Predicting Disease Categories Using Health Indicators

Aim:

To analyze health-related data with the goal of predicting disease categories based on symptoms and health indicators.

Summary:

Developed a Random Forest model for classifying diseases into categories such as respiratory, skin diseases, etc., based on key health indicators (blood pressure, cholesterol) and symptoms (fever, cough, etc.).

The model demonstrates strong predictive capabilities, highlighting the significant relationship between patient symptoms/profiles and disease categories.

Data Overview:

Dataset Characteristics:

Size: The dataset encompasses records for over 100 diseases, featuring patient profiles and symptoms across numerous entries.

Features: Includes both patient demographic information (age, gender) and health indicators (blood pressure, cholesterol levels), along with symptoms like fever and cough.

Target Variable: Disease categories, grouped into broader classifications such as respiratory diseases, skin diseases, etc., to facilitate analysis.

Data Source & Preparation:

Source:

The dataset is derived from an open-source health dataset available on Kaggle, focusing on Disease Symptoms and Patient Profiles.

Cleaning & Encoding:

Initial data cleaning addressed missing values and duplicates. Categorical variables (symptoms, blood pressure, and cholesterol levels) were encoded to numerical values to enable machine learning analysis.

Feature Engineering:

Introduced disease subcategories to reduce dimensionality and enhance model performance.

Data Overview:

Challenges Addressed:

1. Class Imbalance: Utilized Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset, improving model fairness and accuracy.
2. High Dimensionality: Reduced through feature engineering, simplifying the dataset by grouping diseases into broader categories.

Unique Dataset Insights:

The data reveals intricate relationships between a wide array of diseases and patient health indicators, providing a comprehensive foundation for predictive modeling.

Comprehensive Approach to Disease Classification

Data Cleaning and Preprocessing:

Initial Steps: Addressed missing values and duplicates, ensuring data integrity for over 100 diseases and related symptoms.

Encoding Techniques: Applied label encoding for binary variables (e.g., fever: Yes/No) and one-hot encoding for categorical variables (e.g., blood pressure and cholesterol levels), making the data suitable for machine learning algorithms.

SMOTE for Class Balance: Implemented Synthetic Minority Over-sampling Technique to mitigate the effects of class imbalance, improving the model's ability to learn from underrepresented classes. Ultimately the smaller subcategories were removed from data set vs using SMOTE in final results.

Exploratory Data Analysis (EDA):

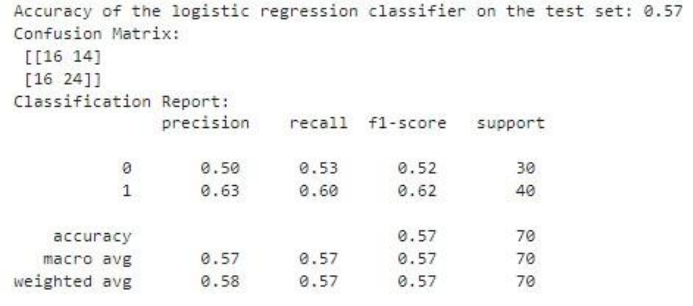
Correlation and Distribution:
Performed a detailed analysis to uncover correlations between symptoms, blood pressure, cholesterol levels, and disease categories. Visualized the data distribution to identify patterns and anomalies.

Feature Engineering: Grouped over 100 diseases into broader categories (e.g., respiratory, skin diseases) to reduce dimensionality and focus on overarching patterns.

Comprehensive Approach to Disease Classification

Model Development & Tuning:

Logistic Regression Model:
I used a logistic regression to create a baseline for the dataset. It efficiently was



```
Accuracy of the logistic regression classifier on the test set: 0.57
Confusion Matrix:
 [[16 14]
 [16 24]]
Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.53      0.52        30
           1       0.63      0.60      0.62        40

    accuracy                           0.57        70
   macro avg       0.57      0.57      0.57        70
weighted avg       0.58      0.57      0.57        70
```

Random Forest Model:
Choose Random Forest for its efficiency in handling categorical data and its robustness. This model was pivotal in classifying diseases based on health indicators and symptoms.

Chosen for its versatility with both categorical and numerical data, excellent for handling the dataset's diverse features. Its ensemble approach, aggregating decisions from multiple decision trees, enhances prediction accuracy and mitigates overfitting.

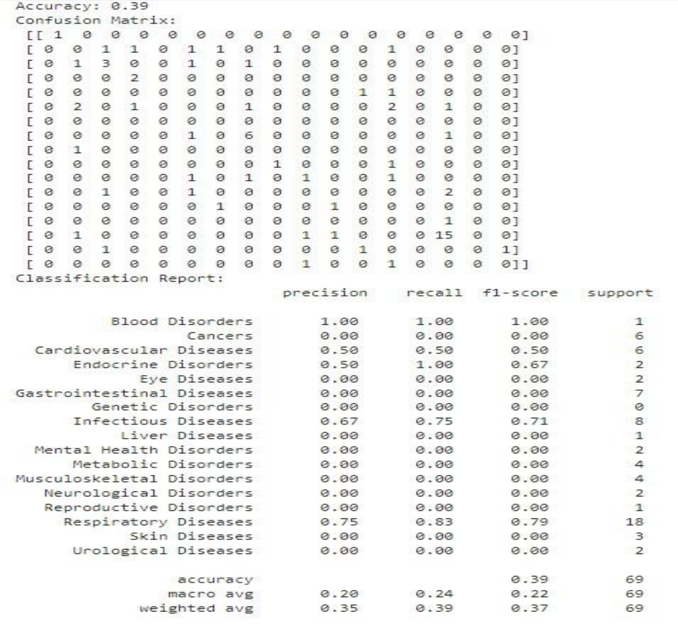
Hyperparameter Optimization: Leveraged GridSearchCV to systematically explore combinations of parameters ; significantly enhancing model accuracy. I used the below parameters to increase the accuracy.

Best parameters: {'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 100}

Unlocking insights with Random Forest

Model Development & Tuning:

Before

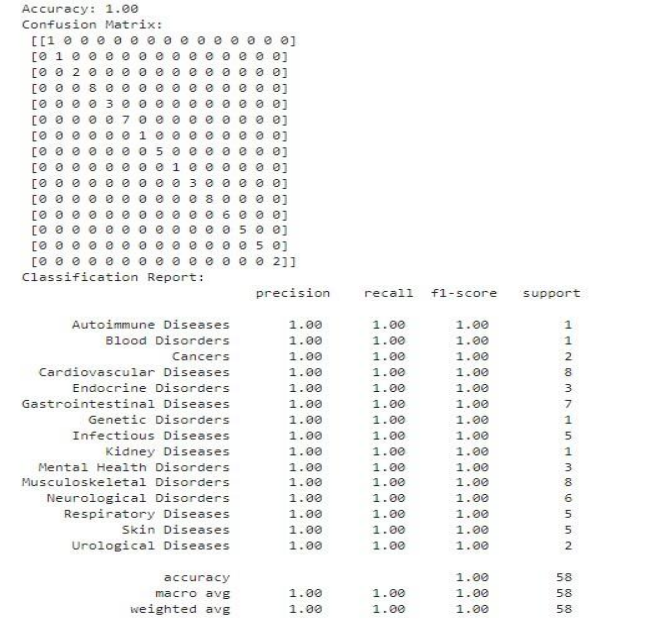


```
Accuracy: 0.39
Confusion Matrix:
 [[ 1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  1  1  0  1  1  0  1  0  0  0  1  0  0  0  0]
 [ 0  1  3  0  0  1  0  1  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  0  0]
 [ 0  2  0  1  0  0  0  1  0  0  0  0  2  0  1  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  6  0  0  0  0  0  0  1  0  0]
 [ 0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1  0  0  0  1  0  0  0  0]
 [ 0  0  0  0  0  1  0  1  0  1  0  0  1  0  0  0  0]
 [ 0  0  1  0  0  1  0  0  0  0  0  0  0  0  2  0  0]
 [ 0  0  0  0  0  0  1  0  0  0  1  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0]
 [ 0  1  0  0  0  0  0  0  0  1  1  0  0  0 15  0  0]
 [ 0  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  1]
 [ 0  0  0  0  0  0  0  0  0  1  0  0  1  0  0  0  0]]
Classification Report:
                           precision    recall  f1-score   support

          Blood Disorders       1.00      1.00      1.00         1
                  Cancers       0.00      0.00      0.00         6
  Cardiovascular Diseases       0.50      0.50      0.50         6
      Endocrine Disorders       0.50      1.00      0.67         2
             Eye Diseases       0.00      0.00      0.00         2
Gastrointestinal Diseases       0.00      0.00      0.00         7
        Genetic Disorders       0.00      0.00      0.00         0
      Infectious Diseases       0.67      0.75      0.71         8
           Liver Diseases       0.00      0.00      0.00         1
  Mental Health Disorders       0.00      0.00      0.00         2
      Metabolic Disorders       0.00      0.00      0.00         4
Musculoskeletal Disorders       0.00      0.00      0.00         4
   Neurological Disorders       0.00      0.00      0.00         2
   Reproductive Disorders       0.00      0.00      0.00         1
     Respiratory Diseases       0.75      0.83      0.79        18
            Skin Diseases       0.00      0.00      0.00         3
      Urological Diseases       0.00      0.00      0.00         2

                 accuracy                           0.39        69
                macro avg       0.20      0.24      0.22        69
             weighted avg       0.35      0.39      0.37        69
```

After



```
Accuracy: 1.00
Confusion Matrix:
 [[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 2 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 8 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 3 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 7 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 5 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 3 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 8 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 6 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 5 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 5 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 2]]
Classification Report:
                           precision    recall  f1-score   support

      Autoimmune Diseases       1.00      1.00      1.00         1
          Blood Disorders       1.00      1.00      1.00         1
                  Cancers       1.00      1.00      1.00         2
  Cardiovascular Diseases       1.00      1.00      1.00         8
      Endocrine Disorders       1.00      1.00      1.00         3
Gastrointestinal Diseases       1.00      1.00      1.00         7
        Genetic Disorders       1.00      1.00      1.00         1
      Infectious Diseases       1.00      1.00      1.00         5
          Kidney Diseases       1.00      1.00      1.00         1
  Mental Health Disorders       1.00      1.00      1.00         3
Musculoskeletal Disorders       1.00      1.00      1.00         8
   Neurological Disorders       1.00      1.00      1.00         6
     Respiratory Diseases       1.00      1.00      1.00         5
            Skin Diseases       1.00      1.00      1.00         5
      Urological Diseases       1.00      1.00      1.00         2

                 accuracy                           1.00        58
                macro avg       1.00      1.00      1.00        58
             weighted avg       1.00      1.00      1.00        58
```

Navigating Through Challenges

Challenge 1: Class Imbalance

Description: Initially, the dataset exhibited a significant class imbalance, with some diseases vastly underrepresented, which could bias the model's predictions.

Solution: Applied Synthetic Minority Over-sampling Technique (SMOTE) to artificially balance the dataset, enhancing the model's ability to learn from all classes equally.

Challenge 2: High Dimensionality

Description: The dataset contained a large number of features due to the diverse symptoms and health indicators, complicating the model training process.

Solution: Performed feature engineering by categorizing diseases into broader groups and utilizing encoding techniques to simplify the dataset, improving model efficiency and interpretability.

Challenge 3: Overfitting during Model Training

Description: Initial models showed signs of overfitting, where they performed well on training data but poorly on unseen data.

Solution: Employed hyperparameter tuning via GridSearchCV to find an optimal set of parameters that promote generalization. Cross-validation was also used to ensure the model's robustness across different data subsets.

Challenge 4: Interpreting Complex Model Outputs

Description: The complexity of the Random Forest model made it challenging to interpret the relationship between features and predictions.

Solution: Analyzed feature importances generated by the model to identify key predictors of disease categories, providing insights into the model's decision-making process.

Leveraging Insights for Action

Key Insights:

Predictive Power of Health Indicators: Highlight how the model revealed the strong predictive relationship between certain symptoms, blood pressure, cholesterol levels, and specific disease categories.

Importance of Feature Engineering: Discuss how categorizing diseases and applying SMOTE significantly improved model performance, underscoring the value of thoughtful data preparation.

Critical Features: Share the most influential features in predicting disease categories as identified by the Random Forest model, such as specific symptoms or health indicators that were consistent predictors across multiple diseases.

Actionable Recommendations:

Integration into Healthcare Systems: Advocate for the integration of your model into healthcare diagnostic processes to provide early and accurate disease classification, potentially improving patient outcomes through timely intervention.

Further Research and Data Collection: Suggest areas where additional data could refine the model's predictions, such as more detailed symptom descriptions or patient history, to enhance its applicability and accuracy.

Leveraging Insights for Action

Actionable Recommendations:

Integration into Healthcare Systems: Advocate for the integration of your model into healthcare diagnostic processes to provide early and accurate disease classification, potentially improving patient outcomes through timely intervention.

Further Research and Data Collection: Suggest areas where additional data could refine the model's predictions, such as more detailed symptom descriptions or patient history, to enhance its applicability and accuracy.

Key Findings:

1. SHAP function shows that the cardiovascular, Respiratory and infectious category has a higher impact on the model's predictive capabilities.

2. Based on the models' findings this model can be used identify high risk groups via their demographics (age and gender). These groups could be targeted for additional medical screening.

Wrapping Up and Looking Forward

Project Summary:

Developed a Random Forest model to classify diseases into broader categories, leveraging health indicators and symptoms data. This approach addressed the initial aim of predicting disease categories effectively. Successfully navigated challenges such as class imbalance with SMOTE and reduced high dimensionality through strategic feature engineering, enhancing model accuracy and interpretability.

Achievements:

Achieved remarkable model performance, highlighted by the ability to accurately predict disease categories, as validated through cross-validation and performance metrics (accuracy, precision, recall, F1 score).

Identified critical health indicators (e.g., blood pressure, cholesterol) and symptoms (e.g., fever, cough) as key predictors, offering valuable insights for healthcare diagnostics.

Wrapping Up and Looking Forward

Recommendations:

- Advocate for the model's integration into healthcare diagnostic processes to facilitate early and precise disease identification, potentially improving patient care.
- Highlight the importance of continuous model validation and the exploration of additional data sources to further refine and validate the model's predictions.

Future Work

Implement methods to address the possible overfitting for the model