**Bayesian Modeling Workshop on US Vehicle Crash Data**

**AIM**

* To Build a Naïve Bayes network and Tree Augmented Naïve Bayes network to predict the likely injury level for vehicle occupants and then compare the results
* To Interact with GeNie Bayesian Network and understand the factors impacting vehicle safety

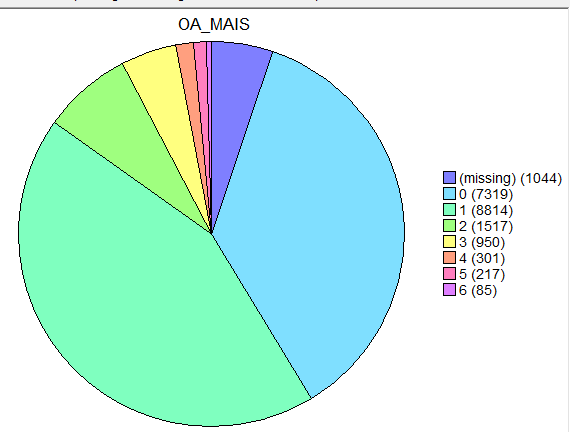
**Data Preprocessing**

Total 20247 observations was considered for building the model. The rows with missing data across the fields were removed and sanitized .The data was assured to have uniqueness , accuracy, consistency and integrity by performing the standard data cleansing methods.

**Data Cleansing Approaches**

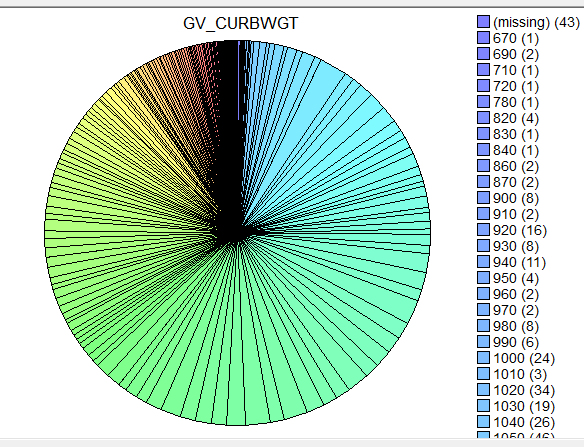
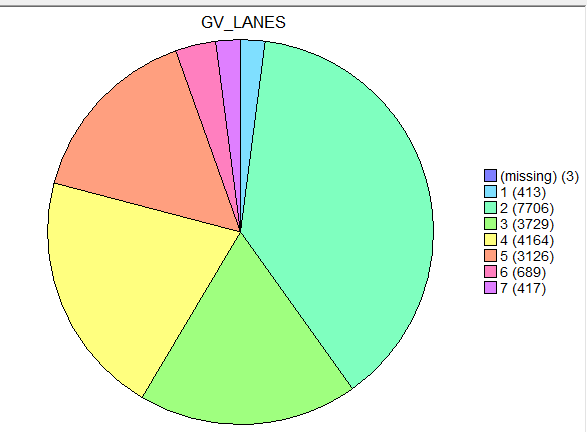
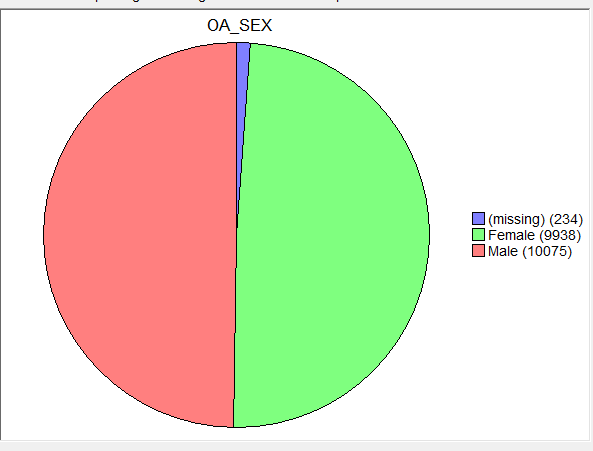
* **Deletion**

The target Prediction variable (OA-MAIS) was observed to contain 1044 missing values out of 20247 records present. List-wise deletion method was used to delete those 1044 missing rows of data since it was only 5% representative of the data set.



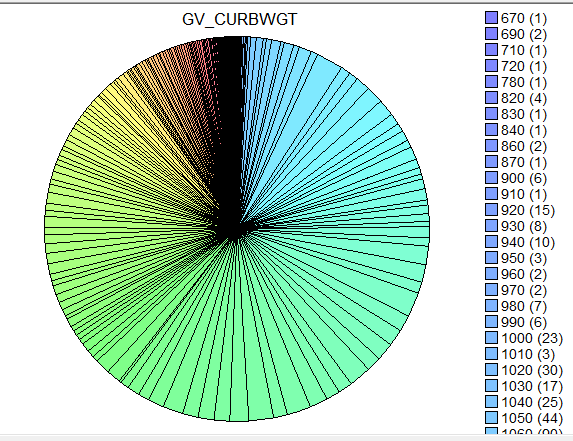
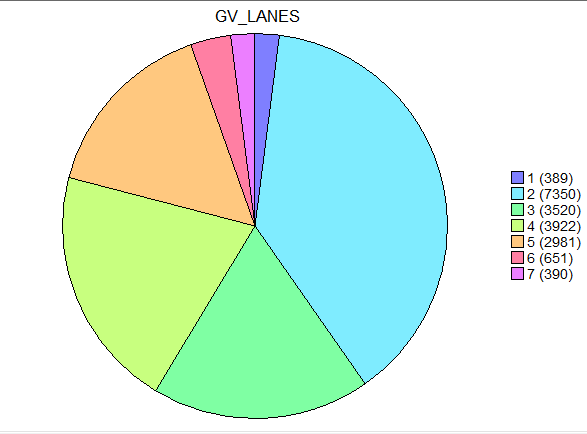
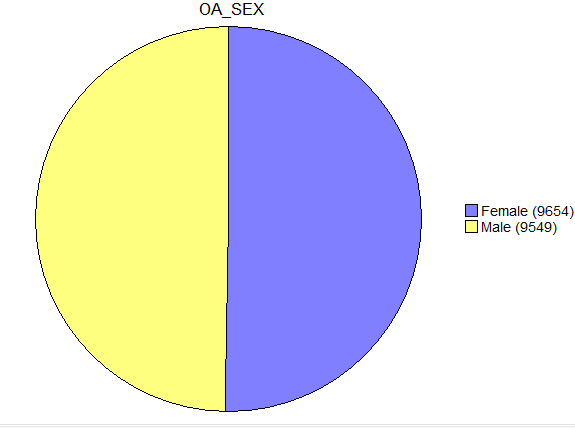
***Fig:*** *Missing data in Target Variable*

* **Imputation**
* Mean Imputation was performed to fill in the missing data for continuous data field columns
* Mode Imputation was performed to fill in the missing data for nominal and ordinal data filled columns
* Maximum Likelihood imputation was done to fill in the missing data for categorical data field   
  columns

***Fig****:Left to Right( continuos data , ordinal data, categorical) data with missing values*

* After Imputation following data distribution was observed.

***Fig****:Left to Right( continuos data , ordinal data, categorical) after data Imputation*.

**Data Binning**

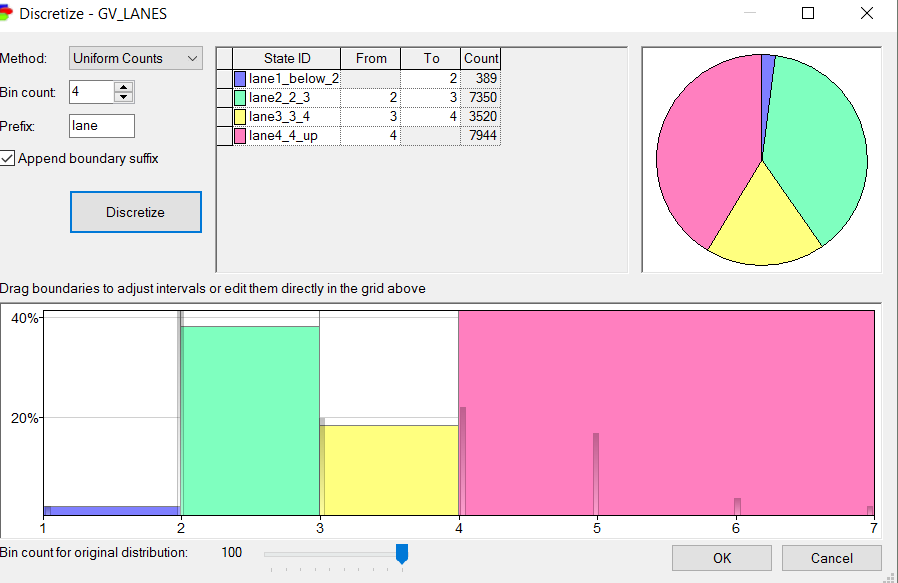
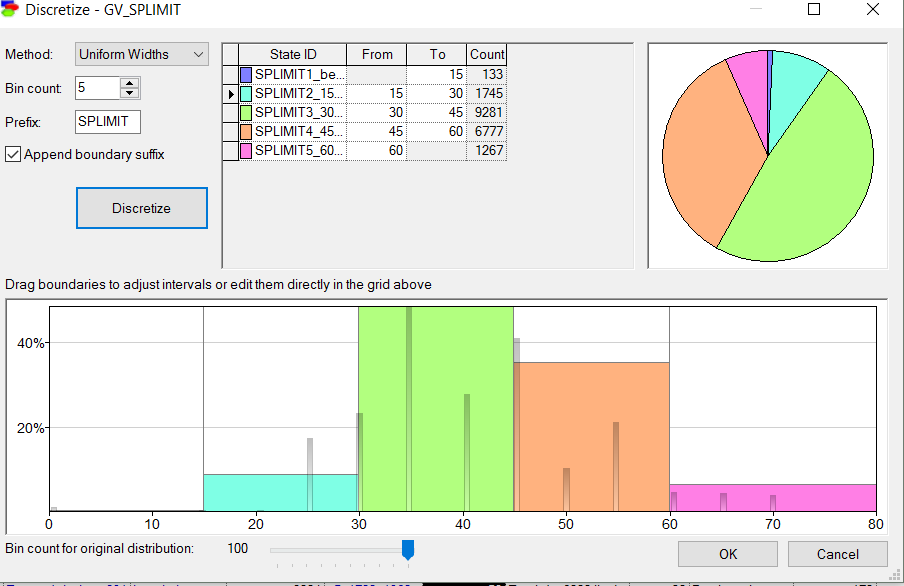
We utilized the knowledge about cars in the best possible ways to discretize the dataset using appropriate binning methods that helped us to get the best predictive model performance.

* The US crash dataset contained 21 variables out of which 14 variables were continuous variables while others were categorical.
* Data Binning technique (It is a way to group more or lesser number of continuous values into a smaller no of bins) was used for pre-processing the continuous data.
* By the use of unsupervised binning methods and some subject knowledge about the cars in hand, numerical variables were converted into their categorical counterparts
* In addition to this outliers were identified and invalid or missing values was removed .Thus, improving accuracy of the model by reducing noise or non-linearity.

.

**Binning Method Used**

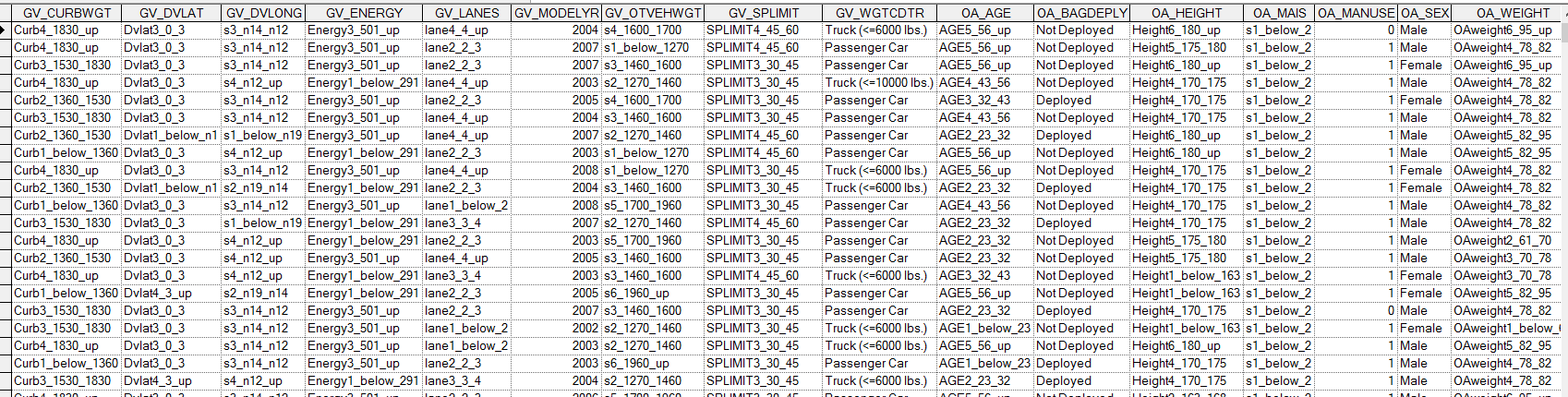
* Uniform Width Binning
* Uniform Count Binning

***Fig****: Illustration of uniform count method and uniform width method*

In the above figure, road\_lanes have been divided using uniform count binning method and car\_speed\_limit using uniform width method.

Final dataset after applying all data cleaning techniques



***Fig:*** *Final data set for model building*

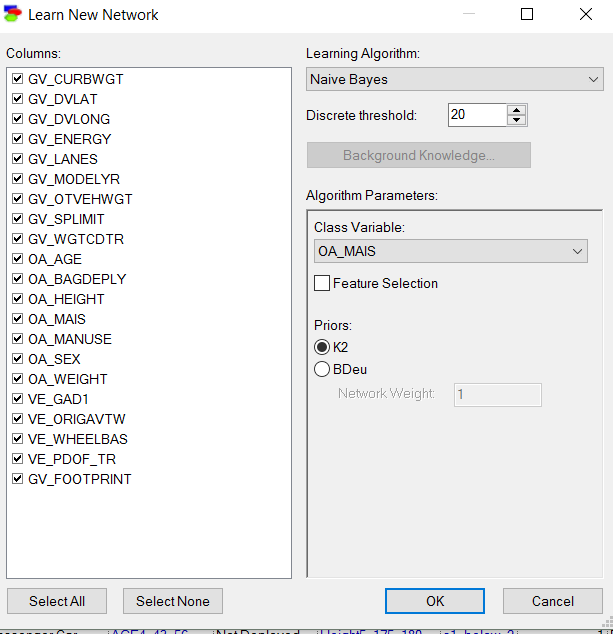
**Preparing the data for Test and Validation**

The data was split into Test and Train data in the ratio of 70:30 so as to project proper representation of the entire occupants.

Since the data set was of moderate size we use K fold cross validation method for separating the test and train data

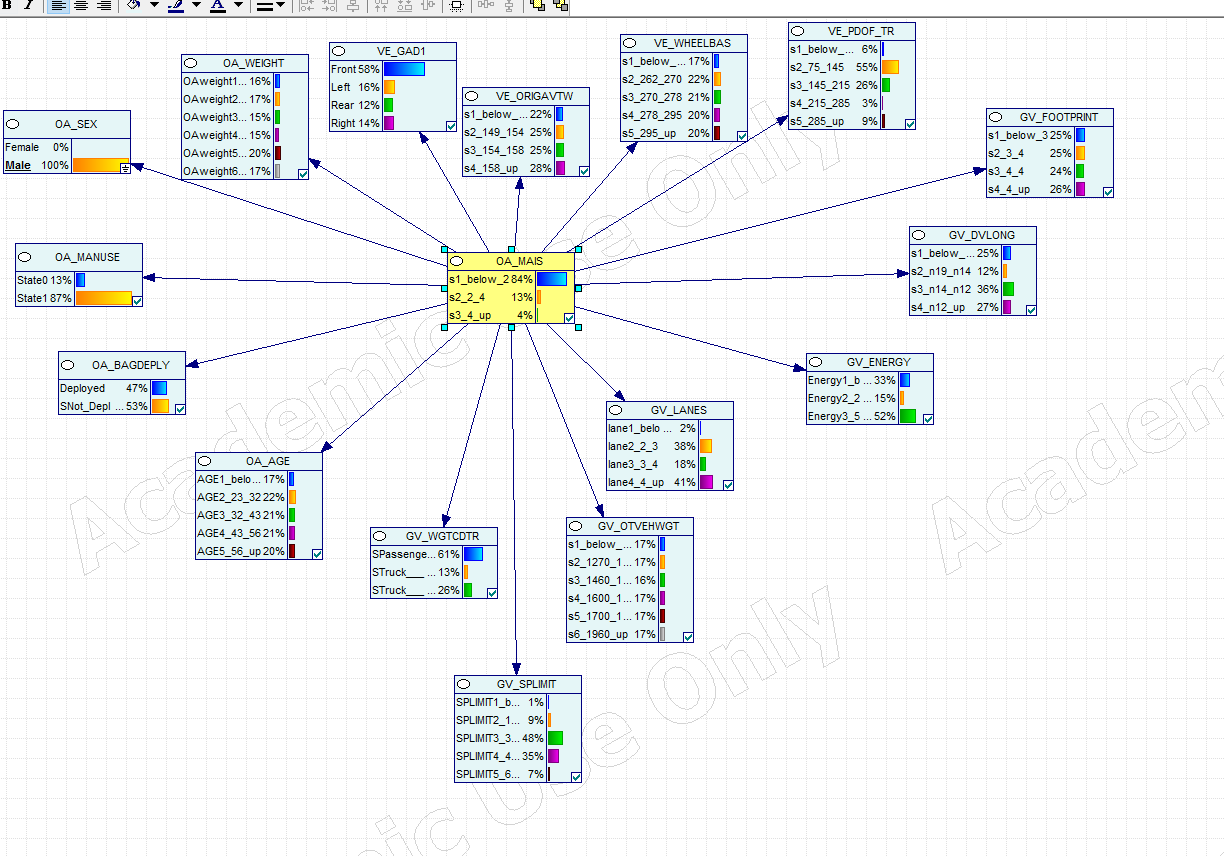
**Model Building**

After having sanitized dataset for prediction of “Injury levels” of vehicle occupants, model was built using Naïve Bayes prediction.



***Fig:***  *selection of Naïve Bayes as prediction model in GeNie Software*

Naïve Bayes Model was built after setting the trigger as male for the variable OA\_SEX. The following network was observed to have probability of “minor or no injuries” as 84%.



***Fig:*** *Detailed Naïve Bayes network diagram for predicting injury level of vehicle occupants.*

***Inference***

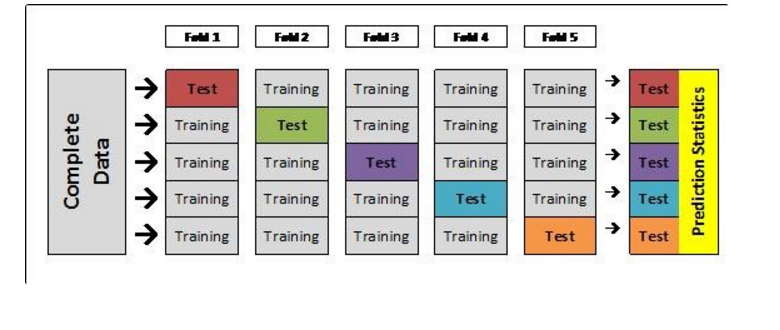
The following inferences can be made from the model:

* When the trigger for the category “sex” was set as “male” ~ 84% of the vehicle occupants didn’t suffer any injury or sustained minor /moderate injuries, 13% suffered serious or critical injuries and remaining could not survive.
* Metrics also show that 58% of vehicle collisions happen in the vehicle frontal region.
* It also shows that 87% of occupants didn’t use seat belts.

**Model Testing**

We have used k fold cross validation method to validate our model. We selected fold count of 5 and seed count of 250 for our validation. During this validation test, data set was divided into 5 subsets, and the holdout method was repeated for 5 times. Each time, one of the 5 subsets was used as the test set and the other 4 subsets were put together to form a training set. Every data point gets to be in a test set exactly once, and gets to be in a training set 4 times. Then the average error across all 5 trials was computed.

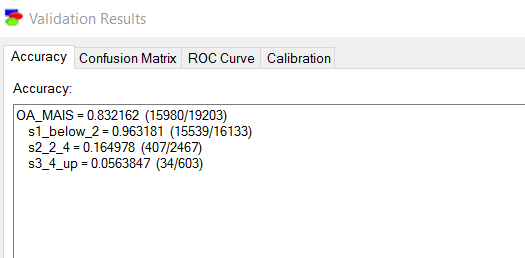
The advantage of using this method is that it matters less how the data gets divided.



***Fig:*** *Schematic representation of 5-fold cross validation.*

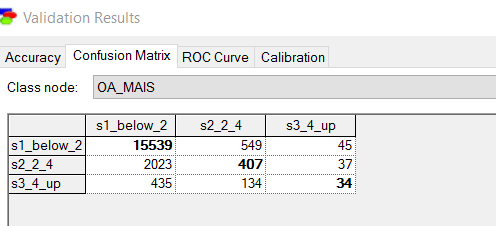
**Validation Results**

**Model Accuracy**



***Fig:*** *Accuracy of the model when trigger set as male*

**Confusion Matrix**

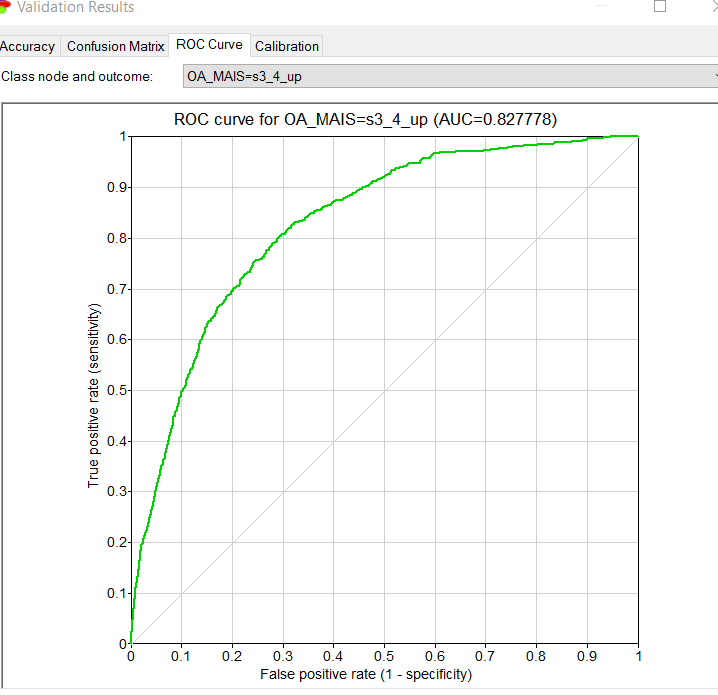
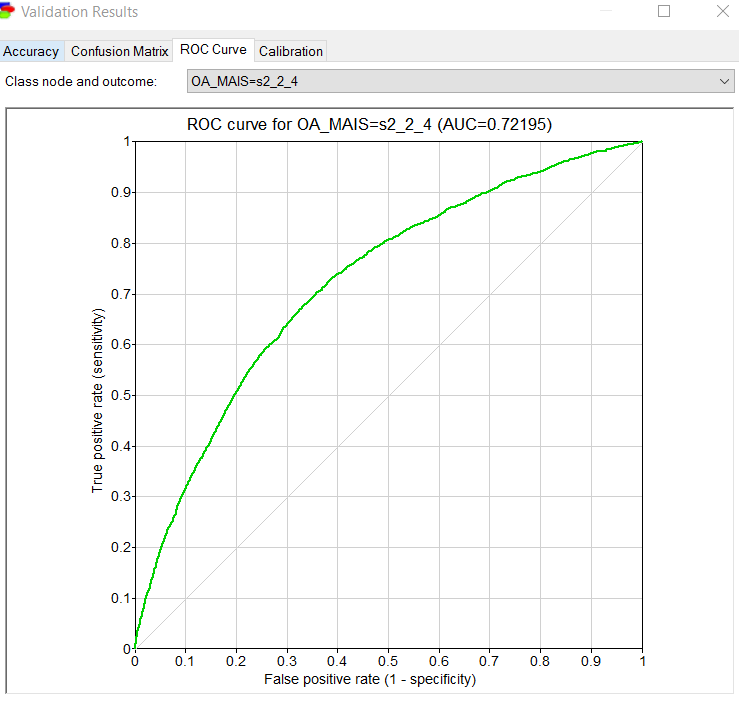
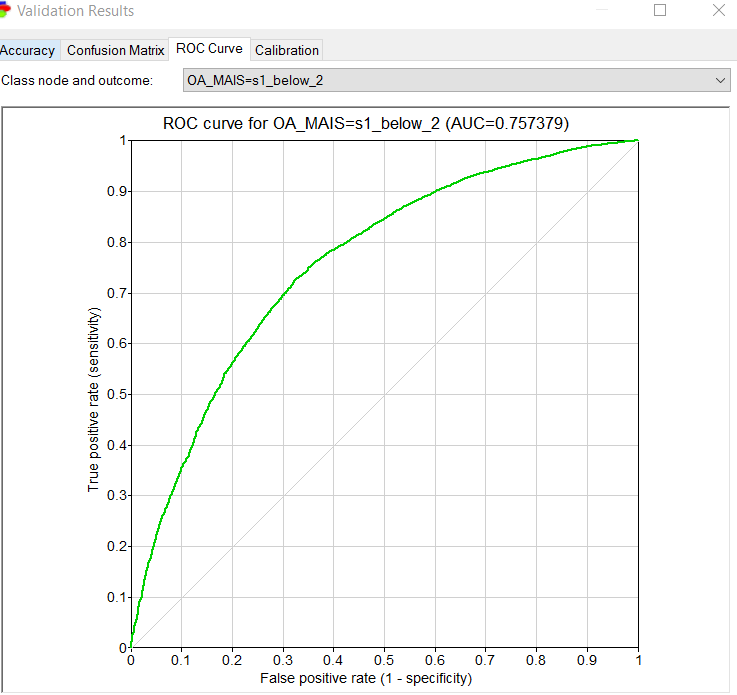


***Fig:*** *Confusion matrix derived after modelling*

***Accuracy***

The Accuracy of the model was found out to be **83.216%** out of which **99.6%** of “s1\_below\_2” class, **16.498%** of “s2\_2\_4” class ,and **5.638%** of class “s3\_4\_up “are captured by the model.

**ROC (Receiver operating characteristic) Curve and AUC**



***Fig****: Area under**ROC Curve and AUC Curve*

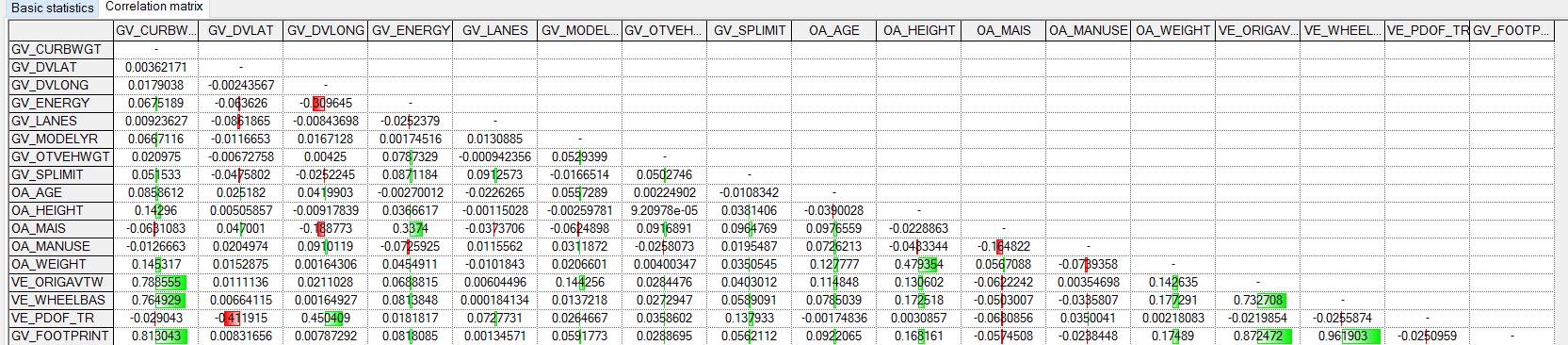
The Area under the ROC Curve as shown in the figure above has

* **“75.74% “**for class “s1\_below\_2”
* “**72.20%**” for class “s2\_2\_4”
* **“82.78%**”forclass“s3\_4\_up”

So, from the above observations, it can be concluded that model is accurate.

**Model Improvement**

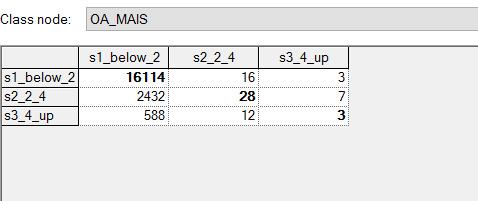
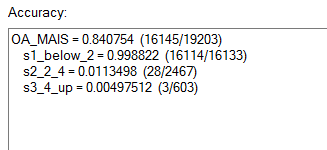
The model was improved by removing the variables having weak correlation with the target variable using the correlation matrix and rebuilding the model again using Naïve Bayes Algorithm.

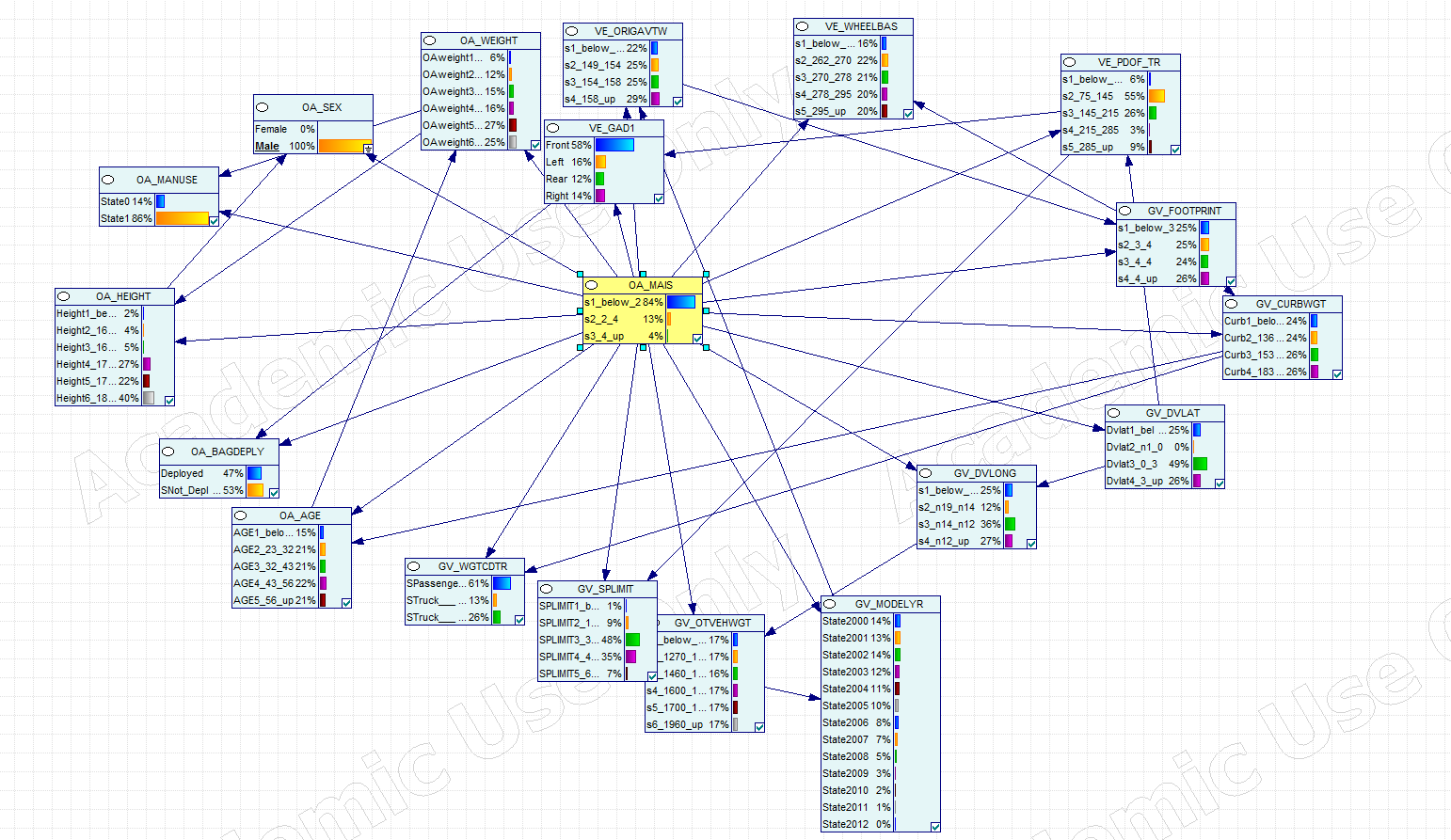


***Fig:*** *Correlation matrix showing the correlation between different variables.*

The new model gave us as an overall accuracy of **“84 .1%”** which suggests model predication improved by **1%.**

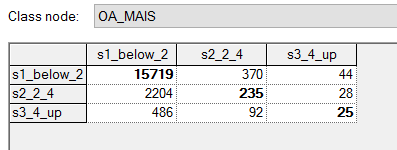
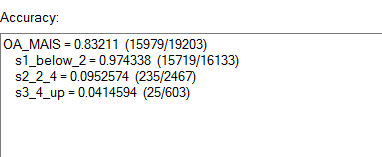
**New Accuracy and Confusion Matrix**





***Fig:***Tree Augmented Naïve Bayes Model for the same dataset

**Accuracy and Confusion Matrix of Tree Augmented Naïve Bayes Model**



**Naïve Bayes Model Vs Tree Augmented Naive Bayes Model Comparison**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Error Rate | Sensitivity |  | Specificity | Precision | Recall | F |
| Naïve Bayes (A) | 83.216% | 16.784% | 86.34% |  | 5.6% | 96.318% | 86.34% | 91.05 |
| Naïve Bayes after removing weakly co-related variables (A1) | 84.075% | 15.925% | 99.88% |  | 4.9% | 97.88% | 84.21% | 90.53 |
| Tree Augmented Naïve Bayes (B) | 81.042% | 18.958% | 86.44% |  | 8.12% | 96.48% | 86.41% | 91.16 |
| Tree Augmented Naïve Bayes after removing weakly co-related variables (B1) | 84.107% | 15.893% | 84.962% |  | 7.13% | 99.145% | 84.96% | 91.50 |

**Conclusion**

From the above table it can be Inferred that amongst model **A (**Naïve Bayes**), A1(**Naïve Bayes after removing weakly co-related variables**),B(**Tree Augmented Naïve Bayes **),B1(**Tree Augmented Naïve Bayes after removing weakly co-related variables**). B1 is the best model as the accuracy is 84.107% and precision is 99.145%.**