Heart Failure Prediction using Data Mining Techniques

|  |  |
| --- | --- |
| Kkshitij Kapadia *Dept. of CSE School of Technology Pandit Deendayal Energy University Gandhinagar, Gujarat Email: kkshitij.kce@sot.pdpu.ac.in* | Pranav Sirodaria *Dept. of CSE School of Technology Pandit Deendayal Energy University Gandhinagar, Gujarat Email:pranav.sce@sot.pdpu.ac.in* |
| Suyamoon Pathak *Dept. of CSE School of Technology Pandit Deendayal Energy University Gandhinagar, Gujarat Email:suyamoon.pce@sot.pdpu.ac.in* | Kshitij Raj Burnwal *Dept. of CSE School of Technology Pandit Deendayal Energy University Gandhinagar, Gujarat Email: kshitij.bce@sot.pdpu.ac.in* |

**Abstract:** Congestive heart failure has been one of the primary causes of death in the world today, estimated to take 17.9 million lives yearly,. The health care sectors gather enormous amounts of data that contain some hidden information and are useful for making effictive decisions. The use of data mining techniques to predict heart disease is only partially explored in a number of studies.Using machine learning techniques, we propose a novel approach in this paper that aims to identify significant features, thereby increasing the accuracy of cardiovascular disease prediction.

**Introduction**

Because of numerous contributing risk factors, including diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors, it is challenging to diagnose heart disease.The severity of heart disease in humans has been determined using various data mining and neural network techniques.

Data mining is the practice of getting useful insights out of numerous databases.Due to the nontrivial information found in large amounts of data, data mining is most beneficial in exploratory analyses.The process of extracting data in order to discover hidden patterns that can be transformed into significant is known as data mining.Knowledge of data mining enables a user-oriented approach to new and hidden patterns in the data.

The suggested method may extract patterns and correlations related with heart disease from a historical heart disease database.

It can also respond to difficult questions about heart disease diagnosis, which can help medical professionals make wise clinical judgments.

Results indicated that the suggested system is particularly effective at achieving the specified mining goals.

**Methodology**

In this Exploratory Data Analysis a variety of Model Classifications including Logistic Regression (LR), Support Vector Machine (SVM), Desicion Tree (DT), and Naive Bayes have been used .This study will examine the dataset on heart failure prediction.A total of 302 records make up the dataset, which was split into two sets: a training set (40%) and a testing set (60%).

**Features Used**

**14 attributes used:**   
1. Age: age in years  
2. Sex: sex (1 = male; 0 = female)  
3. cp: chest pain type   
-- Value 1: typical angina   
-- Value 2: atypical angina   
-- Value 3: non-anginal pain   
-- Value 4: asymptomatic  
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)  
5. chol: serum cholestoral in mg/dl  
6 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)  
7. restecg: resting electrocardiographic results   
-- Value 0: normal   
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)   
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria  
8. thalach: maximum heart rate achieved  
9. exang: exercise induced angina (1 = yes; 0 = no)  
10. oldpeak = ST depression induced by exercise relative to rest  
11. slope: the slope of the peak exercise ST segment   
-- Value 1: upsloping   
-- Value 2: flat   
-- Value 3: downsloping  
12. ca: number of major vessels (0-3) colored by flourosopy  
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect  
14 num: diagnosis of heart disease (angiographic disease status)   
-- Value 0: < 50% diameter narrowing   
-- Value 1: > 50% diameter narrowing

**DATA PRE-PROCESSING**

After gathering various records, preprocessing of heart disease data occurs.There are 302 patient records in the dataset overall, 7 of which have some missing data. For the attributes of the given dataset, multiclass variables and binary classification are introduced.The presence or absence of heart disease is determined using the multi-class variable.If the patient has heart disease, the value is set to 1, otherwise it is set to 0 to indicate that the patient is heart disease-free.Pre-processing of data is done by converting diagnosis values from medical records.The results of 297 patient records' pre-processing show that 137 of the records reflect a value of 1, indicating the presence of heart disease, while the other 160 records reflected a value of 0, indicating the absence of heart disease.

**FEATURE SELECTION AND REDUCTION**

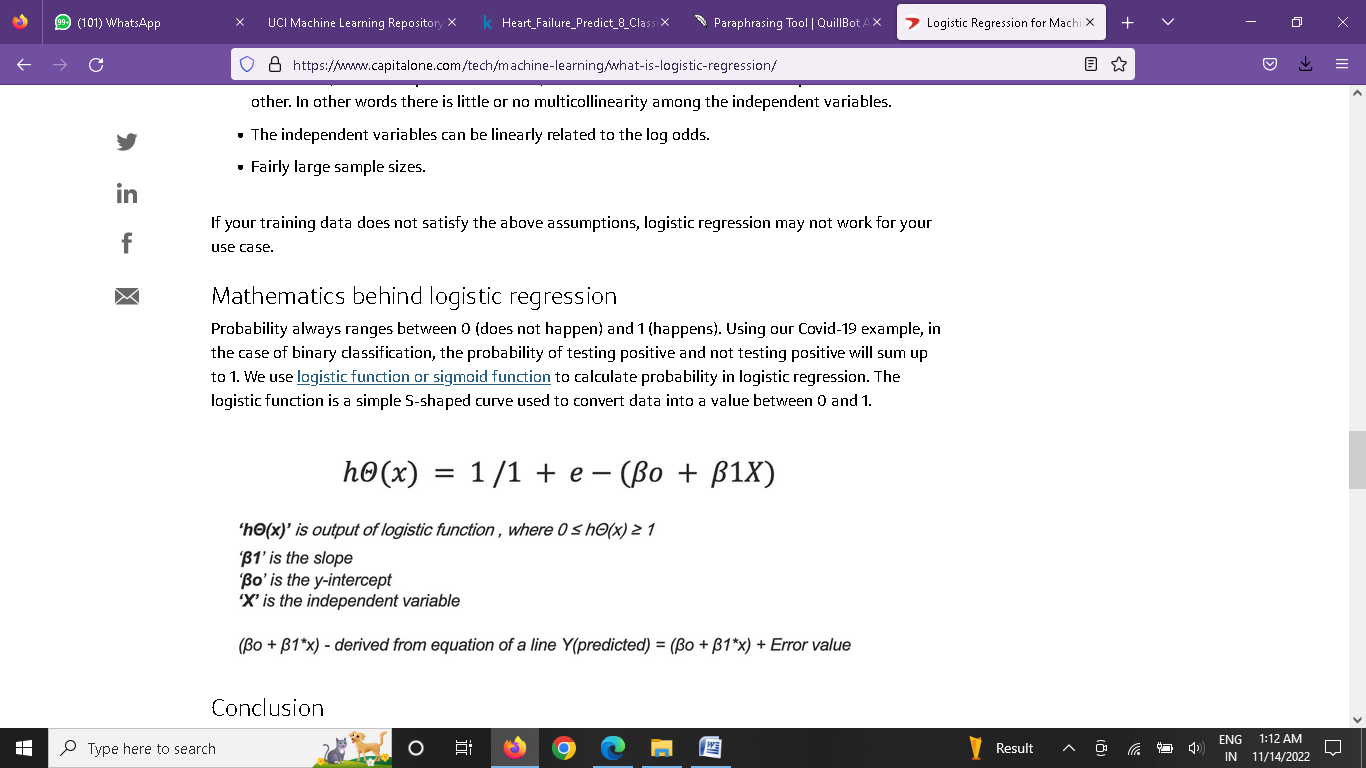
From among the 76 attributes of the data set, total of 14 attributes have been selected which contain vital clinical records. Clinical records are essential for heart disease diagnosis and severity assessment.

**DATA MINING CLASSIFIERS**

**LOGISTIC REGRESSION**

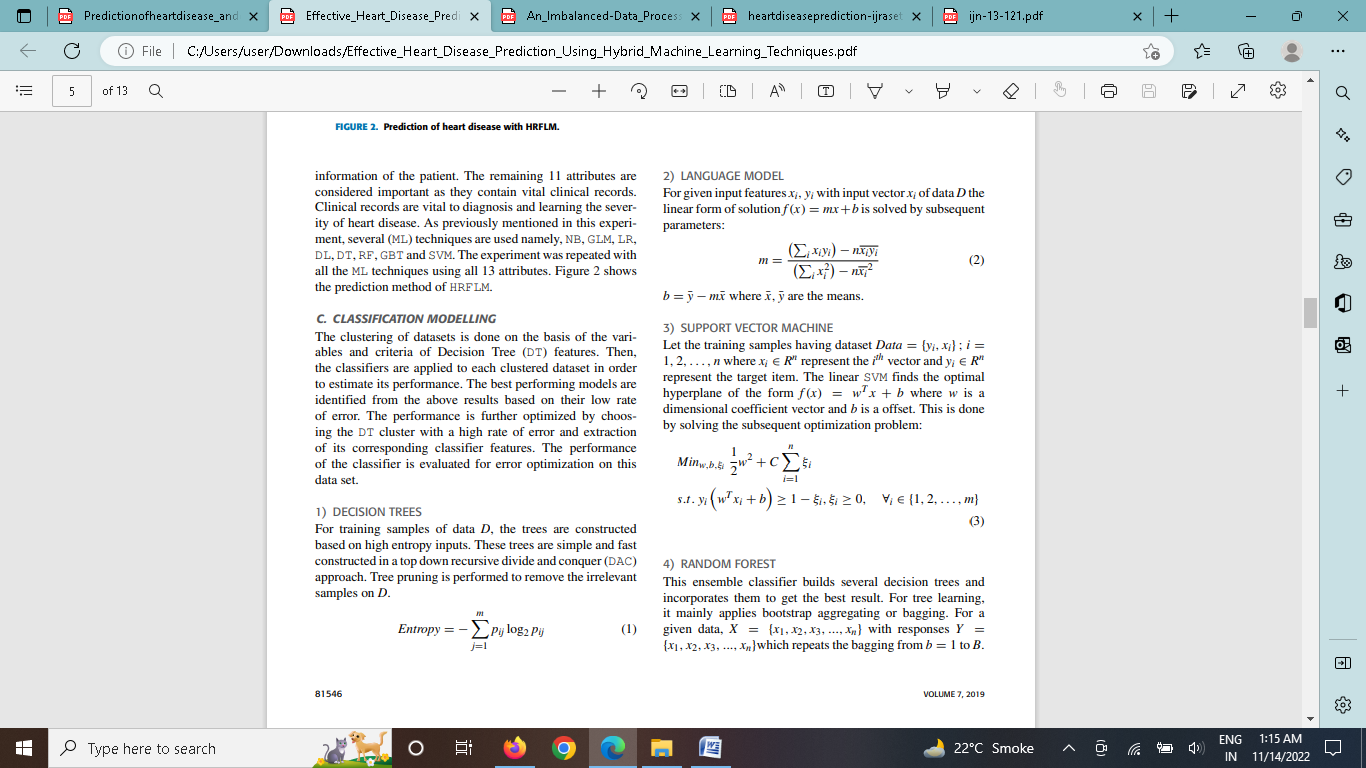
When the output or dependent variable is categorical or dichotomous, logistic regression is used for classification problems.When using logistic regressions, there are a few presumptions to be aware of, including the numerous forms of logistic regression, the various kinds of independent variables, and the readily available training data.

Probability is always between 0 and 1.In logistic regression, probability is computed using the sigmoid function.The logistic function is a straightforward S-shaped curve that transforms data into a value between 0 and 1.

****

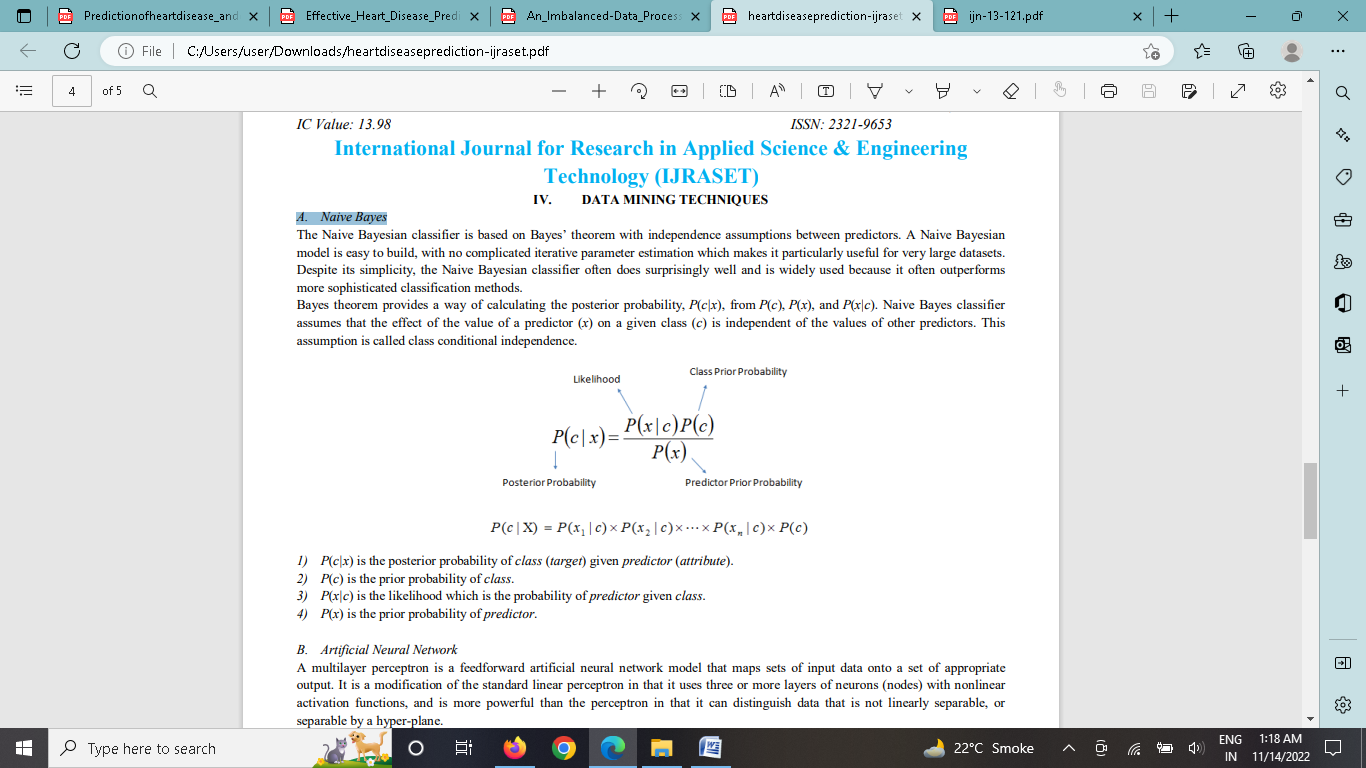
**SUPPORT VECTOR MACHINE**

Let the training samples having dataset Data = {yi, xi}; i = 1, 2, . . . , n where xi ∈ R n represent the i th vector and yi ∈ R n represent the target item. The linear SVM finds the optimal hyperplane of the form f (x) = w T x + b where w is a dimensional coefficient vector and b is an offset. This is done by solving the subsequent optimization problem:



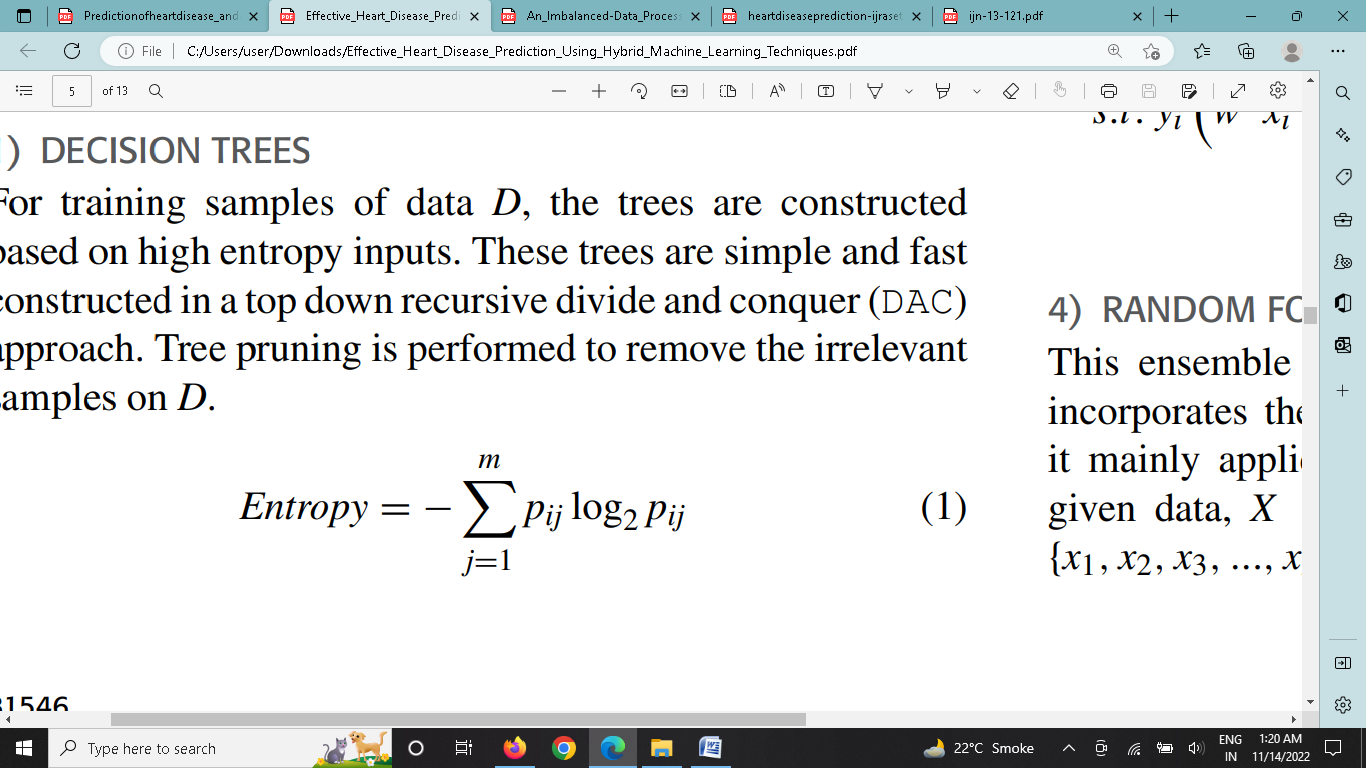
**Naive Bayes**

The Naive Bayesian classifier is based on Bayes’ theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

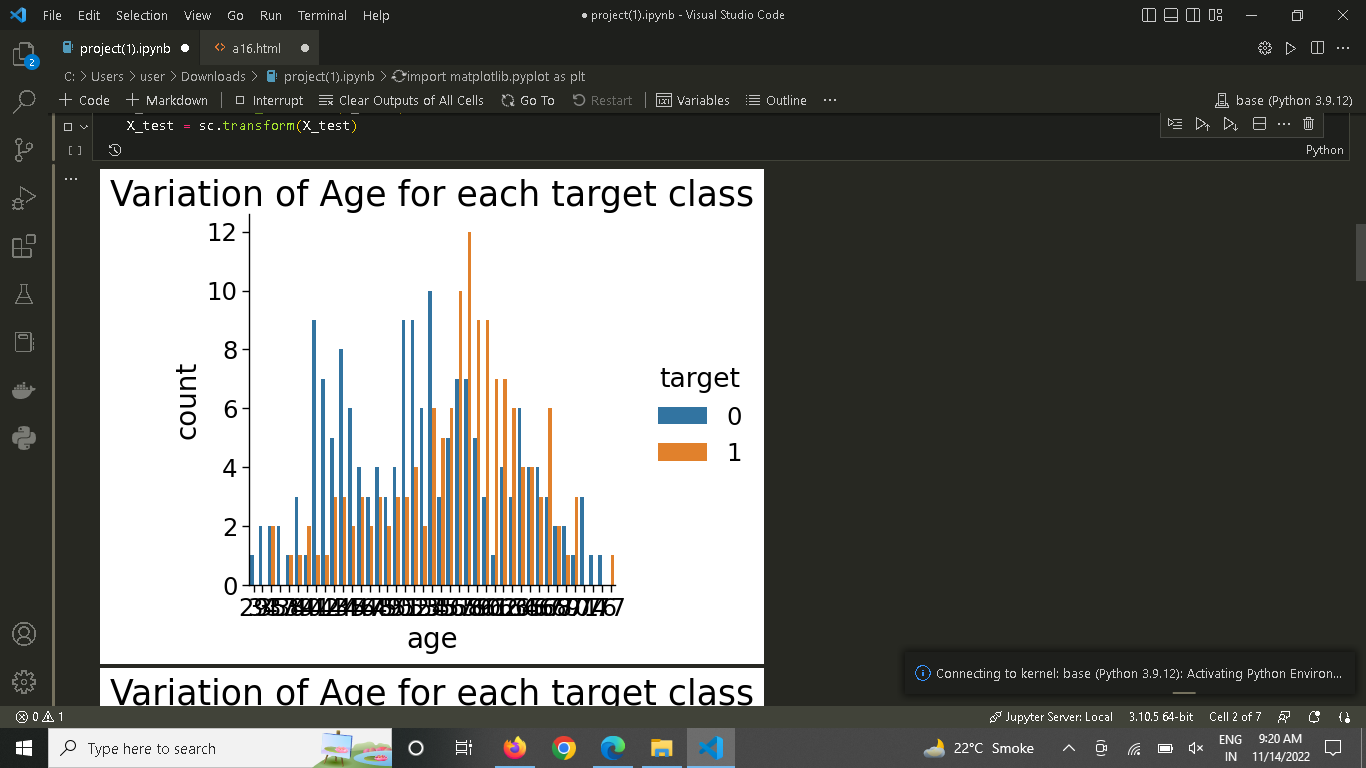
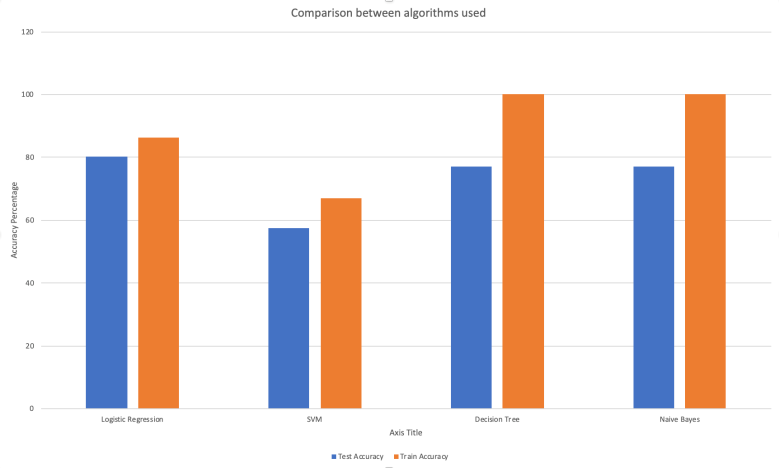


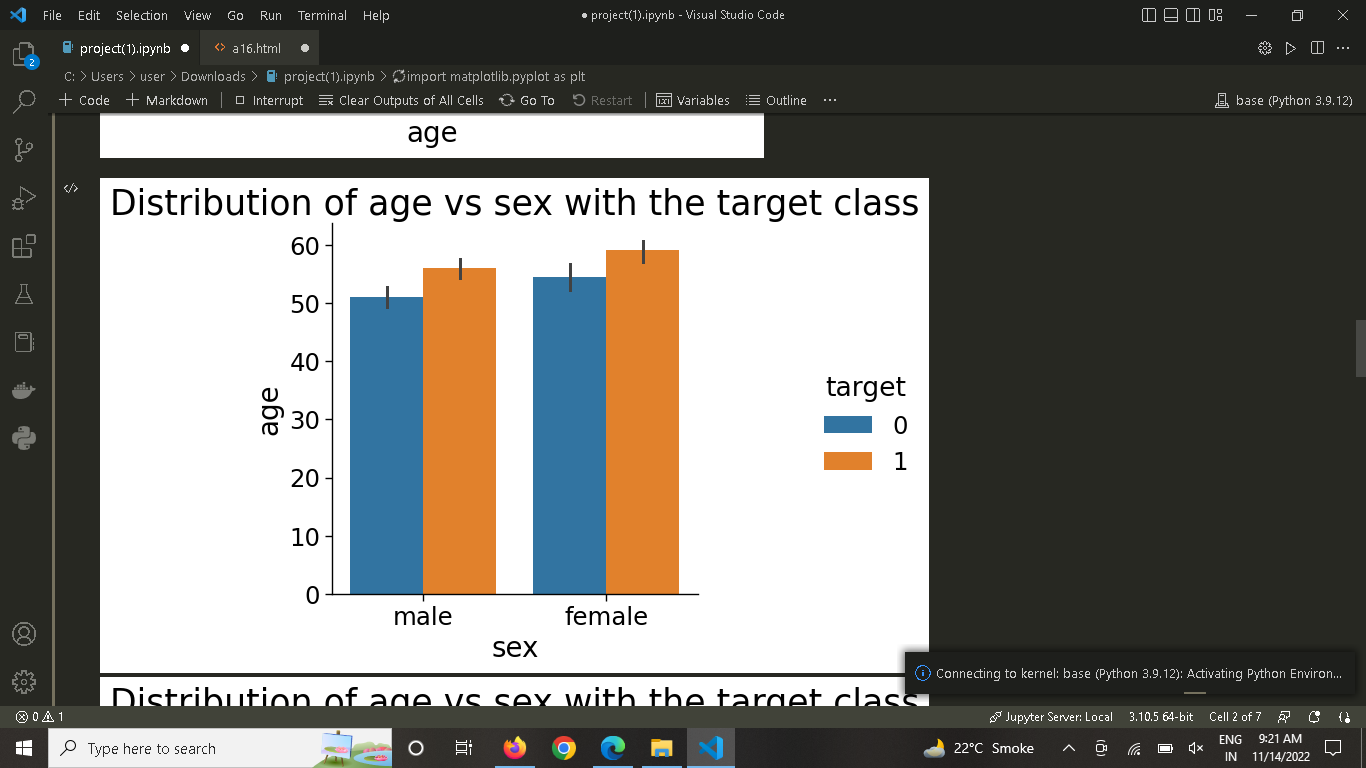
**DECISION TREES**

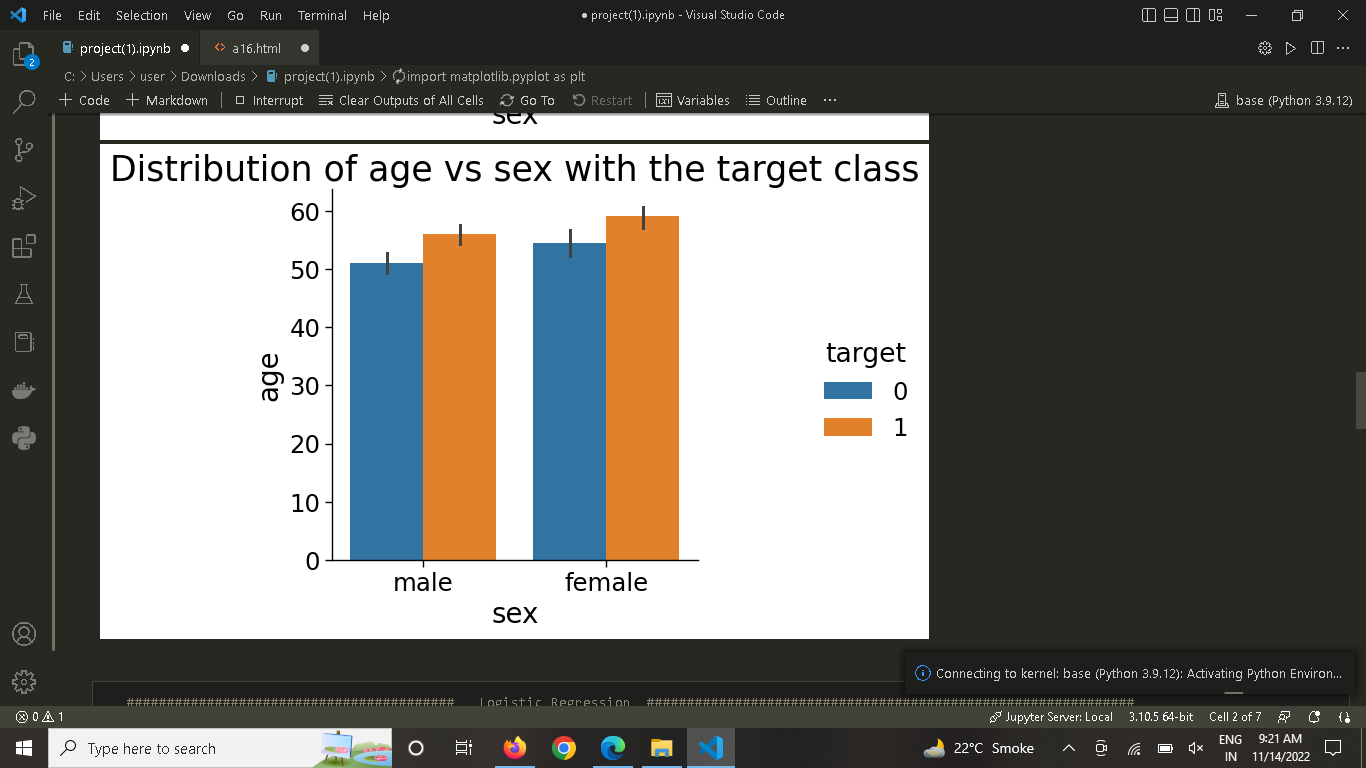
For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on D.



**EXPLORATRY DATA ANALYSIS AND VISUALIZATION**

** **

****

****

**Result of various models with proposed model**

The prediction models are developed using 13 features and the accuracy is calculated for modeling techniques. The best classification methods are given below in Table 3. This table compares the accuracy, classification error, precision, F-measure, sensitivity and specificity. The highest accuracy is achieved by HRFLM classification method in comparison with existing methods.