Submitted By,

Kedar Sunil Desai

kedardesai9005@gmail.com

LinkedIn | Portfolio | GitHub

Liver Disease Prediction

Overview:

This project is a full-stack machine learning web application that predicts the likelihood of liver disease based on patient health and lifestyle data. It uses a high-accuracy XGBoost classifier trained on a clean clinical dataset with features like age, BMI, liver function test results, alcohol consumption, and genetic risk. Extensive exploratory data analysis and SHAP-based interpretation ensure model transparency. The backend is built with FastAPI, and the frontend uses HTML, CSS, and JavaScript to provide a responsive and user-friendly interface for real-time predictions.

Dataset Information :

The dataset consists of 1,700 patient records with 10 input features and 1 target label (Diagnosis). Dataset is downloaded from Kaggle here is the link to the dataset:

https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset

The input features include:

- **Age** (20–80 years)
- **Gender** (0 = Male, 1 = Female)
- **BMI** (15–40)
- **Alcohol Consumption** (0–20 units per week)
- **Smoking** (0 = No, 1 = Yes)
- **Genetic Risk** (0 = Low, 1 = Medium, 2 = High)
- **Physical Activity** (0–10 hours per week)
- **Diabetes** (0 = No, 1 = Yes)
- **Hypertension** (0 = No, 1 = Yes)

- **Liver Function Test** (20–100)

The target variable, Diagnosis, is binary: **0** for no liver disease and **1** for liver disease.

➢ **Exploratory Data Analysis (EDA):**

1. **Dataset Overview:**
   - Shape: 1700 rows × 11 columns
   - Target variable: Diagnosis (0 = No disease, 1 = Liver Disease)
   - **Feature types:**
     - Numerical: Age, BMI, AlcoholConsumption, PhysicalActivity, LiverFunctionTest
     - Categorical/Binary: Gender, Smoking, GeneticRisk, Diabetes, Hypertension

2. Missing Values:
   - Checked for missing values (null values), no missing values(null values) present in data

3. **Class Balance Check:**
   - Plotted value counts of Diagnosis column.
   - Slight class imbalance was detected.

4. **Univariate Analysis:**
   - Histograms and boxplots used for each numerical column
   - Most features were within defined medical ranges.
   - Minor outliers detected (but not severe).
   - BMI and AlcoholConsumption are right-skewed.

5. **Bivariate Analysis with Target:**
   - Used boxplots and violin plots to compare features by Diagnosis.
   - Liver Function Test and Alcohol Consumption higher in diseased patients.
   - Physical Activity lower in diseased patients.

6. **Categorical Feature Analysis:**
   - Used count plots grouped by Diagnosis for Gender, Smoking, etc
   - **Chi-square tests** performed to assess significance.

- **Result**: Gender, Smoking, GeneticRisk, Diabetes, and Hypertension were statistically significant ($p < 0.05$).

7. **Correlation Matrix:**
    - Created correlation heatmap to check multicollinearity
- Result: No strong multicollinearity and LiverFunctionTest and GeneticRisk highly correlated with target

8. **Outlier Detection**:

- Boxplots used to detect outliers
- **Result**: No critical outliers — data within acceptable clinical ranges.
- Key Insights from EDA:
    - LiverFunctionTest, GeneticRisk, and AlcoholConsumption are top features influencing liver disease
    - Dataset is clean, balanced enough for modeling, and well-suited for tree-based algorithms.
    - All features contribute useful variance — no redundant columns.

➢ **Data Preprocessing:**

1. **Feature Selection:**

- Removed no features — all 10 input features were found relevant from EDA.

2. **Feature Scaling:**

- Data Scaled Using Standard scaling technique.

3. **Train-Test Split:**

- Dataset split using train_test_split()
- 80% for training, 20% for evaluation

- **Model Training:**

1. **Model Selection:**

   - XGBoost Classifier model used for classification.
   - Model trained on the train data

2. **Hyperparameter Tuning :**

   - Used GridSearchCV for hyperparameter tuning to improve model performance

3. **Cross Validation**:

   - Used K-Fold Cross Validation to check model stability.
   - Results: Accuracy ~ 91% and Standard Deviation ~ 0.0198.

- **Model Evaluation:**

After training the XGBoost classifier and optimizing hyperparameters, the model was evaluated using a combination of **classification metrics**, **cross-validation**, and **visual tools** to ensure reliability and interpretability.

1. Evaluation Metrics:

   - Accuracy: 90.88%
   - Precision: 0.9483
   - Recall: 0.8824
   - F1 Score: 0.9141
   - AUC ROC Score: 0.9591

2. **Confusion Matrix**:

   [144 9

   22 165]

3. **SHAP-Based Model Interpretation**:

   - **Gender** (value = 1, likely Male) had the strongest positive influence: it pushed the prediction **toward "Disease"**.
   - **PhysicalActivity** and **AlcoholConsumption** had a slight **negative impact**, pushing the prediction down

➢ Model Deployment:

1. Model Saving:

- The best-performing pipeline (preprocessing + XGBoost model) was serialized using pickle

2. Backend – FastAPI:

- FastAPI was used to create a lightweight, asynchronous RESTful API for model inference.
- **POST endpoint**: /predict
- **Input validation**: via Pydantic BaseModel
- **Response**: JSON with prediction result

3. Frontend – HTML, CSS, JavaScript

- Clean form with 10 input fields
- Uses dropdowns and radio buttons for categorical data
- Displays prediction dynamically
- Professional UI with card-style container
- JavaScript sends JSON to /predict via fetch()

➢ Conclusion:

This project successfully delivers a complete machine learning solution for predicting liver disease using clinical and lifestyle data. Through comprehensive exploratory data analysis, model training with XGBoost, and SHAP-based interpretation, the system achieves high accuracy and transparency. The integration of FastAPI for backend deployment and a clean, responsive frontend ensures a user-friendly experience for real-time predictions. This project demonstrates the effective application of data science and web technologies in a healthcare context, with potential for real-world impact and future scalability.