Report on Human Disease Prediction Model

Overview

The Human Disease Prediction Model was built using **Logistic Regression** to classify diseases based on 20 selected features related to various symptoms. The primary purpose of this project was to explore Logistic Regression as an algorithm and understand its practical application in a multiclass classification problem. This was undertaken as part of a learning exercise, acknowledging that Logistic Regression may not be the most suitable algorithm for such a complex problem.

Challenges Faced

1. **High Dimensionality of Data**:
    - The original dataset contained 134 features, which made the model prone to overfitting and computational inefficiency. Feature selection was performed to reduce the number of input features to 20 based on correlation analysis.
2. **Multiclass Classification**:
    - Logistic Regression, being inherently binary, required adaptation for multiclass classification using techniques like one-vs-rest (OvR). This increased computational complexity and reduced interpretability.
3. **Imbalanced Dataset**:
    - The dataset had imbalances in disease classes, with some classes having significantly fewer samples than others. To address this, class weights were balanced during model training, but this approach had limited impact on performance.
4. **Low Model Performance**:
    - Despite feature selection and class balancing, the model struggled to achieve satisfactory metrics:
        - **Accuracy**: ~24%
        - **Precision** and **Recall**: Both were below acceptable thresholds for many disease classes.
        - **ROC-AUC Score**: Demonstrated weak discrimination power for certain classes.
    - These results highlight that Logistic Regression, a linear model, struggles with complex, nonlinear relationships typical in medical datasets.
5. **Feature Compatibility**:
    - Ensuring that input features for new predictions matched the training features in both order and scaling presented challenges. This was resolved by saving and loading the scaler and ensuring feature alignment during preprocessing.
6. **Scalability**:

- Logistic Regression may not scale well when dealing with large datasets or highly complex relationships, as seen in this problem.

Key Steps Taken

1. **Feature Selection**:
   - Correlation analysis was used to select the 20 most relevant features based on their correlation with the target variable.
2. **Data Preprocessing**:
   - Scaling was applied to the features to standardize them and improve model training and performance.
3. **Model Evaluation**:
   - Metrics such as Accuracy, Precision, Recall, and ROC-AUC were computed to evaluate the model. Visualizations such as the ROC curve were generated for better interpretability.
4. **Model Deployment Simulation**:
   - The trained model, scaler, and other preprocessing steps were saved and reloaded for making predictions on new data.

Learning Outcomes

This project demonstrated the importance of selecting the right algorithm for the problem at hand. Logistic Regression, being linear, is not well-suited for complex, nonlinear relationships as seen in disease prediction. However, it provided a valuable opportunity to:

- Understand feature selection techniques.
- Learn about multiclass classification using Logistic Regression.
- Handle practical challenges like scaling, feature alignment, and data imbalance.

Conclusion

While the model's performance was suboptimal, this project was a significant step in learning machine learning concepts and working with real-world data. Future improvements could include exploring non-linear algorithms such as Random Forests, Gradient Boosting Machines, or Neural Networks to better capture the complexity of disease prediction.