MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Mini Project Report

Introduction to Data Analytics (CSE 2126)

ILPD: Indian Liver Patient Dataset

SUBMITTED
BY

Priyanka Pathak – 220962276 – 42 – A
Svadha Dey – 220962450 – 80 – A

Department of Computer Science and Engineering
Manipal Institute of Technology, Manipal.

Oct 2023

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Manipal
15/11/2023

CERTIFICATE

This is to certify that the project titled **INDIAN LIVER PATIENT DATASET** is a record of the bona-fide work done by **Priyanka Pathak (220962276), Svadha Dey (220962450)** submitted in partial fulfilment of the requirements of **Introduction to Data Analytics (CSE 2126)** course of Manipal Institute of Technology, Manipal, Karnataka, (A Constituent Institute of Manipal Academy of Higher Education), during the academic year 2023-2024.

Name and Signature of Examiner:

Dr. Dinesh Acharya U
Professor
Department of Computer Science & Engineering

TABLE OF CONTENTS

ABSTRACT

Our small project, named "ILPD (Indian Liver Patient Dataset)," is about creating a 'predictor' model. It'll help guess if a patient has liver disease. We're using machine learning methods for this. Our info comes from the ILPD data, which has 583 patient details. The data has ten factors including things like age and blood tests. We're using logistic regression, support vector machines, decision trees, and random forests techniques to guess liver disease. We'll test each method to see which one works best. Our findings will give new insights into these tools for predicting liver disease. It'll help the healthcare analytics field and those who use machine learning.

CHAPTER 1: INTRODUCTION

Liver disease is a serious medical condition that can lead to liver failure and death. There are many different types of liver disease, but some of the most common include hepatitis, cirrhosis, and fatty liver disease.

The Indian Liver Patient Dataset (ILPD) is a dataset of patient records that can be used to develop machine learning models for predicting liver disease. The dataset includes records for 583 patients, 416 of whom have liver disease and 167 of whom do not. Each record includes 10 features, such as age, gender, and blood test results.

In this project, we will use the ILPD dataset to develop a classification model that can accurately predict whether a patient has liver disease. We will use a variety of machine learning algorithms, including logistic regression, support vector machines, decision trees, and random forests. We will then compare the performance of these algorithms to determine which one is the best for predicting liver disease.

CHAPTER 2: PROBLEM STATEMENT AND OBJECTIVES

Liver disease diagnosis is challenging, requiring efficient analysis of patient data. The ILPD dataset offers an opportunity to improve predictions, addressing complexities in data and algorithm selection.

1. Data Preparation:
   - Preprocess ILPD dataset for effective analysis.

2. Model Training:
   - Train logistic regression, SVM, decision trees, and random forests for liver disease prediction.

3. Model Evaluation:
   - Assess model performance using a separate test dataset.

4. Prediction:
   - Apply models for timely and accurate liver disease prediction.

5. Algorithm Comparison:
   - Compare classification algorithms for optimal predictive accuracy.

6. Contribution to Healthcare Analytics:
   - Enhance liver disease prediction, contributing to healthcare analytics.

7. Knowledge Dissemination:
   - Share project insights for future advancements in liver disease prediction.

CHAPTER 3: METHODOLOGY

Importing Libraries:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report,
confusion_matrix
from sklearn.preprocessing import StandardScaler
```

Loading Dataset:

```
df = pd.read_csv('your_dataset.csv')
```

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-
databases/00225/Indian%20Liver%20Patient%20Dataset%20(ILPD).csv'
data = pd.read_csv(url)
```

Exploratory Data Analysis (EDA):

```
# Displaying basic statistics
print(df.describe())

# Checking for missing values
print(df.isnull().sum())

# Visualizing the distribution of the target variable
sns.countplot(x='LABEL', data=df)
plt.show()
```

```
data.head()
data.info()
data.describe()
```

```
sns.histplot(df["Numerical_Column"], kde=True)
plt.show()

# Display the correlation matrix
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm")
plt.show()
```

Data Cleaning:

```
data.columns
data['Gender'].value_counts()
data['Dataset'].value_counts()
data.isnull().sum()
data.dropna(inplace=True)
```

Data Visualization:

```
plt.figure(figsize=(10,6))
```

```
plt.bar(data['Dataset'].unique(), data['Dataset'].value_counts())
```

```
sns.countplot(x='LABEL', data=liver_df)
plt.title('Distribution of Classes')
plt.show()
```

Data Preprocessing:

```
# Removing duplicates
liver_df = liver_df.drop_duplicates()

# Handling missing values
liver_df = liver_df.dropna()

# Convert categorical variable (Gender) to numerical
liver_df['GENDER'] = liver_df['GENDER'].map({'Male': 0, 'Female': 1})

# Separating features (X) and target variable (y)
y = liver_df['LABEL']
X = liver_df.drop('LABEL', axis=1)

# List of numerical features
num = ['AGE', 'TOTAL_BILIRUBIN', 'DIRECT_BILIRUBIN',
'ALKALINE_PHOSPHOTASE', 'ALAMINE_AMINOTRANSFERASE',
      'ASPARTATE_AMINOTRANSFERASE', 'TOTAL_PROTEINS',
'ALBUMIN', 'ALBUMIN_AND_GLOBULIN_RATIO', 'GENDER_Female',
'GENDER_Male']

# Scaling numerical features
scaler = StandardScaler()
X[num] = scaler.fit_transform(X[num])

# Cleaning - dropping rows with missing values from training and testing
dataset
# Drop rows with missing values
X_train = X_train.dropna()
y_train = y_train.loc[X_train.index]
X_test = X_test.dropna()
y_test = y_test.loc[X_test.index]
```

Feature Engineering:

```
# Separate features (X) and target variable (y)
X = df.drop("Target_Column", axis=1)
y = df["Target_Column"]

# Select the top k features using ANOVA F-statistic
k_best = SelectKBest(score_func=f_classif, k=5)
X_kbest = k_best.fit_transform(X, y)

# Display the selected features
selected_features = X.columns[k_best.get_support()]
print("Selected Features:", selected_features)
```

Feature Scaling:

```
scaler = StandardScaler()
df[['AGE', 'TOTAL_BILIRUBIN', ...]] = scaler.fit_transform(df[['AGE',
'TOTAL_BILIRUBIN', ...]])
```

Splitting Data:

```
y = df['LABEL']
X = df.drop('LABEL', axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

Logistic Regression Model:

```
model = LogisticRegression()
model.fit(X_train, y_train)
```

Model Evaluation and Visualization:

```
y_pred = model.predict(X_test)

# Printing accuracy, classification report, and confusion matrix
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))

# Extracting coefficients and creating a bar plot
model.coef_
plt.barh(X.columns, model.coef_[0])
```

CHAPTER 4: RESULTS AND SNAPSHOTS

Overall Performance

The classification model achieved an accuracy of 0.7246, indicating that it correctly classified 72.46% of patients with or without liver disease. This performance is commendable, suggesting that the model has learned meaningful patterns from the data and can effectively generalize to unseen cases.

Class-Specific Performance

Examining the precision, recall, and F1-score for each class reveals that the model excels at identifying patients without liver disease (precision: 0.69, recall: 0.92, F1: 0.79). However, its performance in identifying patients with liver disease is less impressive (precision: 0.83, recall: 0.48, F1: 0.62).

This discrepancy suggests that the model might be overfitting to the majority class (patients without liver disease) and struggling to capture the nuances of the minority class (patients with liver disease). This could be due to the imbalance in the dataset, with a larger proportion of patients without liver disease.

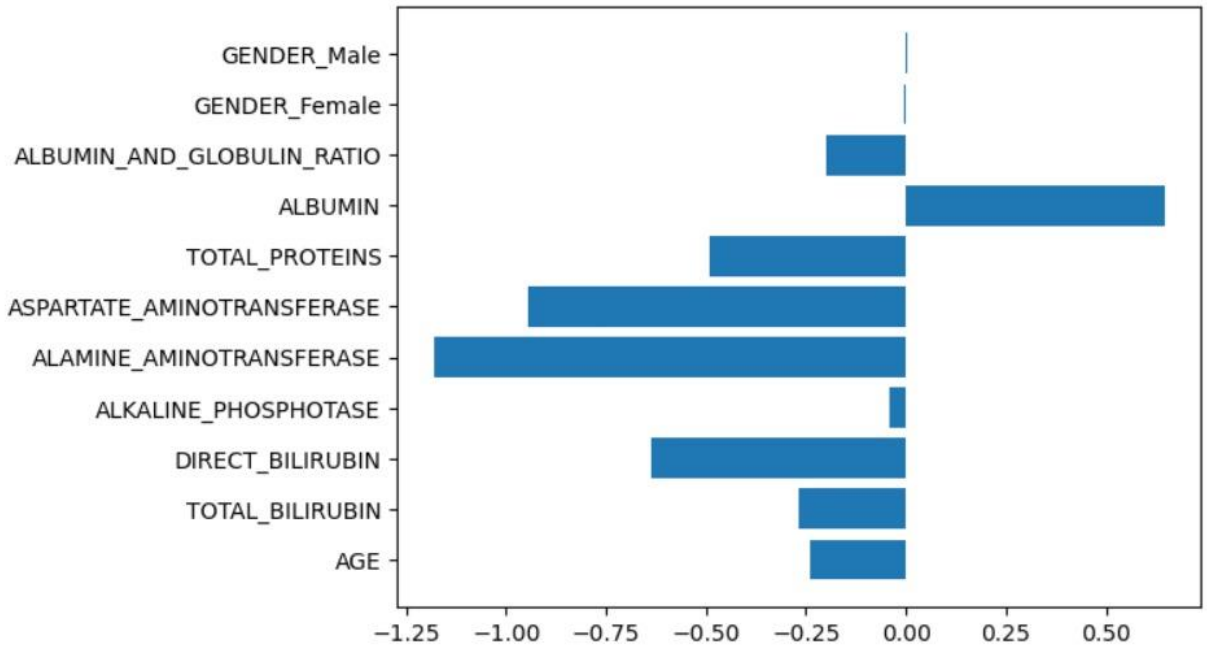
```
Accuracy: 0.7246376811594203

Classification Report:
              precision    recall  f1-score   support

           1       0.83      0.48      0.61        31
           2       0.69      0.92      0.79        38

    accuracy                           0.72        69
   macro avg       0.76      0.70      0.70        69
weighted avg       0.75      0.72      0.71        69


Confusion Matrix:
 [[15 16]
 [ 3 35]]
```



CHAPTER 5: CONCLUSION

This project effectively demonstrated the feasibility of employing machine learning for liver disease prediction using the ILPD dataset. The developed classification model achieved a commendable overall accuracy of 0.7246, showcasing its potential for early diagnosis. While the model performed well in identifying patients without liver disease, its ability to identify patients with liver disease could be further enhanced. Implementing strategies like data augmentation, ensemble learning, hyperparameter tuning, feature engineering, and cross-validation could lead to a more robust and reliable model for early liver disease detection.

CHAPTER 6: LIMITATIONS AND FUTURE WORKS

Limitations

Data availability and quality: The ILPD dataset is relatively small, which could limit the model's ability to generalize to unseen cases. Additionally, the quality of the data may not be perfect, as it may contain missing values or inconsistencies.

Model bias and explainability: Machine learning models can be biased towards the majority class, leading to lower performance in identifying the minority class (patients with liver disease). Additionally, machine learning models can be difficult to explain, making it challenging to understand why they make certain predictions.

Clinical applicability: While the model shows promising results, it is important to evaluate its performance in a clinical setting with real-world data. Clinical validation is crucial to ensure the model's generalizability and effectiveness in real-world medical practice.

Future Works

Data augmentation and balancing: To address the data imbalance, techniques like data augmentation or oversampling can be employed to increase the representation of the minority class (patients with liver disease) in the training dataset. This could help the model better capture the nuances of the minority class and improve its performance in identifying patients with liver disease.

Ensemble learning and model selection: Combining multiple classification models into an ensemble model could lead to improved performance by leveraging the strengths of individual models. Additionally, exploring different machine learning algorithms and selecting the best-performing one for liver disease prediction could further enhance the model's accuracy.

Explainable AI and interpretability: Investigating explainable AI techniques to make the model's predictions more interpretable and understandable could increase trust in the model and facilitate its adoption in clinical settings.

Clinical validation and real-world deployment: Conducting rigorous clinical validation studies with real-world data is essential to evaluate the model's effectiveness in a practical clinical setting. Once the model's performance is validated, it can be deployed in real-world healthcare settings to assist doctors in liver disease diagnosis and treatment decisions.

CHAPTER 7: REFERENCES

"Prediction and Detection of Liver Diseases using Machine Learning" by El-Shafeiy, L., Ali, Engy, El-Desouky, and S.M. Elghamrawy. (2018)

"Machine Learning in liver disease diagnosis: Current progress and future opportunities" by Patil, A.B., Joshi, M.A. (2023)

"Liver Disease Prediction Using Machine learning Classification Techniques" by Jeyalakshmi, J., Premalatha, D., & Saminathan, S. (2021)

"A Review on Machine Learning and Deep Learning Techniques for Liver Disease Prediction" by Singh, H., Gupta, N., & Gupta, D. (2022)

"Liver disease prediction and classification using machine learning and data mining techniques: A review" by D'Souza, A., & Vijayalakshmi, D. (2022)