**MACHINE LEARNING FOR DATA SCIENCE**

**ASSIGNMENT NO – 7**

**INSTRUCTIONS:**

Apply at least two different machine learning techniques (e.g., clustering, classification, regression) with appropriate algorithms to investigate how the student overdrawing from the checking account relates to age, sex and days drink. Justify your choice of analysis, techniques, and algorithms.

**SOLUTION:**

**LOGISTIC REGRESSION:**

First Machine learning technique, I am using is the Logistic regression technique. I choose this technique as the output variable “Overdrawn” is a dichotomous variable which takes only two possible outcomes.

The Logistic regression algorithm estimates the best coefficient with the given training data using the maximum likelihood estimation technique.

The intuition for maximum-likelihood for logistic regression is that the algorithm searches the best values for the coefficients (Beta values) that minimizes the error in the probabilities predicted by the model. The best coefficients would result in a model that would predict a value very close to 1.

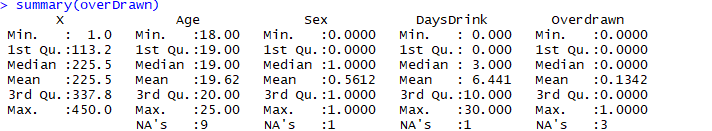
The better model should have the maximum likelihood which in turn refers to the accuracy. The higher the accuracy, the better the model is.

Now let us use the Logistic regression to the given dataset to investigate how the student overdrawing from the checking account relates to the predictor variables.

**STEPS FOLLOWED:**

1. Load the dataset into the R studio and check for the statistical summary and the structure of the dataset.

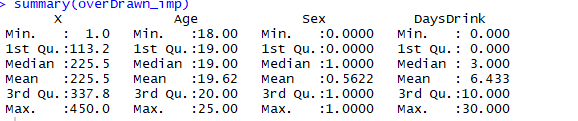
2. We can see that there are missing values in the datasets. Statistical analysis with the missing values will not give a good model. So, let us impute the missing values.



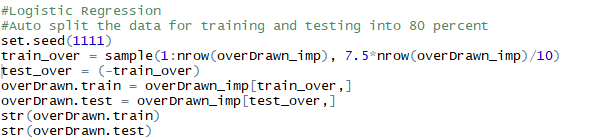
3. In-order to impute the missing values, we can use the kNN imputation technique which will look for the k nearest neighbor values and will do the imputation of the missing values based on the majority votes of the neighborhood values. For this dataset, we are looking for the 3 nearest neighbors.



4. After the imputation, we can see that there are no more missing values and we can proceed with our Logistic regression.



5. Let us split the dataset into 75% training and remaining testing to evaluate the performance of the model.



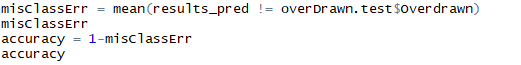
6. Perform Logistic regression with the training data

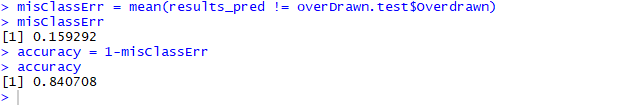


7. Prediction with the test data on the training model



8. Once the test data prediction is performed, the performance of the model can be evaluated by its accuracy metric.





9. Here we can see that we have obtained an accuracy of 84% which is a good model. So, we can say that this model with 84% accuracy have the maximum likelihood of estimating how the student overdrawing from the checking account relates to the given predictor variables.

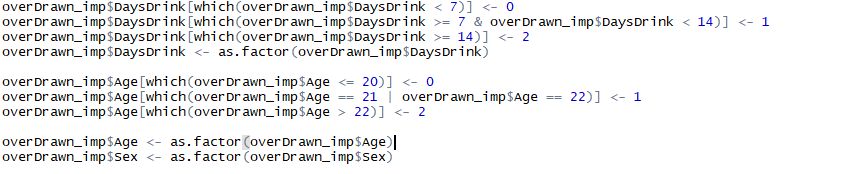
**NAÏVE BAYES:**

Second machine learning technique used is the Naïve Bayes classifier

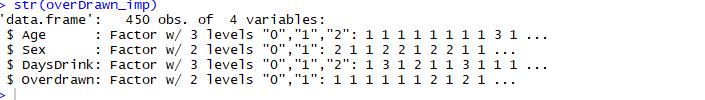
**Steps followed:**

1. Loading the dataset into R studio and doing the missing value imputation with the kNN imputation technique.

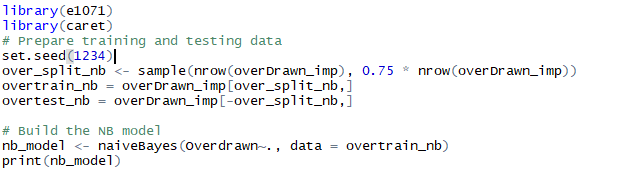
2. Now I have converted all my predictor variables as categorical variables for my classification procedure.



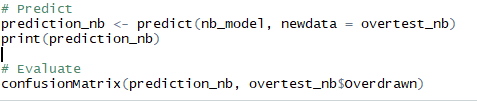
3. Now the structure of the data looks like below output –

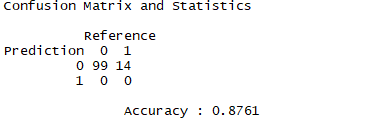


4. After splitting the dataset into training and testing, let us perform the naïve bayes classification



5. Now the prediction and its accuracy measure with the help of confusion matrix

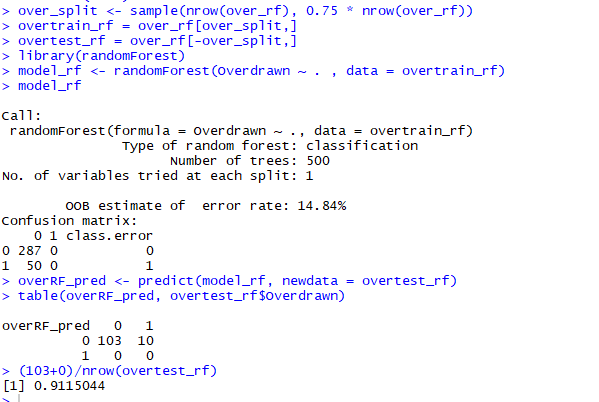




6. We can see the accuracy of this naïve bayes classifier model is 88% which is better than our first logistic regression model. Also, the confusion matrix gives the details about how many times the predictions were made correct and incorrect.

**RANDOM FOREST:**

Note: I have performed the classification with Random forest as well and got the accuracy nearly to be around 91%. May be with the proper cleaning and scaling of data, Random forest being one of the ensemble methods can give a better accuracy without overfitting unlike the Decision Tree technique. This classification technique plays a key role in terms of accuracy driven results.



**Reason behind choosing Logistic and Naïve Bayes techniques:**

1. Even though, my random forest has given the better model with high accuracy and can be one of the appropriate methods, I focus on the Logistic regression and Naïve Bayes classification here (as part of the assignment) as I would like to compare and understand the difference between the outcome of the Discriminative and the generative models with the given set of input data.

Discriminative model – Logistic regression

Generative model – Naïve Bayes

2. As compared to the real-time problems, the given dataset has less amount of data and we can see that the Naïve bayes gives better accuracy with the lesser data. With my model comparison, I find that the Naïve Bayes classifier gives better model with the higher accuracy level.

3. Also, accuracy does not play a key role in most of the real-world problems. When we wanted to get best accuracy for our model, then we can choose to go for complex algorithms like ensemble models which are sure to give the accuracy at its best.

4. But when accuracy is not our area of concern and if we like to deal with the simple and easy to interpret procedures, then we can follow the simple algorithms like the Naïve Bayes classifier algorithm.