**Comparison of Several Machine Learning Models for Early Detection and Prediction of Liver Disease**

Davy Viriya Chowa, Felicia Nataniaa, Zaafira Sekar Safitria, Jason Susantoa, Hansen Andersen Sutiknoa

*aComputer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

**Abstract**

The liver is essential for several metabolic activities, including protein synthesis, the creation of enzymes, and detoxification. Specific proteins such as bilirubin, albumin, proteins, and alkaline phosphatase which are associated with liver function and dysfunction are some crucial measures of the health of the liver. Cirrhosis and Chronic Liver Disease (CLD) are the main causes of morbidity and mortality in the world. Liver disease can be predicted and detected early, which can greatly enhance patient outcomes. This study uses machine learning techniques to detect liver illness in its early stages and forecast its course. The study makes use of a dataset of liver patients from India that contains numerous clinical factors. To improve data quality, pre-processing techniques including label encoding, imputation of missing values, and standardization are used. The dataset's dimensionality is decreased by using Principal Component Analysis. For classification and regression problems, a variety of machine learning techniques are utilized, such as Gradient Boosting, Stochastic Gradient Descent, K-Nearest Neighbor, Random Forest, Light Gradient Boosting Machine Classifier, Gaussian Naive Bayes, Decision Tree, and Logistic Regression. The findings of this study shed light on prospective biomarkers for early liver disease indications as well as proteomic fingerprints. The prediction model shows the potential to enhance early identification and management of liver disorders by employing machine learning methodologies.

*Keywords:* Liver Disease, Early Prediction, Machine Learning, Data visualization

1. **Introduction**

The major cause of mortality and morbidity worldwide is Chronic Liver Disease (CLD) and Cirrhosis. It is the 9th leading cause of death, having the crude rate death of Chronic liver disease and cirrhosis (K70, K73-K74) is reported to be 17.0 per 100,000 population [1]. Alcohol-related liver disease (ALD), non-alcoholic (metabolic) fatty liver disease (NAFLD), and viral hepatitis are the most common causes of CLD. These may be avoided, and hepatitis C (HCV) is now treatable. Hepatocellular carcinoma (HCC), the eighth most common cause of cancer death in the UK, will develop in between 5% to 10% of people with liver cirrhosis [2].

The liver, the body's largest organ, is responsible for a wide range of essential metabolic processes. The liver, which is responsible for multiple processes including protein synthesis, the production of digestive enzymes, and primary detoxification of different metabolites, is situated in the right upper quadrant of the body and lies below the diaphragm [3]. The regulation of red blood cells (RBCs), the production of glucose, and its storage are all important functions of the liver. Alkaline phosphatase, gamma-glutamyl transferase, 5'-nucleotidase, total bilirubin, conjugated (direct) bilirubin, unconjugated (indirect) bilirubin, prothrombin time, the international normalized ratio (INR), lactate dehydrogenase, total protein, globulins, and albumin are frequently discussed when discussing liver function tests. With the aid of these tests, the location of the liver damage can be identified, and the elevation pattern can assist in organizing a differential diagnosis [4].

The phrase "liver function tests" is frequently misleading because many of the tests identify the cause of the damage rather than assessing how well the liver is functioning. Hepatocellular illness is indicated by ALT, AST, and bilirubin increases that are out of proportion to ALP. A cholestatic pattern would be characterized by an increase in ALP and bilirubin that is out of proportion to ALT and AST. Alkaline phosphatase and AST/ALT values rising indicate a mixed damage pattern. An increase in bilirubin with normal alkaline phosphatase and AST/ALT values is referred to as isolated hyperbilirubinemia [5].

Despite this, CLD is still identified only after the onset of the disease. There is plenty of chance for earlier diagnosis because of the disease's protracted natural history, which involves progressive fibrosis leading to cirrhosis, along with known risk factors (alcohol abuse, obesity, and metabolic syndrome). Ultimately, 90% of CLD (ARLD, NAFLD, and viral hepatitis) is preventable [6]. Our aim in this paper is to identify early phase and prediction of liver disease by involving the analysis of specific proteins such as bilirubin, albumin, proteins, and alkaline phosphatase which are associated with liver function and dysfunction using comparison of several machine learning methods.

* 1. **Differential Diagnosis Based on Elevated LFTs**

Hepatocellular damage or cholestasis are the two most common mechanisms in which liver enzyme abnormalities occur. A detailed explanation of the next two mechanisms is provided below [7].

* + 1. Hepatocellular pattern

The levels of aminotransferases (ALT and/or AST) are significantly increased compared to the levels of alkaline phosphatase.

* + - 1. ALT-predominant

Acute or chronic viral hepatitis, steatohepatitis, acute Budd-Chiari syndrome, ischemic hepatitis, autoimmune, hemochromatosis, medications/toxins, autoimmune, alpha1-antitrypsin deficiency, Wilson disease, Celiac disease.

* + - 1. AST-predominan

Alcohol-related, steatohepatitis, cirrhosis, non-hepatic (hemolysis, myopathy, thyroid disease, exercise).

* + 1. Cholestatic pattern

The levels of alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), and bilirubin, which are out of proportion to the levels of aspartate aminotransferase (AST) and alanine aminotransferase (ALT).

* + - 1. Hepatobiliary causes

Bile duct obstruction, Primary biliary cirrhosis, Primary sclerosing cholangitis, Medication-induced, Infiltrating diseases of the liver (sarcoidosis, amyloidosis, lymphoma, among others), Cystic fibrosis, Hepatic metastasis, Cholestasisalpha1-antitrypsin deficiency, Wilson disease, Celiac disease.

* + - 1. Non-Hepatic causes

Bone disease, pregnancy, chronic renal failure, lymphoma or other malignancies, congestive heart failure, childhood growth, infection, or inflammation.

1. **Data**

The Indian Liver Patient Records dataset utilized in this study came from the machine learning repository at the University of California, Irvine (UCI) [8]. The dataset consists of medical records from Andhra Pradesh's North East region in India. There are 167 entries for patients who don't have liver disease and 416 data for liver patients. The dataset consists of 10 parameters and 2 marker groups, where one group represents liver disease and the other represents liver disease absence.

Table 1. INDIAN LIVER PATIENT DATASET

| Attributes | | Attributes Description | |
| --- | --- | --- | --- |
| Age  Gender  Total Bilirubin  Direct Bilirubin  Alkaline Phosphatase  Alanine Aminotransferase  Aspartate Aminotransferase  Total Proteins  Albumin  Albumin and Globulin Ratio | | Age of the patients, from 4 to 90.  Sex of the patients.  The sum of direct and indirect bilirubin in the patient’s blood, from 0.4 to 75 in mg/dL  Conjugated Bilirubin in mg/dL  ALP in IU/L  ALT in IU/L  AST in IU/L  Total Proteins g/dL  Albumin in g/dL  A/G ratio | |

* 1. **Dataset Description**

As shown in Table 1, the dataset utilized in this study includes a variety of key traits that provide insight into the variables that influence liver disease. The parameters of the dataset that we utilize are explained in more detail below.

* + 1. Age

A person gradually loses the ability to maintain homeostasis as they age because of anatomical changes or dysfunction, making them more susceptible to stress or damage from external factors. Studies on the effects of aging on the healthy liver and liver disorders have been done. In addition to raising death rates, aging also raises chances for a number of liver illnesses and acts as a poor prognosticator [9].

* + 1. Gender

Cirrhosis and primary liver cancer, particularly hepatocellular carcinoma, are more common in men. Additionally, they have higher mortality rates from cirrhosis and chronic liver disease, particularly alcohol-related liver disease, which is more prevalent in men due to greater rates of binge drinking. The prevalence of non-alcoholic fatty liver disease (NAFLD), on the other hand, increases in women after menopause but decreases in males before menopause, probably as a result of hormonal factors. Women are more likely than males to die from NAFLD, and drinking too much alcohol can harm their liver more easily. Additionally, they have a higher risk of developing

noncancerous liver tumors, primary biliary cholangitis, and autoimmune liver illnesses including autoimmune hepatitis. the intricate interactions between different risk factors, hormones, and gender [10].

* + 1. Total and Direct Bilirubin

Bilirubin is a pigment that becomes yellow as hemoglobin, red blood cells are broken down. Red blood cells decompose after completing their life cycles in your body and travel through your bloodstream to your liver where they are processed. In order to create bile, your liver separates bilirubin from other waste materials. Bilirubin will not be secreted effectively if the liver is not operating normally. Therefore, if the level of bilirubin is higher than expected, it could indicate that the liver isn't functioning properly [11]. Based on the study, they demonstrate that in patients with liver cirrhosis, DB level predicted outcome more accurately than TB level. Additionally, MELD score in patients with liver cirrhosis should be improved and replaced with DB-based prediction models (DB-MELD and DiBIC score) [12].

* + 1. Alkaline Phosphatase

Our body contains an enzyme called alkaline phosphatase (ALP). An enzyme is a specific kind of protein found in cells that functions as a catalyst and enables specific physiological processes. Due to its predominant presence in the liver, ALP is frequently referred to as a liver enzyme. The bile duct, bones, kidneys, intestines, and placenta in pregnant women all have it, though. ALP levels that are abnormal can signify tissue injury or disturbances in regular body functions. High levels of alkaline phosphatase (ALP) could mean that your liver is damaged or that you have a specific bone problem [13].

* + 1. Alanine Aminotransferase

Alanine aminotransferase (ALT) is an enzyme that is primarily found in liver cells (hepatocytes) and to an extent in other tissues such as the kidneys, heart, and skeletal muscles. When liver cells are damaged or undergo significant necrosis (cell death), ALT is released into the bloodstream, leading to elevated levels of ALT in serum. Concentration of ALT is quantified using the NADH oxidation method [14]. It is crucial to keep in mind that any liver cell damage might elevate ALT serum levels. There is no connection between the absolute peak of the ALT elevation and the degree of hepatic injury, despite the fact that certain hepatic disorders are linked to an increase in ALT levels. Consider acute viral hepatitis, severe drug-induced liver injury, or acute ischemic liver injury if your ALT levels are over 1000 U/L. Hepatitis E infection and common bile duct stones are additional explanations [15].

* + 1. Aspartate Aminotransferase

Aspartate Aminotransferase (AST) has a crucial role in our metabolic processes, as an enzyme in the liver. The elevation of AST in the blood can mean the damage of a liver. During liver disease such as alcoholic liver disease and hepatitis, the cells inside the liver may become injured, inflamed or die. AST is thus discharged into the bloodstream, raising AST levels. Important diagnostic and prognostic data for liver disease can be obtained from monitoring AST levels [16].

* + 1. Total Proteins

The liver has an important role in producing proteins and blood clotting inside the body. It serves as one of the main organs inside the body that helps keep it alive. The measurement of total proteins inside the body is capable of providing valuable information regarding the overall health and function of the liver. It is also supported by the fact that the treatment of nonalcoholic fatty liver disease (NAFLD), shows the significant improvement of total proteins inside the body [17].

* + 1. Albumin and Globulin Ratio

Albumin is a component which is usually associated with liver cirrhosis. The treatment of liver cirrhosis using albumin [18]. A low serum level of albumin (normal range 35–50 g/L) indicates either impaired liver function or poor nutritional condition in the patient. In patients with chronic inflammatory disorders such chronic hepatitis, a high level of globulin (normal range 20–35 g/L) denotes immune system overactivity. The albumin-to-globulin ratio, a surrogate index created by combining albumin and globulin, is frequently used in clinical settings to measure liver function.9 The albumin-to-globulin ratio has a typical threshold value between 1.0 and 2.0. The albumin-to-globulin ratio declines as a result of either insufficient albumin production or excessive globulin production. Patients with advanced cirrhosis or severe inflammatory liver disorders frequently have an inverted albumin-to-globulin ratio (IAGR, 1.0) [19].

1. **Methodology**

This study proposes a machine learning approach in identifying proteomic signatures for early liver disease indications. By analyzing proteomic data, the changes in protein expressions and modification patterns can be able to find out the potential biomarkers for liver disease progression. The machine learning approaches which are utilized in the study are classification and regression types such as, Gradient Boosting, Stochastic Gradient Descent, K-Nearest Neighbour, Random Forest,, Light Gradient Boosting Machine Classifier, Gaussian Naive Bayes, Decision Tree, and Logistic Regression.

***3.1* *Pre-Processing***

The data is preprocessed to remove discrepancies, enhance data quality, and guarantee compatibility with the chosen algorithms before further analysis. This calls for actions like managing missing values and label encoding. The efficiency and precision of the models used for analysis can be improved by improving the data format. The correlations between the parameters are depicted in Fig. 1's heatmap, which reveals both strong and weak relationships. In order to concentrate on the most important predictors for an accurate prediction of liver disease, many variables are excluded from the study due to the low correlations seen in some columns. The goal of this focused strategy is to increase the effectiveness and efficiency of the prediction model.0



* + 1. Label Encoding



The gender attribute in the Liver Dataset consists of categorical values, including "Male" and "Female." To facilitate analysis, these categorical values are transformed into numerical labels using a label encoder. Specifically, the label encoder assigns the value 0 for "Male" and 1 for "Female." From the label encoding process, we can observe that out of the 583 liver patients in the dataset, there are 441 male samples and 112 female samples (Fig. 2).

* + 1. Impute Missing Value

There are various attributes in the Liver Patients Record Dataset that have missing values. An imputation technique is used to fill in these missing values and guarantee that the dataset is full. The mean imputation technique is utilized because the dataset has already been converted to numerical data. Mean imputation includes substituting the mean value of the associated attribute for any missing values. The dataset is improved for further analysis and modeling by imputing the missing values, enabling a more thorough knowledge of the patterns of liver disease.

* + 1. Standardization

Standardization shifts the distribution to have a mean of zero and a standard deviation of one by scaling each input variable independently by removing the mean (a process known as centering) and dividing by the standard deviation. By deducting the mean from a feature and scaling it to unit variance, StandardScaler standardizes it. Divide all the values by the standard deviation to get the unit variance. When the data has a Gaussian distribution (or Normal distribution), standardization might be useful.

* + 1. Principal Component Analysis

Principal component analysis, or PCA, is a technique for reducing the number of dimensions in big data sets by condensing a large collection of variables into a smaller set that retains the majority of the large set's information. Machine learning algorithms can analyze data points considerably more quickly and easily with smaller data sets since there are less irrelevant factors to evaluate.

* 1. ***Basic Machine Learning Algorithm***
     1. Gradient Boosting

Gradient Boosting is an ensemble approach that combines a number of models to obtain a better prediction. Gradient boosting machines, or just GBMs, employ a learning process that sequentially fits new models to fresh data to produce a more precise estimate of the response variable. The fundamental concept of this technique is to build the new base-learners to have a maximum correlation with the ensemble's overall negative gradient of the loss function [20].

* + 1. Stochastic Gradient Descent (SGD)

This classification algorithm is a member of the linear model family and is used in machine learning. The linear classifiers' best coefficients are discovered via the fitting technique Stochastic Gradient Descent (SGD), which modifies the model's parameters based on the gradient of the loss function [21].

* + 1. K-Nearest Neighbour (KNN)

The supervised machine learning technique known as the K-Nearest Neighbors (KNN) may be used to handle classification and regression issues. It is straightforward and simple to implement. The KNN algorithm believes that related objects are located nearby. In other words, related objects are located close to one another. In order for the KNN algorithm to be effective, this assumption must be true. KNN uses the arithmetic we may have learned as children—calculating the distance between points on a graph—to encapsulate the notion of similarity, also known as distance, proximity, or closeness [22].

* + 1. Random Forest (RF)

The supervised machine learning technique known as the K-Nearest Neighbors (KNN) may be used to handle classification and regression issues. It is straightforward and simple to implement. The KNN algorithm believes that related objects are located nearby. In other words, related objects are located close to one another. In order for the KNN algorithm to be effective, this assumption must be true. KNN uses the arithmetic we may have learned as children—calculating the distance between points on a graph—to encapsulate the notion of similarity, also known as distance, proximity, or closeness [23].

* + 1. Light Gradient Boosting Machine Classifier (LGBM)

Light Gradient Boosting Machine Classifier uses a leaf-wise tree approach, the algorithm includes a new sampling method from one side that is based on gradients and unique characteristics. This strategy hastens the adoption of Gradient Boost [24].

* + 1. Gaussian Naive Bayes

Naive Bayes is an algorithm that is used for classification tasks from calculating the group probability and sum of the combination of values from the dataset. The processed data is numeric data, for that in calculating class probability values ​​can use the Probability Density Function (PDF). This approach is also called Gaussian Naive Bayes [24].

* + 1. Decision Tree (DT)

The goal of a decision tree is to resolve classification and regression-type issues within a tree's structural framework [24]. The dataset used to calculate attribute selection measures' best feature, the Information gain or the Gini index, is referred to as the tree's root. Each leaf node relates to the result, class, or dependent variable, whereas each inner node refers to a characteristic or feature. This method can handle category and numerical data.

* + 1. Logistic Regression

A predictive machine learning approach called Logistic Regression (LR) is based on the idea of probability. It is intended to categorize situations with two classes (binary), where the class can either be 0 or 1. To compress the output between 0 and 1, LR uses the logistic function.

* 1. ***Evaluation Metrics***
     1. Accuracy

Accuracy is calculated as the proportion of accurate predictions to all other predictions. It gives a broad indication of how well the model does at accurately identifying situations.

|  |
| --- |

* + 1. Precision

The fraction of positive predictions that were accurate is determined by precision. It may be measured as the proportion of True Positives, or forecasts, to all positive predictions (True Positive and False Positive), which are truly accurate.

* + 1. Recall

Recall is used to determine the percentage of real positives that were mistakenly detected. It can be measured as True Positive, or forecasts that really match the overall number of positives, either properly forecasted as positive or wrongly anticipated as negative (true Positive and false negative).

* + 1. F-1 Score

F1 Score metric based on the predictions provided for the positive class. With the use of Precision and Recall, it is computed. It is a particular kind of score that combines Precision and Recall. As a result, the F1 Score may be determined by taking the harmonic mean of both accuracy and recall and giving each variable equal weight.

* + 1. Confusion Matrix

When real values are known, a confusion matrix—a tabular representation of the predicted outcomes of any binary classifier—is used to indicate how well the classification model performed on the set of test data. Using a confusion matrix, the prediction values made by the model and the ground truth values mapped into a matrix containing True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. These values are used to calculate the evaluation metrics.

1. **Result & Analysis** 
   1. ***Measurements of Classification***

The results and analysis of the comparison of machine learning models are presented in this section. The performance metrics of each model include accuracy, precision, recall, F1 score, and confusion matrix. The resulting metric reported in Table 2, provides a thorough comparison of the model's performance.

Table 2. CLASSIFICATION PERFORMANCE OF MACHINE LEARNING MODELS

| Models | | Accuracy | | Precision | | Recall | F1-Score |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gradient Boosting  Stochastic Gradient Descent  K-Nearest Neighbors  Random Forest  LGBM  Gaussian Naive Bayes  Decision Tree  Logistic Regression | | 0.74  0.69  0.75  0.74  0.75  0.73  0.74  0.76 | | 0.71  0.78  0.77  0.75  0.76  0.76  0.76  0.69 | | 0.74  0.69  0.75  0.74  0.75  0.73  0.74  0.76 | 0.72  0.72  0.76  0.75  0.75  0.74  0.75  0.71 |

|  |  |
| --- | --- |
|  |  |
| **Fig. 4.** Gradient Boosting Confusion Matrix | **Fig. 5.** Stochastic Gradient Descent Confusion Matrix |
|  |  |
| **Fig. 6.** K-Nearest Neighbors Confusion Matrix | **Fig. 7.** Random Forest Confusion Matrix |
| **Fig. 8.** Logistic Regression Confusion Matrix | **Fig. 9.** LGBM Confusion Matrix |
| **Fig. 10.** Gaussian Naive Bayes Confusion Matrix | **Fig. 11.** Decision Tree Confusion Matrix |

Berikut adalah hasil confusion matrix dari 9 model yang digunakan. Pada kasus Liver Disease Detection ini, perlu memperhatikan value False Negative agar seminimal mungkin karena akan sangat fatal apabila terdapat kasus yang seharusnya Positif Liver Disease namun terdeteksi sebagai Negatif. Oleh karena itu, metric Recall menjadi salah satu penilaian yang paling diperhatikan pada riset ini. Dari 9 model ini, dapat terlihat bahwa model Logistic Regression merupakan model yang memiliki Recall yang paling tinggi sehingga akan cocok untuk penelitian pada kasus ini. Namun, peningkatan recall dalam kasus ini memberikan dampak pada nilai precision yang menjadi rendah. Nilai precision yang rendah menunjukkan adanya banyak kasus prediksi False Positive.

* 1. ***Analysis Of The Results***

Based on the data presented in Table 2, it is evident that the machine learning algorithm Logistic Regression achieved the highest scores in terms of Accuracy, Recall, and F1-Score, with values of 0.76, 0.76, and 0.71, respectively. This indicates that Logistic Regression performed the best among the evaluated models in predicting liver disease based on these metrics. Additionally, the K-Nearest Neighbors (KNN) algorithm achieved the highest Precision score of 0.77.

Among the evaluated models, Logistic Regression had the highest overall performance, followed closely by KNN and LGBM (Light Gradient Boosting Machine), which achieved similar accuracy, precision, recall, and F1-Score values of 0.75. Gaussian Naive Bayes, Decision Tree, and Random Forest performed slightly lower with accuracy, precision, recall, and F1-Score values of 0.73, 0.74, and 0.75, respectively. Gradient Boosting and Stochastic Gradient Descent had the lowest performance among the models, with accuracy values of 0.74 and 0.69, respectively.

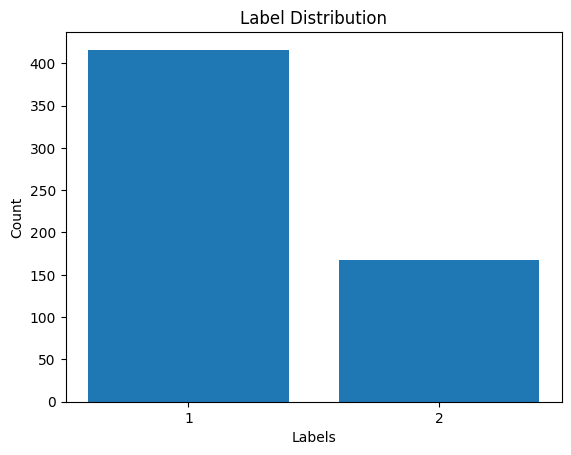
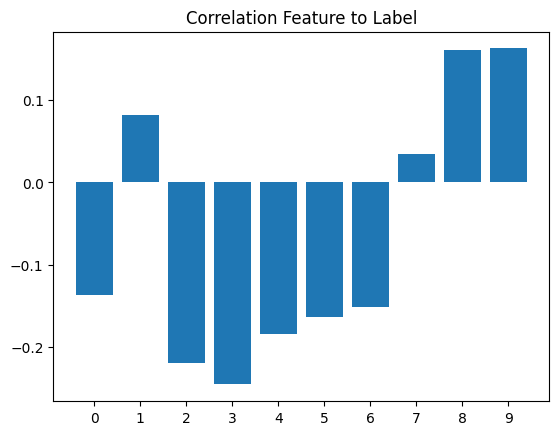
Therefore, based on the provided results, Logistic Regression can be considered the best-performing model in terms of Accuracy, Recall, and F1-Score, while KNN excelled in terms of Precision.

1. **Conclusion**

Based on this research, we have successfully developed an effective diagnostic system for patients with liver infection using 8 different machine learning models. In this study, we utilized all patient parameter information obtained through liver function tests as input for the machine learning models. The results of the research show that the K-Nearest Neighbors model outperforms the other 7 models with an accuracy rate of 0.75, precision of 0.77, recall of 0.75, and F1-score of 0.76.

Therefore, these findings contribute significantly to early detection of liver disease and predicting the presence of liver disease based on important biomarker information. Additionally, they can assist in making more accurate decisions in the management of patients with liver disease. Based on our research findings, there are several aspects that could be explored in future work:

1. To develop, research can be carried out with a larger dataset coverage so that it can provide even better abilities to the developed model in analyzing liver disease.
2. Conducting research using datasets sourced from diverse populations to test the effectiveness and accuracy of the models.
3. Expanding the utilization of relevant demographic information and biomarkers as inputs into the model can help enhance the model's ability to diagnose liver disease.
4. By performing hyperparameter tuning, the performance of machine learning models in predicting liver disease can be optimized.
5. Lastly, exploring the implementation of deep learning models, such as neural networks, can open up new opportunities in the analysis and prediction of liver disease.



# Reference

| [1] | Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Mortality 2018-2021 on CDC WONDER Online Database, released in 2021. *Data are from the Multiple Cause of Death Files, 2018-2021,* as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/ucd-icd10-expanded.html on Jun 11, 2023 10:58:49 PM |
| --- | --- |
| [2] | *Statistics. (n.d.). British Liver Trust. https://britishlivertrust.org.uk/about-us/media-centre/statistics/* |
| [3] | *The relative expression of hepatocellular and cholestatic liver enzymes in adult patients with liver disease - PubMed*. (2020, April 1). PubMed. https://doi.org/10.1016/j.aohep.2019.08.004 |
| [4] | S. Ribeiro, A. J., Yang, X., Patel, V., Madabushi, R., & Strauss, D. G. (2019, June 4). *Liver Microphysiological Systems for Predicting and Evaluating Drug Effects*. PubMed Central (PMC). https://doi.org/10.1002/cpt.1458 |
| [5] | Vagvala, S. H., & O’Connor, S. D. (2018, June 5). *Imaging of abnormal liver function tests*. PubMed Central (PMC). https://doi.org/10.1002/cld.704 |
| [6] | Macpherson, I., M Abeysekera, K. W., Harris, R., Mansour, D., McPherson, S., Rowe, I., Rosenberg, W., Dillon, J. F., & Yeoman, A. (2022, August 5). Identification of liver disease: why and how. Frontline Gastroenterology. https://doi.org/10.1136/flgastro-2021-101833 |
| [7] | Lala, V., Zubair, M., & Minter, D. A. (2023, April 7). Liver Function Tests - StatPearls - NCBI Bookshelf. Liver Function Tests - StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK482489/ |
| [8] | UCI Machine Learning Repository. ILPD (Indian Liver Patient Dataset) Data Set. https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver +Patient+Dataset) |
| [9] | Kim, H., Kisseleva, T., & Brenner, D. A. (n.d.). Aging and liver disease. PubMed Central (PMC). https://doi.org/10.1097/MOG.0000000000000176 |
| [10] | Integration, C. (2023, January 11). Are Men More Susceptible to Liver Disease Than Women? | UPMC HealthBeat. UPMC HealthBeat. https://share.upmc.com/2023/01/gender-and-liver-disease/ |
| [11] | C. (n.d.). Bilirubin Test: Understanding High vs. Low Levels & Its Causes. Cleveland Clinic. https://my.clevelandclinic.org/health/diagnostics/17845-bilirubin |
| [12] | Lee, H. A., Jung, J. Y., Lee, Y. S., Jung, Y. K., Kim, J. H., An, H., Yim, H. J., Jeen, Y. T., Yeon, J. E., Byun, K. S., Um, S. H., & Seo, Y. S. (2020, December 11). Direct Bilirubin Is More Valuable than Total Bilirubin for Predicting Prognosis in Patients with Liver Cirrhosis. PubMed Central (PMC). https://doi.org/10.5009/gnl20171 |
| [13] | C. (n.d.). Alkaline Phosphatase (ALP): What It Is, Causes & Treatment. Cleveland Clinic. https://my.clevelandclinic.org/health/diagnostics/22029-alkaline-phosphatase-alp |
| [14] | Washington, I. M., & Van Hoosier, G. (2012). Clinical Biochemistry and Hematology. The Laboratory Rabbit, Guinea Pig, Hamster, and Other Rodents, 57–116. https://doi.org/10.1016/b978-0-12-380920-9.00003-1 |
| [15] | Moriles, K. E., & Azer, S. A. (2022, December 10). Alanine Aminotransferase - StatPearls - NCBI Bookshelf. Alanine Aminotransferase - StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK559278/ |
| [16] | Z. Li et al., "Aspartate aminotransferase-to-platelet ratio index as a noninvasive predictor of liver fibrosis in patients with chronic hepatitis B," Journal of Gastroenterology and Hepatology, 2006. DOI: 10.1111/j.1440-1746.2005.04013.x. |
| [17] | Nagai, M., Sho, M., Satoi, S., Toyokawa, H., Akahori, T., Yanagimoto, H., Yamamoto, T., Hirooka, S., Yamaki, S., Kinoshita, S., Nishiwada, S., Ikeda, N., Kwon, A. H., & Nakajima, Y. (2013, June 20). Effects of pancrelipase on nonalcoholic fatty liver disease after pancreaticoduodenectomy. Journal of Hepato-Biliary-Pancreatic Sciences, 21(3), 186–192. https://doi.org/10.1002/jhbp.14 |
| [18] | New Insights About Albumin and Liver Disease. (2018, July 1). New Insights About Albumin and Liver Disease - ScienceDirect. https://doi.org/10.5604/01.3001.0012.0916 |
| [19] | Zhang, C. C., Zhang, C. W., Xing, H., Wang, Y., Liang, L., Diao, Y. K., Chen, T. H., Lau, W. Y., Bie, P., Chen, Z. Y., & Yang, T. (2020, October 12). Preoperative Inverted Albumin-to-Globulin Ratio Predicts Worse Oncologic Prognosis Following Curative Hepatectomy for Hepatocellular Carcinoma. PubMed Central (PMC). https://doi.org/10.2147/CMAR.S275307 |
| [20] | Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021 |
| [21] | Du, J. (2019, May 1). The Frontier of SGD and Its Variants in Machine Learning. Journal of Physics: Conference Series, 1229(1), 012046. https://doi.org/10.1088/1742-6596/1229/1/012046 |
| [22] | Rakshith D B , Mrigank Srivastava , Ashwani Kumar, Gururaj S P, 2021, Liver Disease Prediction System using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 06 (June 2021). |
| [23] | V. Singh, M. K. Gourisaria and H. Das, "Performance Analysis of Machine Learning Algorithms for Prediction of Liver Disease," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 2021, pp. 1-7, doi: 10.1109/GUCON50781.2021.9573803. |
| [24] | F. Alzamzami, M. Hoda and A. E. Saddik, "Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation," in IEEE Access, vol. 8, pp. 101840-101858, 2020, doi: 10.1109/ACCESS.2020.2997330. |

**SUMMARY Literature Review**

Davy (1 - 6)

1. Source : <https://www.researchgate.net/publication/371262536_Alzr-Net_A_Novel_Approach_to_Detect_Alzheimer_Disease>

Summary : This study employs transfer learning to effectively detect Alzheimer's disease using MRI scans. A customized version of Inception v3 named Alzr-Net is being tested to determine how well it can identify Alzheimer's disease.

1. Source :

<https://www.researchgate.net/publication/352390100_Alzheimer's_Disease_Classification_Using_Deep_CNN>

Summary : This research paper uses several approaches to classify the Alzheimers Disease. First method Voxel-Based that use neuroimaging. For picture characteristics, the region of interest-based method is employed, and it will categorize different parts of the brain, including the corpus callosum, better longitudinal fascia, hippocampi, and singulam. A patch-based method is suggested for image representation and recovery. VGG16 and Inception are the two popular architectures that are used in image classification techniques in this research. Several extraction feature are used in this research, including SVM, Random Forest,Decision Tree and KNN.

1. Source : <https://www.researchgate.net/publication/366169073_Transfer_Learning_Approaches_for_Alzheimer_disease_Classification_A_Review>

Summary : Purpose of this research is to guide future research on Alzheimer's stage prediction by discussing several Machine Learning and Deep Learning approaches and their merits. Alzheimer's Disease Stages diagnosis or categorization requires an understanding of the disease's various features. However, identifying defining patterns in characteristics can be difficult, especially when the data is large.

1. Source :

<https://www.researchgate.net/publication/349623421_A_Transfer_Learning_Approach_for_Early_Diagnosis_of_Alzheimer's_Disease_on_MRI_Images>

Summary : The diagnosis of mild cognitive impairment (MCI) using magnetic resonance imaging (MRI) is important in the early treatment of dementia disease. In cases like this, the deep learning architecture produces positive results. This study also provided a comparison with other studies, demonstrating that the proposed model exceeded state-of-the-art models in terms of testing accuracy.

1. Source : <https://www.researchgate.net/publication/339069756_A_Deep_Siamese_Convolution_Neural_Network_for_Multi-Class_Classification_of_Alzheimer_Disease>

Summary : The early detection of Alzheimer's disease is a difficult issue for researchers. To handle various issues related to brain image data analysis, machine learning and deep convolutional neural network (CNN)-based technologies are accessible. This research developed a Siamese convolutional neural network (SCNN) model based on the VGG-16 which outperformed the state-of-the-art models in terms of performance, efficiency, and accuracy.

1. Source : <https://www.researchgate.net/publication/355174632_Alzheimer_Disease_Detection_Empowered_with_Transfer_Learning>

Summary : The proposed Alzheimer's Disease detection method employs transfer learning on multi-class using brain MRI working to classify the images in four stages, Mild demented (MD), Moderate demented (MOD), Non-demented (ND), Very mild demented (VMD). When compared to previous research, the proposed system produces more accurate result.

Jasons [🐱‍🏍🥠Mr. Ironman] → 7-12

1. Source : <https://www.researchgate.net/profile/F-M-Shamrat/publication/338924601_A_Comparative_Study_On_Liver_Disease_Prediction_Using_Supervised_Machine_Learning_Algorithms/links/5fcfbde8299bf188d403df8a/A-Comparative-Study-On-Liver-Disease-Prediction-Using-Supervised-Machine-Learning-Algorithms.pdf>

**Judul:** A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms

**Hasil:**

**Method:** KNN, Naive Bayes, Random Forest, Support Vector Machine, Decision Tree Classifier, Logistic Regression

**Data used:** UCI Machine, such as age, gender, TB, DB, Alkaline Phosphotase, Alamine Aminotransferase, Asparatate Aminotransferase, Total Proteins, Albumin, Albumin and Globulin Ratio.

1. Source : <https://ieeexplore.ieee.org/document/9952144>

**Judul:** Liver Disease Prediction using Semi Supervised based Machine Learning Algorithm

**Hasil:** Using a hybrid model like SVM and K-Means Model showed high accuracy in detecting liver diseases.

**Method:** Hybrid model: SVM & K-Means algorithm

**Data used:** taken from online. Information like family history, consuming alcoholic beverages, smoking, and ingesting, sugar level, body mass index, obesity

1. Source : <https://ieeexplore.ieee.org/document/9759201>

**Judul:** Liver disease prediction using W-LR-XGB Algorithm

**Hasil:**

**Method:** Logistic Regression, SVM, Naive Bayes, XGBoost.

**Data used:** 416 liver patient records and 167 non liver patients collected from North East of Andhra Pradesh, India. The “Dataset” contains age, gender, total bilirubin, alkaline, alanine, albumin, and globulin ratio..

1. Source : <https://ieeexplore.ieee.org/document/10101221>

**Judul:** Classification and Prediction of Liver Disease Diagnosis Using Machine Learning Algorithms.

**Hasil:** In comparison to other algorithms Parameter tunned Logistic regression performed well with training and testing accuracy along with no overfitting, parameter-tuned gradient boosting, KNN, and the normal random forest is getting overfitted so not recommended, Age is the most influencing Factor for Liver Disease, as age increases those people get affected, but Cannot be sure due to insufficient data.

**Method:** Logistic Regression, SVM, KNN, Decision Trees, Random Forest Algorithm, Gradient Boosting, Hyper-parameter Tuning with LR,KNN, and Gradient boosting also called Hyper-boosting.

**Data used:** 583 Patients with liver datasets taken from blood tests and CT scans image. It contains age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total protein, albumin, albumin and globulin ratio.

1. Source : <https://ieeexplore.ieee.org/document/9716312>

**Judul:** Diagnosing for Liver Disease Prediction in Patients Using Combined Machine Learning Models

**Hasil:** The combined model of prediction liver disease performs better than individual machine learning models. the maximum accuracy obtained in this combined method is 96%

**Method:** ANN, Decision Tree, K-Nearest Neighbors (KNN), combined models (KNN, DT, Neural Network)

**Data used:** 583 patients’ records are present in which 167 persons are free from liver disease and 416 persons are affected with liver disease. this dataset is contain age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total protein, albumin, albumin and globulin ratio. Taken from the UCI Machine Learning Repository, namely the Indian Liver Patient Dataset (ILPD)

1. Source : <https://ieeexplore.ieee.org/document/9808059>

**Judul:** Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms

**Hasil:** Logistic regression performance the best model with the maximum accuracy of 75% when compared against the other various algorithms.

**Method:** LR, NB, KNN

**Data used:** taken from website Kaggle. 116 records of people with ten input variables. The data was separated into two categories.

Hansen (13-18)

1. Prediction of Fatty Liver Disease using Machine Learning Algorithm ->

Summary:

Method:

Data Used:

1. Summary:

Method:

Data Used:

1. Z. Li et al., "Aspartate aminotransferase-to-platelet ratio index as a noninvasive predictor of liver fibrosis in patients with chronic hepatitis B," Journal of Gastroenterology and Hepatology, 2006. DOI: 10.1111/j.1440-1746.2005.04013.x.

Summary:

Method:

Data Used:

1. Summary:

Method:

Data Used:

1. Summary:

Method:

Data Used:

1. Summary:

Method:

Data Used:

1. Summary:

Method:

Data Used:

**Fira (21-26)**

1. Source : <https://ieeexplore.ieee.org/document/9885411>

Judul : Predictive Analysis for Hepatitis and Cirrhosis Liver Disease using Machine Learning Algorithms

Summary :

1. Source :

Judul :

Summary :

1. Source :

Judul :

Summary :

1. Source :

Judul :

Summary :

1. Source :

Judul :

Summary :

1. Source :

Judul :

Summary :

**FN (27-32)**

1. <https://sci-hub.se/10.1016/j.neucom.2021.08.138>

Judul: Machine learning based liver disease diagnosis: A systematic review

Summary:

1. <https://sci-hub.se/10.1109/icoei.2018.8553682>

Judul: Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease

Summary: DT, NB, RF, SVM, ANN (DT achieved the best accuracy of 81%)

1. <https://d1wqtxts1xzle7.cloudfront.net/55848081/IRJET-V5I142-libre.pdf?1519107590=&response-content-disposition=inline%3B+filename%3DPerformance_Analysis_of_Liver_Disease_Pr.pdf&Expires=1686465696&Signature=alEPNuWN1p8o7s9Ncco8cc5FMA6Z7fgiCTnKQuoCSVIT0ipJnradeGBPRw7ZgfkOEsUznWkXLGegJdbAkySJs4vGDZNU0VGoF6yV6IoemM1q5~gPI6kxp-y2Mf8VZRZxdzpbQo-dP7yqPllr7EVF3zbQHYrRsvQIQvzEiCYt2NKvca2FZE43cYe6bSiOZ-Toz-qK6Hz5lAYVk0CKoOm8XCh0hGyOLgSxmDPQMU~mx9cI3eGrKVQRUmsJxnu1nwWne9~LKkJlex61U2IYSs81suDTIbRR4Slz2W3nfMHGrMXSh-2TL3hhjRblZtpIPIqsjNU0GdKOkckSpHPID9J~fg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA>

Judul: Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms

Summary: In this thesis the proposed system concludes that PSO feature selection methods for Indian Liver Patient Dataset. This thesis analyzed the liver disease using algorithms such as J48, MLP, SVM, Random Forest, and Bayesnet Classification. These algorithm gives various result based on PSO feature selection model .It has been seen that bayes net and J48 Classification gives better results compare to other classification algorithms. J.48 95.04, Bayesnet 90.33

1. <https://www.ijert.org/research/liver-disease-prediction-system-using-machine-learning-techniques-IJERTV10IS060460.pdf>

Judul: Liver Disease Prediction System using Machine Learning Techniques

Summary: With the dataset that we used for this project, we got 100 % accuracy for SVM model, and though it might be difficult to get such accuracies with very large datasets, from this project’s results, one can clearly conclude that we can predict the risk of liver diseases with accuracy of 90 % or more. .

1. <https://sci-hub.se/10.1109/icaict51780.2020.9333528>

Judul: Prediction of Liver Disorders using Machine Learning Algorithms: A Comparative Study

Summary: As the diagnosis of liver disease is expensive

and sophisticated, numerous researches have been performed

using Machine Learning (ML) methods for classifying liver

disorder cases. In this paper, we have compared four different

ML algorithms such as Logistic Regression (LR), Decision Tree

(DT), Random Forest (RF), and Extra Trees (ET) for

classifying Indian Liver Patient Dataset (ILPD). Pearson

Correlation Coefficient based feature selection (PCC-FS) is

applied to eliminate irrelevant features from the dataset. Also,

a boosting algorithm (AdaBoost) is utilized to enhance the

predictive performance of those algorithms. The comparative

analysis is evaluated in terms of accuracy, ROC, F-1 score,

precision, and recall. After comparing experimental results, we

have found that boosting on ET provides the highest accuracy

of 92.19%.

1. <https://www.researchgate.net/publication/363857071_Liver_disease_detection_using_machine_learning_techniques/link/63338b665f6370520dfe8ed1/download>

Judul: Liver disease detection using machine learning techniques

Summary: This present study introduces the liver disease prediction (LDP) method in predicting liver disease that can be utilised by health professionals, stakeholders, students and researchers. Five algorithms, namely Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (K-NN), Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART), are selected. The accuracy is compared to uncover the best classification method for predicting liver disease using R and Python. From the results, K-NN obtains the best accuracy with 91.7%, and the autoencoder network achieved 92.1% accuracy, which is above the acceptable level of accuracy and can be considered for liver disease prediction.

Gradient Boosting Classifier

Accuracy : 71.7948717948718

precision recall f1-score support

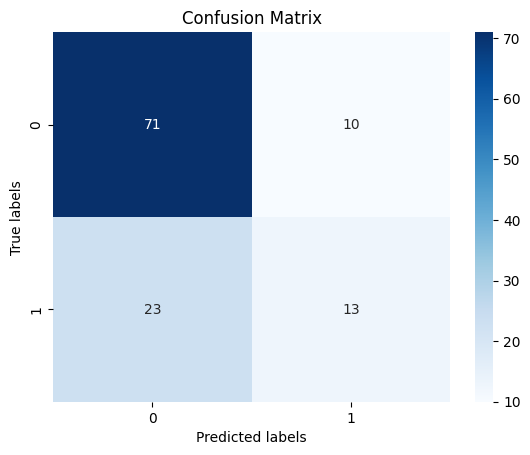
1 0.76 0.88 0.81 81

2 0.57 0.36 0.44 36

accuracy 0.72 117

macro avg 0.66 0.62 0.63 117

weighted avg 0.70 0.72 0.70 117



Random Forest

Accuracy : 70.08547008547008

precision recall f1-score support

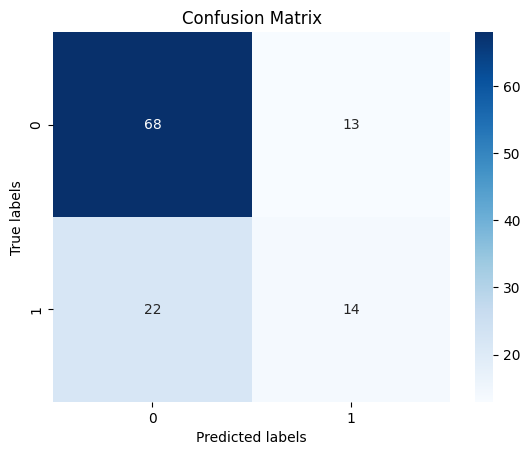
1 0.76 0.84 0.80 81

2 0.52 0.39 0.44 36

accuracy 0.70 117

macro avg 0.64 0.61 0.62 117

weighted avg 0.68 0.70 0.69 117



K-Nearest Neighbours

Accuracy : 70.08547008547008

precision recall f1-score support

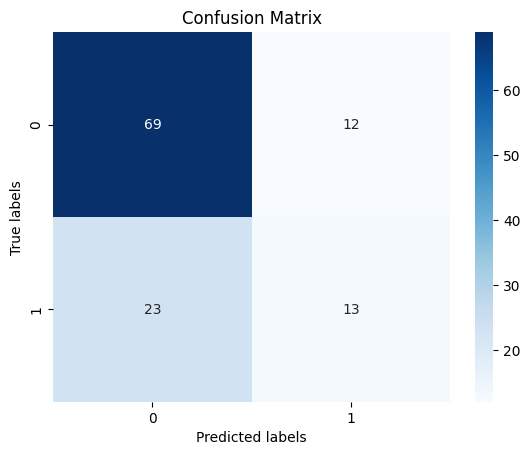
1 0.75 0.85 0.80 81

2 0.52 0.36 0.43 36

accuracy 0.70 117

macro avg 0.64 0.61 0.61 117

weighted avg 0.68 0.70 0.68 117



SGD Classifier

Accuracy : 71.7948717948718

precision recall f1-score support

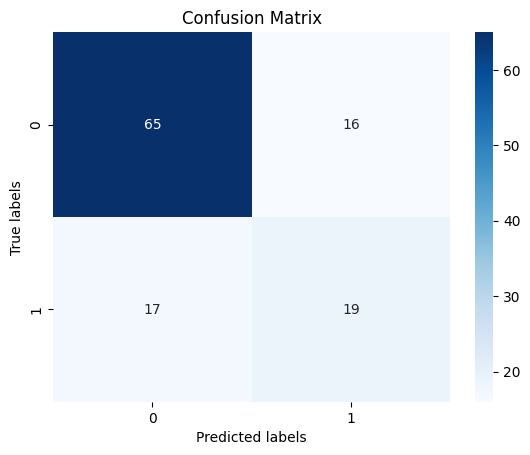
1 0.79 0.80 0.80 81

2 0.54 0.53 0.54 36

accuracy 0.72 117

macro avg 0.67 0.67 0.67 117

weighted avg 0.72 0.72 0.72 117



SVC

Accuracy : 69.23076923076923

precision recall f1-score support

1 0.69 1.00 0.82 81

2 0.00 0.00 0.00 36

accuracy 0.69 117

macro avg 0.35 0.50 0.41 117

weighted avg 0.48 0.69 0.57 117

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/\_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

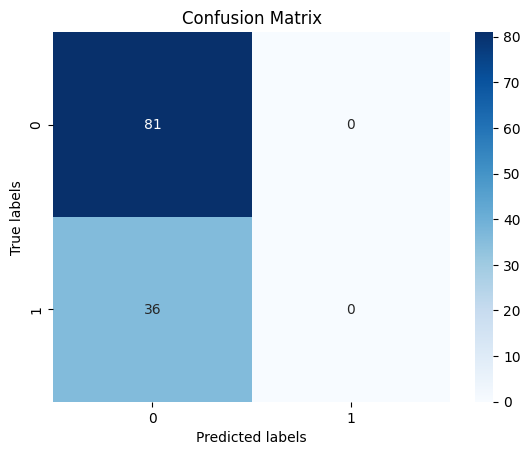
\_warn\_prf(average, modifier, msg\_start, len(result))

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/\_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/\_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))



LGBM Classifier

Accuracy : 70.08547008547008

precision recall f1-score support

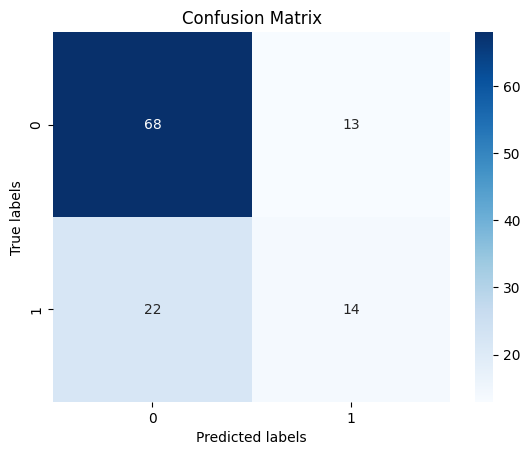
1 0.76 0.84 0.80 81

2 0.52 0.39 0.44 36

accuracy 0.70 117

macro avg 0.64 0.61 0.62 117

weighted avg 0.68 0.70 0.69 117



GaussianNB

Accuracy : 47.008547008547005

precision recall f1-score support

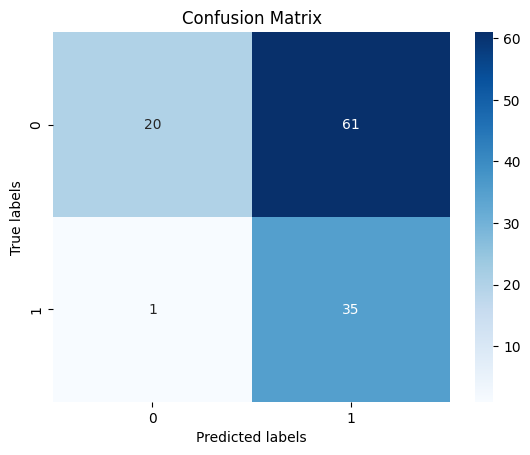
1 0.95 0.25 0.39 81

2 0.36 0.97 0.53 36

accuracy 0.47 117

macro avg 0.66 0.61 0.46 117

weighted avg 0.77 0.47 0.43 117



Decision Tree Classifier

Accuracy : 64.1025641025641

precision recall f1-score support

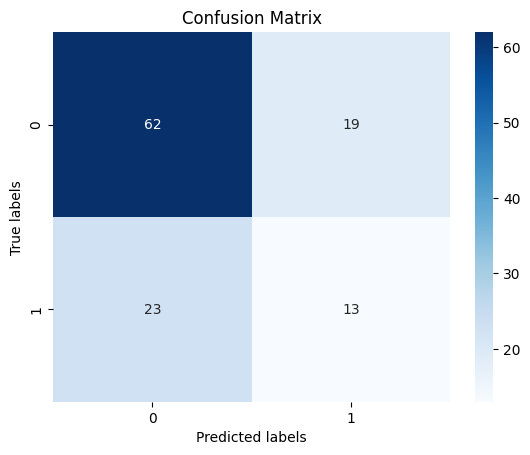
1 0.73 0.77 0.75 81

2 0.41 0.36 0.38 36

accuracy 0.64 117

macro avg 0.57 0.56 0.56 117

weighted avg 0.63 0.64 0.63 117



Logistic Regression

Accuracy : 71.7948717948718

precision recall f1-score support

1 0.74 0.90 0.82 81

2 0.58 0.31 0.40 36

accuracy 0.72 117

macro avg 0.66 0.60 0.61 117

weighted avg 0.69 0.72 0.69 117

