Bushra Naeemi

ID:  29064

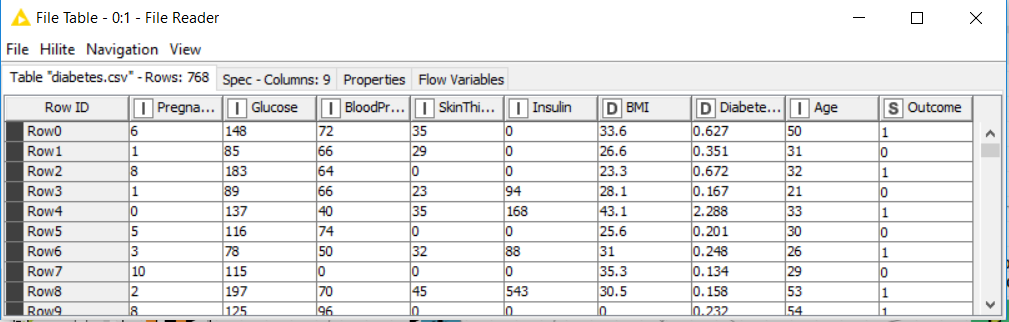
data analytics using knime

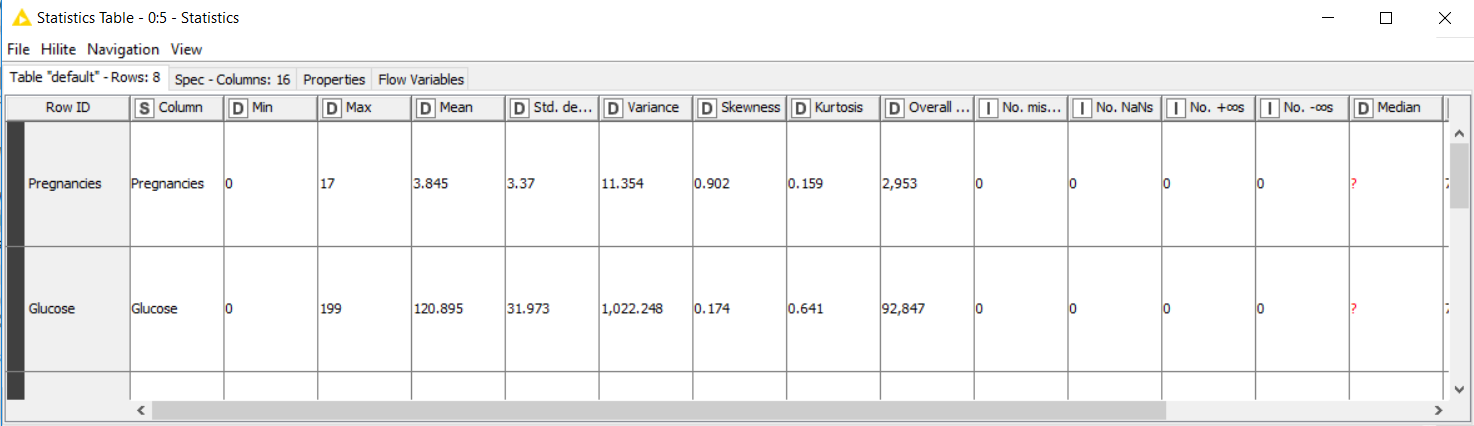
Introduction:

The project is about working or analysis of a dataset, which is about diabetes rate by considering their factors: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome. The whole point of the project is about predicting whether a person has diabetes or not based on the factors.

We used KNIME for the whole process. KNIME Analytics Platform is the open source software for creating data science. KNIME makes understanding data and designing data science workflows and reusable components accessible to everyone.

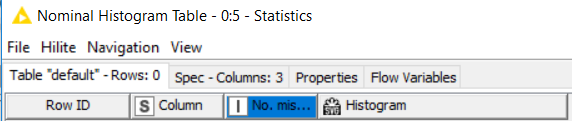
Result of 12 Steps:

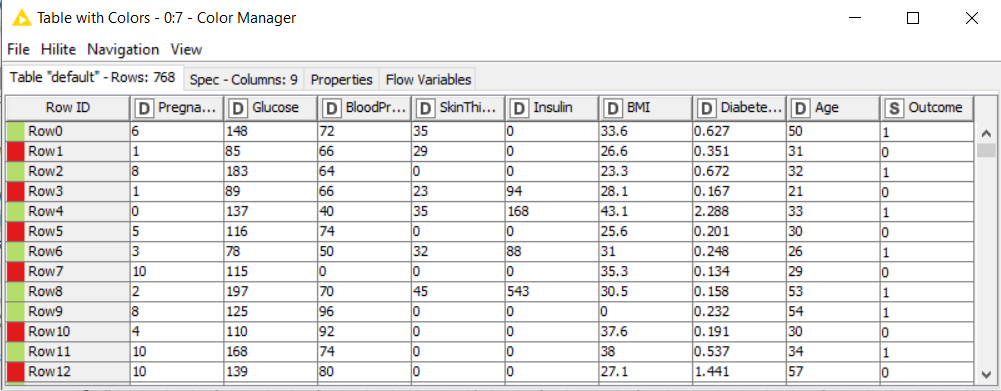
1. File Reader: It reads the diabetes.csv file and we can see our data in it.
2. We created a statistical analysis for raw data using statistics node, that shows the statistical results like mean, standard deviation and etc.



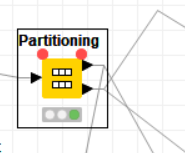
1. Missing Values: It checks weather our data has missing values, if yes then we can deal with it. But in our data we didn’t have any missing value. Our data was clean.

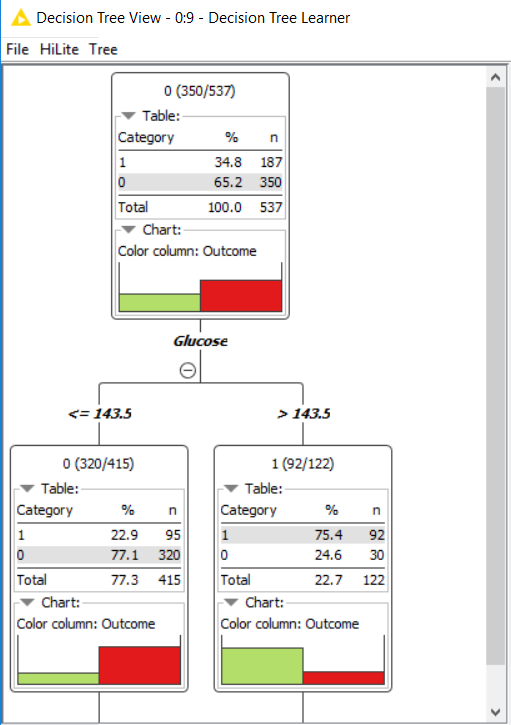


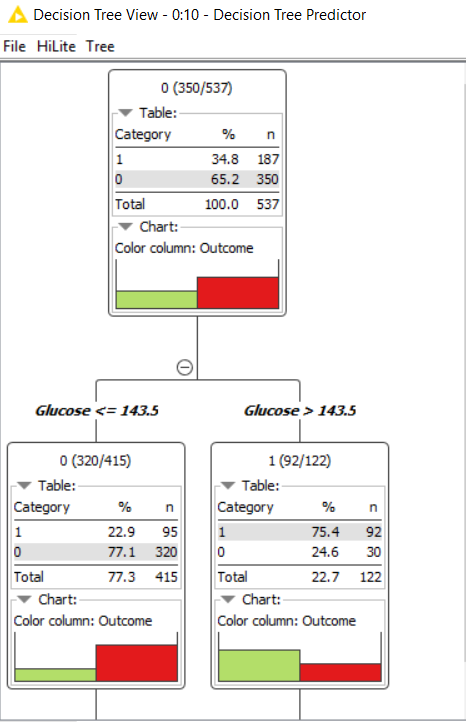
1. Statistics after removing missing value: in here it shows that we don’t have any missing value.
2. Color Manager: we used the color manager to colorize our target variable (Outcome):



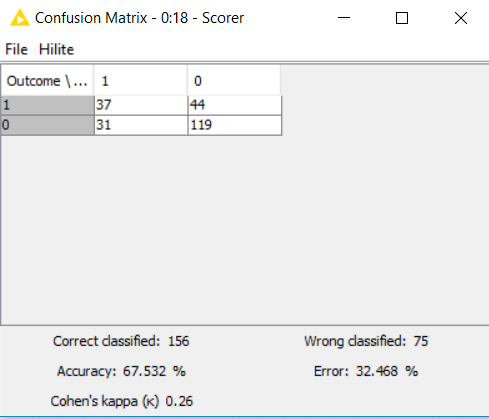
1. Data Partition: In this step we are partitioning our data to training and testing, we allocated 30% of the entire dataset for testing and the remaining for training. When we see 537 rows are for training, and 231 rows for the testing.

.

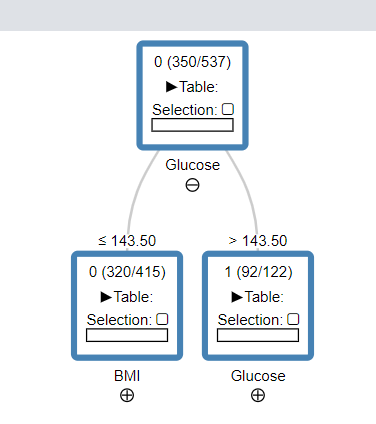
1. Decision three algorithm learner: it shows the whole prediction and data in the form of a tree with colors.
2. Decision tree algorithm predictor: it shows the prediction using the tree.



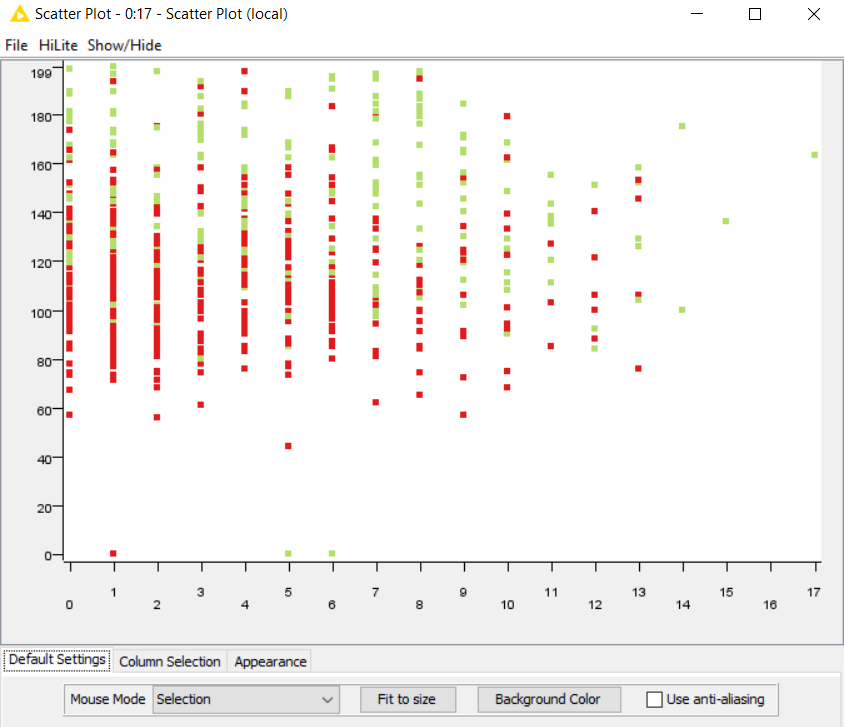
1. Score of the decision tree: it shows the accuracy of the prediction.



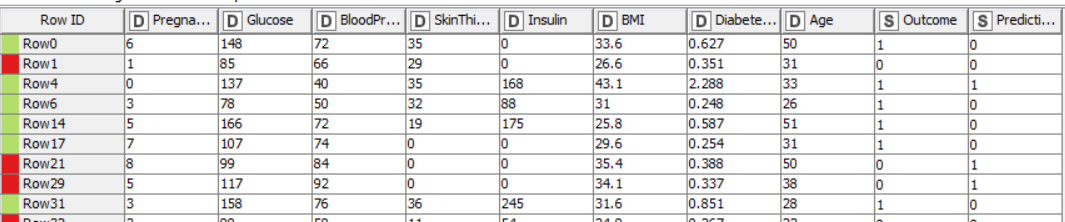
1. Decision tree view simple:



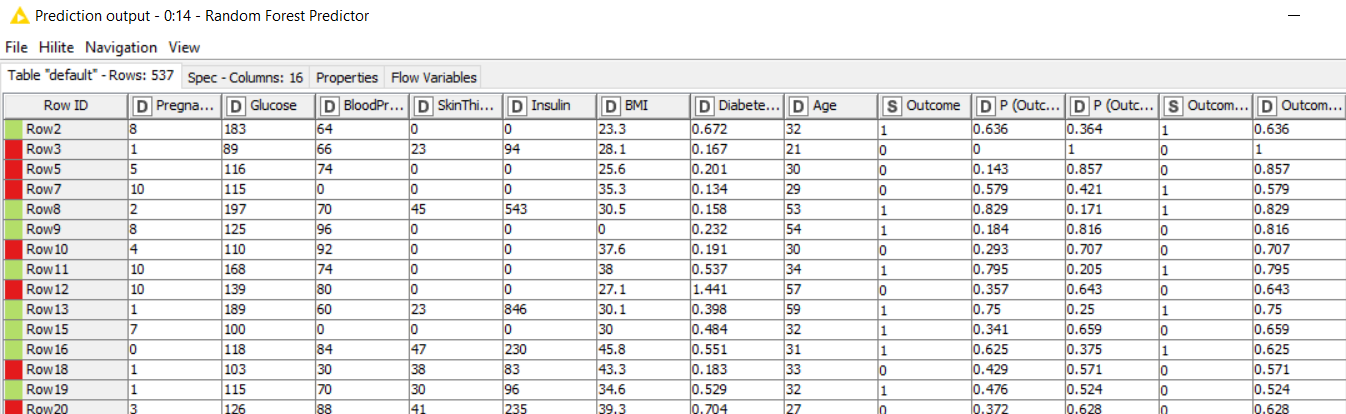
1. Scatter plot of the decision tree. It mainly used for comparison of the data’s.

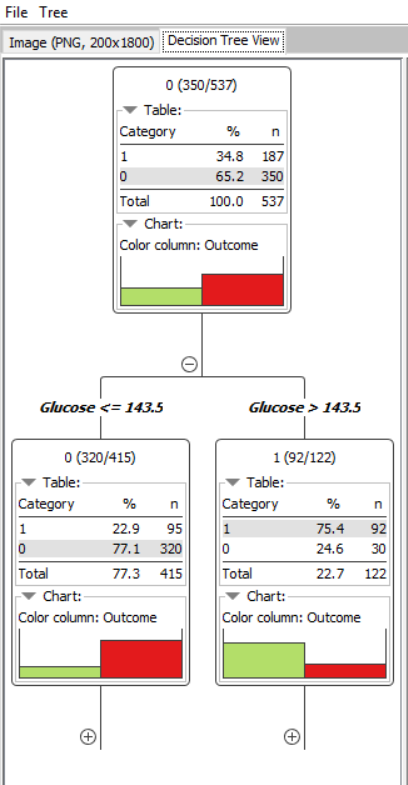


1. Data Report: It creates a good report for our data that explains everything so clear by using the colors.



d. we can compare the results of our data by using random forest learner and predictor and result of decision tree.





Conclusion:

since we sliced the dataset into two parts of 70% (main) and the remaining 30% for testing, at first we changed the type of the Outcome from integer to string. the results after implementing functions such as confusion matrix, data report, scatter plot, we can differentiate their records and observations which have diabetes and diabetes free. Our Outcome