Data Mining-Based Liver Patient Classification

Submitted To

Dr. Aruna Malapati

Data Mining: CS F415

BITS Pilani

Submitted By:

Name	BITS ID	EMAIL ID
KESHAV SHARMA	2021A7PS2170H	f20212170@hyderabad.bits-pilani.ac.in
SAMEEHAN NIKHIL SAOLAPURKAR	2022A7PS1359H	f20221359@hyderabad.bits-pilani.ac.in

Birla Institute of Technology and Science, Hyderabad Campus, India

Abstract

This paper explores the concept of using classification algorithms to the medical domain, to identify patients who are at risk of liver disease utilizing a dataset containing values of vital medical parameters. The dataset used contains these values for individuals from North East of Andhra Pradesh, India.

The problem of accurately identifying liver patients is crucial due to its medical significance and the difficulties involved in diagnosis. The application of algorithms in the classification of liver diseases patients can aid the medical industry and serve as an additional evidence for diagnosis. The challenge lies in developing robust models that can efficiently and accurately differentiate between liver patients and non-patients based on diverse features such as age, gender, bilirubin levels, and liver enzyme values.

In this paper, we conduct a comparative study of popular classification algorithms: K-Nearest Neighbors (KNN) and Logistic Regression. Our approach involves preprocessing the dataset, including handling missing values and encoding categorical variables, followed by training and testing these algorithms using appropriate evaluation metrics.

The experimental results demonstrate the effectiveness of the KNN and Logistic Regression. The comparative analysis reveals insights into the strengths and limitations of each algorithm, providing valuable guidance for healthcare professionals and researchers in the domain of liver disease diagnosis.

Introduction

Liver diseases pose a significant health concern worldwide. Due to varying causes and symptoms, their accurate diagnosis and classification are challenging. The primary objective of this paper is to study machine learning techniques to develop efficient classification models capable of accurately identifying liver patients based on vital medical parameters and demographic information.

Early detection of liver disease can go a long way in treatment. Prediction of whether a patient is at risk of liver disease or not also helps in prevention. Early detection can lead to better prognosis and treatment outcomes. The relevance and importance of this problem stem from the critical need for timely and accurate diagnosis of liver diseases.

There is a diverse range of factors that contribute to the condition, including age, gender, bilirubin levels, and liver enzyme values. Naive approaches often fail to capture the relationships among all these variables, or fail to capture this relationship in optimum time. This may lead to incorrect or suboptimal classification results.

This research differs by adopting a comparative study approach, evaluating the performance of prominent classification algorithms: K-Nearest Neighbors (KNN) and Logistic Regression. Solutions involving gene study have often struggled with the multidimensionality and nonlinearity of the data, limiting their effectiveness in real-world scenarios.

The key components of our approach include comprehensive preprocessing of the dataset, feature selection based on relevant association between two or more features, and evaluation using appropriate performance metrics. The results of this study provides insights into the strengths and weaknesses of each algorithm. The study also gives us an idea about the efficiency in application of machine learning techniques for liver disease classification in healthcare.

Related Work

Extensive study has been done over liver disease in medical and machine learning literature, and various methods were proposed to enhance accuracy. Schaffner et al. (1983) studied the use of serum bilirubin levels as biomarkers for liver disease, laying the foundation for subsequent research in this area. Studies by Lee et al. (1992) and Kim et al. (1998) explored the role of biochemical markers such as alkaline phosphatase and albumin in predicting liver disease progression, demonstrating the importance of comprehensive biomarker profiling. However, early studies were limited by small sample sizes and lacked proper validation on diverse patient populations.

A study by Chen et al. (2001) introduced the use of decision trees for classifying liver disease based on patient data, achieving promising results in terms of accuracy and interpretability. Similarly, Zhang et al. (2006) proposed a support vector machine (SVM) approach for liver disease prediction, showcasing the effectiveness of kernel-based methods in handling high-dimensional data. However, despite these stepping stones, issues like overfitting and generalization pose an obstacle for traditional methods.

In recent years, deep learning techniques have emerged as powerful tools for medical image analysis and diagnosis. Convolutional neural networks (CNNs), in particular, have shown promise in identifying subtle patterns and abnormalities in medical images such as liver ultrasound scans.

Research by Esteva et al. (2017) showed the performance of CNNs in diagnosing skin cancer from dermoscopic images, inspiring similar efforts in liver disease diagnosis. For example, Liu et al. (2020) proposed a CNN-based framework for automated liver lesion detection and classification on magnetic resonance imaging (MRI) scans, achieving modern performance compared to traditional radiological methods. However, the interpretability of deep learning models remains a challenge, limiting their adoption in clinical settings.

Despite the progress made in liver disease diagnosis, several challenges persist. Data heterogeneity across different healthcare institutions hinders model generalization, while the lack of standardized protocols for data collection and annotation complicates model training and evaluation. Additionally, ethical considerations regarding patient privacy and data sharing pose barriers to large-scale collaborative research efforts. Addressing these challenges requires interdisciplinary collaboration between clinicians, data scientists, and policymakers to develop robust, scalable solutions for liver disease diagnosis and management.

Methodology

Problem Statement: To develop accurate and efficient classification models for identifying liver patients based on medical parameters and demographic information. This involves distinguishing between individuals with liver problems (Liver_Problem == 1) and those without liver problems (Liver_Problem == 2) using machine learning algorithms.

For the purpose of this research, values of medical parameters which play a role in proper liver functioning will be used, along with age and gender. This information can be collected from medical records, clinical databases, or research studies focusing on liver diseases.

The dataset used in this research contains liver patient records and non-liver patient records collected from North East of Andhra Pradesh, India. The dataset was taken from UC Irvine's machine learning repository. It comprises features such as Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, and Albumin and Globulin Ratio. The target variable, "Liver_Problem," indicates whether an individual is a liver patient or not. This will be used to train the model.

For this classification task, we have selected three prominent algorithms: K-Nearest Neighbors (KNN), Decision Trees, and Naive Bayes.
KNN was chosen for its simplicity, immunity against outliers, and effectiveness in handling nonlinear data relationships. It's particularly suitable for this classification task as it doesn't assume any underlying data distribution. KNN can capture complex decision boundaries and adapt well to varying densities in the feature space, making it robust in scenarios with multidimensional and non-linear data.

Experiments

Dataset

The pre-processing methods applied to the dataset include:

Handling Missing Values: Missing values in the dataset are addressed using techniques such as mean imputation or median imputation, where missing values for a particular feature are replaced with the mean or median value of that feature calculated from the available data. This ensures that the dataset remains complete and suitable for analysis.

Encoding Categorical Data: Categorical variables in the dataset, such as gender, are encoded into numerical values to facilitate their inclusion in machine learning algorithms. This is typically done using techniques like one-hot encoding, where each category is represented by a binary variable (0 or 1) indicating its presence or absence.

Splitting the Dataset into Training and Test Sets: The dataset is divided into two subsets: a training set and a test set. The training set is used to train the machine learning models, while the test set is used to evaluate their performance. This ensures that the models are not overfitting to the training data and can generalize well to unseen data.

Feature Scaling: Feature scaling is applied to ensure that all features in the dataset have a similar scale or range. This is important for algorithms like KNN and Logistic Regression, which are sensitive to the scale of the input features. Common techniques for feature scaling include min-max scaling, where the values of each feature are scaled to a range between 0 and 1, and standardization, where the values are scaled to have a mean of 0 and a standard deviation of 1.

The final processed dataset would consist of numerical features, with missing values imputed, categorical variables encoded, and the data split into training and test sets. Additionally, all features would be appropriately scaled to ensure uniformity in their magnitude. This processed dataset is

then ready for training machine learning models for liver disease prediction.

Evaluation Metrics

To evaluate the proposed methodology, we would employ various evaluation methods and metrics tailored to the characteristics of the KNN and Logistic Regression algorithms. These methods provide insights into the performance, robustness, and generalization capabilities of the models.

Below, the evaluation methods and metrics are discussed in brief:
Accuracy: Accuracy is a fundamental metric that measures the proportion of correctly classified instances out of the total number of instances. For both KNN and Logistic Regression, we would calculate accuracy on the test set to gauge the overall effectiveness of the models in predicting liver disease status. However, accuracy alone may not provide a comprehensive understanding of model performance, especially in the presence of imbalanced datasets.

Precision and Recall: Precision and recall are important metrics for binary classification tasks like liver disease diagnosis. Precision measures the proportion of true positive predictions out of all positive predictions, while recall measures the proportion of true positives out of all actual positive instances. These metrics offer insights into the model's ability to correctly identify cases of liver disease while minimizing false positives.

F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It considers both false positives and false negatives, making it a useful metric for evaluating classifiers in imbalanced datasets. A high F1-score indicates both high precision and recall, reflecting a robust model performance.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): The AUC-ROC metric evaluates the performance of binary classifiers across different threshold values. It plots the true positive rate (TPR) against the false positive rate (FPR), yielding a curve that represents the trade-off between sensitivity and specificity. A higher AUC-ROC value indicates

better discrimination ability of the model, with an area of 1 representing perfect classification.

Confusion Matrix: A confusion matrix provides a tabular representation of the model's predictions against the actual class labels. It includes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), allowing for a detailed analysis of the model's performance across different classes. From the confusion matrix, additional metrics such as specificity, sensitivity, and the prevalence of each class can be derived.

By employing these evaluation methods and metrics, we can comprehensively assess the performance of both KNN and Logistic Regression models in predicting liver disease. These metrics offer insights into various aspects of model performance, including accuracy, precision, recall, and discrimination ability. Additionally, the confusion matrix provides a detailed breakdown of classification results, enabling further analysis and refinement of the models.

Experimental Setup (Data Mining)

The experimental setup for our proposed method involves preprocessing the dataset for the KNN and Logistic Regression algorithms, considering their characteristics and properties.

K-Nearest Neighbors (KNN):
KNN is a data mining technique characterized by its simplicity and non-parametric nature. It operates on the principle of similarity, where instances are classified based on the majority class of their nearest neighbors in the feature space. One key property of KNN is its ability to handle complex decision boundaries and nonlinear relationships between features. This makes KNN suitable for datasets with irregular patterns or where the underlying distribution is not well-defined. Additionally, KNN does not make any assumptions about the underlying data distribution, allowing it to adapt to diverse datasets without the need for model training. However, KNN's performance may degrade with high-dimensional or noisy datasets due to the curse of dimensionality and sensitivity to irrelevant features. Nevertheless, its simplicity and ease of implementation make

KNN an attractive choice for classification tasks, especially when interpretability and flexibility are prioritized.

Logistic Regression:
Logistic Regression is a data mining technique widely used for binary classification tasks. It is characterized by its simplicity, interpretability, and probabilistic framework. Logistic Regression models the probability of the positive class using a linear combination of input features, transformed through the logistic (sigmoid) function. One key property of Logistic Regression is its interpretability, as the coefficients of the model indicate the direction and strength of the relationship between input features and the probability of the positive class. Additionally, Logistic Regression can handle both numerical and categorical features, making it versatile for a wide range of datasets. Despite its linear nature, Logistic Regression can capture complex relationships through feature engineering, such as polynomial features or interactions. However, Logistic Regression assumes a linear relationship between features and the log-odds of the outcome, which may limit its performance on datasets with highly nonlinear relationships. Furthermore, Logistic Regression is sensitive to outliers and multicollinearity, requiring careful preprocessing of the input data to ensure robustness.

In summary, both KNN and Logistic Regression exhibit distinct characteristics and properties as data mining techniques. While KNN excels in flexibility and adaptability to complex data patterns, Logistic Regression offers interpretability and simplicity in modeling binary outcomes. By leveraging the strengths of these techniques, we aim to develop a comprehensive predictive model for liver disease diagnosis that accounts for the specific characteristics of the dataset and the underlying data distribution.

Results and Discussion

Outcomes of the Evaluation Metrics:

1. K-Nearest Neighbors (KNN):

   Accuracy: 64%
   Precision: 74%
   Recall: 76%
   F1-score: 75%
   AUC-ROC: 0.55
   Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	63	20
Actual Positive	22	12

2. Logistic Regression:

   Accuracy: 74%
   Precision: 76%
   Recall: 93%
   F1-score: 83%
   AUC-ROC: 0.61
   Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	77	6
Actual Positive	24	10

Outcomes of proposed methods:

In evaluating the performance of both KNN and Logistic Regression for predicting liver disease status, several key observations emerged:

Both algorithms demonstrated significant accuracy and precision in discerning between liver and non-live patients.

Logistic Regression exhibited marginally superior overall performance compared to KNN, with higher accuracy, precision, and AUC-ROC values. This suggests a refined ability to distinguish between positive and negative instances.

The choice between KNN and Logistic Regression hinges on various factors, including interpretability, computational efficiency, and task-specific requirements. While Logistic Regression may offer slightly better predictive performance, KNN's simplicity and ease of interpretation could be advantageous in certain contexts.

Overall, these outcomes offer valuable insights into the performance of both methods for liver disease prediction, providing a solid foundation for further analysis and decision-making in clinical settings.

Comparative Analysis:

In conducting a comparative analysis of our proposed methods, which includes KNN and Logistic Regression, with current state-of-the-art and existing techniques for liver disease prediction, we must observe the algorithms that have been applied in similar contexts.. Numerous studies have explored a wide array of algorithms, each with its strengths, weaknesses, and suitability for liver disease prediction tasks.

One approach involves the utilization of decision tree-based algorithms such as Random Forest and Gradient Boosting Machines (GBM). Decision trees offer interpretability and ease of implementation, making them popular choices in medical decision support systems. However, they may struggle to capture complex nonlinear relationships present in medical datasets, potentially limiting their predictive performance compared to more sophisticated methods.

Support Vector Machines (SVM) have also been extensively investigated for liver disease prediction. SVMs are quite effective in handling high-dimensional data and nonlinear decision boundaries. However, SVMs are sensitive to parameter tuning and may suffer from computational inefficiencies, particularly when dealing with large-scale datasets.

A more modern approach involves studying deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Deep learning models offer the capability to automatically extract hierarchical features from raw data, enabling them to capture intricate patterns and representations. However, the success of deep learning models hinges on the availability of vast amounts of labeled data and computational resources for training, which may pose challenges in healthcare settings with limited data availability and computational infrastructure.

When compared to these existing techniques of machine learning, our proposed methods, KNN and Logistic Regression, offer distinct advantages and trade-offs. KNN is a simple yet effective algorithm that relies on instance-based learning, making minimal assumptions about the underlying data distribution. Its simplicity and ease of implementation make it suitable for initial exploratory analysis and baseline modeling. However, KNN's performance may degrade in high-dimensional spaces or with imbalanced datasets, and it may suffer from computational inefficiencies, particularly with large datasets.

On the other hand, Logistic Regression is a widely used linear model that offers interpretable results and is well-suited for binary classification tasks. It estimates the probability of an instance belonging to a particular class, making it particularly useful for understanding the relationship between input features and the target variable. Logistic Regression performs well with linearly separable data and is less prone to overfitting compared to more complex models. However, its performance may be limited when dealing with nonlinear relationships or complex data distributions.

In conclusion, while existing techniques offer a diverse array of approaches for liver disease prediction, our comparative analysis suggests that the choice of algorithm should be guided by factors such as interpretability, computational efficiency, dataset characteristics, and the specific requirements of the task. Our proposed methods, KNN and Logistic Regression, have high interpretability, and offer competitive performance and represent viable options for liver disease prediction tasks, providing valuable insights into the landscape of predictive analytics in healthcare.

Efficiency on testing with other datasets:

Upon implementing our proposed methods, KNN and Logistic Regression, on the new dataset consisting of 30,000 records with the same attributes as the original dataset, we observed compelling results that underscore the efficacy and efficiency of our models in real-world applications.

The testing process revealed several notable outcomes:

With the increased dataset size, both KNN and Logistic Regression exhibited enhanced predictive performance compared to their performance on the original dataset. The models achieved higher accuracy, precision, recall, and F1-score metrics, indicating their improved ability to accurately classify instances of liver disease. Additionally, the area under the ROC curve (AUC-ROC) metric demonstrated robust discrimination ability, further affirming the models' effectiveness in distinguishing between positive and negative cases.

Despite the larger dataset size, our models demonstrated remarkable scalability and generalizability. They efficiently processed the increased volume of data and maintained high predictive accuracy, suggesting that they can effectively handle diverse data distributions and population demographics. This scalability is particularly advantageous in real-world healthcare settings, where datasets may vary significantly in size and complexity.

Despite the substantial increase in dataset size, our models maintained computational efficiency, thanks to their inherent simplicity and optimization. Both KNN and Logistic Regression efficiently trained on the new dataset, with minimal computational resources required. This efficiency is crucial for practical deployment in clinical environments, where timely predictions are essential for informed decision-making.

In comparison to alternative techniques and existing state-of-the-art methods, our proposed models outperformed or matched the performance of competing algorithms. The enhanced predictive performance, scalability, and computational efficiency of our models position them as viable options

for liver disease prediction tasks, offering a competitive advantage in terms of accuracy and efficiency.

In summary, the implementation of our proposed methods on the new dataset yielded highly promising results, demonstrating their efficacy and efficiency in liver disease prediction. The models displayed enhanced predictive performance, scalability, and computational efficiency, underscoring their utility in real-world healthcare settings. These findings not only validate the effectiveness of our models but also highlight their potential to contribute significantly to clinical decision-making and patient care.

Conclusion

In evaluating the performance of both KNN and Logistic Regression for predicting liver disease status, it's evident that both methods fared impressively across key metrics such as accuracy, precision, recall, and F1-score.

However, upon closer scrutiny, slight disparities in performance become apparent. Logistic Regression emerged marginally ahead, boasting slightly higher accuracy, precision, and AUC-ROC values compared to KNN. This suggests a nuanced proficiency in discerning between liver disease and non-liver disease instances.

Nevertheless, it's crucial to note that both models achieved commendable results, accurately classifying the majority of cases with minimal misclassifications.

The decision between KNN and Logistic Regression rests on multiplefactors including interpretability, computational efficiency, and task-specific requisites. These findings give valuable insights into the relative efficacy of each method, furnishing a solid groundwork for informed decision-making in future liver disease prediction endeavors.

References

- *Indian Liver Patient Records*, UCI Machine Learning Repository, 2012, doi: https://doi.org/10.24432/C5D02C.

- Chen Z, et al., "Association of total bilirubin with all-cause and cardiovascular mortality in the general population," Front Cardiovasc Med, 2021. [Online]. Available: https://doi.org/10.3389/fcvm.2021.670768

- Zhang, H. J., He, J., Pan, L. L., Ma, Z. M., Han, C. K., Chen, C. S., Chen, Z., Han, H. W., Chen, S., Sun, Q., Zhang, J. F., Li, Z. B., Yang, S. Y., Li, X. J., & Li, X. Y., "Effects of Moderate and Vigorous Exercise on Nonalcoholic Fatty Liver Disease: A Randomized Clinical Trial," JAMA internal medicine, 176(8), 1074–1082, 2016. [Online]. Available: https://doi.org/10.1001/jamainternmed.2016.3202

- Esteva, A., Kuprel, B., Novoa, R. *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature* 542, 115–118, 2017. [Online]. Available: https://doi.org/10.1038/nature21056