Liver Patient Predication And Analysis Of Data Science Using IBM Machine Learning Service

By M.Nithesh,P.Reddisri,K.Renuka.

Branch : Computer Science And Engineering

1.INTRODUCTION:

1.1 overview:

The notion behind Machine Learning is to apprentice a machine through knowledge, experience and precedents so to make it proficient in making decisions asgood as humans or even surpass the human experts.With the growing modernization, the rate at which the diseases are increasing is a cause of concern andthe most critical aspect of human lives, their health is at a great risk.Rigorous cataloging or labeling of a disease in its early stage is a prerequisite for proper treatment.In many cases,the methodolgy of carrying out several diagnostic tests is complicated followed by skeptical results.In my work,I have particularly focused on the precise prediction of Liver Disease, the tenth most common cause of death in India as per the World Health Organization. Many recent disease diagnostic methods use clinical decision support system(CDSS),better than clinical diagnosis and therefore increases the disease prediction rate and human well being.For the present study,the dataset is taken from UCI vault.It is often difficult to choose the best classifier.The study aims to evaluate seven well known Supervised MachineLearning classification algorithms many a times with varying feature selection techniques on each.The classifiers that have been addressed. Liver diseases averts the normal function of the liver. Mainly due to the large amount of alcohol consumption liver disease arises. Early prediction of liver disease using classification algorithms is an efficacious task that can help the doctors to diagnose the disease within a short duration of time. Discovering the existence of liver disease at an early stage is a complex task for the doctors. The main objective of this project is to analyse the parameters of various classification algorithms and compare their predictive accuracies so as to find out the best classifier for determining the liver disease.

1.2 purpose:

Machine Learning today in data analysis field robotize interperative model framework. Classification algorithms in Machine learning have grown to be among the leading research topics and its utilization in therapeutic datasets are in discussion all over. Acknowledging the fact that combining multiple predictions leads to more accurate results than merely depending on a single prediction, a single dataset has been trained on various algorithms and the highest voted class is predicted as the result. Liver disease is the only major instigation of death still perennial, hence early detection and treatment of not very symptomatic liver disease is must which can significantly reduce the chances of death. In the proposed work, the dataset of Indian Liver Patient has been utilized and it clearly states that grouping classification algorithms efficiently improves the rate of prediction of illnesses.

2. LITERATURE SURVEY:

2.1 Existing Problem:

Trends with respects to various features in the dataset have also been observed. proposed various parameters such as sensitivity, specificity, etc. in order to evaluate the model performance. Mere accuracy is not a sufficient measure for model performance when the data is non-uniform. Classifiers used are SVM ,Gradient Boosting, Random Forest, Naive Bayes and Logistic Regression.
We Used Pearson Correlation along with Decision Trees, Naive Bayes ,Random Forest and ANN as machine learning classifiers with decision tree outperforming the rest of the algorithms. Various model performance parameters have been evaluated.
Applied and analyzed Genetic algorithm, Particle Swam optimization and Artificial Neural Network for predicting cardio diseases.
Applied Principal Component Analysis for Feature Selection .Finding the best features in the dataset is necessary because it helps prevent overfitting, reduces training time and many a times provides better accuracy. Associative classification techniques and various datamining algorithms for foretelling heart diseases we utilized Principal Component Analysis to determine the important features followed by Alternate Decision Trees in order to find heart diseases.
frequently higher for pregnant women or those with gallstone conditions. Total protein reveals number of diverse proteins in the blood.
It is further branched into Albumin and Globulin fractions. Minor levels of total protein in the blood developers impaired utility of the liver. The A/G ratio normal range0.8 - 2.0.All the records were obtained from UCI

Machine Learning repository for this study. It represents a sample of the unified Indian population collected from Andhra Pradesh region and has the data of 583 patients amongst which 416 are the liver patient records and 167 non liver patient records. Machine learning achieves best result when we have plenitude of data and hence as the dataset for non liver patient is comparatively less, it has been oversampled and made equal to the liver patient record class. The dataset is named as The Indian Liver Patient Dataset(ILPD) and has 11 attributes. The last attribute is the target
class, i.e. 1 represents diseased patients and 2 represents non diseased.It isolates the disease patterns. The Table 1 shows the list of all the attributes
on which methodology is used.
This Project examines data from liver patients concentrating on relationships between a key list of liver enzymes, proteins, age and gender using them to try and predict the likeliness of liver disease. Here we are building a model by applying various machine learning algorithms find the best accurate model. And integrate to flask based web application. User can predict the disease by entering parameters in the web application.

s.no	Attributes	Attribute TYPE
1	Age	Numeric
2	Gender	Nominal
3	Total Bilirubin	numeric
4	Direct Bilirubin	numeric
5	Alkaline phosphatase	numeric
6	Alamine Aminotransferase	numeric
7	Total proteins	numeric

2.2 Proposed Solutions:

There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms that you can choose according to the

objective that you might have may be Classification algorithms are Regression algorithms.

Example:

1. Random Forest Classification.

2. Support Vector Machine

3. KNN Classification

You will need to train the datasets to run smoothly and see an incremental improvement in the prediction rate.

Now we apply classification algorithms on our dataset.

Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. Support Vectors are simply the co-ordinates of individual observation. The goal of a support vector machine is not only to draw hyperplanes and divide data points, but to draw the hyperplane the separates data points with the largest margin, or with the most space between the dividing line and any given data point.

Random Forest Regression: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

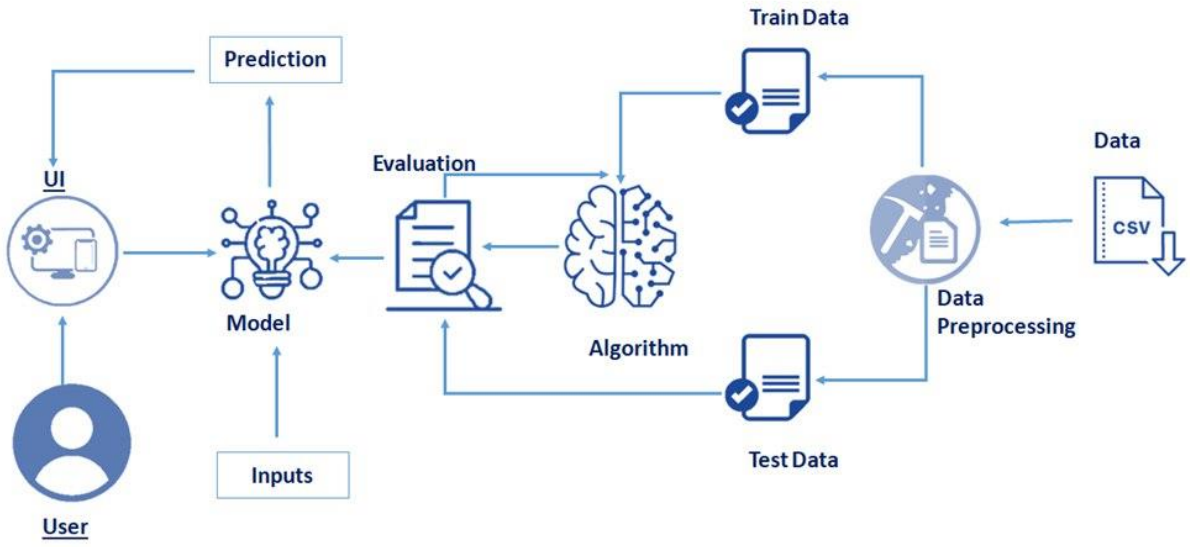
K-Nearest Neighbors algorithm: K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

3. Theoritical Analysis:

3.1 Block diagram –

- User interacts with the UI (User Interface) to upload the input features.
- Uploaded features/input is analyzed by the model which is integrated.
- Once a model analyses the uploaded inputs, the prediction is showcased on the UI.

To accomplish this, we have to complete all the activities and tasks listed below



3.2 Hardware/Software Designing :

In order to develop this project we need to install following software/packages

Anaconda Navigator:

Anaconda Navigator is a free and open-source distribution of the Python and R programming languages for data science and machine learning related applications. It can be installed on Windows, Linux, and macOS.Conda is an open-source, cross-platform, package management system. Anaconda comes with so very nice tools like JupyterLab, Jupyter Notebook,

QtConsole, Spyder, Glueviz, Orange, Rstudio, Visual Studio Code. For this project, we will be using Jupyter notebook and Spyder

To build Machine learning models you must require the following packages

- **Numpy:**
  - It is an open-source numerical Python library. It contains a multidimensional array and matrix data structures and can be used to perform mathematical operations

- **Scikit-learn:**
    - It is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy
- **Matplotlib and Seaborn:**
    - Matplotlib is mainly deployed for basic plotting. Visualization using Matplotlib generally consists of bars, pies, lines, scatter plots and so on. Seaborn: Seaborn, on the other hand, provides a variety of visualization patterns. It uses fewer syntax and has easily interesting default themes.
- **Pandas:**
    - It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

- **Pickle:** The pickle module implements serialization protocol, which provides an ability to save and later load Python objects using special binary format.

If you are using **anaconda navigator**, follow below steps to download required packages:

- Open the anaconda prompt.
- Type “pip install jupyter notebook” and click enter.
- Type “pip install spyder” and click enter.
- Type “pip install numpy” and click enter.
- Type “pip install pandas” and click enter.
- Type “pip install matplotlib” and click enter.
- Type “pip install seaborn” and click enter.
- Type “pip install sklearn” and click enter.
- Type “pip install Flask” and click enter.

If you are using Pycharm IDE, you can install the packages through the command prompt and follow the same syntax as above.

4. Experimental Investigations:

- Data visualization is where a given data set is presented in a graphical format. It helps the detection of patterns, trends and correlations that might go undetected in text-based data.
- Understanding your data and the relationship present within it is just as important as any algorithm used to train your machine learning model. In fact, even the most sophisticated machine learning models will perform poorly on data that wasn't visualized and understood properly.
- To visualize the dataset we need libraries called Matplotlib and Seaborn.
- The Matplotlib library is a Python 2D plotting library which allows you to generate plots, scatter plots, histograms, bar charts etc.

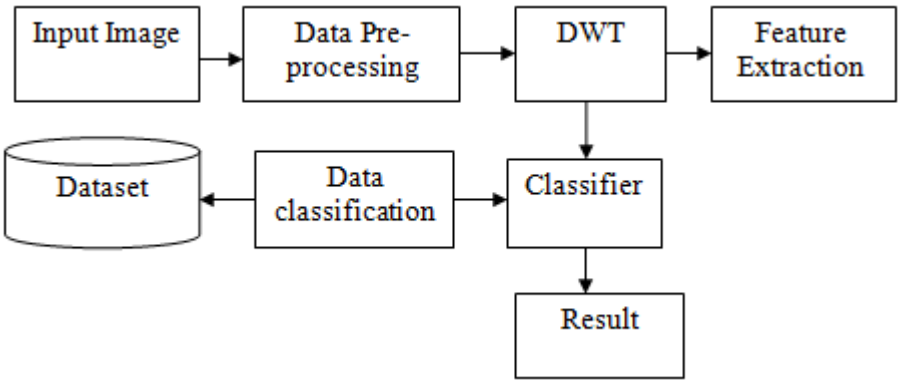
Univariate Analysis

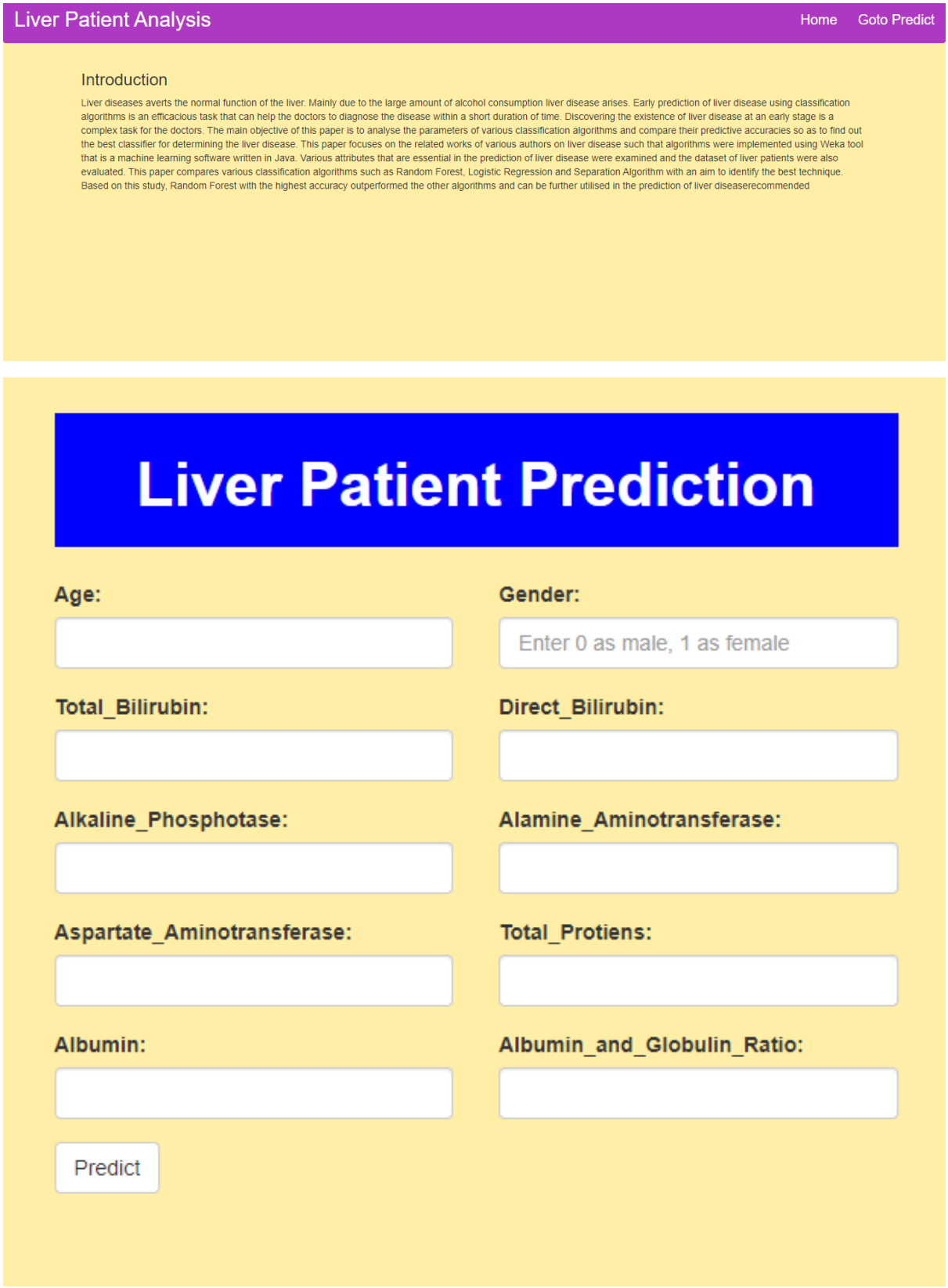
Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable.

Bivariate Analysis

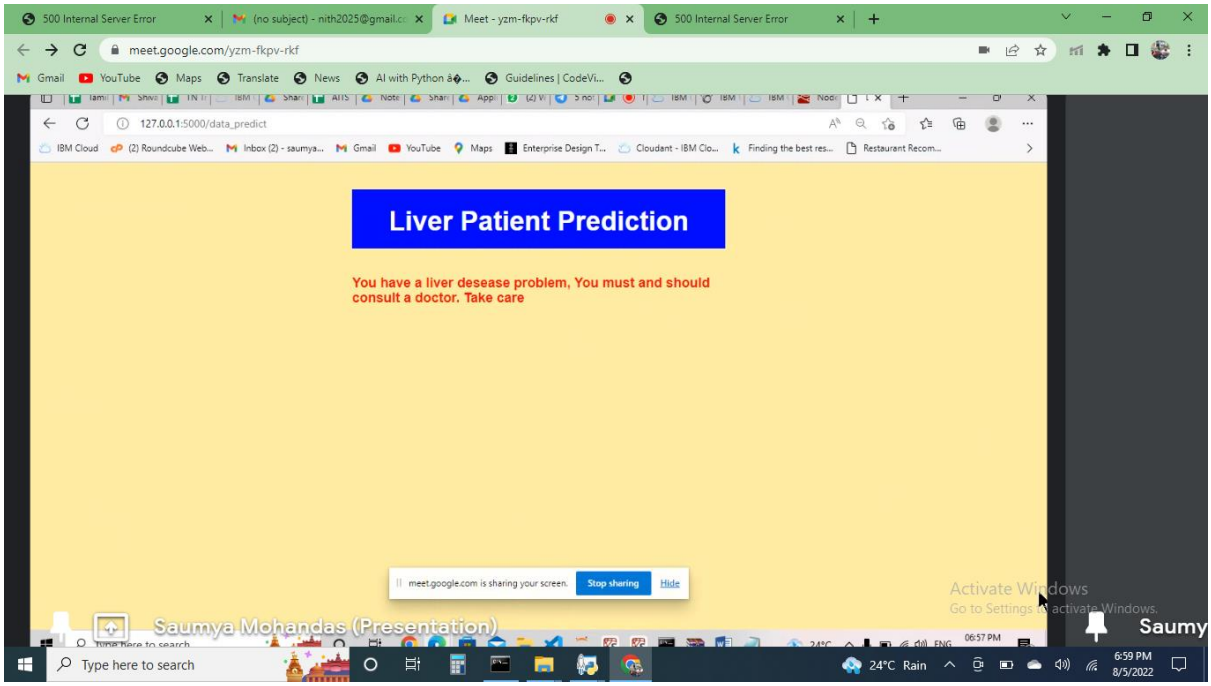
It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.

5. Flow chart:




Liver Patient Analysis
Home
Goto Predict
Introduction
Liver diseases averts the normal function of the liver. Mainly due to the large amount of alcohol consumption liver disease arises. Early prediction of liver disease using classification algorithms is an efficacious task that can help the doctors to diagnose the disease within a short duration of time. Discovering the existence of liver disease at an early stage is a complex task for the doctors. The main objective of this paper is to analyse the parameters of various classification algorithms and compare their predictive accuracies so as to find out the best classifier for determining the liver disease. This paper focuses on the related works of various authors on liver disease such that algorithms were implemented using Weka tool that is a machine learning software written in Java. Various attributes that are essential in the prediction of liver disease were examined and the dataset of liver patients were also evaluated. This paper compares various classification algorithms such as Random Forest, Logistic Regression and Separation Algorithm with an aim to identify the best technique. Based on this study, Random Forest with the highest accuracy outperformed the other algorithms and can be further utilised in the prediction of liver diseaserecommended
Liver Patient Prediction
Age:
Gender:
Enter 0 as male, 1 as female
Total_Bilirubin:
Direct_Bilirubin:
Alkaline_Phosphotase:
Alamine_Aminotransferase:
Aspartate_Aminotransferase:
Total_Protiens:
Albumin:
Albumin_and_Globulin_Ratio:
Predict

6. Result:



7. Advantages & Disadvantages :

ADVANTAGES:

1. **Economical benefit** – Elimination of CAPEX model, low operating cost
2. **Speed** – It is faster to deploy a project without the need of setup
3. **Distributedness**
4. **Reliability through redundancy**
5. **Scalability**
6. **Least burden of server sysadmin works**
7. **Least one-time investment**
8. **Least vendor lock-in**
9. **Least downtime unlike normal server maintenance**

DISADVANTAGES :

1. **Communication problems** are possible resulting in poorer performance
2. Ambivalence between **security** and ease of access to resources
3. **Higher cost** for too bigger projects
4. **Lesser reach of product offerings**, features to the client resulting in wrong product selection

8. Applications:

Patients in India for liver disease are continuously increasing because of

excessive consumption of alcohol, inhale of
harmful gases, intake of contaminated food, pickles and drugs. It is expected that by 2025 India may become the World Capital
for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased
alcohol consumption and smoking. There are about 100 types of liver infections. Therefore, building a model that will help
doctors to predict whether a patient is likely to have liver diseases, at an early stage will be a great advantage. Diagnosis of liver
disease at a preliminary stage is important for better treatment. We also compare different algorithms for the better accuracy.

9.Conclusion:

Initially, the dataset was explored and made ready to be fed into the classifiers. This
was achieved by removing some rows containing null values, transforming some columns
which were showing skewness and using appropriate methods (one-hot encoding) to convert
the labels so that they can be useful for classification purposes. Performance metrics on
which the models would be evaluated were decided. The dataset was then split into a training
and testing set.
Firstly, a naive predictor and a benchmark model ('Logistic Regression') were run on
the dataset to determine the benchmark value of accuracy. The greatest difficulty in the
execution of this project was faced in two areas- determining the algorithms for training and
choosing proper parameters for fine-tuning. Initially, I found it very vexing to decide upon 3
or 4 techniques out of the numerous options available in sklearn.
This exercise made me realize that parameter tuning is not only a very interesting but

also a very important part of machine learning. I think this area can warrant further
improvement, if we are willing to invest a greater amount of time as well as computing
power**.**

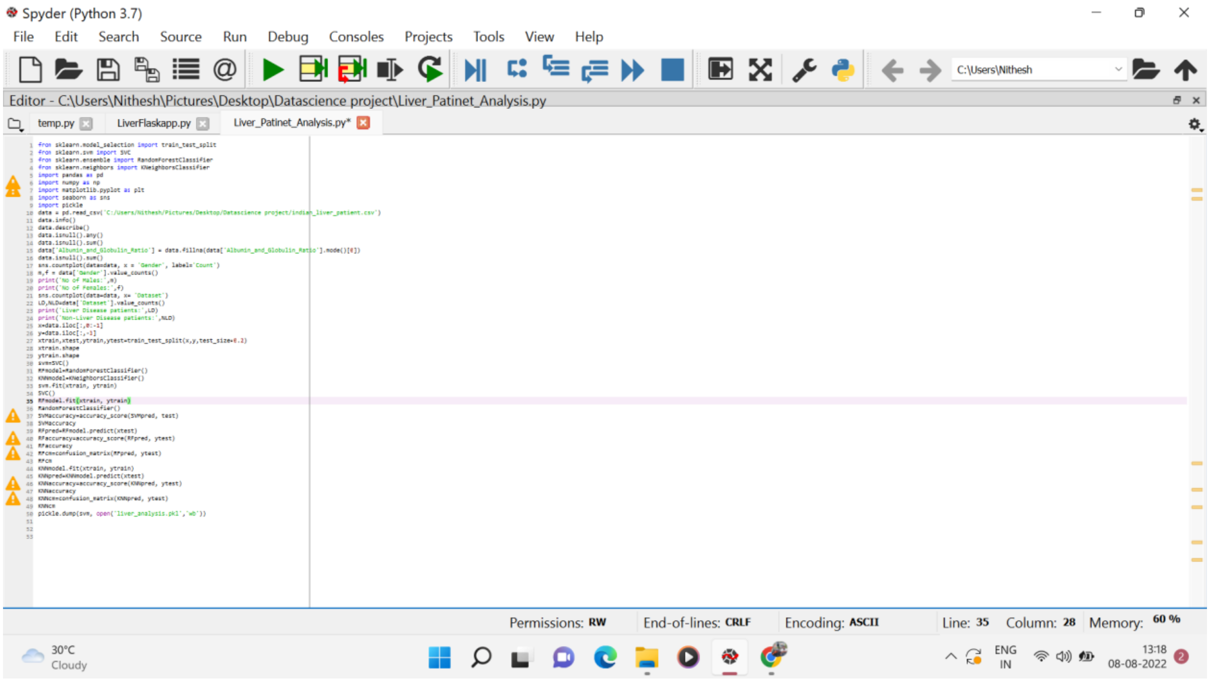
10.Future scope :

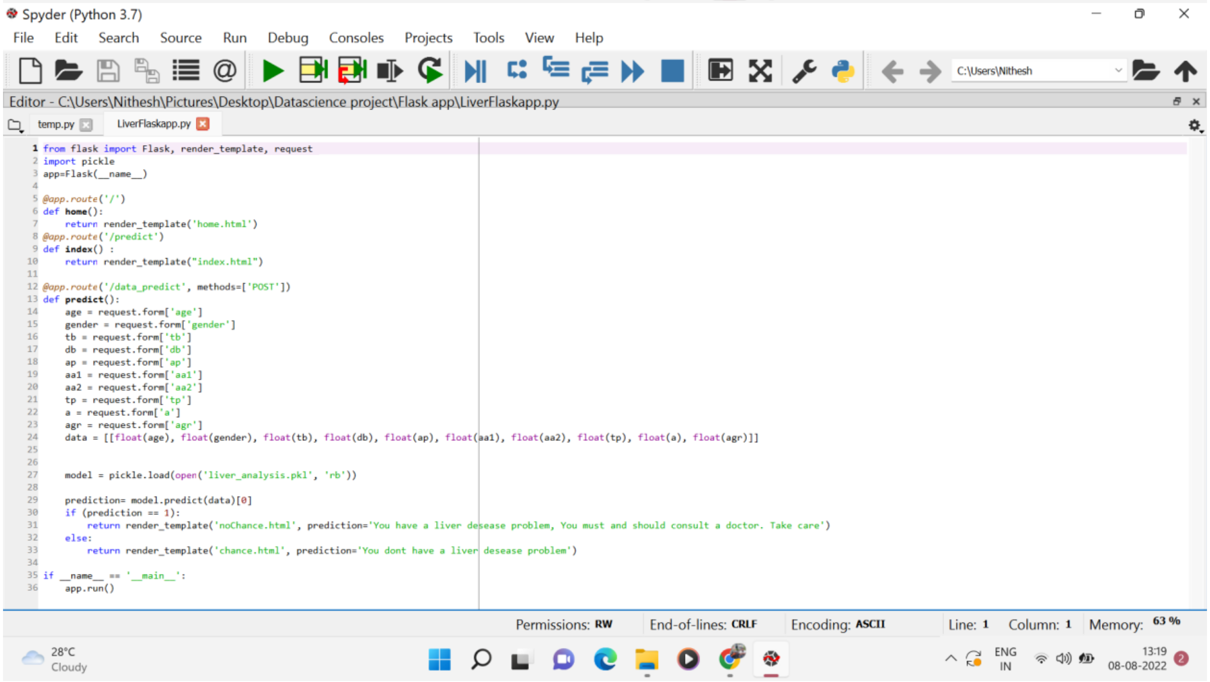
The work done propose an approach for correct prediction of Liver Disease at preparatory stage using various Classifiers with the adoption of Feature
Selection techniques.Future work includes setting more diversity in the model.
Similar approach can benefit diagnosis of other diseases too obligatory for improving ailing condition of patients.

11.BIBILIOGRAPHY :

1. Han C.W. et al., ”Deep Learning for Risk Analysis of Specific CardiovascularDiseases Using Environmental Data and Outpatient Records”
2. Dinesh et al., ”Prediction of Cardiovascular Disease Using Machine Learning Al-gorithms”
3. L. Alice Auxilia, ”Accuracy Prediction using Machine Learning Techniques forIndian Patient Liver Disease”
4. Igor Machine learning for medical diagnosis: history, state of art & perspective”
5. Sana et al.” Analytical study of heart disease comparing with different algorithms”
6. B.Dhomse Kanchan, M.Mahale Kishore Study of Machine learning algorithms forspecial disease prediction using principal of component analysis
7. Matjaz et al.” Analysing and improving the diagnosis of ischaemic heart diseasewith machine learning”
8. Jagdeep et al. ”Prediction of heart diseases using associative classification”
9. M. A. Jabbar, B.L.Deekshatulu, Priti Chandra, ”Alternating decision trees forearly diagnosis of heart diseases”

APPENDIX :


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\Nithesh
Editor - C:\Users\Nithesh\Pictures\Desktop\Datascience project\Liver_Patinet_Analysis.py
temp.py
LiverFlaskapp.py
Liver_Patinet_Analysis.py*
Permissions: RW
End-of-lines: CRLF
Encoding: ASCII
Line: 35
Column: 28
Memory: 60 %
30°C
Cloudy
ENG
IN
13:18
08-08-2022


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\Nithesh
Editor - C:\Users\Nithesh\Pictures\Desktop\Datascience project\Flask app\LiverFlaskapp.py
temp.py
LiverFlaskapp.py
from flask import Flask, render_template, request
import pickle
app=Flask(__name__)
@app.route('/')
def home():
return render_template('home.html')
@app.route('/predict')
def index() :
return render_template("index.html")
@app.route('/data_predict', methods=['POST'])
def predict():
age = request.form['age']
gender = request.form['gender']
tb = request.form['tb']
db = request.form['db']
ap = request.form['ap']
aa1 = request.form['aa1']
aa2 = request.form['aa2']
tp = request.form['tp']
a = request.form['a']
agr = request.form['agr']
data = [[float(age), float(gender), float(tb), float(db), float(ap), float(aa1), float(aa2), float(tp), float(a), float(agr)]]
model = pickle.load(open('liver_analysis.pkl', 'rb'))
prediction= model.predict(data)[0]
if (prediction == 1):
return render_template('noChance.html', prediction='You have a liver desease problem, You must and should consult a doctor. Take care')
else:
return render_template('chance.html', prediction='You dont have a liver desease problem')
if __name__ == '__main__':
app.run()
Permissions: RW
End-of-lines: CRLF
Encoding: ASCII
Line: 1
Column: 1
Memory: 63 %
28°C
Cloudy
ENG
IN
13:19
08-08-2022