JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Heart & Liver Disease Prediction Using Hybrid Machine Learning Model

[1]Prof Ashish Sawant , [2]Prachi Awandekar , [3]Dhawal Saratkar , [4]Viraj Thak

[1]Assistant Professor, [234]Student.
[1]Computer Science and Engineering
[1]Prof. Ram Meghe College of Engineering and Management , Amravati , India

Abstract :

Machine learning, powered by vast healthcare data, helps diagnose heart and liver diseases early, saving lives. This research investigates the use of hybrid machine learning approaches for accurate prediction of heart and liver diseases. We propose a novel model combining the strengths of Support Vector Machines (SVM), Logistic Regression (LR), and Decision Tree Classifier (DTC) to improve upon individual algorithms' limitations. The model leverages SVM's non-linearity handling, LR's interpretability, and DTC's feature extraction capabilities for robust disease prediction.
This research contributes to the advancement of disease prediction with hybrid machine learning techniques. The proposed model offers a promising approach for improved diagnosis and patient care, while the individual component analysis provides valuable insights for future model development and refinement.

Introduction :

This research aims to build an improved model for predicting heart and liver diseases using a hybrid machine learning approach. This model combines the strengths of three powerful algorithms: Support Vector Machines (SVM), Logistic Regression (LR), and Decision Tree Classifier (DTC) and Random Forest Classifier (RFC). Each of these algorithms brings unique advantages to the table:

SVM: Efficient in handling non-linear data patterns and complex relationships.

LR: Provides interpretability, allowing us to understand which features contribute the most to the predictions.

DTC: Offers strong feature extraction capabilities and handles categorical data effectively.

RFC: Random Forest is an ensemble of decision trees, combining the strengths of multiple trees to improve overall performance.

By combining these algorithms, we hope to achieve:

Improved accuracy: Higher prediction accuracy for both heart and liver diseases.

Enhanced generalizability: The model should perform well on unseen data, ensuring reliable predictions in real-world scenarios.

Interpretability and insights: Understanding which features have the most significant impact on the predictions can guide further research and clinical decision-making.

Approach :

➔ Support Vector Machine (SVM):

- Machine learning algorithm for analyzing and identifying patterns in data.
- Used for classification and regression tasks in various applications like image recognition, forecasting, etc.

➔ Logistic Regression Model:

- Use the scikit-learn LogisticRegression class to build your model.
- Split your data into training and testing sets (e.g., 80/20 split).
- Train the model on the training set, potentially tuning hyperparameters like
- regularization strength for optimal performance.

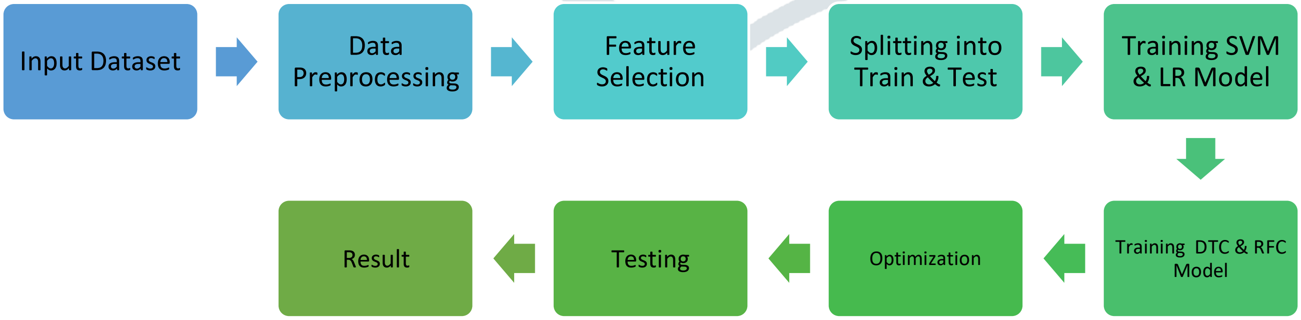
➔ Decision Tree Model:

- Use the scikit-learn DecisionTreeClassifier class to build your model.
- Experiment with different hyperparameters, such as tree depth,
- maximum features, and pruning strategies, to control model complexity and avoid overfitting.

➔ Random Forest Model :

- Combines multiple decision trees for improved accuracy and reduced overfitting.
- Handles missing data and works well with various data types (categorical, continuous).
- Provides insights into the significance of features for prediction.
- Enables efficient training on multi-core processors due to independent tree creation.

Block Diagram



Proposed Methodology :

- **Data Preprocessing :**

1. Tackled both categorical and numeric data: Missing values were imputed, outliers addressed, and features scaled, using techniques like imputation and standardization.
2. Enhanced data quality: Standardization ensured features on the same scale, while feature engineering created new features for better predictive power.
3. Prepared for analysis: By resolving data inconsistencies and enriching features, the dataset is now ready for robust analysis and model training.

- **Data Visualization :**

1. Transforming Data into Insights: After careful cleaning, the dataset's stories are being unveiled through a visual tapestry of bar graphs and diverse charts.
2. Bar Graphs Uncover Patterns: Bar graphs ably reveal trends and comparisons within categories, illuminating patterns that previously lay hidden in numbers.
3. Charts Paint a Multifaceted Picture: Various charts, each with their unique strengths, join the chorus—pie charts slice data into proportions, line charts trace journeys over time, and scatter plots reveal relationships between variables.
4. Unlocking Hidden Gems: Each visualization acts as a key, unlocking insights that might have remained concealed within the raw data.
5. Illuminating the Path Ahead: These visual insights will guide model development, feature engineering, and analysis, ensuring decisions are grounded in a deep understanding of the data's essence.

- **Model Building :**

1. Leveraging multiple strengths: Combining decision power of Support Vector Machines (SVM) , Logistic Regression (LR) ,Decision Tree Classifier and Random Forest Classifier for robust disease prediction.
2. SVM handles non-linear data: Catches complex relationships not easily captured by LR, improving accuracy for specific disease patterns.
3. LR adds interpretability: Provides insights into influential features contributing to the disease outcome, valuable for medical understanding.
4. DTC : Decision trees are inherently interpretable. The tree structure allows clinicians to easily understand and explain the decision-making process.
5. RFC : Random Forest is an ensemble of decision trees, combining the strengths of multiple trees to improve overall performance.
6. Feature selection and weighting: Utilize feature engineering techniques to optimize prediction, potentially including dimensionality reduction for improved efficiency.
7. Cross-validation and tuning: Ensure generalizability and optimize model hyperparameters for accurate and reliable disease prediction.

- **Model Deployment :**

1. Streamlit integration: Embed your hybrid disease prediction model into a user-friendly Streamlit app for web deployment.
2. Interactive interface: Design input forms for users to enter data (e.g., symptoms, demographic info) and display model predictions in real-time.
3. Visualization and insights: Leverage Streamlit's capabilities to visualize risk factors and model performance, enhancing user understanding.
4. Deployment options: Choose from cloud platforms like Heroku or Vercel for easy app hosting and accessibility from any device.
5. Scalability and maintenance: Consider containerization and version control for streamlined deployment updates and future enhancements.

- **Evaluation:**

Evaluate the performance of the hybrid model on the testing data using metrics like accuracy, precision, recall, and F1-score. Compare these results to those of individual algorithms to assess the effectiveness of the hybrid approach.

- **Analysis and Interpretation:**

Analyze the individual contributions of each component within the hybrid model. Identify which features have the most significant impact on the predictions and interpret how the model arrives at its conclusions.

Result Analysis :

Evaluation Metrics:

Evaluating your hybrid model's performance requires a comprehensive set of metrics. Here are some crucial ones:

F1-score: A harmonic mean of precision and recall, balancing both metrics and providing a single measure of model effectiveness. Higher F1 indicates better overall performance.
Recall: Measures the model's ability to correctly identify true positive cases (e.g., correctly diagnosing diseased patients). High recall is crucial for avoiding missed diagnoses.
Precision: Measures the model's ability to avoid false positives (e.g., incorrectly diagnosing healthy patients). High precision ensures accurate positive predictions.
Confusion Matrix: A visual representation of the model's predictions, showing true positives, true negatives, false positives, and false negatives. It aids in understanding classification errors and identifying potential biases.
Interpreting Results:

F1-score: Consider the context of your problem. A high F1 might be ideal for critical situations where both precision and recall are crucial. In less critical contexts, a slightly lower F1 with higher precision or recall might be acceptable depending on the specific priorities.

Recall vs. Precision Trade-off: Analyzing the relationship between recall and precision can reveal valuable insights. For example, a model with high recall but low precision might be overfitting to the training data, while one with high precision but low recall might be too conservative in its predictions.

Confusion Matrix Analysis: Identify patterns in the confusion matrix. High false positives might indicate a need for adjusting the model's thresholds or considering alternative algorithms. High false negatives might suggest data imbalances or the presence of confounding factors.

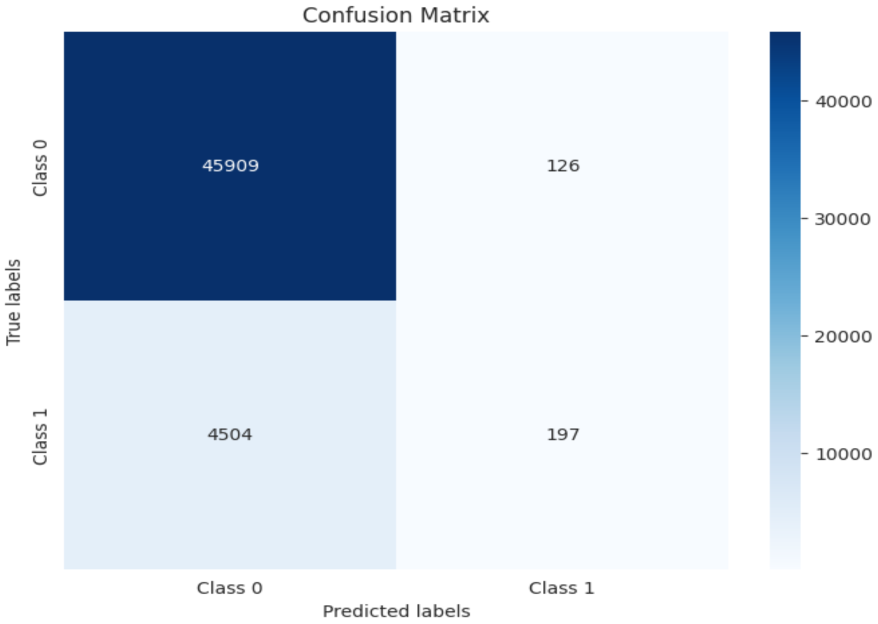
Additional Considerations:

Accuracy: While it's a common metric, it can be misleading in imbalanced datasets. Analyze accuracy alongside other metrics for a more nuanced understanding.

Cross-validation: Don't just rely on test set results. Perform cross-validation to ensure generalizability and robustness of your model's performance.

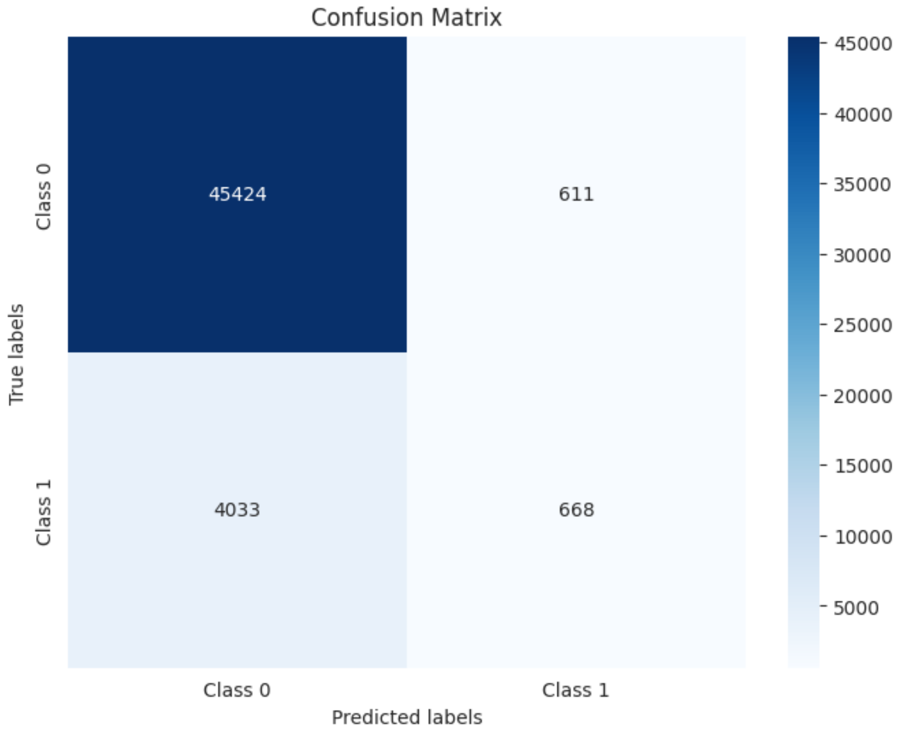
Statistical significance: Use statistical tests to compare the performance of your hybrid model against individual algorithms and assess the improvement achieved.

- **Confusion matrices for heart disease :**



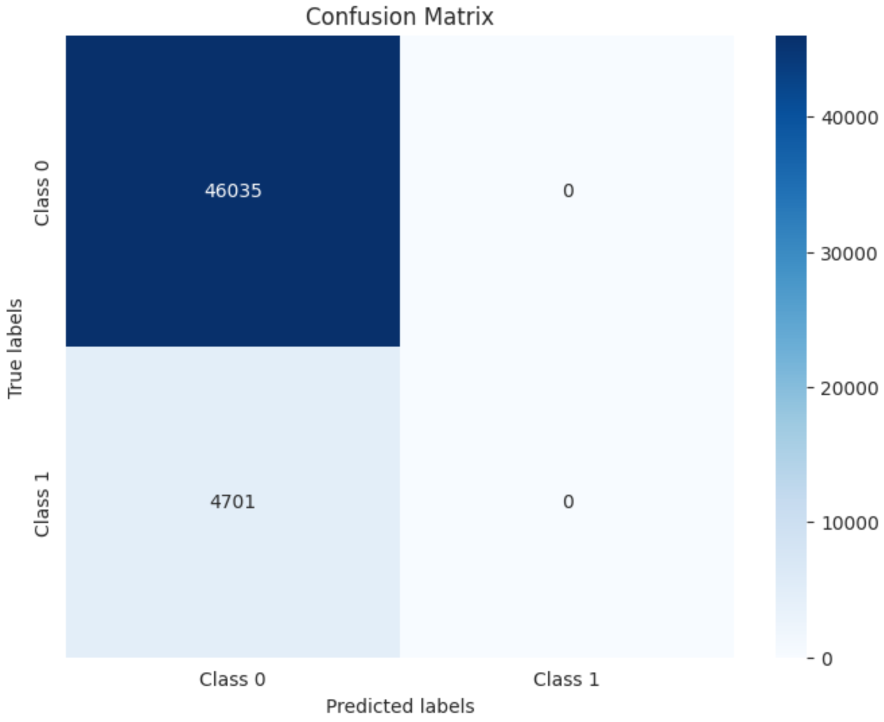
Decision Tree Classifier

The Decision Tree Classifier correctly predicts 45909 instances of the first class and 197 instances of the second class. It misclassifies 126 instances of the firstclass as the second class and 4504 instances of the second class as the first class.(Accuracy : 0.9087432986439609)



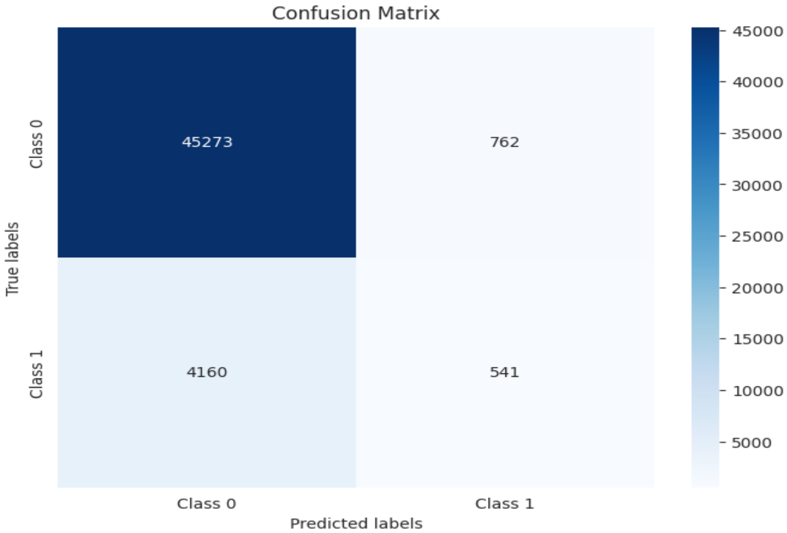
Support Vector Machine

The SVM model predicts two classes. In this case, it correctly identifies the first class (46035 instances) but fails to predict any instances of the second class (4701 instances). It shows no True Positives for the second class.(Accuracy : 0.9073438978240302)



Logistic Regression

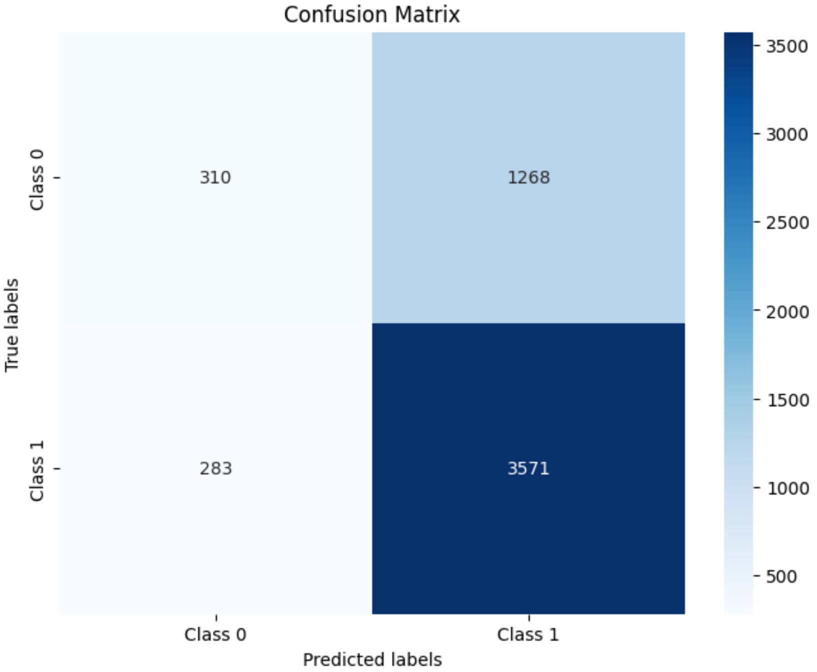
The Logistic Regression model correctly predicts 45424 instances of the first class but misclassifies 611 instances of the first class as the second class. Additionally, it correctly predicts 668 instances of the second class and misclassified 4033 instances of the second class as the first class.(Accuracy : 0.9084673604541155)



Random Forest Classifier

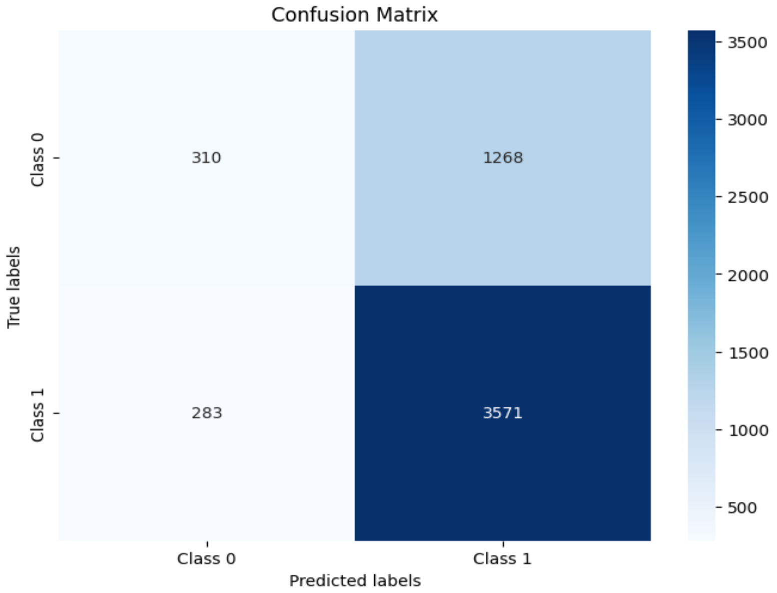
The model correctly predicted 45273 instances of the first class and 541 instances of the second class. However, there were 762 false positives, where instances of the second class were incorrectly predicted as the first class, and 4160 false negatives, where instances of the first class were incorrectly predicted as the second class. This confusion matrix provides a detailed breakdown of the Random Forest Classifier's performance, highlighting both correct and incorrect predictions for each class.

- **Confusion matrices for liver disease :**



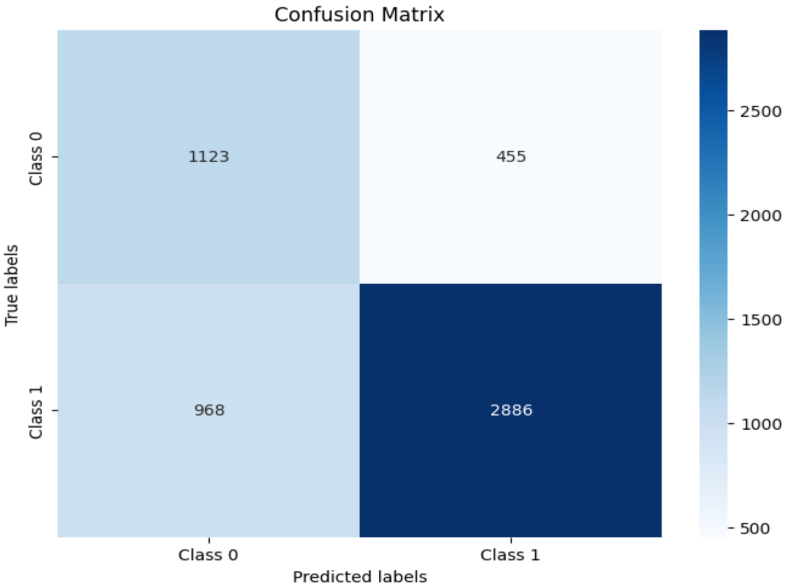
Support Vector Machine (SVM):

The SVM confusion matrix indicates that the model predicted all instances in the first class (0) as belonging to the second class (1), resulting in a large number of false positives (1578). There were no correct predictions for the first class, while all instances of the second class were correctly classified (3854).



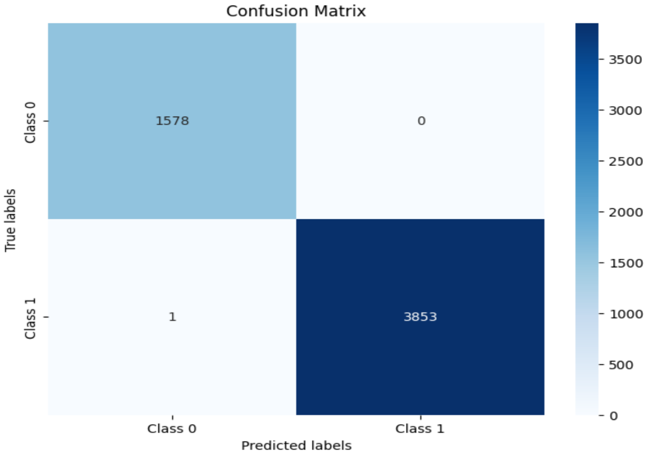
Logistic Regression:

The Logistic Regression confusion matrix demonstrates a balanced prediction for both classes. The model correctly predicted 310 instances of the first class and 3571 instances of the second class. However, there were some misclassifications, with 1268 instances of the first class being incorrectly predicted as the second class and 283 instances of the second class being incorrect.



Decision Tree Classifier:

The confusion matrix for the Decision Tree Classifier shows a mixed performance. The model achieved correct predictions for 1123 instances of the first class and 2886 instances of the second class. However, there were misclassifications, with 455 instances of the first class being predicted as the second class and 968 instances of the second class being predicted as the first class.



Random Forest Classifier:

The Random Forest Classifier performed well, correctly predicting all instances of the first class (1578) and almost all instances of the second class (3853), with only one misclassification. This indicates a robust performance of the model with a high accuracy and minimal false predictions.

Table 1: Comparison of Various Metrics of different models.

Diseases	**Heart**				**Liver**			
Algorithms	SVM	LR	DTC	RFC	SVM	LR	DTC	RFC
Accuracy	90.7	90.8	90.8	90.29	70.9	71.4	73.8	99.9
F1 Score	89.55	22.34	07.84	18.02	83.01	82.16	80.22	99.99
Precision	90.91	52.23	60.99	41.52	70.95	73.80	86.38	100
Recall	88.24	14.21	04.19	11.51	100	92.66	74.88	99.97

Conclusion :

The conclusion of the research study suggests that four algorithms, namely SVM, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, were applied to predict heart and liver diseases. Through comprehensive performance analysis, considering metrics such as confusion matrix, classification accuracy, classification error rate, precision, recall, and F1 score, it was determined that the Random Forest Classifier outperformed the other algorithms. The Random Forest Classifier demonstrated superior classification results, achieving the highest classification rate and the lowest error rate for heart and liver disease prediction. The experimental findings support the effectiveness of the proposed Random Forest Classifier in disease prediction.The study acknowledges the need for future research to extend the proposed hybrid algorithm to forecast various diseases, with plans to utilize additional parameters on larger datasets covering a wider range of diseases to enhance accuracy in subsequent studies.

- ## References :

[1] Vijayashree, J., Sultana, H.P. A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier. Program Comput Soft 44, 388–397 (2018). https://doi.org/10.1134/S0361768818060129

[2] Raj, S., Ray, K. C., & Shankar, O. (2016). Cardiac arrhythmia beat classification using DOST and PSO tuned SVM. Computer methods and programs in biomedicine, 136, 163-177.

[3] Vijayarani, S., &Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. International Journal of Science, Engineering and Technology Research (IJSETR), 4(4), 816-820.

[4] k. Thirunavukkarasu, A. S. Singh, M. Irfan and A. Chowdhury, "Prediction of Liver Disease using Classification Algorithms," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-3, doi: 10.1109/CCAA.2018.8777655.

[5] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[6]Senbagavalli, M., & Arasu, G. T. (2016). "Opinion Mining for Cardiovascular Disease using Decision Tree based Feature Selection." Asian Journal of Research in Social Sciences and Humanities, 6(8), 891-897.

[7] Valli, M. S., and G. T. Arasu. "An Efficient Feature Selection Technique of Unsupervised Learning Approach for Analyzing Web Opinions." (2016).

[8] D. R, K. R, P. G, P. S and S. R. A K, "Breast Cancer Classification using the Supervised Learning Algorithms," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1492-1498, doi: 10.1109/ICICCS51141.2021.9432293.

[9] J. Kennedy and R. Eberhart, "Particle swarm optimization," Proceedings of ICNN'95 - International Conference on Neural Networks, 1995, pp. 1942-1948 vol.4, doi: 10.1109/ICNN.1995.488968.

[10] A. Kumar, A. Ashok and M. A. Ansari, "Brain Tumor Classification Using Hybrid Model of PSO And SVM Classifier," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 1022-1026, doi: 10.1109/ICACCCN.2018.8748787.

[11] S. K. Sarangi, R. Panda and A. Sarangi, "Crazy firefly algorithm for function optimization," 2017 2nd International Conference on Man and Machine Interfacing (MAMI), 2017, pp. 1-5, doi: 10.1109/MAMI.2017.8307875.

[12] H. Wang, C. Li, Y. Liu and S. Zeng, "A Hybrid Particle Swarm Algorithm with Cauchy Mutation," 2007 IEEE Swarm Intelligence Symposium, 2007, pp. 356-360, doi: 10.1109/SIS.2007.367959.