PROJECT REPORT

TEAM MEMBERS

A.MAHESH(22781A3308)

C.ESWARI(22781A3112)

S.KARTHIK (21781A33E0)

Project Documentation: Liver Cirrhosis Prediction Using Machine Learning

Table of Contents

1. Introduction

Liver cirrhosis is a serious health condition where the liver tissue becomes scarred and affects the organ's functionality. Predicting the status of liver cirrhosis patients using machine learning can assist healthcare professionals in making data-driven decisions and providing timely treatment.

This project leverages Random Forest Classifier along with various preprocessing and visualization techniques to build a predictive model for liver cirrhosis.

2. Objectives

The main goals of this project are:

- To clean and preprocess the given dataset.
- To address class imbalance using SMOTE.
- To train a machine learning model using Random Forest Classifier.
- To build an interactive interface for real-time predictions.

---

3. Dataset Description

The dataset used in this project is 'cirrhosis.csv'.

Key Columns

- **ID**: Unique identifier for each patient.
- **N_Days**: Number of days between registration and end of study.
- **Status**: Patient status (Censored, Liver Transplant, Death).
- **Drug**: Drug given (D-penicillamine or placebo).
- **Age**: Age in years.
- **Sex**: Male or Female.
- **Ascites**: Presence of ascites (yes/no).
- **Hepatomegaly**: Presence of hepatomegaly (yes/no).
- **Spiders**: Presence of spider angiomata (yes/no).
- **Edema**: Presence of edema (yes/no/with diuretics).
- **Bilirubin**: Serum bilirubin level.

- **Cholesterol**: Serum cholesterol level.
- **Albumin**: Albumin level in blood.
- **Copper**: Urine copper level.
- **Alk_Phos**: Alkaline phosphatase level.
- **SGOT**: Serum glutamic-oxaloacetic transaminase.
- **Tryglicerides**: Triglyceride levels.
- **Platelets**: Platelet count.
- **Prothrombin**: Prothrombin time.

Stage: Histological stage of disease (1-4).

4. Data Preprocessing

-

Steps Involved

1. **Handling Missing Values**
   - Numerical columns: Median imputation.

    - Categorical columns: Mode imputation.

2. **Encoding Categorical Variables**
    - Used LabelEncoder for columns containing object types.
3. **Feature Selection**
    - Dropped irrelevant columns: 'ID' and 'Status'.
4. **Handling Class Imbalance**
    - Used **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the dataset.
5. **Train-Test Split**
    - **80%** of data for training and **20%** for testing.
6. **Feature Scaling**
    - Used **StandardScaler** for normalizing numerical data.

5. Model Training

The **Random Forest Classifier** was chosen due to its high accuracy and ability to handle imbalanced datasets.

Model Parameters

- n_estimators: 100
- class_weight: 'balanced'
- random_state: 42

6. Model Evaluation

The following metrics were used to assess the model:

- **Accuracy Score**
- **Confusion Matrix**
- **Classification Report**

- **ROC-AUC Score**
- **ROC Curve**

---

7. Interactive Prediction Interface

A user-friendly prediction interface was created using Jupyter widgets. The interface allows users to input clinical parameters using sliders and text fields, and get real-time predictions by clicking a 'Predict' button.

---

8. Conclusion

This project successfully demonstrates the use of machine learning in predicting liver cirrhosis status. By preprocessing the data, handling imbalances with SMOTE, and training a Random Forest model, we achieved reliable predictions. The interactive interface adds further usability for healthcare professionals.

9. Future Scope

- Integrate the model into a **web-based application**.
- Experiment with deep learning models to enhance accuracy.
- Use additional medical datasets to improve model robustness.
- Add visualizations for better explainability of model predictions.

10. References

- Research articles on liver cirrhosis.
- Scikit-learn documentation.

- Jupyter widgets official guide.
- Python libraries: Pandas, Numpy, Seaborn, Matplotlib, Imbalanced-learn.

End of Documentation

Prediction of Liver Cirrhosis Using Machine Learning

1. Introduction

Liver cirrhosis is a chronic disease characterized by the replacement of healthy liver tissue with scar tissue, leading to liver dysfunction. Early detection and prediction of liver cirrhosis can help in managing the disease and improving patient outcomes. This project leverages machine learning techniques to build a predictive model using a Random Forest Classifier to identify the status of liver cirrhosis patients based on various clinical attributes.

2. Objectives

The main objectives of this project are:

- To preprocess and clean the dataset for effective modeling.
- To address class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).
- To build a predictive model using Random Forest Classifier.

- To design an interactive interface using Jupyter widgets for real-time prediction.

3. Dataset Description

The dataset used in this project is the 'cirrhosis.csv' file, containing the following key columns:

- **ID**: Unique identifier for each patient.
- **N_Days**: Number of days between registration and end of study.
- **Status**: Patient status ('C' = censored, 'CL' = liver transplant, 'D' = death).
- **Drug**: Type of drug given (D-penicillamine or placebo).
- **Age**: Age of the patient in years.

- **Sex**: Gender of the patient (male or female).
- **Ascites**: Presence of ascites (yes or no).
- **Hepatomegaly**: Presence of hepatomegaly (yes or no).
- **Spiders**: Presence of spider angiomata (yes or no).
- **Edema**: Presence of edema (yes, no, or with diuretics).
- **Bilirubin**: Serum bilirubin level.
- **Cholesterol**: Serum cholesterol level.
- **Albumin**: Albumin level in the blood.
- **Copper**: Urine copper level.

- **Alk_Phos**: Alkaline phosphatase level.
- **SGOT**: Serum glutamic-oxaloacetic transaminase.
- **Tryglicerides**: Triglyceride levels.
- **Platelets**: Platelet count.
- **Prothrombin**: Prothrombin time.
- **Stage**: Histological stage of disease (1–4).

4. Data Preprocessing

The preprocessing steps involved:

- **Handling Missing Values**: Median imputation for numerical variables and mode imputation for categorical variables.

- **Encoding Categorical Variables**: Label Encoding for all object-type columns.
- **Feature Selection**: Dropping irrelevant columns such as 'ID' and 'Status'.
- **Handling Class Imbalance**: Using SMOTE to balance the dataset.
- **Train-Test Split**: Splitting data into training (80%) and testing (20%) sets.
- **Feature Scaling**: Standardizing the numerical features using StandardScaler.

5. Model Training

A Random Forest Classifier was used for model training with the following parameters:

- **n_estimators**: 100
- **class_weight**: 'balanced'
- **random_state**: 42

The model was trained on the resampled data and evaluated using accuracy, classification report, confusion matrix, and ROC curve.

6. Model Evaluation

The model's performance was assessed using:

- **Accuracy Score**
- **Confusion Matrix**
- **Classification Report**

- **ROC-AUC Score**
- **ROC Curve**

7. Interactive Prediction Interface

To make the model user-friendly, an interactive interface was created using Jupyter widgets. Users can input patient data via sliders and text fields and obtain real-time predictions by clicking a 'Predict' button.

8. Conclusion

The project successfully built a machine learning model to predict liver cirrhosis status based on clinical features. The use of SMOTE balanced the dataset, and Random Forest Classifier provided reliable predictions. The interactive widget interface enhances user experience, allowing for real-time analysis and decision-making.

9. Future Scope

- Integrate additional medical datasets for more robust predictions.
- Deploy the model as a web application for broader accessibility.
- Experiment with deep learning models to further improve accuracy.

10. References

- Research articles on liver cirrhosis and its clinical indicators.

- Documentation for Scikit-learn, Pandas, and Seaborn.
- Official guides for Jupyter widgets.

This project demonstrates the power of machine learning in healthcare and highlights the importance of early detection in managing chronic diseases like liver cirrhosis.

Conclusion

In this project, we successfully developed a machine learning-based solution to predict liver cirrhosis status using clinical data. By leveraging Random Forest Classifier and addressing class imbalance with SMOTE, we built a robust predictive model capable of assisting healthcare professionals in early detection and decision-making. The interactive interface using Jupyter widgets adds a practical layer, allowing real-time predictions based on patient data inputs.

The project highlights the potential of machine learning in healthcare, emphasizing how data-driven approaches can enhance disease management. While the current model performs well, there is scope for further improvement by integrating more advanced algorithms and expanding the dataset. Ultimately, this project lays the foundation for developing a reliable and accessible liver cirrhosis prediction tool.