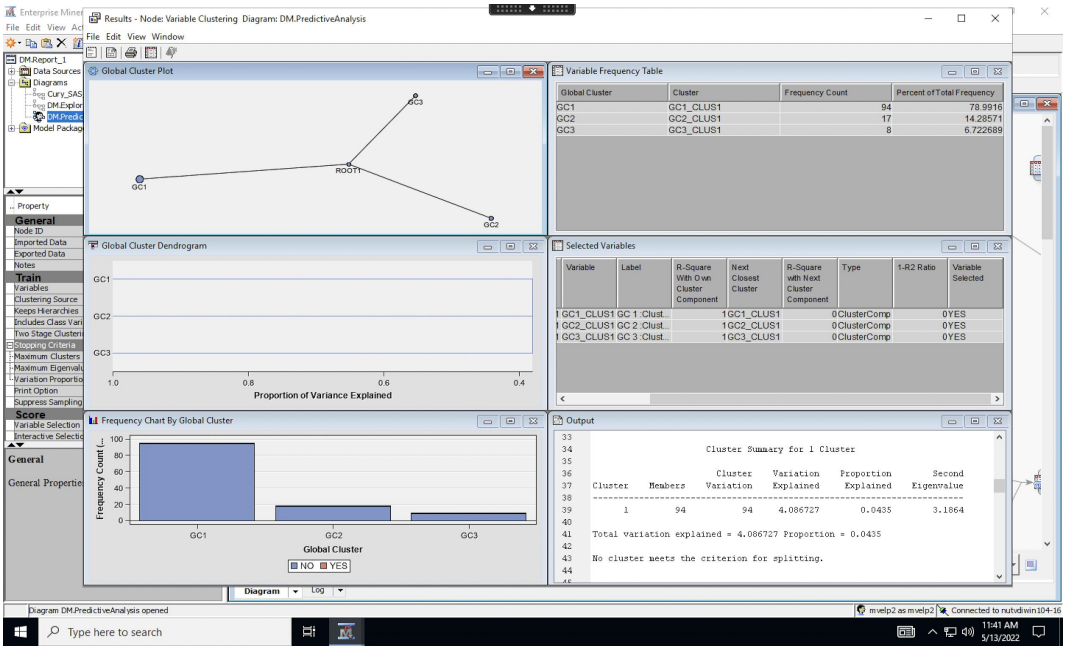
Data cleaning, preparation, and modification:

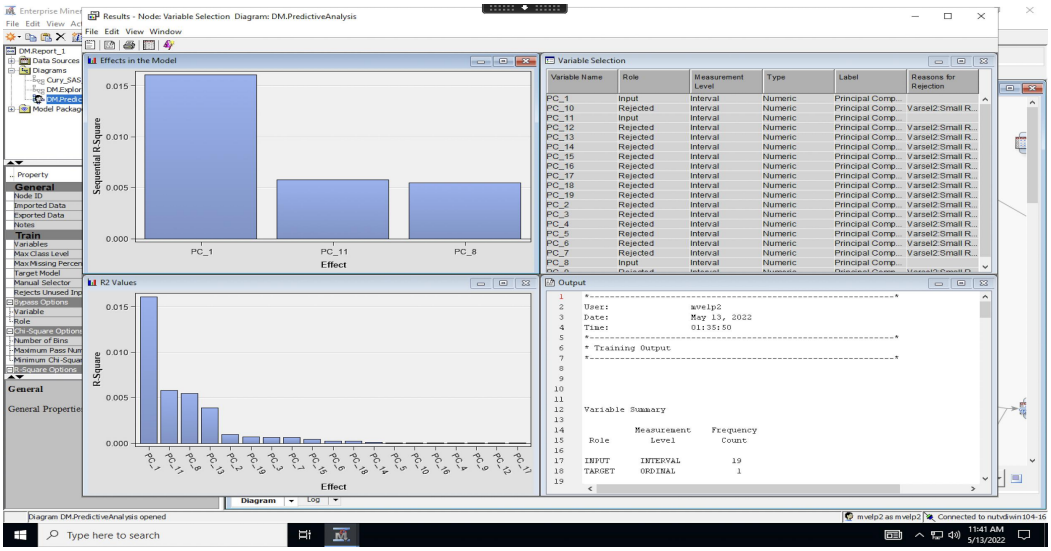
When we obtained the original dataset, we generally skimmed through it to check for any abnormalities. We performed the correlation analysis on the numeric columns and found no issues and no close correlation between any two variables. We then formatted all the columns to either number or text to data whichever applicable after which we transformed the excel data to SAS format and started performing data exploration using various nodes like statexplore, optimal, best, maximum normal, default, tree, principal components, variable clustering and variable selection.

When we started data exploration, we used the advanced settings when adding the data source and customized the values which led to rejection of a few variables automatically, like accident_index, accident_reference, date, accident_year. We then deleted these columns from

our data to make it simpler whenever dealing with the dataset on SAS. After examining the results from all the nodes, we realized that some rows data from latitude and longitude columns were missing, hence we decided to delete these columns on the whole as we felt that these variables would not contribute much to our analysis anyway.

With respect to variable clustering, there were three clusters generated in this node. After running the principal components node and variable selection node, we realized that the system was suggesting only three variables as usable variables and rejected the rest of them. We reviewed the data set along with our research questions and realized that we need more variables than the suggested number, hence we moved forward with the list of variables decided after data cleaning.





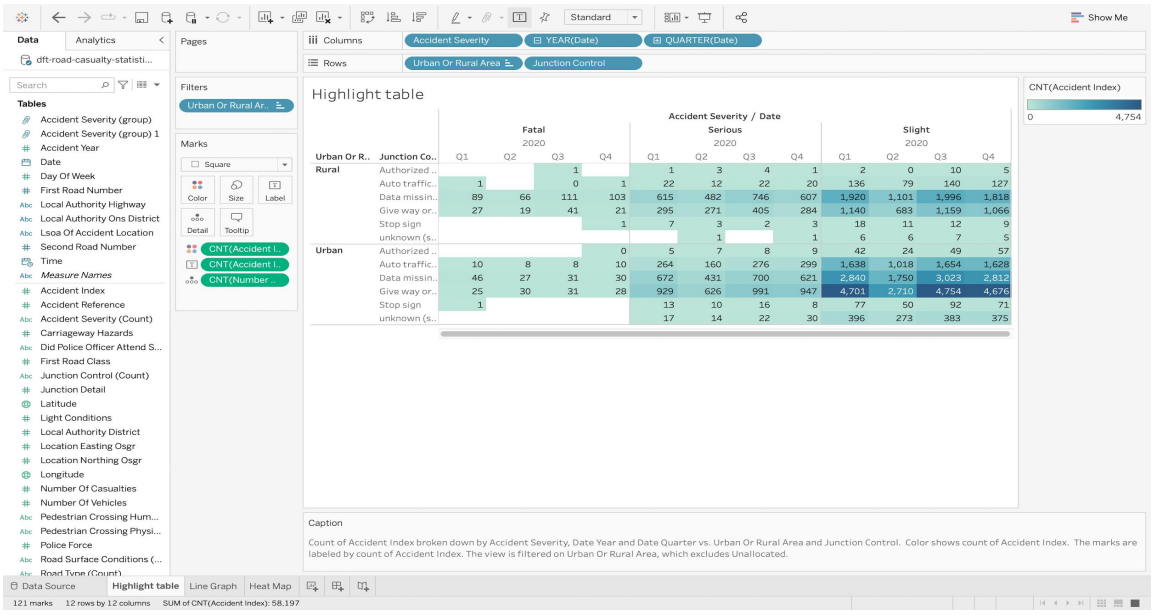
The reasoning for Models/techniques Selection

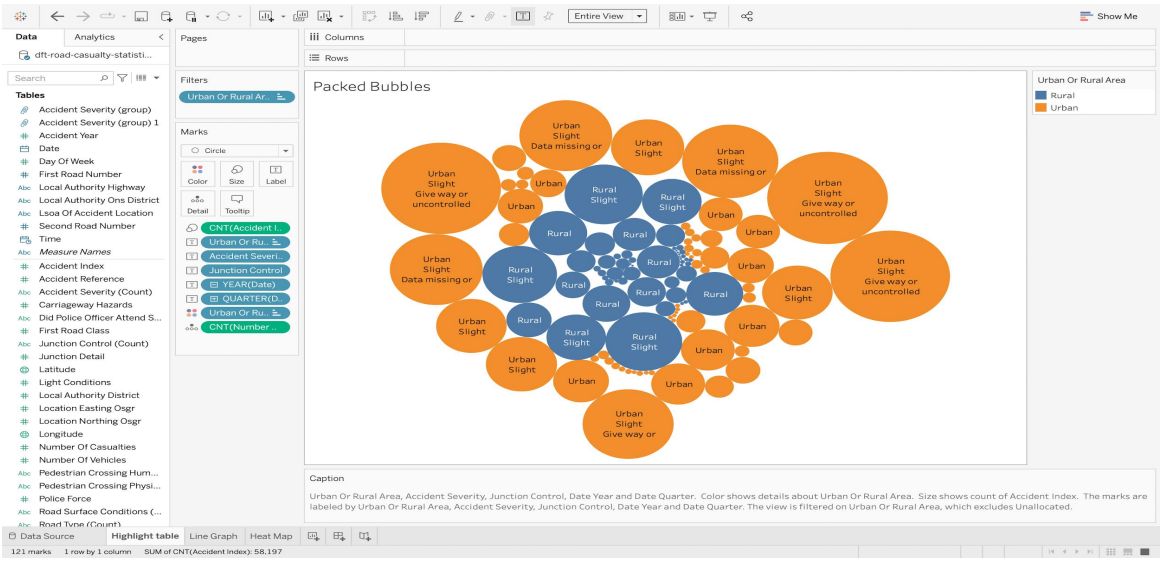
In our project, we used classification decision tree, logistic regression and neural networks for specific reasons. Classification decision tree was chosen as our target variable, accident_severity is a categorical variable and the recursive partition algorithm of decision tree that develops the classification tree eventually will help us in categorizing the accident's severity levels. We used logistic regression and not linear regression as our output was a categorical one and not continuous output. We used neural networks as we wanted to explore into the possible complicated relationships between the variables and also observe how the hidden layers can affect the results.

Data Visualization:

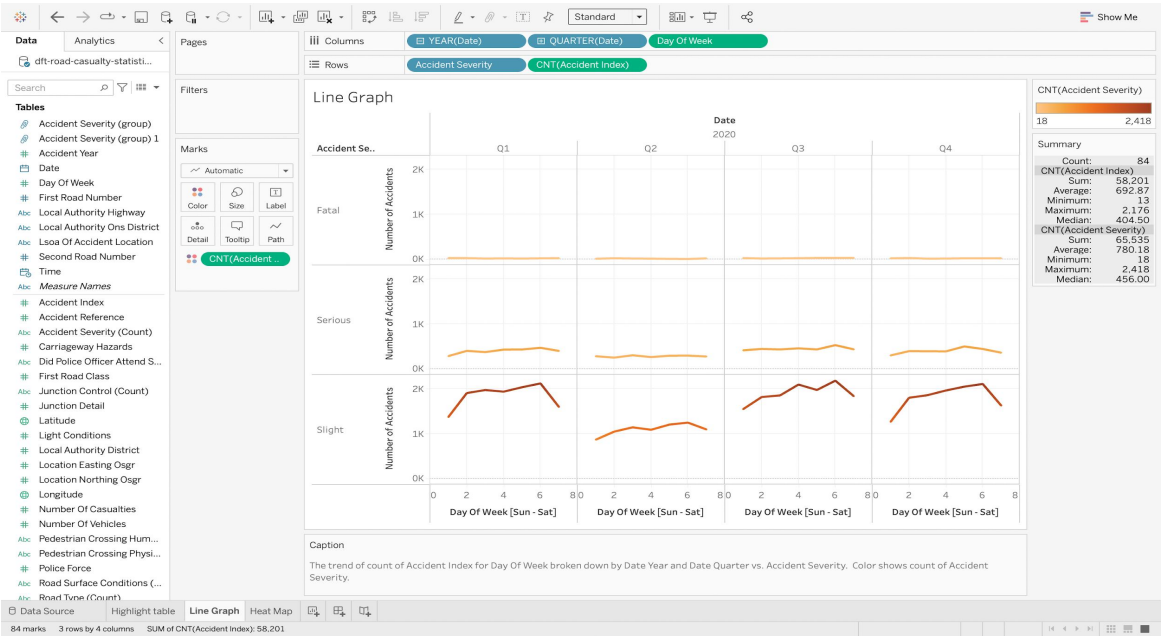
We visualized our data set using tableau. We tried having 3 different perspectives when coming to creating the visualizations.

The first visualization is based on looking at the number of accidents classified by the accident severity in the region - urban or rural on a quarterly basis. We used the highlight table in this case so as to identify the high and low intensity regions easily. We found that comparatively, less number of accidents happened in the quarter 2 of 2020 in terms of months. With respect to accident severity, most accidents were of slight level of severity on the scale of slight, serious and fatal. On the other hand, the majority number of accidents took place in the urban regions compared to the rural regions. We then used packed bubbles to have a better visualization that can easily depict the ratios.

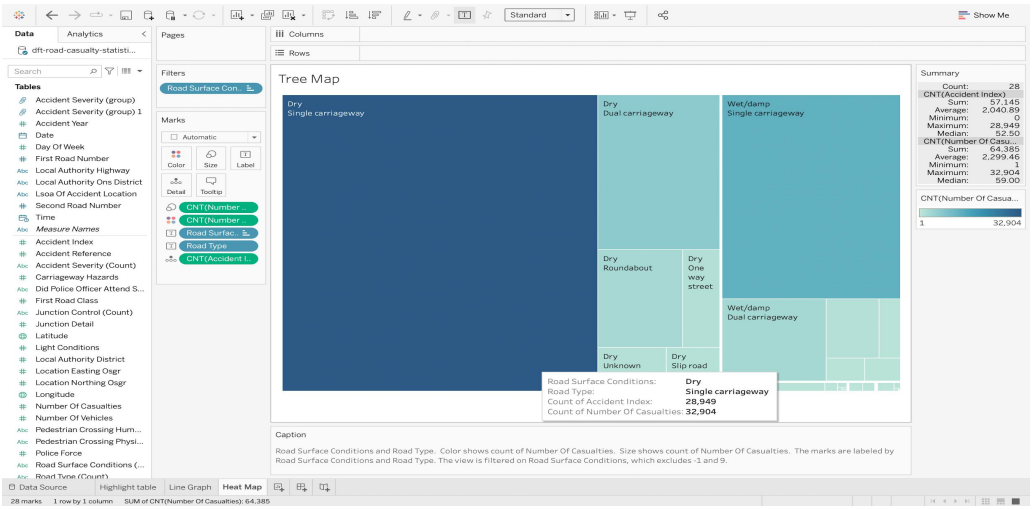




The second visualization is focused on knowing how the number of accidents has ranged in terms of severity on a daily basis of a week in different quarters. We found that the number of accidents, irrespective of the quarter, have seen an increase in general from the beginning of the week and plummeted by the end of the week. On the whole, as we mentioned above, quarter 2 has the lowest number of accidents on a quarterly basis. We used line graphs to visualize this case.

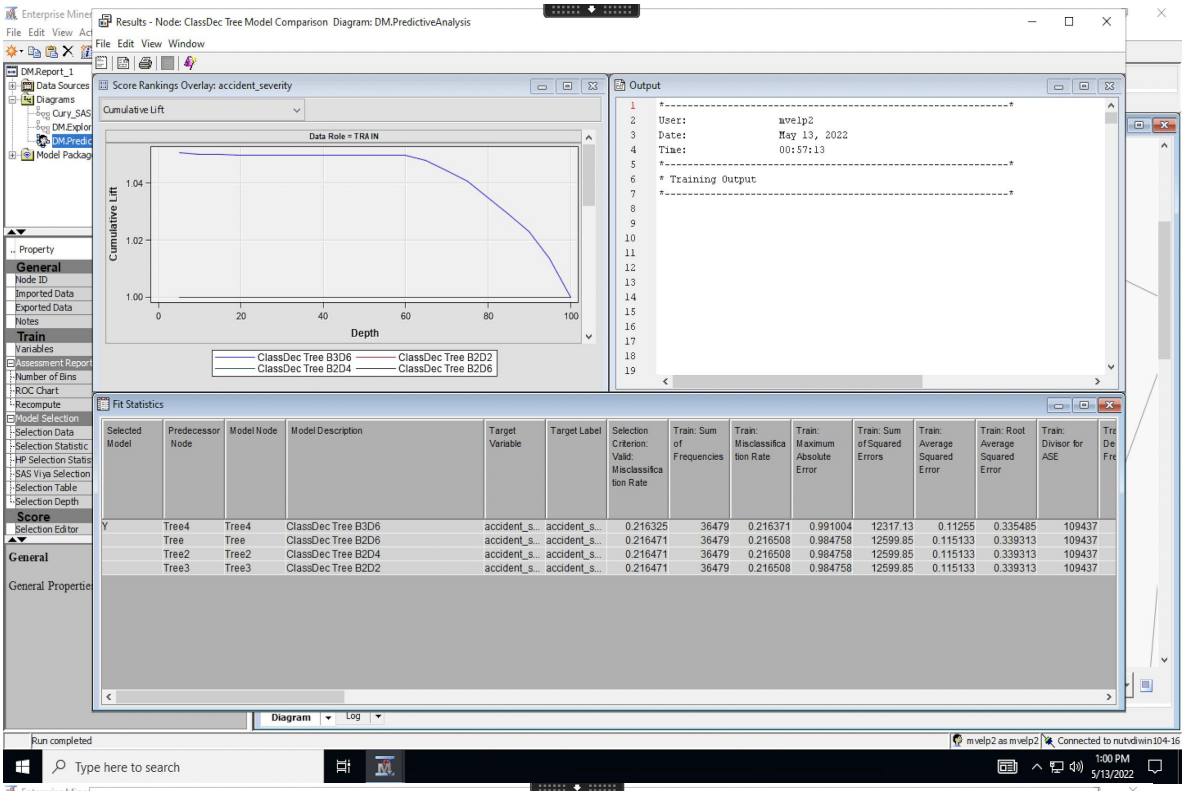


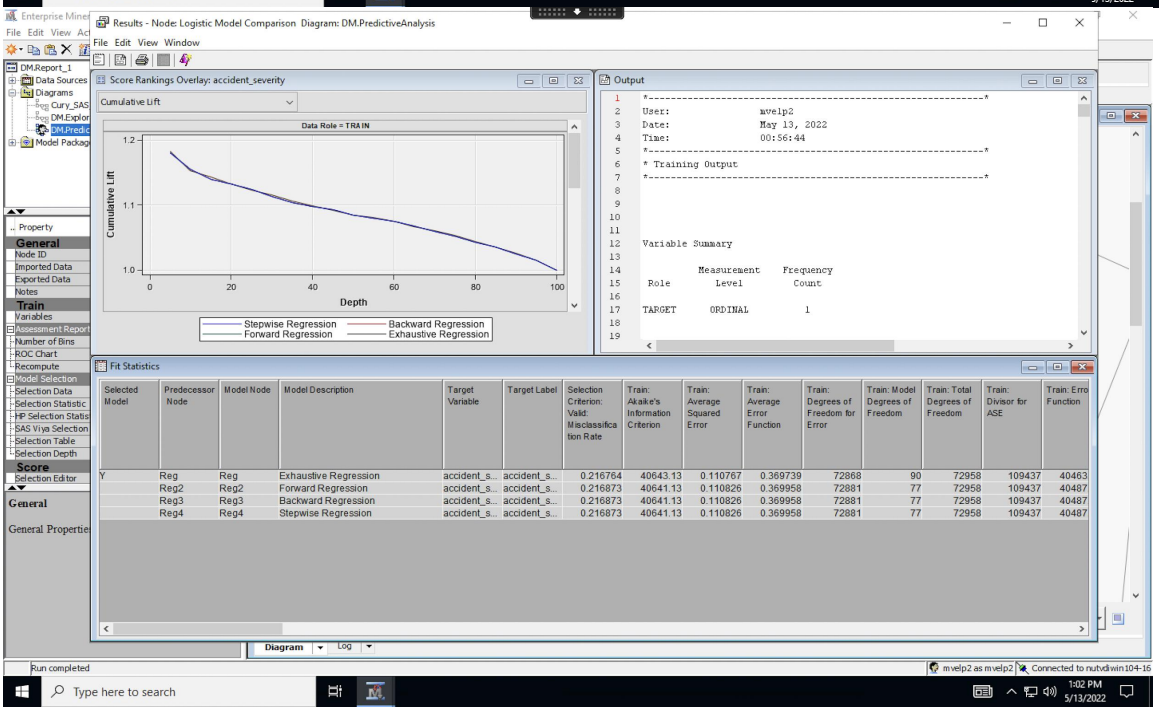
The third visualization is based on scaling the number of accidents based on two different variables. We wanted to see how road surface conditions and road type have contributed to the number of accidents. We used a tree map to visualize this case and found that major causes of accidents in terms of road type were - single carriageway, dual carriageway and roundabout and in terms of road surface conditions were - dry, wet/damp and frost/ice road surfaces.

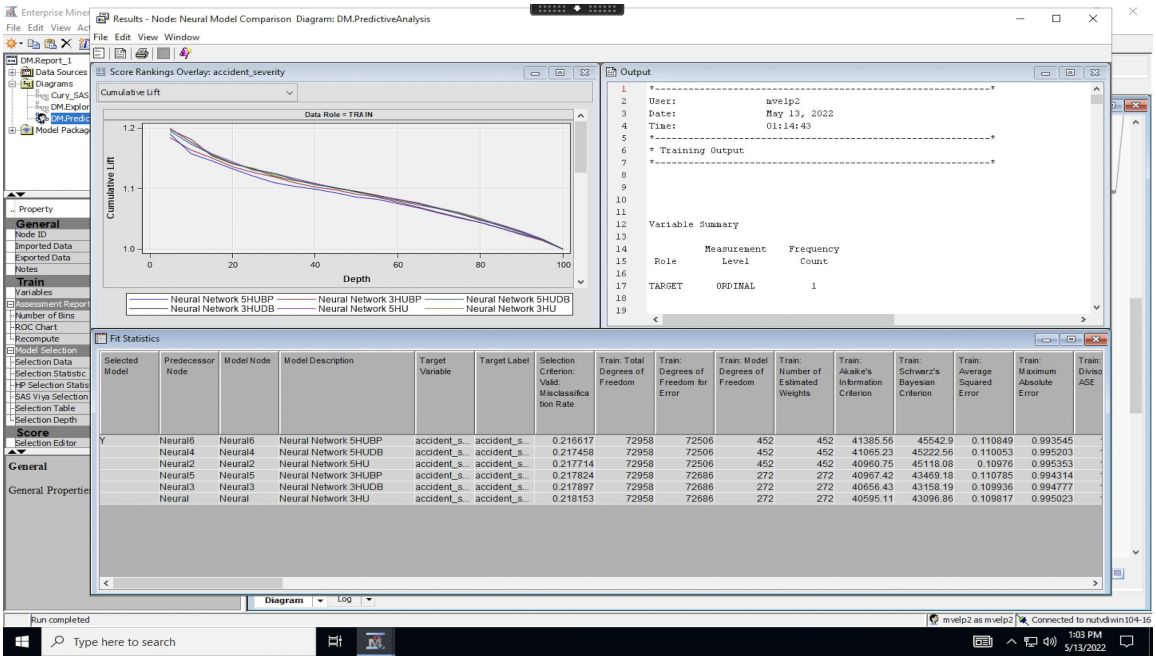


Comparison of at least 3 supervised data mining techniques used for data analysis and developing models for prediction and/or classification:

The supervised techniques that were used for the prediction analysis were the Classification Decision Trees which include B2D6, B2D4, B2D2 and B3D6, Regression Trees which include Exhaustive, Forward, Backward and Stepwise and Neural Networks which include 3HU (Default), 5HU (Default), 3HU-DB, 5HU-DB, 3HU-BP, 5HU-BP. The results of the best models within each data mining technique are attached below. From the results, we can understand that almost all the supervised data techniques gave similar results with very close misclassification rates and overfitting. Apart from slight deviations, even the score rankings matrix for the target variable accident_severity look similar in all the three cases. In terms of average mean square error, classification decision tree and regression have almost same values while neural networks has a very minute difference in terms of the default setting but has similar values with other optimizations.


Results - Node: ClassDec Tree Model Comparison Diagram: DM.PredictiveAnalysis
File Edit View Window
Score Rankings Overlay: accident_severity
Cumulative Lift
Data Role = TRAIN
Cumulative Lift
Depth
ClassDec Tree B3D6
ClassDec Tree B2D2
ClassDec Tree B2D4
ClassDec Tree B2D6
Output
User: mvelp2
Date: May 13, 2022
Time: 00:57:13
* Training Output
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Misclassification Rate
Train: Sum of Frequencies
Train: Misclassification Rate
Train: Maximum Absolute Error
Train: Sum of Squared Errors
Train: Average Squared Error
Train: Root Average Squared Error
Train: Divisor for ASE
Y Tree4 Tree4 ClassDec Tree B3D6 accident_s... accident_s... 0.216325 36479 0.216371 0.991004 12317.13 0.11255 0.335485 109437
Tree Tree ClassDec Tree B2D6 accident_s... accident_s... 0.216471 36479 0.216508 0.984758 12599.85 0.115133 0.339313 109437
Tree2 Tree2 ClassDec Tree B2D4 accident_s... accident_s... 0.216471 36479 0.216508 0.984758 12599.85 0.115133 0.339313 109437
Tree3 Tree3 ClassDec Tree B2D2 accident_s... accident_s... 0.216471 36479 0.216508 0.984758 12599.85 0.115133 0.339313 109437
Diagram
Log
Run completed
mvelp2 as mvelp2
Connected to nutvdiwin104-16
Type here to search
1:00 PM
5/13/2022

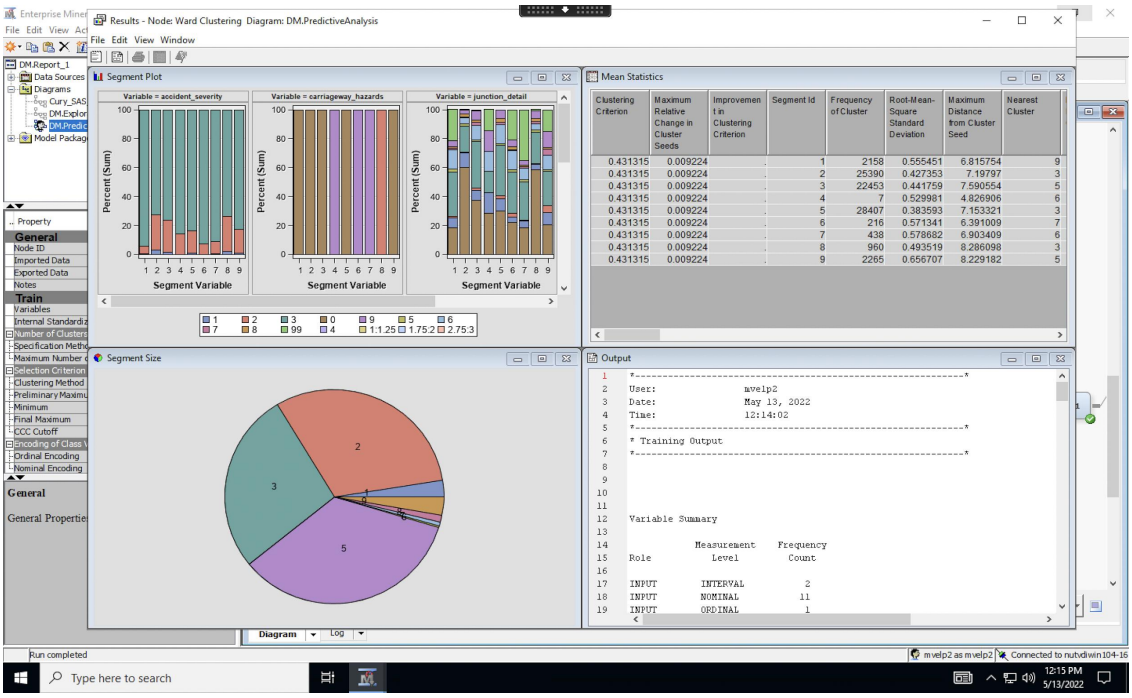

Results - Node: Logistic Model Comparison Diagram: DM.PredictiveAnalysis
File Edit View Window
Score Rankings Overlay: accident_severity
Cumulative Lift
Data Role = TRAIN
Cumulative Lift
Depth
Stepwise Regression
Backward Regression
Forward Regression
Exhaustive Regression
Output
User: mvelp2
Date: May 13, 2022
Time: 00:56:44
* Training Output
Variable Summary
Role Measurement Level Frequency Count
TARGET ORDINAL 1
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Misclassification Rate
Train: Akaike's Information Criterion
Train: Average Squared Error
Train: Average Error Function
Train: Degrees of Freedom for Error
Train: Model Degrees of Freedom
Train: Total Degrees of Freedom
Train: Divisor for ASE
Y Reg Reg Exhaustive Regression accident_s... accident_s... 0.216764 40643.13 0.110767 0.369739 72868 90 72958 109437 40463
Reg2 Reg2 Forward Regression accident_s... accident_s... 0.216873 40641.13 0.110826 0.369958 72881 77 72958 109437 40487
Reg3 Reg3 Backward Regression accident_s... accident_s... 0.216873 40641.13 0.110826 0.369958 72881 77 72958 109437 40487
Reg4 Reg4 Stepwise Regression accident_s... accident_s... 0.216873 40641.13 0.110826 0.369958 72881 77 72958 109437 40487
Diagram
Log
Run completed
mvelp2 as mvelp2
Connected to nutvdiwin104-16
Type here to search
1:02 PM
5/13/2022

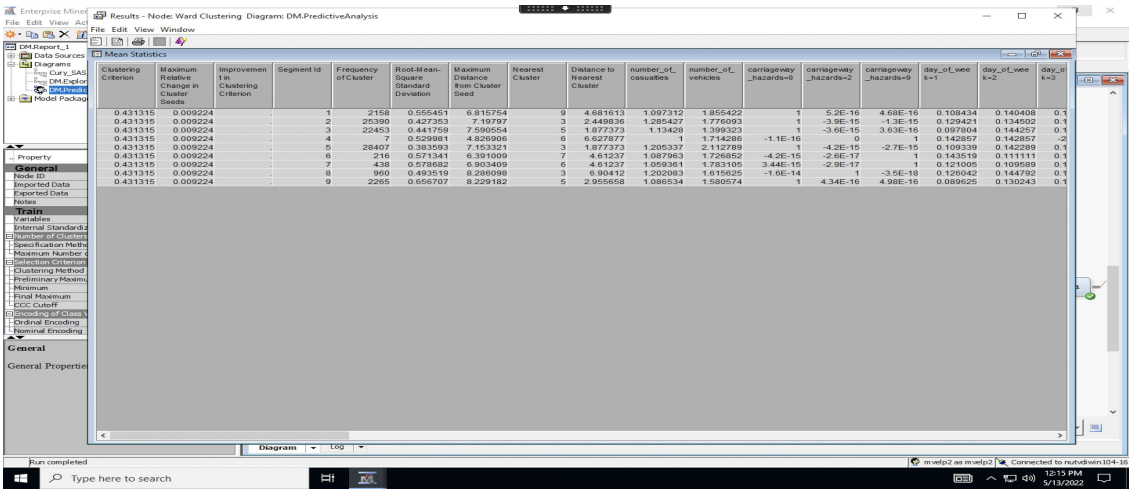


Exploratory analysis using one of the unsupervised data mining techniques:

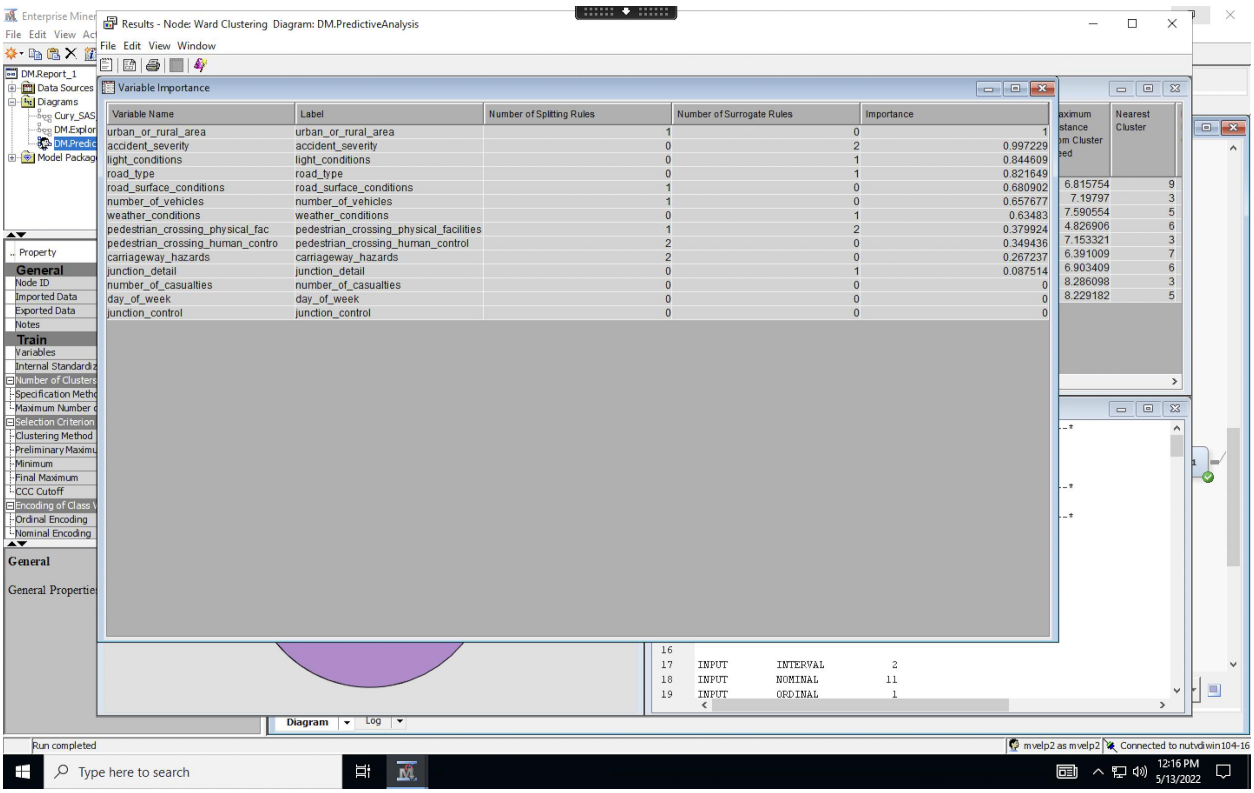
When starting off with exploratory analysis, we rejected a few variables after trying out clustering a few times as a few of the variables were not helping much in a better clustering. We either got 2 or 20 clusters, which is why, we rejected the variables first_road_class, first_road_number, speed_limit, second_road_class, second_road_number, special_conditions to have a better clustering model but also a model which has some scope for improvement.

Initially with ward clustering, the results came out with 9 clusters. With cluster 5, cluster 2 and cluster 3 being the major clusters. Cluster 5 is the biggest one with a percentage of 34.52% of the records which come to 28407 records. Cluster 4 is the smallest cluster with just 7 records. We rejected the variables - first_road_class, first_road_number, second_road_class and second_road_number, speed_limit, time, special_conditions_onsite based on our domain knowledge of the dataset.





Variable Importance - In this case, the most important variable came out to be urban_or_rural_region (variable importance 1). While the three variables - number_of_casualties, day_of_week, junction_control do not have any importance (variable importance 0), other top two important variables are accident_severity and light_conditions. The two least important variables are carriageway_hazards and junction_detail with variable importance of 0.26, 0.08 respectively.



Best Chosen Model for Each Supervised Data Mining Technique

Out of all the models that were used for predictive analysis, classification tree model (B3D6) was identified as the best one by SAS Enterprise Miner. The best models of each supervised technique with the respective misclassification rates are as follows:
Classification Decision Tree - **B3D6** with a misclassification rate of 21.63%
Regression Tree – **Exhaustive Regression** with a misclassification rate of 21.67%
Neural Networks - **5HUBP** with a misclassification rate of 21.66%

Analysis of results of the best model of each supervised data mining technique based on performance evaluation metrics:

Classification Tree

Data Partition	Misclassification rate			
	ClassDec Tree B2D6	ClassDec Tree B2D4	ClassDec Tree B2D2	ClassDec Tree B3D6
Training Set	21.65	21.65	21.65	21.63
Validation Set	21.64	21.64	21.64	21.63
Test Set	21.65	21.65	21.65	21.6

Regression Tree

Data partition	Misclassification rate			
	Exhaustive	Forward	Backward	Stepwise
Training set	21.631	21.639	21.639	21.639
Validation Set	21.676	21.68	21.68	21.68
Test Set	21.64	21.66	21.66	21.66

Neural Networks

Data Partition	Misclassification rate					
	3HU	5HU	3HUDB	5HUDB	3HUBP	5HUBP
Training Set	21.656	21.691	21.620	21.659	21.661	21.650
Validation Set	21.81	21.77	21.78	21.74	21.78	21.66
Test Set	21.686	21.67	21.682	21.70	21.75	21.689

From the details above, we can see that the best models are the ones with the minimum deviation between training set, validation set and test set data along with the one having the least misclassification rate. They turn out to be B3D6 in case of Classification decision tree, Exhaustive in case of regression and 5HU-BP in case of Neural networks all of which coincide with the ones selected as best models by SAS Miner too.

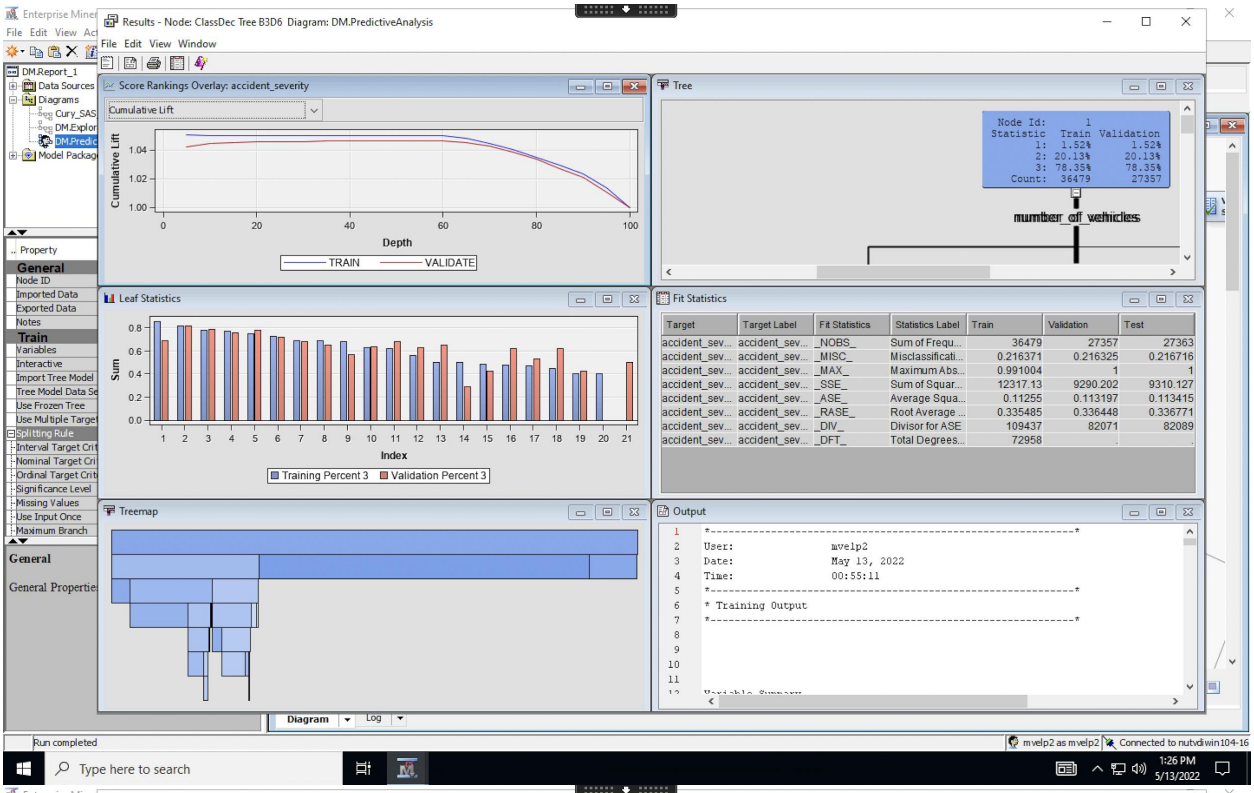
Best Chosen Model To Answer Our Research Questions

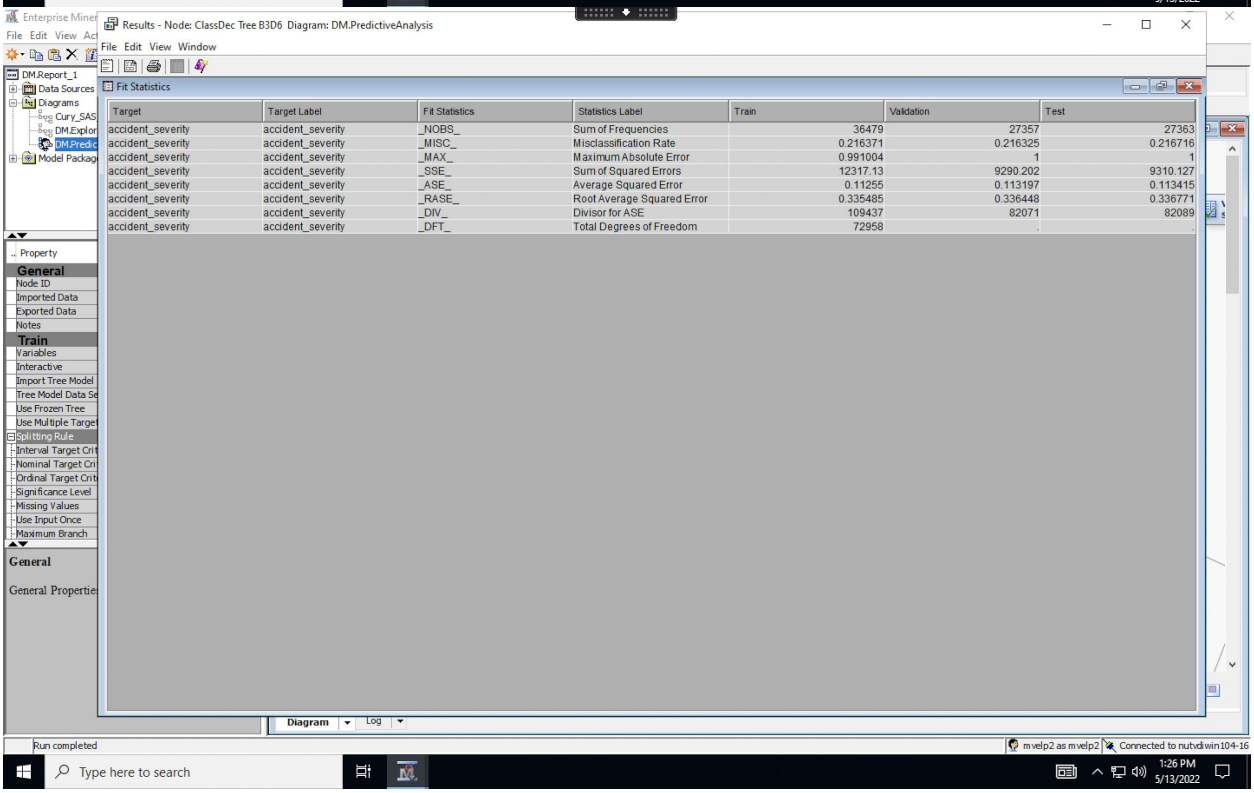
The chosen model B3D6 is the best one for analysis of the project because it has the least misclassification rate and overfitting comparing all other models. The maximum level represents the maximum depth of the tree, and the default value is 6. Also, the accuracy increases with increase in depth of the tree.

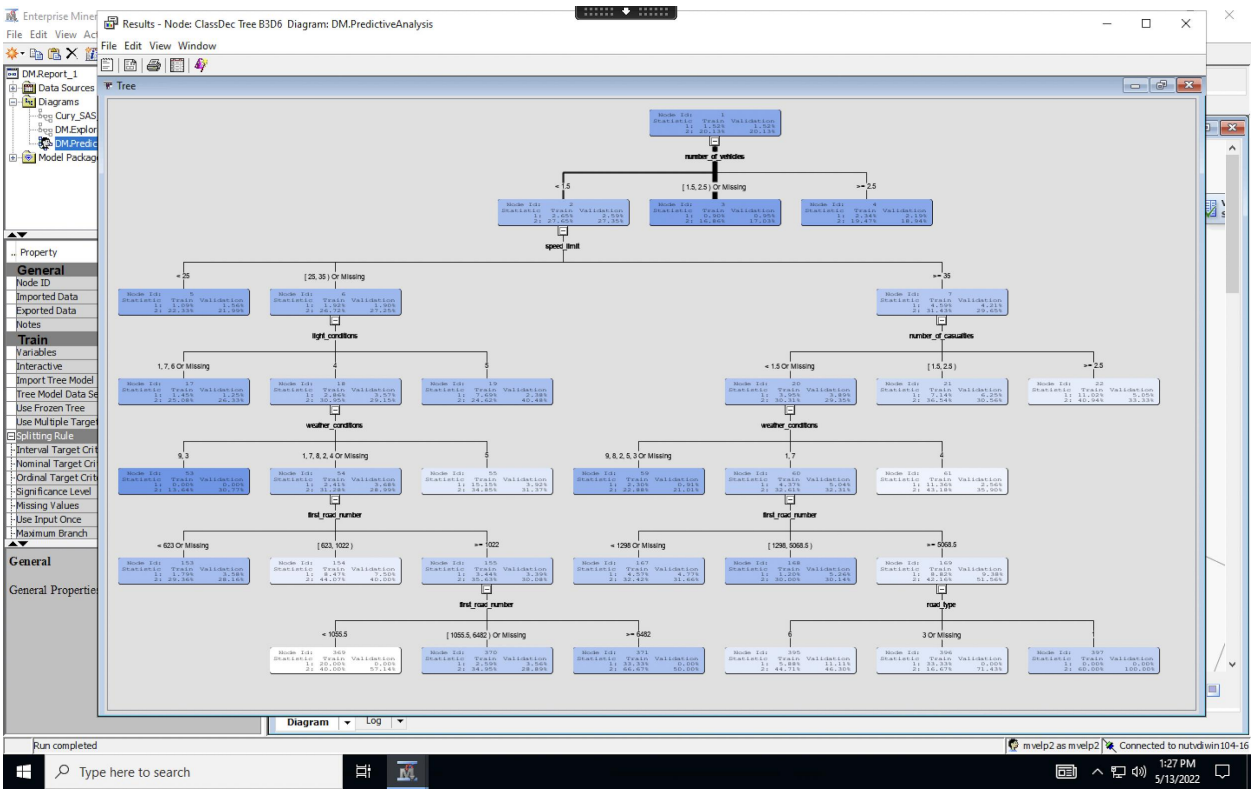
Detailed Analysis (for supervised techniques) of results derived using the best model:

We can start with analyzing the cumulative lift which tells us that there is very less overfitting, which is a good result. From the tree window, we can see that the variables number_of_vehicles, speed_limit, light_conditions, number_of_casualties, weather_conditions, first_road_number and road_type acted as decision nodes. We tried pruning the tree further by decreasing the depth to view much more interesting results. The leaf statistics show that the training and validate percent are similar or close in most of the cases apart from one or two

deviations, which is also a good result. The fit statistics show us the significant misclassification rate which is 21.63 while the average mean square error is around 11.3. This brings our accuracy of the model to 78.37 which is considered satisfactory by the industry standards.







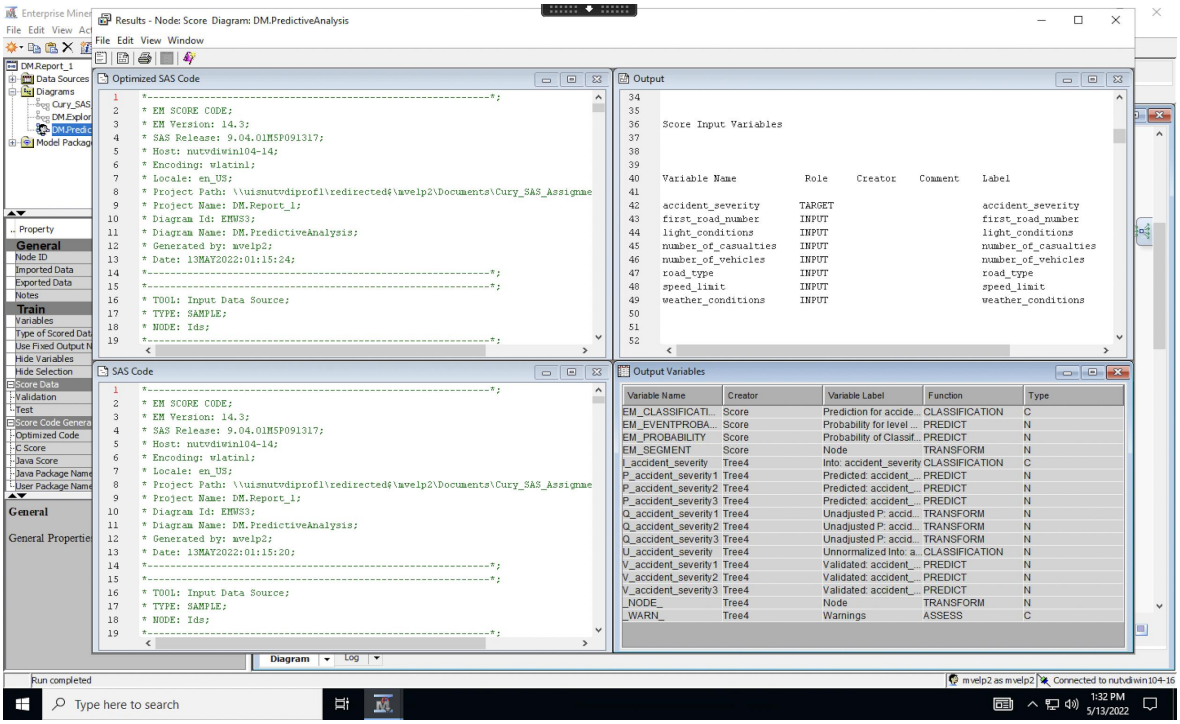
Scoring the model on new data and discussing the results of the score node:
Scored data analysis:

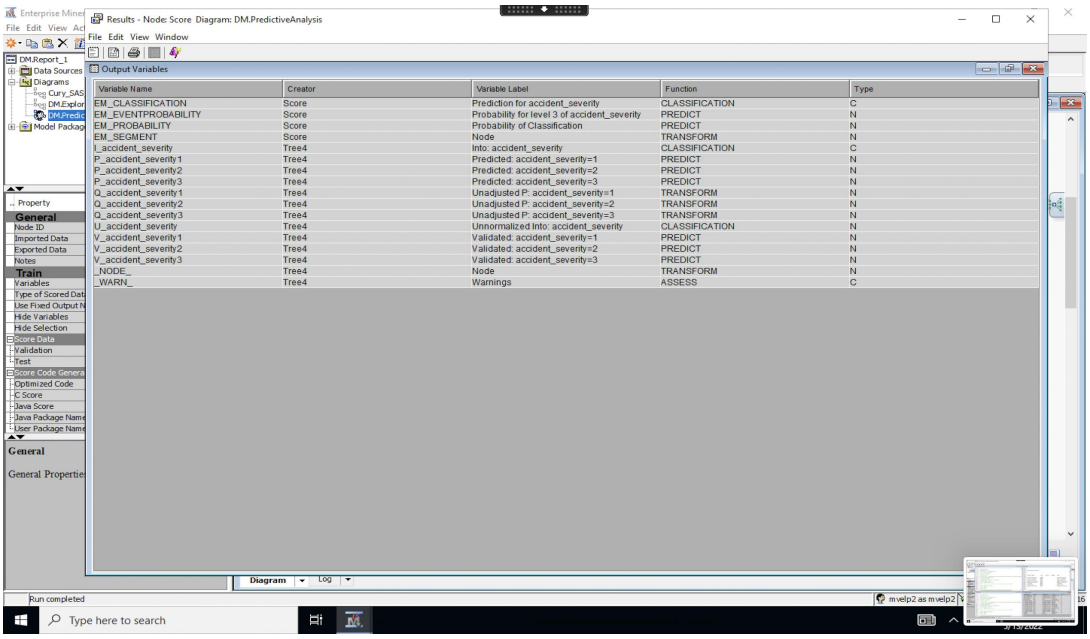
Accident severity 1 - 278/18240
True 1 – 0 (True Positives)
False 2 – 2
False 3 – 276

Accident severity 2 - 3671/18240
True 2 – 30 (True Positives)
False 3 - 3640
False 1 - 1

Accident Severity 3 - 14291/18240
True 3 – 14270 (True Positives)
False 2 - 20
False 1 - 1

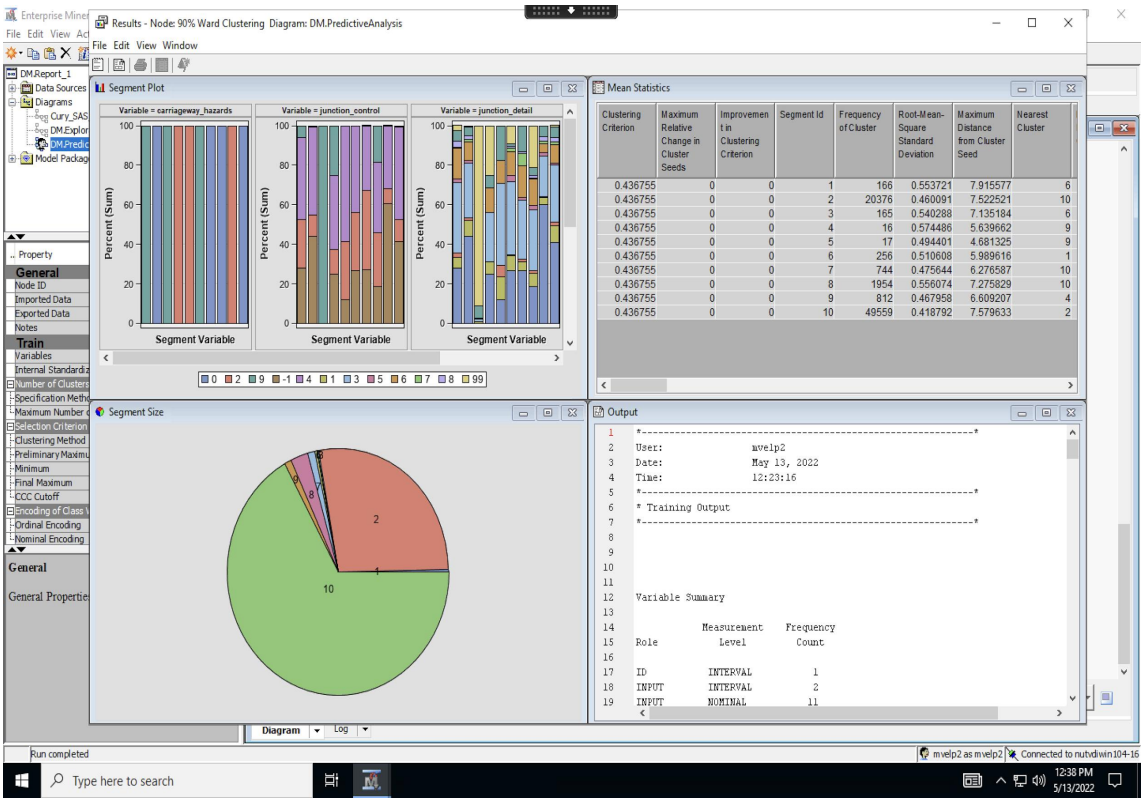
After the calculations from the above results, we obtained the accuracy to be around 78% too. From the scoring results, we saw that the variables first_road_number, light_condtions, number_of_vehicles, number_of_casualties, road_type, speed_limit and weather_conditions were used as input variables to score the data.





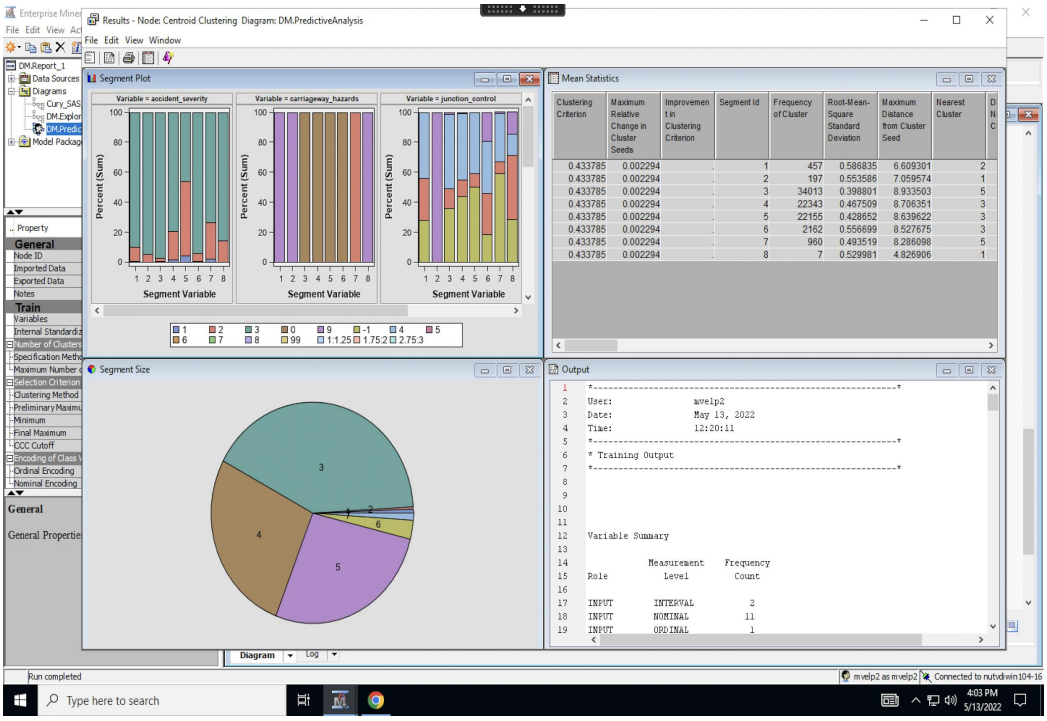
Detailed analysis of exploratory analysis including stability of clusters and/or strength of rules in case of unsupervised data mining techniques:

As mentioned above, the results from ward clustering were not satisfactory as clusters above a number of 6 are hard to interpret. Hence, we tried centroid, average clustering techniques along with checking for stability followed by k-means clustering.

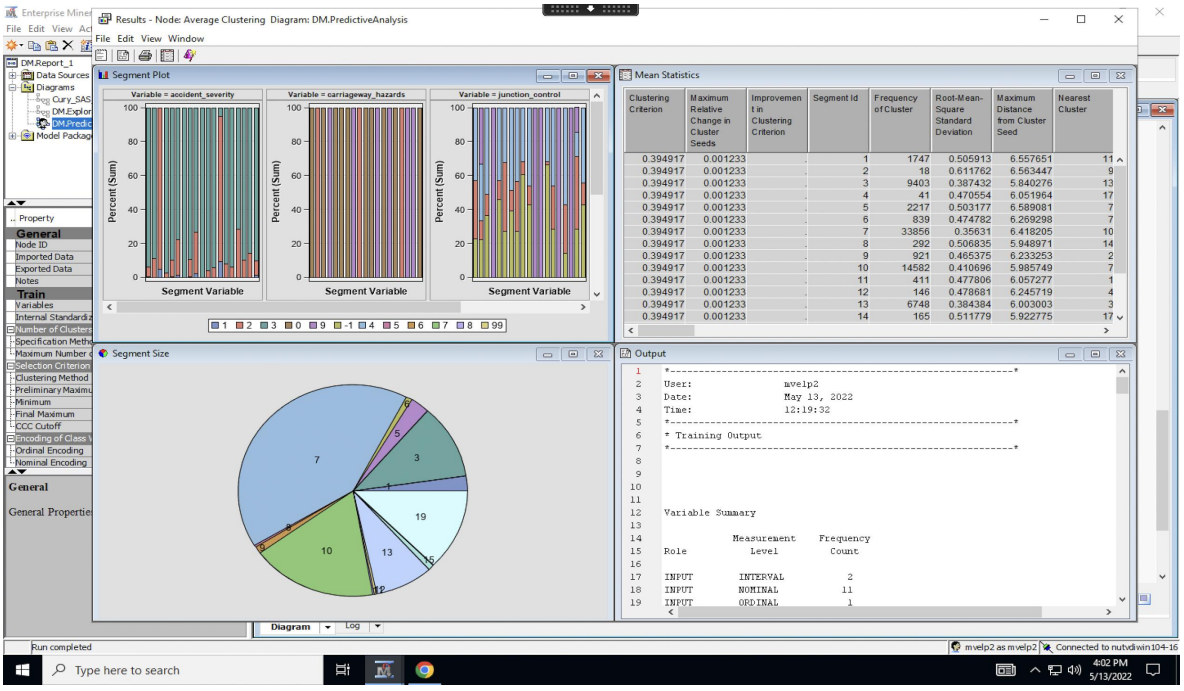


Upon performing ward clustering on the sample data (90% of the original data set), we got 10 clusters in the results. Cluster 10 is the biggest cluster with 49559 records, whereas cluster 4 is the smallest cluster with 16 records. Talking about the stability of the ward clustering model, it could be said that the clusters created by the ward clustering with respect to our dataset are stable based on the number of clusters with the 90% ward clustering(10) is in the range +-2 of the number of clusters from original ward clustering(9) and when looking at the mean statistics of 90% ward clustering, we can see that there is not much divergence from that of the results of the original ward clustering.

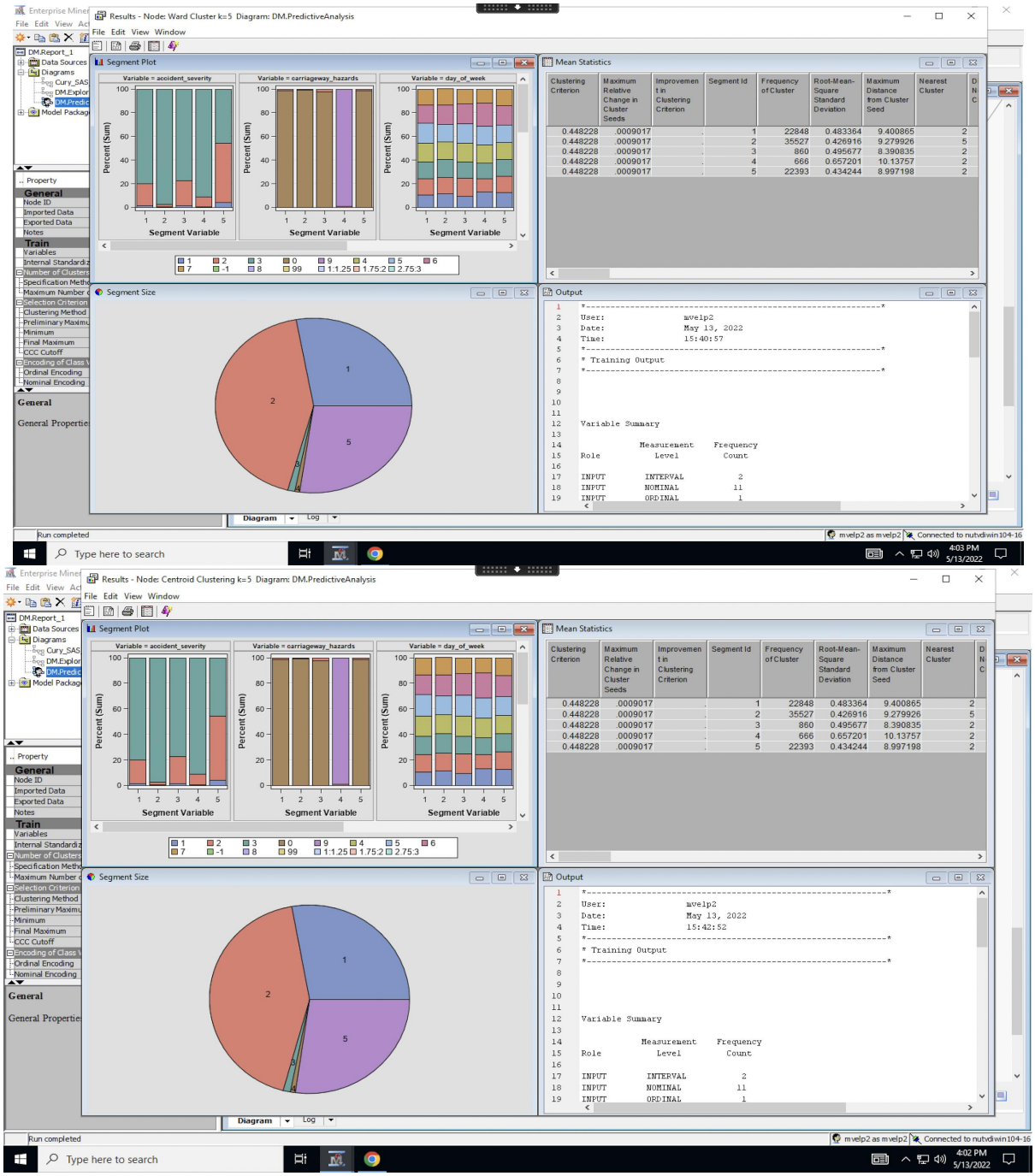
Centroid Clustering and 90% Centroid Clustering - Centroid clustering gave 8 clusters as results, with cluster 3 being the largest having 34013 records, cluster 8 being the smallest with just 7 records. On the other hand, with 90% centroid clustering, the results were very similar to that of 90% ward clustering. 90% Centroid clustering gave 10 clusters which falls in the range of +-8, which makes the clusters generated by Centroid clustering as stable. Upon observation, the mean statistics of both centroid clustering and 90% centroid clustering are closer to each other which makes it easier to call these clusters as stable.



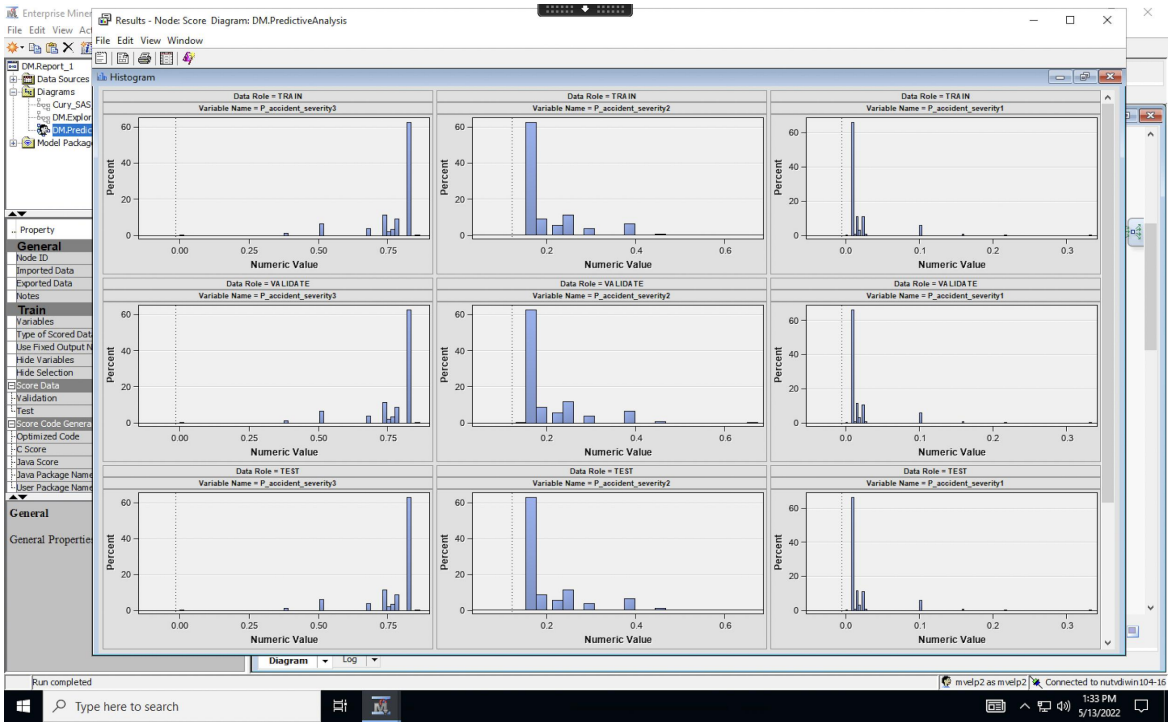
Average Clustering gave 19 clusters in the results, with cluster 7 the largest containing 33856 records, and cluster 18 being the smallest with just 7 records. 20 clusters were the result of 90% average clustering with cluster 3 being the biggest one with 23510 records and cluster 5 being the smallest with 7 records. The difference between the number of clusters from average and 90% average clustering is less than two, which indicates that the clusters generated by average clustering are stable. The mean statistics of both of these clustering algorithms also look decent enough to call these clusters stable.

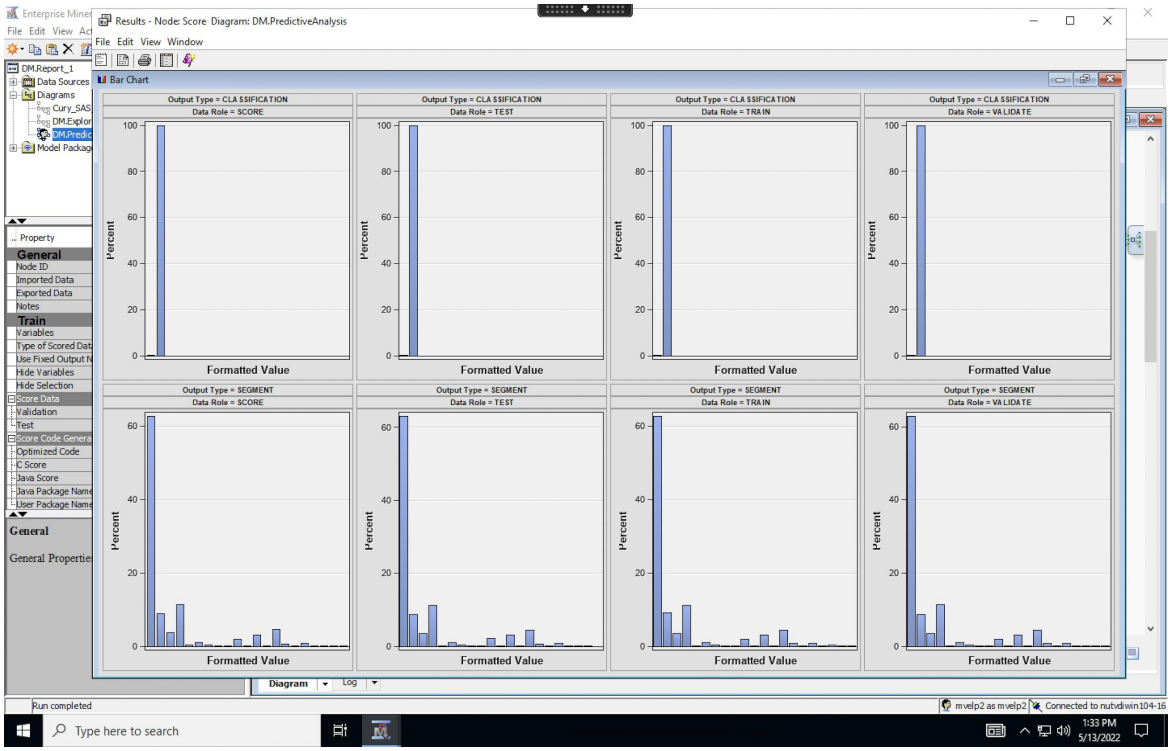


After testing the stability for all the three clustering models, we wanted to try non-hierarchical clustering as we have tried and examined the results of hierarchical clustering. In non-hierarchical clustering (k-means), we initially tried out with different values of k, and finally observed that k=5 gave the best results. Interestingly, all the models, when tried with k=5 resulted in 5 clusters and gave very similar results. Additionally, all the models when tried on sampled data gave similar results to that of 90% ward clustering.



Reporting the final results using data visualization (e.g graphs pf scoring results):





References:

Jayasudha, K., & Chandrasekar, C. (2009). AN OVERVIEW OF DATA MINING IN ROAD TRAFFIC AND ACCIDENT ANALYSIS.

Tavakoli Kashani, A., Rabieyan, R., & Besharati, M. M. (2014). A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. *Journal of safety research*, *51*, 93–98. https://doi.org/10.1016/j.jsr.2014.09.004

Hassan, Q. (2022, April 15). *Road Satefy Data*. Kaggle. Retrieved May 13, 2022, from https://www.kaggle.com/datasets/qasimhassan/reducing-the-number-of-high-fatality-accidents?select=accident-data.csv

Department for transport. (n.d.). Road Safety Data. Retrieved May 13, 2022, from https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data