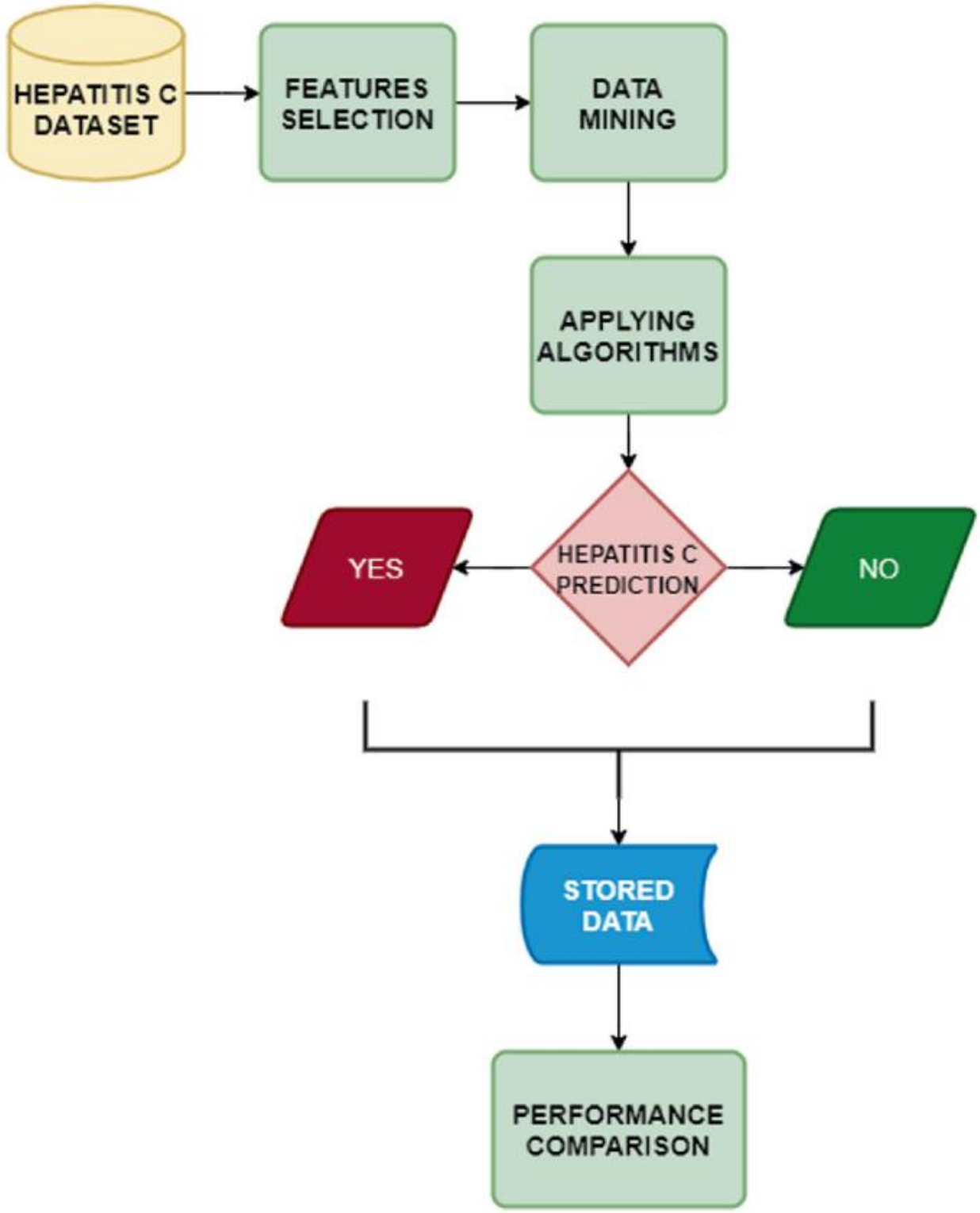
Project Name: Hepatitis C virus data analysis and prediction using machine learning

Abstract: Medical decision support systems have been on the rise with technological advances and they have been subject of studies. Developing an effective medical decision support system requires a high amount of accuracy, precision and sensitivity as well as time efficiency that is inversely proportional to the complexity of the model. Hepatitis C virus infection is one of the most important causes of chronic liver disease worldwide. So we apply data science process and HCV has been estimated with machine learning methods. By analyzing and visualizing the values in the dataset, features that may be important for HCV was determined using machine learning.

Introduction: Hepatitis C is a Ribonucleic acid (RNA) virus with multiple genotypes located in the Hepacivirus genus from the Flaviviridae family [1]. It is a factor that causes hepatitis C disease, chronic liver disease, liver cirrhosis, hepatocellular carcinoma. For these reasons, Hepatitis C virus (HCV) is one of the most important global health problems in the world. Worldwide, 350,000 people out of 185 million patients infected with HCV die

from diseases caused by HCV [2]. HCV is one of the leading causes of chronic liver disease. HCV infections pose serious problems on a global scale. Since there is no vaccine yet to prevent HCV infection, it is essential to prevent infection. Early detection of patients and people at risk is critical to prevent the spread of HCV infection. In this sense, the research question of our study is to develop a model that can predict HCV infection from new features. This research question has been solved by using new feature methods and comparing the performance of various ML algorithms.

Block diagram:



Dataset Description: I choose the dataset from Kaggle. In the dataset row no. 615 and column no. 32 maximum all are numerical value. Target column is Category.I have three classes in the target column.

Result: I successfully create models using logistic regression, decision tree, random forest algorithms. Calculate precision and recall.

Discussion: Machine learning can basically be defined as a basic artificial intelligence approach that supports the creation and development of mathematical models to help understand data [52]. These models, with parameters that can be adapted to the observed data, begin to learn. These models can be used to predict and understand new data to be observed after learning the model based on previously seen data.

Conclusion: Data science is a multidisciplinary science that consists of the combination of scientific problem identification and solving methods, statistics, software development and technology. Data science turns data into useful, meaningful information to solve complex problems. In this study, data was visualized using data science methods, and meaningful information was extracted

from the data. Recall is most important for any medical purpose prediction. I have three algorithms recall value for logistic regression is 0.54, for decision tree is 0.79 and for random forest is 0.66.

So I will be choosing decision tree algorithm.