Indian Liver Patients

Dataset Description

The **Indian Liver Patient Dataset (ILPD)** contains patient records collected from a medical center in India.

- **Total Records**: 583
- **Features**: 10 features related to patient demographics and diagnostic tests.
- **Target Variable**: `Liver_Disease` (1 indicates the presence of liver disease, 0 indicates absence).

Key Features:

1. **Age**: Age of the patient.
2. **Gender**: Male/Female.
3. **Total_Bilirubin**: A diagnostic indicator of liver function.
4. **Direct_Bilirubin**: A more specific measure of bilirubin levels.
5. **Alkaline_Phosphotase**: Enzyme linked to liver and bone health.
6. **Alamine_Aminotransferase**: Enzyme indicating liver inflammation.
7. **Aspartate_Aminotransferase**: Enzyme related to liver and heart damage.
8. **Total_Protiens**: Measure of overall protein in the blood.
9. **Albumin**: Protein produced by the liver.
10. **Albumin_and_Globulin_Ratio**: Ratio of two types of blood proteins.

Steps in the Study

1. Data Preprocessing

- **Handling Missing Values**:
    - Identified and imputed missing values using mean and median for numerical features.
- **Outlier Detection and Treatment**:
    - Used the Interquartile Range (IQR) method to detect and cap outliers in numerical features.

2. Exploratory Data Analysis (EDA)

- Visualized the distribution of numerical variables using histograms and boxplots.
- Examined correlations between features to identify important predictors.

- Gender-wise analysis showed males are more frequently affected by liver disease in this dataset.

3. Model Development

Applied the following machine learning models:

1. **Logistic Regression**: A simple baseline model for binary classification.
2. **Decision Tree Classifier**: A rule-based model that identifies important features.
3. **Random Forest Classifier**: An ensemble method combining multiple decision trees.
4. **Linear Discriminant Analysis (LDA)**: A statistical method for dimensionality reduction and classification.
5. **K-Nearest Neighbors (KNN)**: A distance-based classification method.
6. **Support Vector Machine (SVM)**: A hyperplane-based classifier.

4. Model Evaluation

- Split the dataset into 80% training and 20% testing sets.
- Evaluated models using accuracy, precision, recall, and F1-score.
- Utilized cross-validation to ensure robustness.

Results

- **Best Model**: Random Forest Classifier
- **Accuracy**: 86%
- **Precision**: 85%
- **Recall**: 87%
- **F1-Score**: 86%
- **Feature Importance**: Total_Bilirubin, Direct_Bilirubin, and Aspartate_Aminotransferase were identified as the most significant predictors.

Comparison of Model Performances:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	80%	78%	81%	79%
Decision Tree	82%	83%	84%	83%
Random Forest	**86%**	**85%**	**87%**	**86%**
Linear Discriminant	78%	76%	80%	78%
K-Nearest Neighbors	75%	73%	77%	75%
Support Vector Machine	79%	77%	81%	78%

Conclusion

- **Random Forest Classifier** emerged as the best-performing model due to its high accuracy and robustness.

- The features **Total_Bilirubin**, **Direct_Bilirubin**, and **Aspartate_Aminotransferase** play a crucial role in predicting liver disease.
- Future improvements could involve:
    - **Hyperparameter tuning** for further model optimization.
    - Applying advanced ensemble techniques like **XGBoost** or **LightGBM**.
    - Incorporating additional data or external datasets to enhance model generalizability.

Business Impact

This predictive model can assist healthcare providers in identifying at-risk patients more efficiently, enabling earlier intervention and better allocation of medical resources. By automating the diagnostic process, hospitals can reduce diagnostic errors and improve patient outcomes.