Machine Learning for Sustainable Development Goal 3: Good Health and Well-being

Patient Data Analysis and Chronic Disease Prediction

1. Introduction

Project Objective:
To leverage machine learning for analyzing patient data and predicting the likelihood of chronic diseases such as diabetes, heart disease, or cancer, while providing personalized health recommendations to improve patient outcomes, thereby supporting SDG 3.

Motivation:
Chronic diseases are a leading cause of morbidity and mortality worldwide. By utilizing predictive analytics, this project aims to enhance healthcare outcomes through early detection, enabling healthcare providers to deliver targeted interventions and resource allocation.

2. Data Collection

Data Source: Public health datasets (e.g., CDC, WHO, or healthcare databases) or synthetic datasets if applicable.

Dataset Description:

- **Features:** Patient demographics (age, gender), medical history (previous diagnoses, family history), lifestyle factors (diet, physical activity), and laboratory results (blood pressure, cholesterol levels).
- **Size:** X rows by Y columns (specify actual values).
- **Target Variable:** Presence of chronic disease (binary variable: 0 = no, 1 = yes).

3. Exploratory Data Analysis (EDA)

Summary Statistics:
Calculated the mean, median, and distribution of all features to understand the patient population.

Visualizations:

- Correlation heatmap to identify relationships between lifestyle factors and chronic diseases.
- Boxplots for detecting outliers in clinical measurements.
- Histograms to evaluate the distribution of each feature.

Insights:
Identified significant correlations between factors such as age and blood pressure with the likelihood of chronic diseases, emphasizing the importance of these variables in health outcomes.

4. Data Preprocessing

Handling Missing Values:
Applied median imputation for missing values in numerical features and mode imputation for categorical variables.

Encoding Categorical Variables:
One-hot encoding was used for categorical variables such as gender and lifestyle choices.

Feature Scaling:
Standardized numerical features using StandardScaler to enhance model performance and convergence.

5. Machine Learning Model Selection

- Model Choices: Logistic Regression for its interpretability in binary classification.
- Random Forest Classifier for its ability to capture complex interactions and feature importance.
- Support Vector Machine (SVM) for effective classification in high-dimensional spaces.

Why Scikit-Learn:

Scikit-Learn provides efficient implementations and diverse evaluation metrics, making it ideal for supervised learning tasks.

Evaluation Metrics:

Accuracy, Precision, Recall, F1 Score, and ROC-AUC to assess model performance.

6. Model Implementation

Data Splitting:
Split the dataset into 80% training and 20% testing sets using train_test_split from Scikit-Learn.

Hyperparameter Tuning:

- Used GridSearchCV for the Random Forest model to identify optimal hyperparameters such as the number of estimators and maximum depth.
- Implemented 5-fold cross-validation to improve model generalizability.

Code:

```
from sklearn.model_selection import train_test_split, GridSearchCV

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, classification_report, roc_auc_score

# Splitting the data

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Hyperparameter tuning for Random Forest

param_grid = {'n_estimators': [50, 100, 200], 'max_depth': [10, 20, 30]}

rf = RandomForestClassifier(random_state=42)

grid_search = GridSearchCV(estimator=rf, param_grid=param_grid, cv=5,
scoring='accuracy')

grid_search.fit(X_train, y_train)

# Best model and evaluation

best_model = grid_search.best_estimator_

y_pred = best_model.predict(X_test)

# Evaluation metrics

print("Accuracy:", accuracy_score(y_test, y_pred))

print("Classification Report:\n", classification_report(y_test, y_pred))

print("ROC-AUC Score:", roc_auc_score(y_test, best_model.predict_proba(X_test)[:,
1]))
```

7. Results and Evaluation

Model Performance:
The Random Forest classifier achieved an accuracy of X%, a recall of Y%, and an F1 Score of Z, indicating strong predictive performance for chronic disease classification based on patient data.

Feature Importance:
Identified key features such as age, cholesterol levels, and lifestyle choices that significantly influence disease prediction outcomes.

Confusion Matrix:
Analyzed the confusion matrix to evaluate model performance across different classes and assess false positives and negatives.

8. Conclusion and Future Work

Key Takeaways:
The machine learning model effectively predicts chronic diseases, contributing to proactive healthcare management and improved patient outcomes.

Future Improvements:

- Expand the dataset with more diverse patient demographics for broader applicability.
- Integrate real-time health monitoring data to enhance predictive capabilities.
- Develop a user-friendly interface for healthcare providers to access predictions and recommendations.

9. References

- Public Health Datasets (e.g., CDC, WHO)
- Scikit-Learn Documentation