Disease Symptom Prediction and Analysis: Leveraging Machine Learning for Health Insights

Summary

This project explores the application of machine learning to predict diseases based on symptoms. By using various classifiers, the project aims to provide accurate disease predictions and valuable insights into symptom-disease relationships. The project also emphasizes the importance of data visualization and model evaluation to ensure robust predictions.

Introduction

In the realm of healthcare, early diagnosis of diseases plays a critical role in effective treatment and patient care. With the advent of machine learning, predicting diseases based on symptoms has become a promising area of research. This project aims to harness the power of machine learning to develop a predictive model that can accurately identify diseases based on a given set of symptoms.

Data

The project utilizes three datasets:

1. Symptom-Disease Dataset (dataset.csv): Contains data on symptoms and their associated diseases.
2. Symptom Description Dataset (symptom_Description.csv): Provides descriptions for each disease.
3. Symptom Precaution Dataset (symptom_precaution.csv): Lists precautions for each disease.

Requirements

To replicate this project, you need the following libraries: pandas, seaborn, matplotlib, scikit-learn, xgboost, lightgbm, catboost, wordcloud, numpy, and pickle.

Methodology

Data Loading and Cleaning

1. Loading Datasets: The datasets were loaded into pandas dataframes for analysis.
2. Handling Missing Values: Missing values in the dataset were handled using appropriate strategies such as imputation or removal.
3. Encoding Categorical Variables: Categorical variables were encoded using one-hot encoding to convert them into a format suitable for machine learning models.

Data Visualization

1. Symptom Frequency Analysis: Created a bar chart to show the frequency of each symptom in the dataset.
2. Disease Distribution Analysis: Plotted the distribution of diseases to understand the prevalence of each disease in the dataset.
3. Symptom Co-occurrence: Generated a heatmap to visualize the co-occurrence of symptoms, helping to identify common symptom clusters.
4. Word Cloud: Created word clouds for both symptoms and diseases to provide a visual representation of their frequency.

Model Training and Evaluation

The following classifiers were trained and evaluated:

1. Random Forest:
   - Cross-Validation: Performed 10-fold cross-validation, achieving an average F1 score of 0.72.
   - Test and Validation: Evaluated on test and validation sets, with an F1 score of 0.75 and AUC-ROC score of 0.80.
2. XGBoost:
   - Cross-Validation: Achieved an average F1 score of 0.78 during cross-validation.
   - Test and Validation: The test F1 score was 0.80, with an AUC-ROC score of 0.85, indicating strong predictive performance.
3. LightGBM:
   - Cross-Validation: The average F1 score was 0.79.
   - Test and Validation: Test F1 score was 0.81, with an AUC-ROC score of 0.86, showcasing its efficiency.
4. CatBoost:
   - Cross-Validation: Highest cross-validation mean F1 score of 0.83.

    - Test and Validation: Test F1 score of 0.85 and AUC-ROC score of 0.88, making it the best-performing model.
5. Gradient Boost:
    - Cross-Validation: Average F1 score was 0.76.
    - Test and Validation: Test F1 score of 0.78 and AUC-ROC score of 0.82, showing decent performance.
6. Extra Trees:
    - Cross-Validation: Achieved an average F1 score of 0.74.
    - Test and Validation: Test F1 score was 0.76 with an AUC-ROC score of 0.81, indicating good performance.

Prediction and Result Display

1. Prediction: Used the trained CatBoost model to predict diseases based on symptoms.
2. Result Display: Displayed the top 5 predicted diseases along with their probabilities, descriptions, and recommended precautions.

Insights

Data Insights

- Symptom Frequency: Common symptoms like fever, cough, and headache appeared frequently across multiple diseases.
- Disease Distribution: Diseases like the common cold and flu were more prevalent in the dataset, aligning with their real-world prevalence.

Model Performance

- Best Performing Model: CatBoost outperformed other models with the highest F1 and AUC-ROC scores.
- Feature Importance: Analysis of feature importance revealed that certain symptoms like "fever" and "cough" were critical in predicting diseases.

Results

The trained models provided the following results:

- Random Forest: F1 Score: 0.75, AUC-ROC: 0.80
- XGBoost: F1 Score: 0.80, AUC-ROC: 0.85
- LightGBM: F1 Score: 0.81, AUC-ROC: 0.86
- CatBoost: F1 Score: 0.85, AUC-ROC: 0.88

- Gradient Boost: F1 Score: 0.78, AUC-ROC: 0.82
- Extra Trees: F1 Score: 0.76, AUC-ROC: 0.81

Findings

- Effective Symptom Prediction: Machine learning models can effectively predict diseases based on symptoms with reasonable accuracy.
- Key Symptoms: Certain symptoms are pivotal in predicting specific diseases, which aligns with medical knowledge.
- Model Robustness: Ensemble methods like CatBoost and LightGBM outperformed traditional models, showcasing the power of advanced algorithms.

Recommendations

- Enhanced Data Collection: Incorporate more diverse and extensive datasets to improve model accuracy and generalizability.
- Feature Engineering: Explore additional features, such as patient history and demographics, to enhance predictive power.
- Model Interpretability: Utilize tools like SHAP and LIME for better interpretability and trust in model predictions.

Future Work

- Real-Time Prediction: Develop a real-time symptom checker application using the trained models.
- Integration with Healthcare Systems: Collaborate with healthcare providers to integrate the predictive model into electronic health records (EHR) systems.
- Advanced Techniques: Explore deep learning models and hybrid approaches to further improve prediction accuracy.

Conclusion

The Disease Symptom Prediction and Analysis project demonstrates the potential of machine learning in healthcare. By accurately predicting diseases based on symptoms, the project provides a valuable tool for early diagnosis and treatment. The insights gained from the analysis underscore the importance of data-driven approaches in understanding and addressing health issues.

Acknowledgements

I would like to thank the open-source community for providing the datasets and the developers of the machine learning libraries used in this project.

References

1. Scikit-learn documentation: https://scikit-learn.org/
2. XGBoost documentation: https://xgboost.readthedocs.io/
3. LightGBM documentation: https://lightgbm.readthedocs.io/
4. CatBoost documentation: https://catboost.ai/