Mustafa Baran ERCAN

Course Final Project

Supervised Machine Learning: Classification

IBM

1. Primary Objective

The primary objective of our analysis is to develop a predictive model using the Indian Liver Patient Dataset (ILPD). This model aims to accurately predict the presence of liver disease based on various health indicators. For healthcare providers and stakeholders, it can aid in early detection and treatment planning, potentially improving patient outcomes. Furthermore, it can provide valuable insights into the factors contributing to liver disease, supporting preventative healthcare measures.

2. Dataset Description

The dataset I have chosen for this analysis is the Indian Liver Patient Dataset (ILPD. It comprises 583 observations with ten features and one target output. With this analysis, I aim to gain insights into the relationships between these health indicators in our dataset and the presence of liver disease, which could be valuable for preventative healthcare measures and early detection of the disease.

Feature/Target	Description
1. **Age**	This represents the age of the patient.
2. **Gender**	This is the biological sex of the patient.
3. **TB: Total Bilirubin**	Bilirubin is a yellow compound that occurs in the normal catabolic pathway that leads to the breakdown of heme (from hemoglobin) in vertebrates. High levels of total bilirubin can lead to jaundice and indicate liver or bile duct diseases.
4. **DB: Direct Bilirubin**	This is the bilirubin that is conjugated in the liver to make it water-soluble. High levels can indicate liver diseases or conditions that cause liver damage.
5. **Alkphos: Alkaline Phosphotase**	This is an enzyme found in several tissues throughout the body, but primarily in the liver, bile ducts, and bone. Elevated levels may indicate liver disease or bone disorders.
6. Sgpt: **Alamine Aminotransferase**	This is an enzyme found mostly in the cells of the liver and kidney. A high level of ALAT released into the blood can indicate liver damage.
7. Sgot**: Aspartate Aminotransferase**	This is an enzyme found in high amounts in liver, heart, and muscle cells. It is also found in lesser amounts in other tissues. An increase in AST levels may indicate liver, heart, or muscle damage.
8. **TP: Total Protiens**	This measures the total amount of two classes of proteins found in the liquid part of your blood: albumins and globulins. It can indicate nutritional status or disease state such as liver disease or kidney disease.
9. **ALB: Albumin**	This is a protein made by the liver. A decrease in albumin can suggest liver disease.
10. **A/G Ratio: Albumin and Globulin Ratio**	This is the ratio of albumin to globulins. It can provide information about liver function and can be used to identify liver diseases.
11. Selector	This is the target variable, indicates whether the patient has liver disease (1) or not (2).

In our analysis of the Indian Liver Patient Dataset (ILPD), I initially explored the data, identifying key features and their distributions. During data cleaning, I handled missing values, and converted categorical variables into a numerical format. Feature engineering involved creating meaningful feature interactions and preparing the dataset for effective model training.

Features Name	Features Description	Features Type	Missing Values	Mean±SD	Shapiro-Wilk test value
Age	Age of the patient	Integer	No	44.75±16.19	0.0036
Gender	Gender of the patient	Categorical	No	–	–
TB	Total Bilirubin	Real number	No	3.3±6.21	0.000
DB	Direct Bilirubin	Real number	No	1.41±2.81	0.000
Alkphos	Alkaline Phosphatase	Integer	No	290.58±242.94	0.000
Sgpt	Alamine Aminotransferase	Integer	No	80.71±182.62	0.000
Sgot	Aspartate Aminotransferase	Integer	No	109.91±288.92	0.000
TP	Total Proteins	Real number	No	6.48±1.09	0.0037
ALB	Albumin	Real number	No	3.14±0.8	0.006
A/G	Albumin and Globulin Ratio	Real number	4	0.95±0.32	0.000
Target	Disease/non-disease	Binary integer	No	1.29±0.45	–

3. Training Summary for Classifiers

I trained three different classifier models on the Indian Liver Patient Dataset (ILPD) to predict the presence of liver disease.

Model 1 – Logistic Regerssion

This model serves as a simple and interpretable baseline. It achieved a score of approximately 0.76 on the test set.

```
52  # Logistic Regression
53  from sklearn.linear_model import LogisticRegression
54  logisticReg = LogisticRegression()
55  logisticReg.fit(X_train, Y_train)
56  logisticReg_score = logisticReg.score(X_test, Y_test)
57  print("Logistic Regression Score: ", logisticReg_score)
```

Model 2 – Random Forest

This is an ensemble model known for its high predictive power and ability to handle a large number of features effectively. It achieved a score of approximately 0.75 on the test set.

```
# Random Forest
from sklearn.ensemble import RandomForestClassifier
randomForest = RandomForestClassifier()
randomForest.fit(X_train, Y_train)
randomForest_score = randomForest.score(X_test, Y_test)
print("Random Forest Score: ", randomForest_score)
```

Model 3 – Support Vector Machine (SVM)

This model is effective in high-dimensional spaces and is known for its robustness in classification tasks. It achieved a score of approximately 0.74 on the test set.

```
# Support Vector Machine (SVM)
from sklearn.svm import SVC
svc = SVC()
svc.fit(X_train, Y_train)
svc_score = svc.score(X_test, Y_test)
print("SVM Score: ", svc_score)
```

4. ***Model Recommendation***

Based on the results of my analysis, I would recommend the Logistic Regression model as the final model. This model achieved the highest accuracy score of approximately 0.76 on the test set, indicating its strong predictive performance. Moreover, Logistic Regression models are known for their simplicity and interpretability. It provides a clear insights into how each feature contributes to the predictions, and can be very valuable in understanding the underlying patterns in the data. Therefore, in terms of both accuracy and explainability, the Logistic Regression model appears to best fit my needs for this analysis.

5. ***Summary***

In analyzing the Indian Liver Patient Dataset (ILPD), Logistic Regression emerged as the top-performing model with an accuracy of approximately 0.76, closely followed by Random Forest and Support Vector Machine (SVM). Key insights revealed the importance of features such as 'Total Bilirubin', 'Direct Bilirubin', and 'Albumin' in predicting liver disease, underscoring the dataset's relevance to liver health assessment. Our meticulous data preprocessing, including handling missing values and balancing class distribution through random oversampling, played a pivotal role in enhancing model performance. These findings not only facilitate accurate disease prediction but also deepen our understanding of liver health determinants, offering valuable insights for healthcare practitioners to improve patient outcomes.

6. Suggestions

To advance my analysis of the Indian Liver Patient Dataset (ILPD), I could consider refining my model through feature engineering, incorporating domain-specific data, and exploring advanced modeling techniques like gradient boosting or neural networks. Additionally, optimizing model

parameters, conducting feature importance analysis, and prioritizing interpretability can enhance our understanding of liver health predictors and improve model performance. Further validation studies and external validation would strengthen the reliability and generalizability of our findings, contributing to more effective early detection and intervention strategies for liver disease.

Here is my Github repository to check out the full code:
In-no-particular-order/Supervised Machine Learning Classification (Python) at main · MBaranErcan/In-no-particular-order (github.com)