

**Submitted by:**

Rupali Taneja 18csu182

Saumya Achantani 18csu194

Saumya Gupta 18csu195

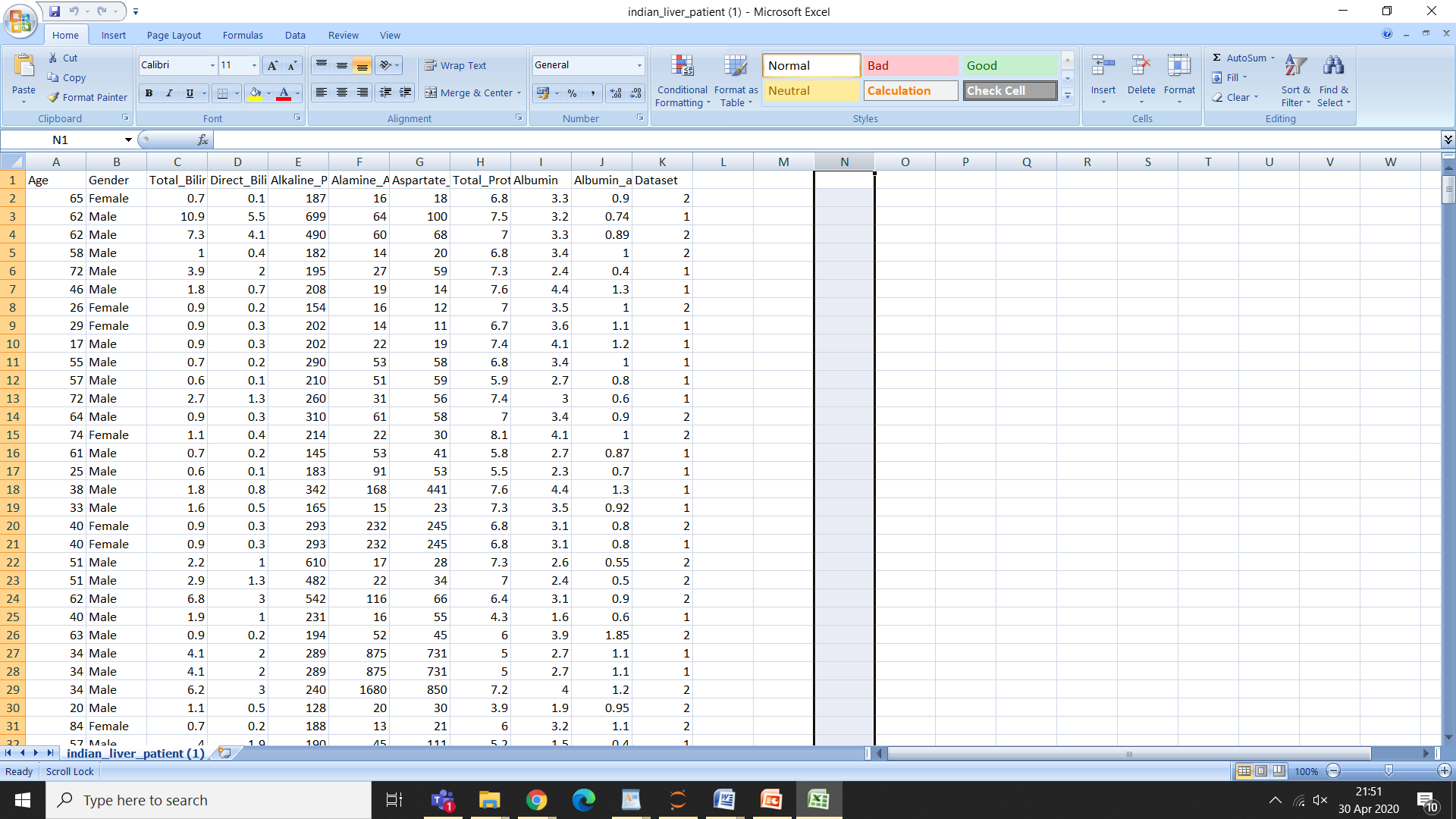
***………………..…..…….STEPS TO CREATE KNIME WORKFLOW……..………..………......***

**STEP 1 :**

**Loading Dataset:-**

We started with loading our “Indian\_liver\_patient.csv” file using filer reader node > csv reader.

The dataset consists of following columns:

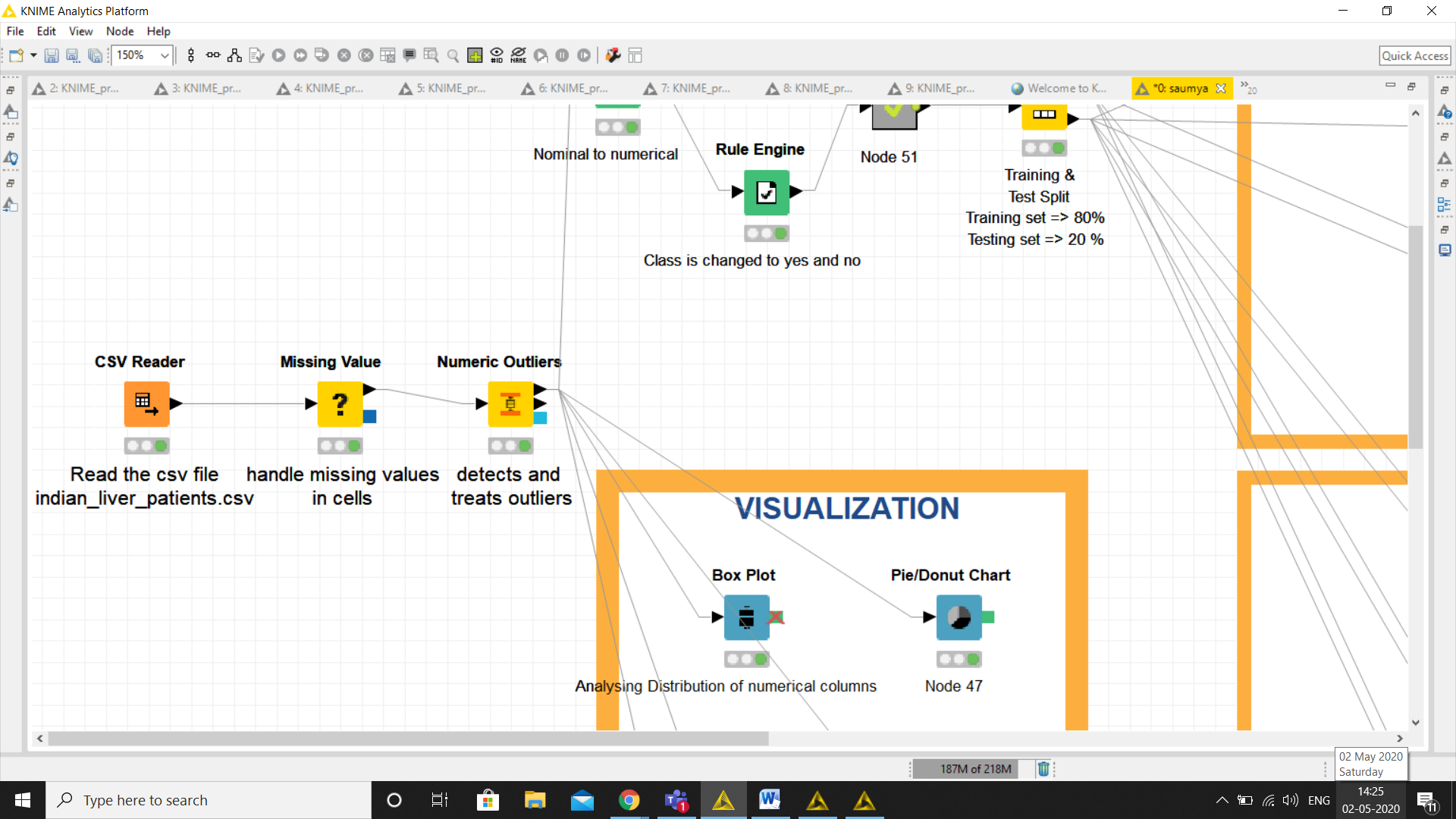


**STEP 2:**

**Data Preparation**

In this step, identification and removal of missing values(if any) is done >Missing Values

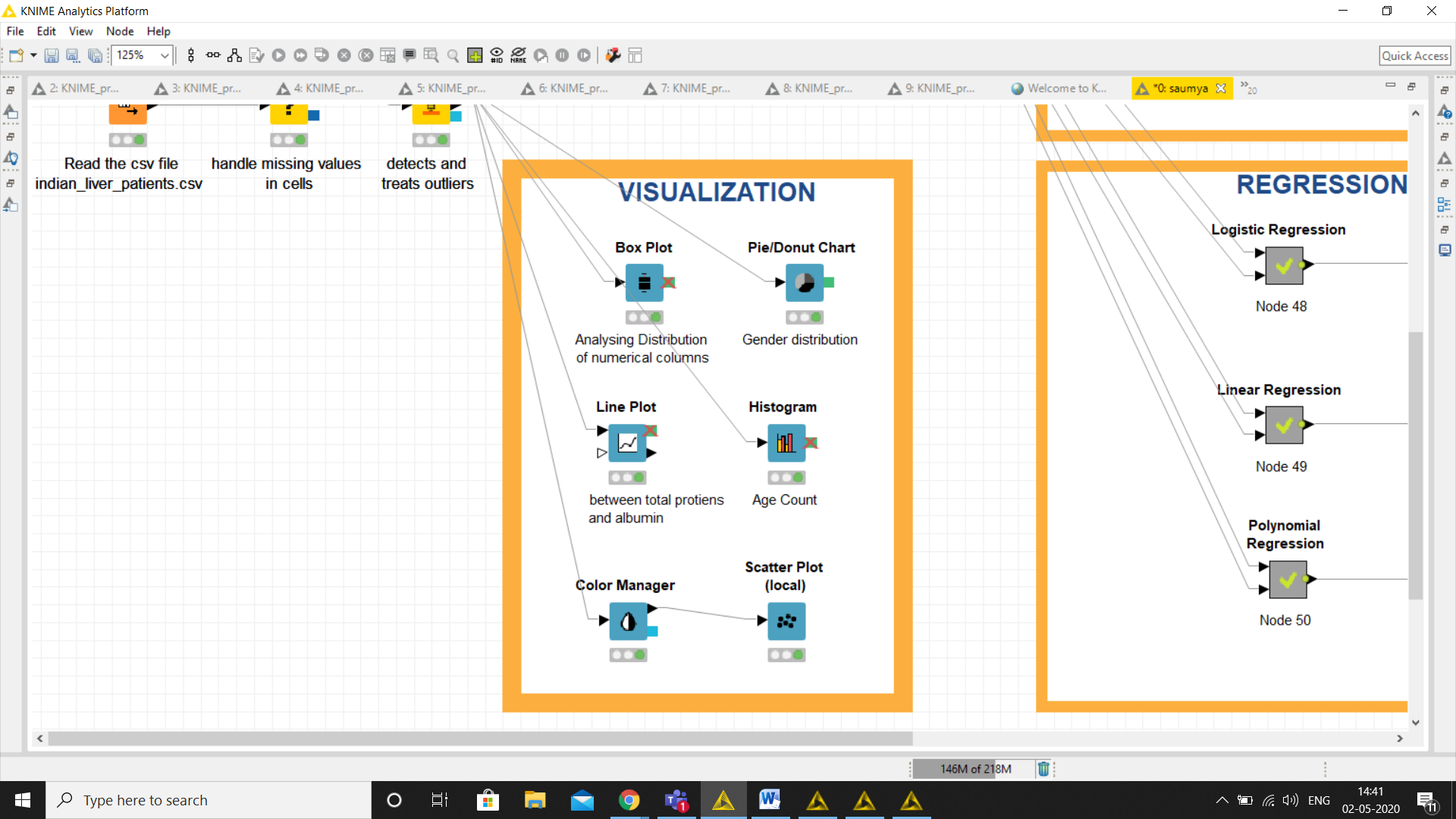
Also, outliers are identified(if any) and removed so that there are no misleading predictions later in out project>Numeric Outliers



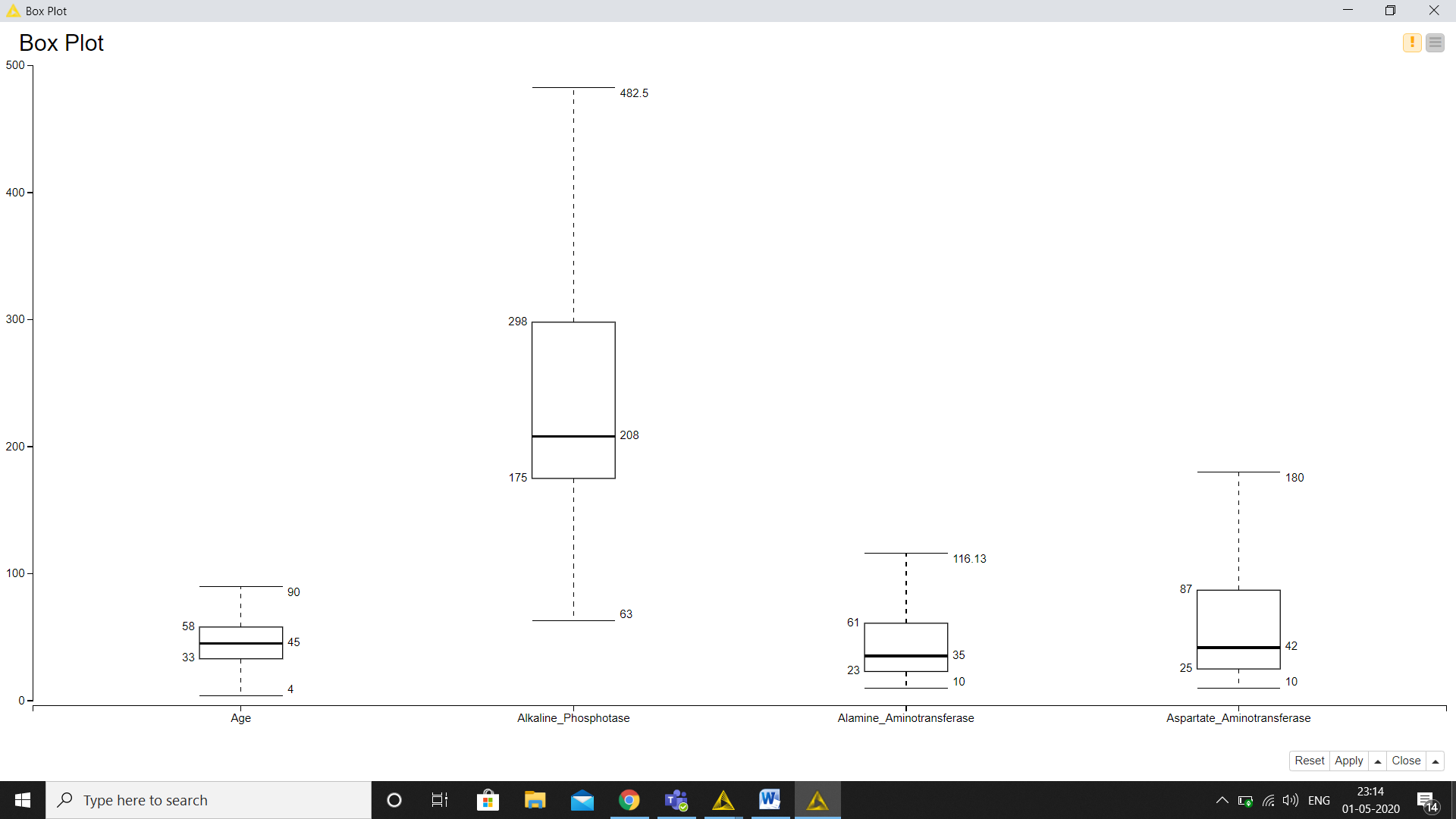
**STEP 3:**

**Data Visualization**

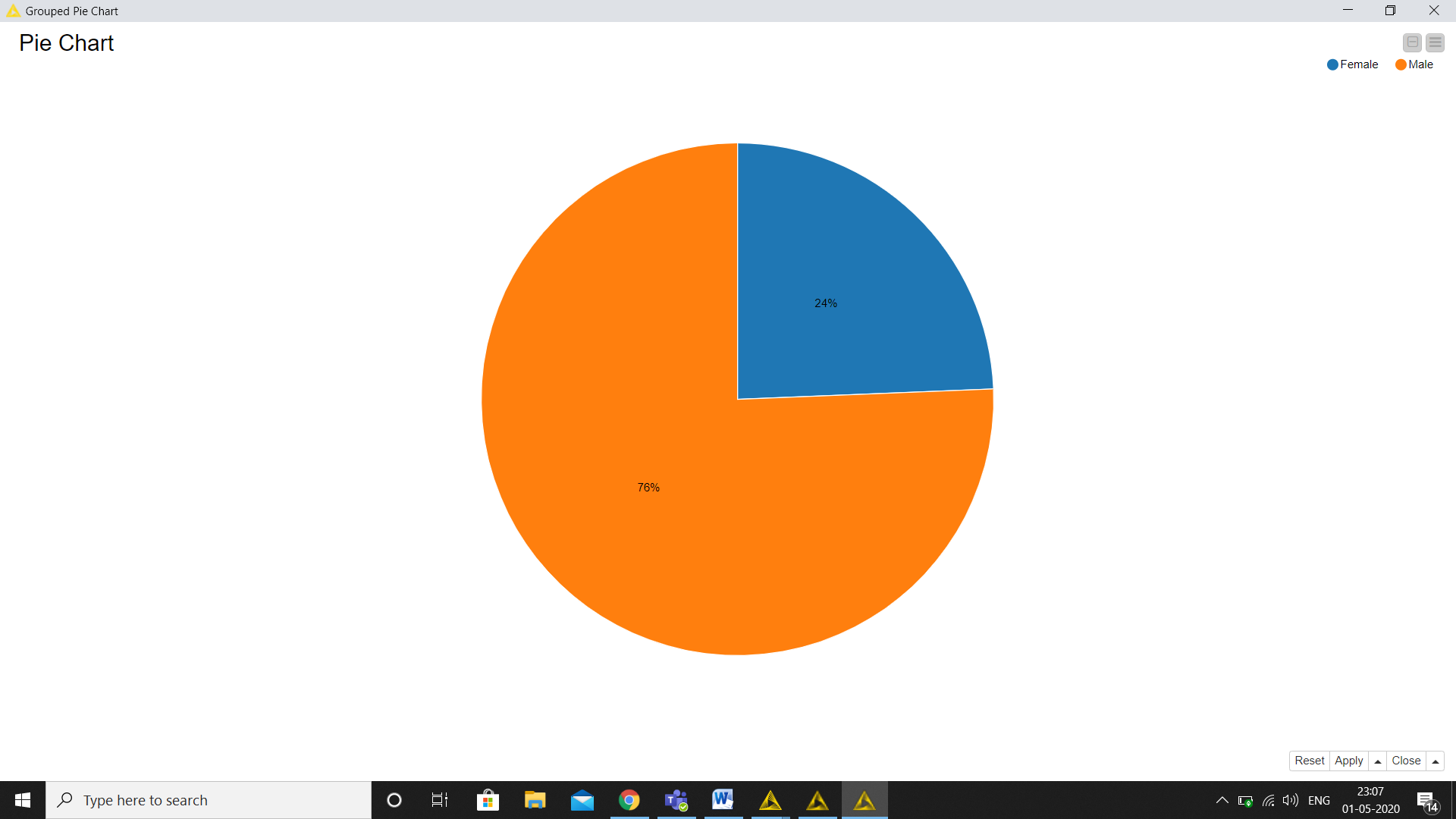
After this, Exploratory data analysis is done using various visualization nodes such as box-plot, bar-plot , histogram, pie chart and scatter plot .



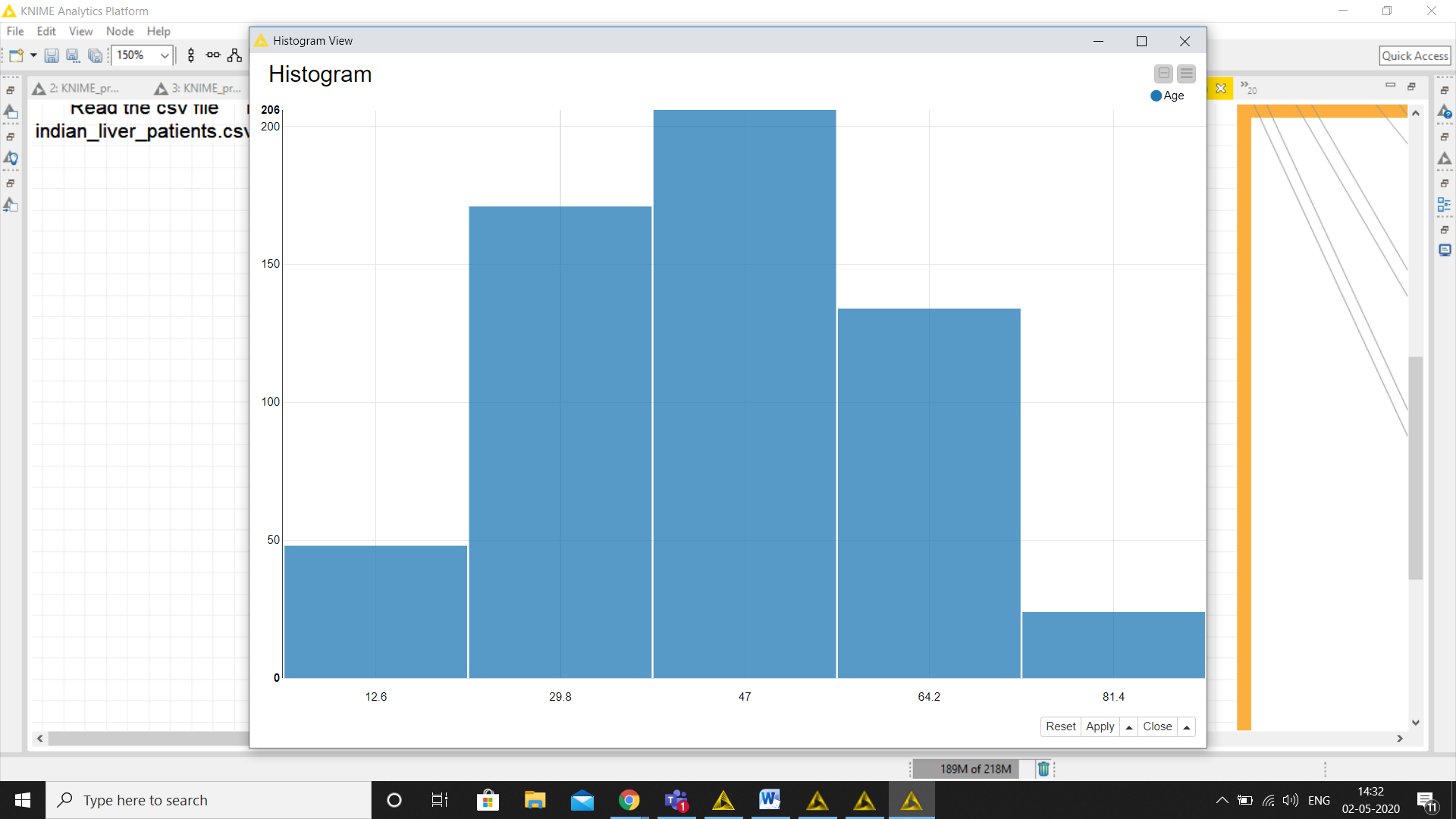
Some of them are shown below:-



**Box Plot showing Distribution of Numerical Columns which affect chances of having liver disease**



**Pie Chart showing gender distribution percentage**

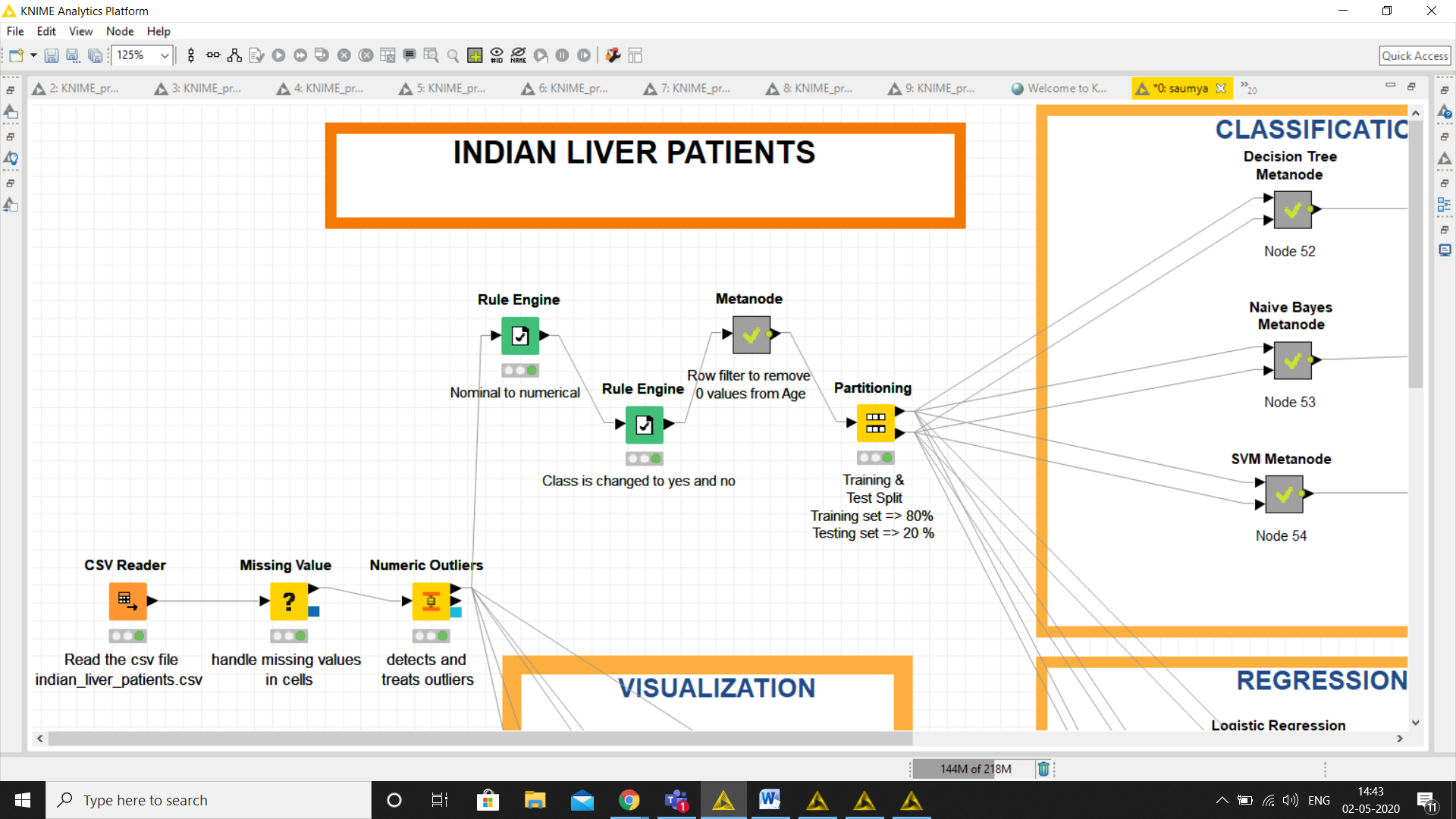


**Histogram of Age(where y axis is the count of people)**

**STEP 4:**

**Data preparation for model training:-**

After that, we have applied rule engine node to reclassify our target value from 1 and 2 to yes and no respectively, so that it can be used later for models.



**STEP 5:**

**Partitioning:-**

This step involves splitting of data into testing and training set by using node partioning.

We have used 80 -20 splitting which means 80% of the data is used for training and rest 20% for testing the model.

Training set is more than test so that our models can learn well and make accurate predictions.

**STEP 6:**

**Classification Models:**

In this step , 3 classification models are used :

1. Decision Tree (A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences)
2. Decision Tree Learner -> to train the model
3. Decision tree predictor -> to test/predict the result

To avoid clutter, we can combine 2 nodes using meta node as done ahead.

Similarly,

1. Naïve Bayes ("probabilistic classifiers" based on applying Bayes theorem with strong independence assumptions between the features)
2. SVM-Support Vector Machine(SVMs are based on the idea of finding a hyper plane that best divides a dataset into two classes )

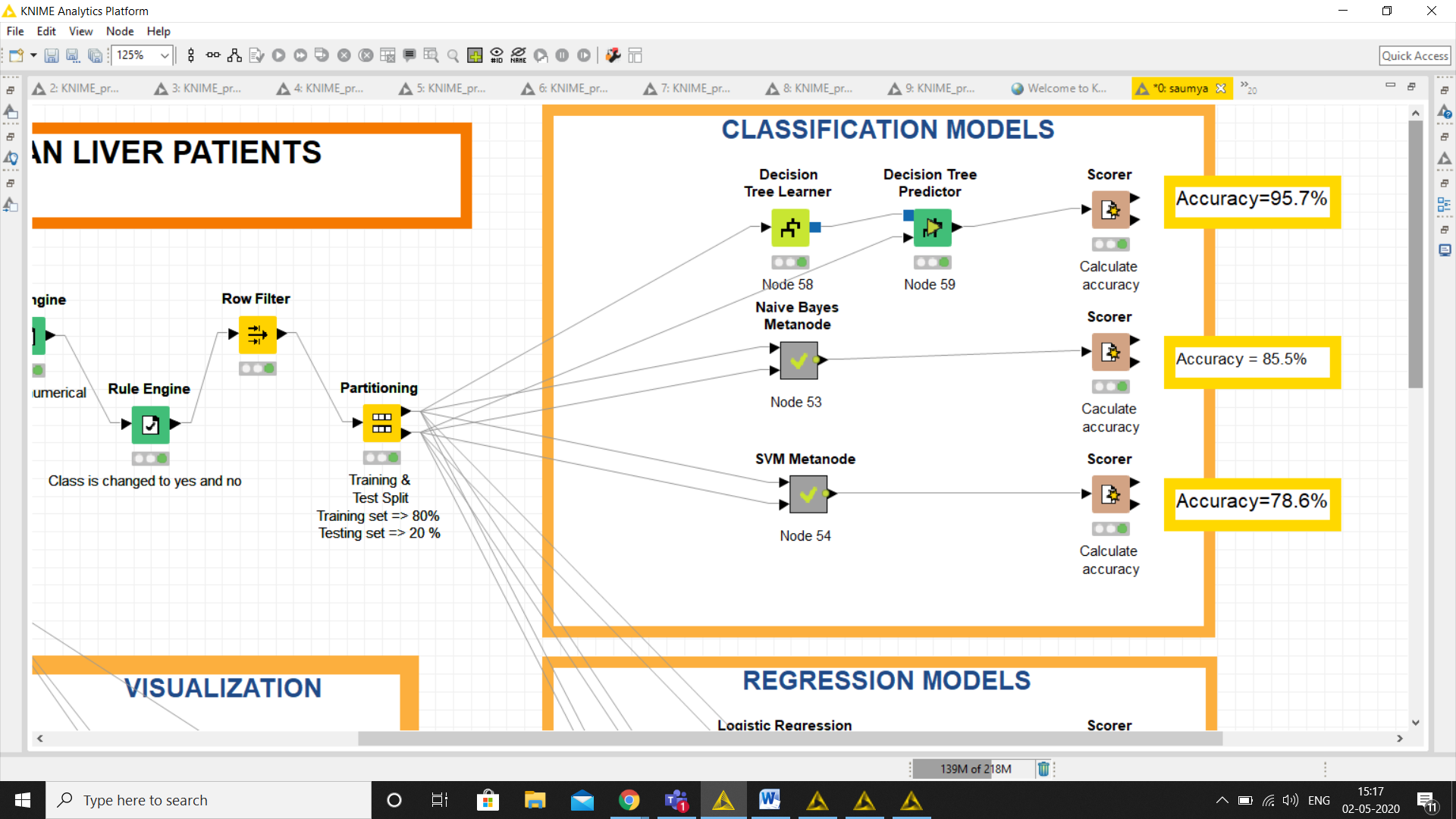
**Accuracy:-**

After training & predicting the model, the accuracy of the model is calculated. Accuracies of the models are as follows:

1. Decision tree -> 95.7%
2. Naïve Bayes -> 85.5%
3. SVM -> 78.6%

**Inference** : Since, Decision Tree has highest accuracy, therefore, it is best model for our dataset.

As our topic is health related, it is very important to get correct classification and best accuracy.



**STEP 7:-**

**Regression models:**

In **Regression,** we have used 3 regression models :-

1.Linear regression (Dependent variable is predicted by using independent (predictor) variable ,modeled on line of best fit)

2.Polynomial regression (relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial)

3.Logistic Regression(uses a logistic function to model a binary dependent variable)

**STEP 8:-**

**R sq value:-**

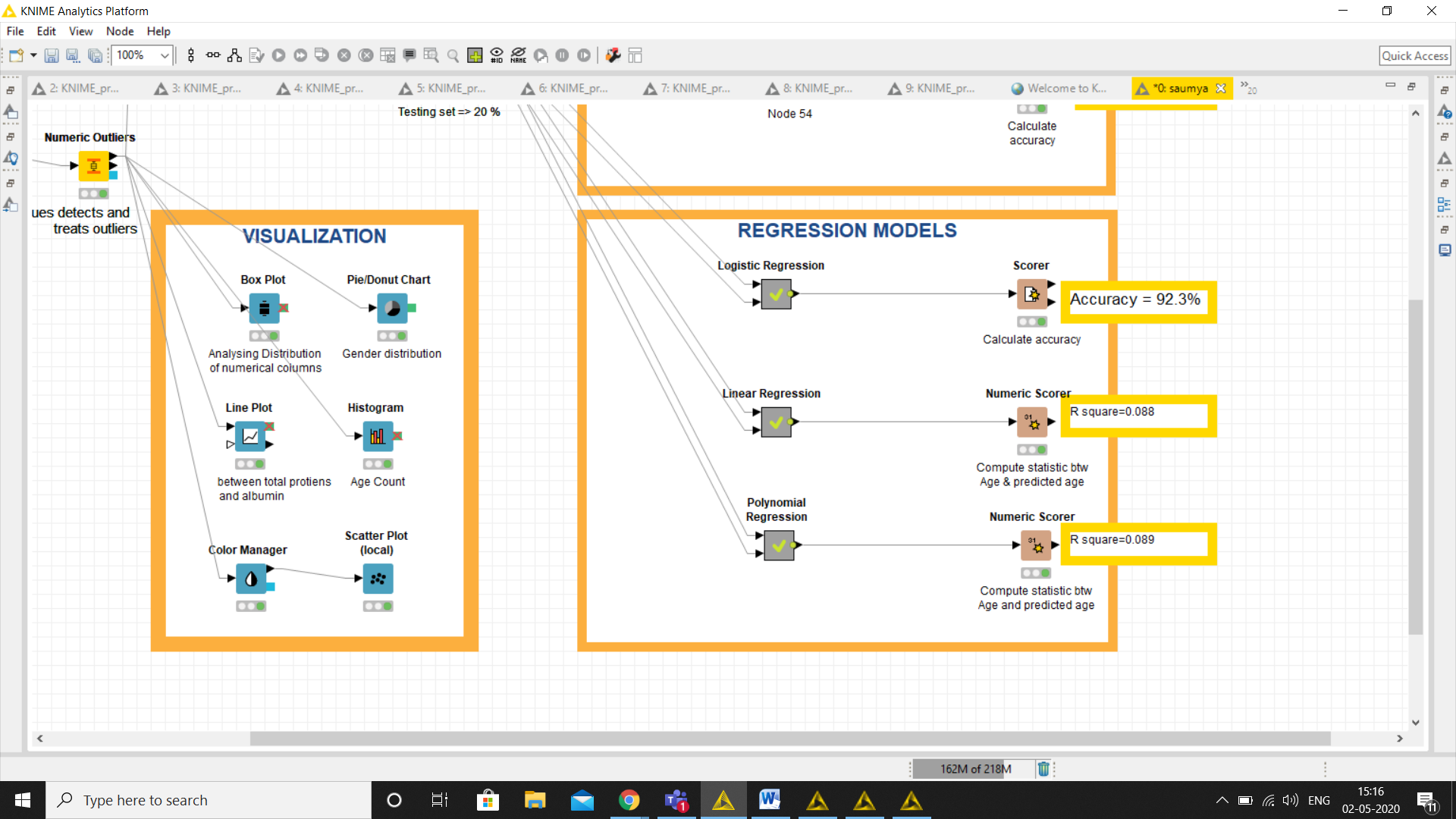
In this we are calculating the **R square** of all the regression models by using **Numeric Scorer.**

**R square values are as follows:**

1. Linear Regression -> 0.088
2. Polynomial Regression ->0.089

Accuracy of Logistic Regression->92.3%

(it is basically classification problem since target is binary)



FINAL WORKFLOW

