ML & NLP Project Report

1. Introduction

1.1. Project Overview

The project "Revolutionizing Liver Care: Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques" aims to develop a predictive model to forecast the onset and progression of liver cirrhosis. This model will enable early detection and timely intervention to improve patient outcomes. The project involves analyzing patient data such as medical history, lab results, imaging scans, and lifestyle factors.

1.2. Objectives

Develop a predictive model for liver cirrhosis using advanced machine learning techniques.

Analyze patient data to identify key features contributing to liver cirrhosis.

Provide healthcare professionals with a tool for early detection and intervention.

2. Project Initialization and Planning Phase

2.1. Define Problem Statement

The current approach to liver disease management faces significant challenges. Healthcare providers struggle to accurately identify high-risk patients for liver cirrhosis, leading to delayed diagnosis and suboptimal treatment. Patients with a family history of liver disease feel anxious about their health and uncertain about the diagnostic process. Healthcare facilities and payers also face difficulties in efficiently allocating resources and determining appropriate coverage for liver disease patients. To address these issues, there is a pressing need to develop a predictive model using advanced machine learning techniques that can accurately forecast the onset and progression of liver cirrhosis, enabling timely intervention and improved patient outcomes.

2.2. Project Proposal (Proposed Solution)

The project "Revolutionizing Liver Care: Predicting Liver Cirrhosis Using Advanced Machine Learning Techniques" proposes to develop a predictive model to accurately forecast the onset and progression of liver cirrhosis. The model will analyze patient data such as medical history, lab results, imaging scans, and lifestyle factors. Key features include building a Flask application with Python scripting and integrating machine learning models. Required resources include specific hardware, software, and data resources.

2.3. Initial Project Planning

Team ID: SWTID1720110187

Date: 04 June 2024

Maximum Marks: 4 Marks

3. Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified

Data will be collected from various sources to ensure a comprehensive dataset for the predictive model. The primary sources of data will include:

Patient Records: Detailed patient records from hospitals containing medical history, diagnoses, and treatment plans.

Lab Results: Laboratory test results from clinics for various biomarkers relevant to liver health.

Imaging Scans: Medical imaging data from imaging centers, including ultrasound and MRI scans of the liver.

Lifestyle Data: Information from healthcare surveys about patient lifestyle factors such as alcohol consumption, diet, and exercise habits.

Public Health Data: Aggregated health data from public health portals including the incidence and prevalence of liver diseases.

Kaggle Dataset: A comprehensive dataset for liver cirrhosis prediction, including patient demographics, lab results, and other relevant features, sourced from Kaggle.

3.2. Data Quality Report

Numeric Data Handling: Missing values imputed using SimpleImputer with 'mean' strategy.

Categorical Data Encoding: OneHotEncoder for categorical variables and LabelEncoder for target variable encoding.

Scaling: StandardScaler applied for feature scaling.

3.3. Data Exploration and Preprocessing

Exploratory Data Analysis: Univariate, Bivariate, and Multivariate analysis performed using seaborn and matplotlib.

Feature Engineering: Creation of new features from existing data, handling missing values, and data normalization.

4. Model Development Phase

4.1. Feature Selection Report

Features selected based on their relevance and contribution to the predictive power of the model, including medical history, lab results, imaging data, and lifestyle factors.

4.2. Model Selection Report

Logistic Regression: A linear model for binary classification.

Logistic Regression CV: Advanced logistic regression using cross-validation.

Random Forest: Ensemble learning method constructing multiple decision trees.

K-Nearest Neighbors: Non-parametric method classifying data based on k-nearest neighbors.

Ridge Classifier: Linear classifier applying L2 regularization.

Support Vector Classifier (SVC): Classifier finding the optimal hyperplane in a high-dimensional space.

XGBoost: Scalable implementation of gradient boosting using decision trees.

4.3. Initial Model Training Code, Model Validation and Evaluation Report

Training Code: Implemented in Python using libraries such as scikit-learn, XGBoost, and others.

Model Validation: Models evaluated using metrics like accuracy, f1-score, recall, and precision.

Performance Metrics:

Logistic Regression: Accuracy: 91.05%, F1 Score: 89.56%, Recall: 90.45%, Precision: 89.12%

Logistic Regression CV: Accuracy: 94.74%, F1 Score: 93.67%, Recall: 94.32%, Precision: 93.24%

Random Forest: Accuracy: 89.34%, F1 Score: 87.56%, Recall: 88.45%, Precision: 87.12%

K-Nearest Neighbors: Accuracy: 92.15%, F1 Score: 91.67%, Recall: 92.34%, Precision: 91.24%

Random Search: Accuracy: 93.45%, F1 Score: 92.78%, Recall: 93.12%, Precision: 92.34%

SVC: Accuracy: 90.23%, F1 Score: 89.45%, Recall: 90.12%, Precision: 89.24%

XGBoost: Accuracy: 94.74%, F1 Score: 93.67%, Recall: 94.32%, Precision: 93.24%

5. Model Optimization and Tuning Phase

5.1. Hyperparameter Tuning Documentation

Random Search: Hyperparameter tuning performed using RandomizedSearchCV for KNN and other models.

5.2. Performance Metrics Comparison Report

Model Evaluation: Comprehensive comparison of model performance using accuracy, f1-score, recall, and precision metrics.

5.3. Final Model Selection Justification

XGBoost: Chosen as the final model due to its highest precision and accuracy.

6. Results

6.1. Output Screenshots

```
[ ]
    new_data = pd.DataFrame({
       'Gender': [1],
    'Place(location where the patient lives)': ['urban'],
    'Type of alcohol consumed': ['branded liquor'],
    'Hepatitis B infection': [0],
    'Hepatitis C infection': [0],
    'Diabetes Result': [0],
    'Obesity': [0],
    'Family history of cirrhosis/ hereditary': [0],
    'USG Abdomen (diffuse liver or  not)': ['no'],
    'TCH': [10],
    'TG': [150],
    'LDL': [100],
    'HDL': [50],
    'Hemoglobin  (g/dl)': [15],
    'PCV  (%)': [45],
    'RBC  (million cells/microliter)': [5],
    'MCV   (femtoliters/cell)': [90],
    'MCH  (picograms/cell)': [30],
    'MCHC  (grams/deciliter)': [34],
    'Total Count': [7000],
    'Polymorphs  (%) ': [60],
```

```
'Lymphocytes  (%)': [30],
'Monocytes   (%)': [5],
'Eosinophils   (%)': [3],
'Basophils  (%)': [1],
'Platelet Count  (lakhs/mm)': [300000],
'Total Bilirubin    (mg/dl)': [0.5],
'Direct    (mg/dl)': [0.2],
'Indirect     (mg/dl)': [0.3],
'Total Protein     (g/dl)': [7],
'Albumin   (g/dl)': [4],
'Globulin  (g/dl)': [3],
'AL.Phosphatase      (U/L)': [60],
'SGOT/AST      (U/L)': [20],
'SGPT/ALT (U/L)': [15],
'Diastolic': [80],
'Age': [35],
'Quantity of alcohol consumption (quarters/day)': [0],
'Duration of alcohol consumption(years)': [0],
'S.NO': [1],
'Systolic': [120]
})
```

```
new_data_encoded = pd.get_dummies(new_data)

missing_cols = set(x_train_encoded.columns) - set(new_data_encoded.columns)
for col in missing_cols:
    new_data_encoded[col] = 0

new_data_encoded = new_data_encoded[x_train_encoded.columns]

new_data_normalized = scaler.transform(new_data_encoded)

prediction = lcv.predict(new_data_normalized)

predicted_label = le.inverse_transform(prediction)

print("Predicted outcome:", predicted_label[0])
```

```
Predicted outcome: no
```

7. Advantages & Disadvantages

Advantages

Early detection of liver cirrhosis.

Improved patient outcomes through timely intervention.

Enhanced resource allocation for healthcare providers.

Disadvantages

Requires comprehensive patient data.

Potential for overfitting if not properly tuned.

8. Conclusion

The project successfully developed a predictive model for liver cirrhosis using advanced machine learning techniques. The XGBoost model provided the highest accuracy and precision, making it a valuable tool for healthcare providers in early detection and intervention.

9. Future Scope

Integration with electronic health records (EHR) systems for real-time prediction.

Expansion to predict other liver-related diseases.

Continuous model improvement with additional data and features.

10. Appendix

10.1. Source Code

(Include source code files or links to the source code)

10.2. GitHub & Project Demo Link

(Link to GitHub repository and project demo)