LIVER FUNCTION ANALYSIS

(Course Name: Introduction to Python Programming Lab)

(Course Code: 20CS3352)

A Python Project Report on Liver Functionality

Submitted by

G. VENKATA SAI RAM (22501A0557)

AAKASH KODALI (22501A0501)

G. RANI (22501A0562)

B.T. THANUSRI (22501A0528)

A. MRUDULA (23505A0501)

II B.Tech I Sem

in

Computer Science and Engineering

Prasad V Potluri Siddhartha Institute of Technology

Accredited with A+ grade by NAAC, NBA Accredited, and

Autonomous ISO 9001:2015 Certified Institute

Permanently Affiliated to JNTUK-Kakinada and approved by AICTE

Kanuru, Vijayawada-520 007

Prasad V Potluri Siddhartha Institute of Technology

Accredited with A+ grade by NAAC, NBA Accredited, and

Autonomous ISO 9001:2015 Certified Institute

Permanently Affiliated to JNTUK-Kakinada and approved by AICTE

Kanuru, Vijayawada-520 007

CERTIFICATE

This is to certify that the python project report titled " **Liver Function Analysis"** of **Mr. G. Venkata Sai Ram(22501A0557) , Mr. Aakash Kodali (22501A0501), Ms. G. Rani (22501A0562) , Ms. B.T. ThanuSri (225010528), Ms. A. Mrudula(23505A0501)** in Computer Science and Engineering during the academic year 2023-2024.

Signature of the Guide **Signature of the H.O.D**

TABLE OF CONTENTS

TITLE	PAGE NOS
Abstract	
1. Introduction	**4 – 5**
1.1 Background	4
1.2 Motivation	4
1.3 Problem Statement	5
2. Literature Review	**5**
3. Methodology	**5 – 6**
3.1 Data Collection	5
3.2 Data Pre-Processing	6
3.3 Data Visualization	6
4. Project Design	**7**
4.1 Data Flow Diagram.	7
5. Implementation	**8 – 9**
5.1 Algorithms Used	8
5.2 Code Development	8
6. Results and Analysis	**9 – 11**
6.1 Performance Evaluation Metrics	9
6.2 Results	11
7. Conclusion	**11**

1. Introduction

1.1 Background

The liver's main job is to filter the blood coming from the digestive tract, before passing it to the rest of the body . The liver also detoxifies chemicals and metabolises drugs. As it does so, the liver secretes bile that ends up back in the intestines. The liver also makes proteins important for blood clotting and other functions.

Liver Failure is a life-threatening condition that demands urgent medical care. Most often, liver failure happens gradually, over many years. It's the final stage of many liver diseases. But a rarer condition known as acute liver failure happens rapidly (in as little as 48 hours) and can be difficult to detect at first.

Liver failure happens when large parts of the Liver become damaged beyond repair and the liver can't work anymore.

There are two types of live failure:

- Acute: This is when your liver stops working within a matter of days or weeks. Most people who get this don't have any type of liver disease or problem before this event.
- Chronic:

  Damage to your liver builds up over time and causes it to stop working.

  Symptoms of Liver Disease and Liver Failure
- Nausea
- Loss of appetite
- Fatigue
- Diarrhoea

1.2 Motivation

Predicting liver functionality is an important application of healthcare analytics and machine learning that can save lives and improve the quality of healthcare services. The liver is a critical organ in the human body that is responsible for an array of functions that help support metabolism, immunity, digestion, detoxification, vitamin storage among other functions. It comprises around 2% of an adult's body weight .Its dysfunctionality can result in severe health complications, long-term disabilities, or even death. Therefore, developing predictive models for early dysfunction detection and prevention is of utmost importance.

Liver Dysfunctionality is a prominent global cause of mortality and disability. Detecting Liver failure risk indicators and symptoms early can help prevent Liver failure.

Predictive models are instrumental in identifying individuals at high risk, enabling timely interventions. Machine learning plays a crucial role in aiding healthcare providers in enhancing diagnostic accuracy and treatment choices.

Some of the Key factors Associated with risk of Liver Dysfunctionality are

- Heavy alcohol use.
- Obesity.
- Type 2 diabetes.
- Tattoos or body piercings.
- Injecting drugs using shared needles.
- Blood transfusion before 1992.
- Exposure to other people's blood and body fluids.
- Unprotected sex.

1.3 Problem Statement

Machine learning models offer the capability to analyse extensive healthcare datasets, uncover elusive patterns, and detect challenging-to-discern risk factors that conventional methods might miss. These algorithms can predict an individual's liver dysfunctionality risk by examining their health records and lifestyle factors, utilising predictive features such as blood pressure, diabetics, smoking habits, age, and medical history. Moreover, machine learning models can extend their analysis to include additional variables such as genetic factors and lifestyle choices.

2. Literature Review:

- Relevant Previous Work

Relevant previous work related to prediction of Liver dysfunctionality encompasses a range of studies and research efforts focused on identifying risk factors, developing predictive models, and improving the accuracy of dysfunctional risk assessments.

3. Methodology:

3.1 Data Collection

The dataset for liver dysfunctionality prediction is from Kaggle [3]. This particular dataset has 11 columns. The columns have ‘Age’, ‘Gender’, ‘Total_Bilirubin’, ‘Direct_Bilirubin’,‘Alkaline_Phosphotase’,‘Alamine_Aminotransferase’,‘Aspartate_Aminotransferase’, ‘Total_Protiens’, ‘Albumin’, ‘Albumin_and_Globulin_Ratio’ and 'Dataset' as the main attributes. The output column 'Dataset' has the value as either '1' or '2'. The value '1' indicates no dysfunctionality detected, whereas the value '2' indicates a possible risk of dysfunctionality. This dataset is highly imbalanced as the possibility of '1' in the output column ('dysfunctionality') outweighs that of '2' in the same column. Only 167 rows have the value '2'. whereas 416 rows with the value '1' in the dataset column.

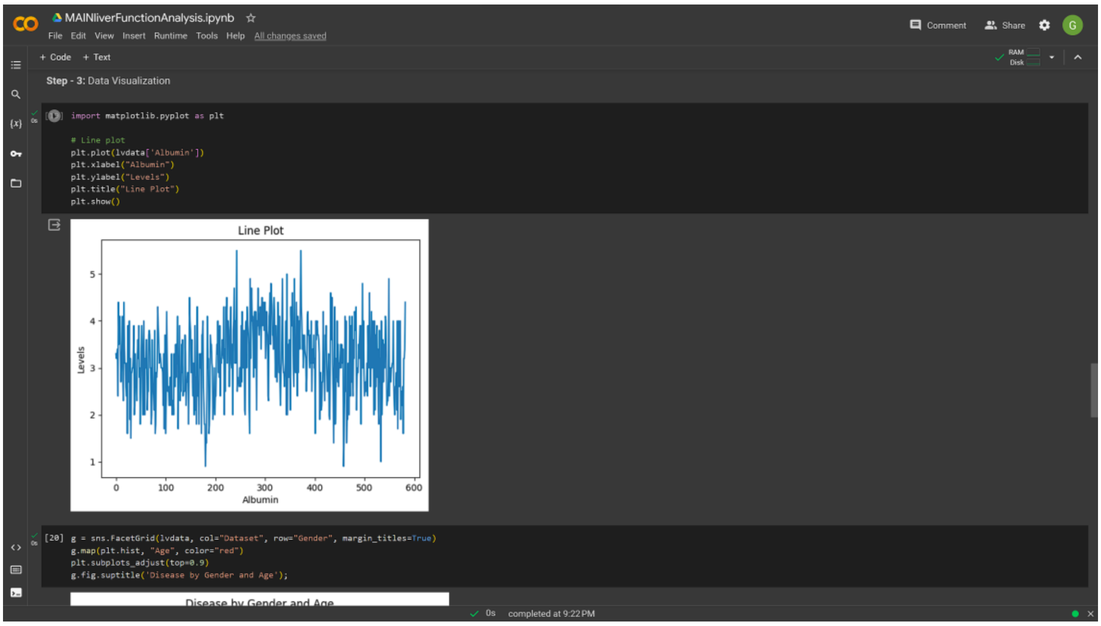
3.2 Data Pre-processing

Data Preprocessing is required before model building to remove the unwanted noise and outliers from the dataset, resulting in a deviation from proper training. Anything that interrupts the model from performing with less efficiency is taken care of in this stage. After collecting the appropriate dataset, the next step lies in cleaning the data and making sure that it is ready for model building. The dataset taken has 11 attributes, as mentioned in Table. Firstly, the dataset is checked for null values and filled if any found. In this case, the column ' Albumin_and_Globulin_Ratio' has null values filled with the 0's or also you can fill with the mean of the column. After removing the null values from the dataset, the next task is Label Encoding.

The dataset chosen for the task of dysfunctionality prediction is slightly imbalanced. The entire dataset has 167 suggesting the occurrence of dysfunctionality and 416 rows having the possibility of no dysfunctionality. The graphical representation of the imbalance is in Fig.

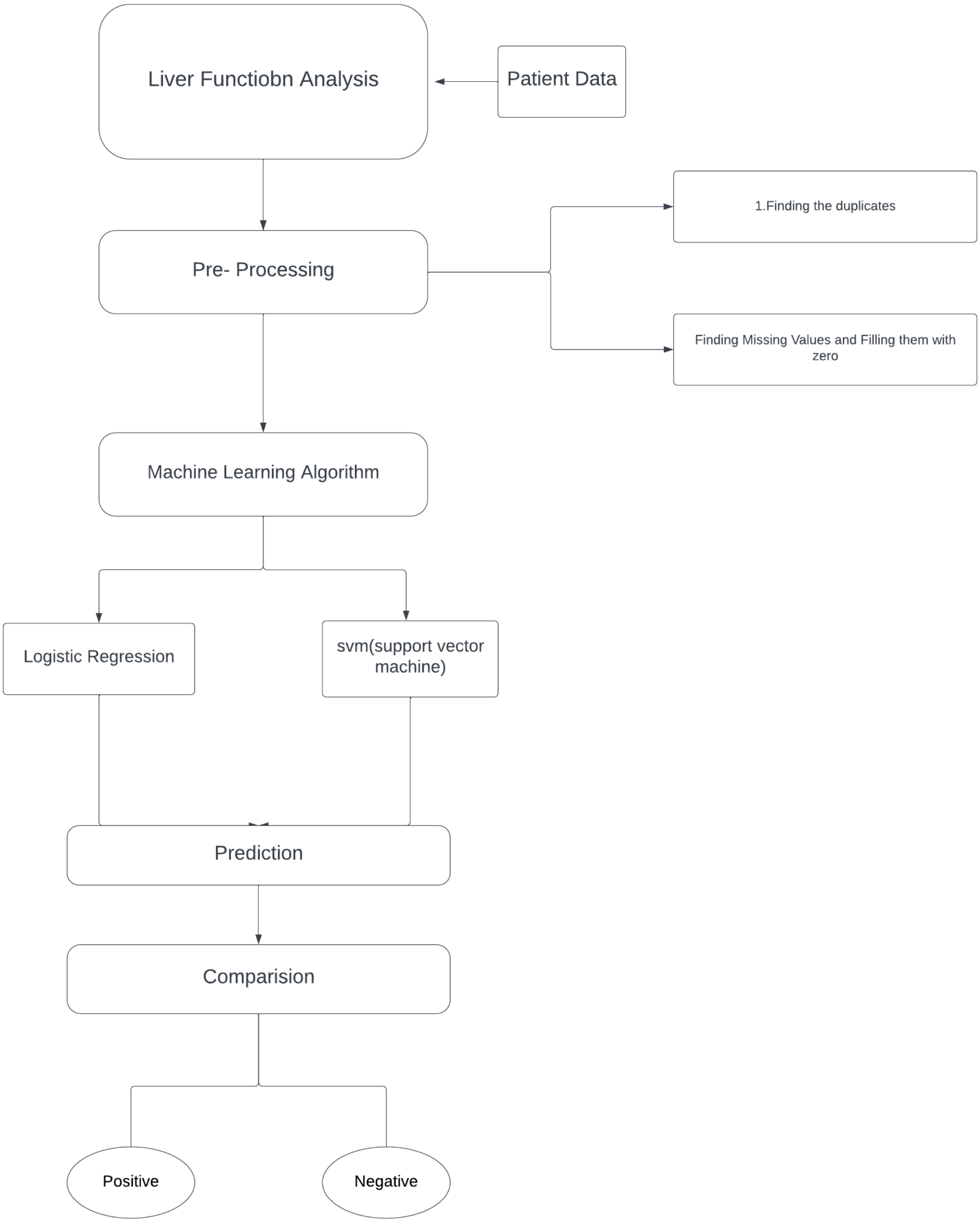
Training a machine-level model with such data might give accuracy, but other accuracy metrics like precision and recall are shallow. Since the difference between the dataset 1 and the dataset 2 is very low, there is no need to do the undersampling for the dataset.

3.3 Data Visualization



4. Project Design:

4.1Data Flow Diagram



5. Implementation:

5.1 Algorithm Used

Logistic regression

Logistic regression is a statistical method and a type of regression analysis used for predicting the probability of a binary outcome (1 / 0, Yes / No, True / False) based on one or more predictor variables. It's particularly useful when the dependent variable is categorical, and it's commonly employed for classification tasks in machine learning.

5.2 Code Development

```
from sklearn.model_selection import train_test_split #training and testing data split
from sklearn import metrics #accuracy measure
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score, classification_report #for confusion matrix
from sklearn.linear_model import LogisticRegression,LinearRegression #logistic regression
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
train, test = train_test_split(lvdata, test_size=0.3, random_state=0, stratify=lvdata['Dataset'])

train_X = train[train.columns[:-1]]
train_Y = train[train.columns[-1]]
test_X = test[test.columns[:-1]]
test_Y = test[test.columns[-1]]

X = lvdata[lvdata.columns[:-1]]
Y = lvdata['Dataset']

len(train_X), len(train_Y), len(test_X), len(test_Y)

# Create a label encoder object
le = LabelEncoder()

# Assuming 'Gender' is the column causing the issue, we transform it
train_X['Gender'] = le.fit_transform(train_X['Gender'])
test_X['Gender'] = le.transform(test_X['Gender'])
```

```
# Now you can scale your features
scaler = StandardScaler()
train_X_scaled = scaler.fit_transform(train_X)
test_X_scaled = scaler.transform(test_X)

# Train the model
model = LogisticRegression()
model.fit(train_X_scaled, train_Y.values.ravel())

# Make predictions
prediction3 = model.predict(test_X_scaled)

# Print the accuracy and classification report
print('The accuracy of the Logistic Regression is', metrics.accuracy_score(prediction3, test_Y.values.ravel()))
report = classification_report(test_Y, prediction3)
print("Classification Report:\n", report)
```

6. Results and Analysis:

6.1 Performance Evaluation metrics

When evaluating a machine learning model for predicting liver functionality, we typically use various performance metrics to assess its effectiveness. Below are some common performance metrics for liver dysfunctionality prediction:

1. **Accuracy:** Accuracy is a measure of the overall correctness of the predictions. It calculates the ratio of correctly predicted instances to the total number of instances. However, accuracy might not be the best metric if the data is imbalanced.
2. **Precision:** Precision is the ratio of true positive predictions to the total number of positive predictions made. It measures how many of the predicted liver dysfunction cases are actual liver dysfunctions.

3. **Recall (Sensitivity or True Positive Rate):** Recall is the ratio of true positive predictions to the total number of actual liver dysfunction cases. It quantifies the model's ability to identify all actual liver dysfunction cases.
4. **F1-Score:** The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall. It is especially useful when you want to find an optimal balance between false positives and false negatives.
5. **Specificity (True Negative Rate):** Specificity is the ratio of true negative predictions to the total number of actual non-liver dysfunction cases. It measures the model's ability to correctly identify non-liver dysfunction cases.
6. **Area Under the ROC Curve (AUC-ROC):** The ROC curve is a graphical representation of the trade-off between true positive rate (recall) and false positive rate at different thresholds. AUC-ROC quantifies the model's ability to distinguish between liver dysfunction and non-liver dysfunction cases.
7. **Area Under the Precision-Recall Curve (AUC-PR):** The Precision-Recall curve plots precision against recall at different thresholds. AUC-PR quantifies the precision-recall trade-off.
8. **Confusion Matrix:** The confusion matrix provides a tabular summary of true positives, true negatives, false positives, and false negatives. It's helpful for a detailed understanding of model performance.
9. **False Positive Rate (FPR):** The FPR is the ratio of false positive predictions to the total number of actual non-liver dysfunction cases. It measures the model's propensity to incorrectly predict liver dysfunction.
10. **True Negative Rate (TNR):** TNR is another term for specificity and measures the model's ability to correctly identify non-liver dysfunction cases.

- **Comparison with different algorithms:**

Logistic Regression is a supervised machine learning algorithm used for classification tasks, predicting the probability of an instance belonging to a given class. It's a statistical algorithm that analyses the relationship between a set of independent variables and the dependent binary variables. Other algorithms for classification tasks include Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbours, Support Vector Machines, and Neural Networks. Each has its own strengths and weaknesses.

6.2 Results

```
The accuracy of the Logistic Regression is 0.7142857142857143
Classification Report:
              precision    recall  f1-score   support

           1       0.75      0.90      0.82       125
           2       0.50      0.26      0.34        50

    accuracy                           0.71       175
   macro avg       0.63      0.58      0.58       175
weighted avg       0.68      0.71      0.68       175
```

7. **Conclusion**: Based on these points we need to frame a conclusion
   - **Summary of Findings**

     The liver function analysis project, implemented in Python, has successfully utilized various data analysis techniques to understand liver functions better. The project has processed and analysed a vast amount of data related to liver functions, leading to several significant findings about liver health and disease.

   - **Achievements**

     The project has created a tool using Python that can understand and make sense of complex liver function data. This tool has been very useful in finding patterns and trends in the data that we didn't know about before. Additionally, through this project, we have gained valuable experience and knowledge in analysing medical data using Python.

   - **Future Work**

     The next steps for this project include refining the analysis tool to incorporate more advanced machine learning algorithms for better predictive capabilities. Additionally, the team plans to extend the project to analyse other organ functions using similar methodologies. The ultimate goal is to create a comprehensive health analysis platform that can provide valuable insights into various aspects of human health.