SciencePG
Science Publishing Group

ANOVA Based Feature Selection Model for Predicting Hepatitis C Virus

Shilpi Bisht[1], Neeraj Bisht[1], Anshul Srivastava[2, *], Anupama Rajesh[2], Sonalika Srivastava[3], Bishwajeet Pandey[4]

[1]Department of Applied Sciences, Birla Institute of Applied Sciences, Bhimtal, India

[2]Amity International Business School, Amity University, Noida, India

[3]King George Medical University, Lucknow, India

[4]Department of Intelligent System and Information Security, Astana IT University, Kazakhstan

Email address:

shilpibisht@yahoo.co.in (Shilpi Bisht), bisneeraj@gmail.com (Neeraj Bisht), anshul_sriv@rediffmail.com (Anshul Srivastava), anupamar@amity.edu (Anupama Rajesh), drsonalikasrivastava@gmail.com (Sonalika Srivastava), bk.pandey@astanait.edu.kz (Bishwajeet Pandey)

[*]Corresponding author

Abstract

The aim of the research is to make precise predictions using machine learning algorithms for HCV detection. Hepatitis C Virus (popularly called HCV) causes various kinds of life-threatening liver cancers. This virus is deadly as more than 80% of patients don’t have any signs or symptoms. Hence, a proper system is necessary which may timely and accurately predict HCV. This paper provides a brief literature review in this field which is the motivation to understand the concepts of machine learning methods in predicting results with higher accuracy in comparatively easier ways. We have applied various machine learning algorithms on the ‘HCV’ dataset, which result in efficient and accurate predictions. The experiments are performed on the ‘HCV’ Dataset which is imported from “UCI Machine Learning Repository”. The dataset contains the laboratory values of Hepatitis C and the blood donors. It has a ‘615’ number of instances, out of which some contain missing values. Out of 615; 533 instances are of blood donors, 7 instances are of suspected cases and the remaining 75 instances contain the data of patients suffering from Hepatitis C. There are a total of ‘13’ attributes present in the dataset. To achieve the best predicting algorithm, the handling of missing data is performed using Linear Regression. Feature selection is the technique that improves the efficiency of a model and reduces model building time. We have implemented the statistical technique Analysis of Variance (ANOVA) for the purpose of selecting important features. The rationale behind choosing ANOVA as a feature selection technique is its efficiency in determining the score of the relationship between two attributes. In our model, the decision tree algorithm predicted with the highest accuracy of ‘0.9878048780487805’ before applying ANOVA. After applying ANOVA, the decision tree algorithm again predicted the results with greatest accuracy i.e. ‘0.9878048780487805’. This research adds to the field by utilizing machine learning methods to enhance HCV prediction, striving for enhanced accuracy and efficiency. The use of Linear Regression to manage missing data and ANOVA for feature selection introduces potentially innovative approaches within HCV detection. Additionally, the comparative assessment of multiple machine learning algorithms aims to pinpoint the optimal model for HCV prediction, potentially providing direction for future research in analogous settings.

Keywords

Prediction, ANOVA, Linear Regression, Random Forest, SVM, Decision Tree, Gaussian NB, KNN