Title:

Personalized Disease Prediction and Health Recommendation System Using Patient Profiles and Symptoms

PRAJWAL SALUTAGI

11 October 2024

Table of Contents

Introduction

In recent years, the importance of health and wellness has garnered significant attention, particularly in the context of personalized healthcare. The prevalence of chronic diseases and lifestyle-related health issues necessitates the need for effective health recommendations tailored to individual patient profiles. Traditional methods of health advice often lack the specificity required to address unique patient needs, leading to suboptimal health outcomes.

Importance of Health Tips and Personalized Recommendations

Personalized health recommendations can empower individuals to make informed decisions about their health, thereby improving overall wellness and potentially reducing healthcare costs. By analyzing a patient's symptoms, demographics, and medical history, healthcare providers can offer targeted advice that aligns with individual health profiles. Research indicates that personalized interventions can lead to better adherence to treatment plans and improved health outcomes, as they consider a person's unique circumstances rather than a one-size-fits-all approach.

Objectives of the Assignment

The primary objective of this assignment is to explore the prediction of diseases using basic medical information through machine learning techniques. By utilizing classification algorithms, we aim to develop models that can effectively predict disease outcomes based on patient data. This analysis seeks to identify the most suitable model for this task, optimize its performance, and ultimately contribute to more personalized health recommendations.

Overview of the Methods Used

In this assignment, we will implement two machine learning algorithms:

K-Nearest Neighbors (K-NN) and Support Vector Machines (SVM). K-NN is a non-parametric method used for classification and regression that relies on the distance between data points to identify the closest neighbors, making it particularly effective for datasets with fewer samples on the other hand, SVM is a powerful classification technique that aims to find the optimal hyperplane separating different classes in high-dimensional spaces, which can be beneficial when dealing with complex decision boundaries. Both methods will be fine-tuned and evaluated based on their performance in predicting disease outcomes, with a focus on maximizing the F1 score to address the multi-class classification challenge effectively.

Data Exploration

Data Description

The dataset consists of 349 entries and 10 features, encompassing various symptoms and demographic information relevant to disease prediction. The features include:

- Disease: The target variable we aim to predict.
- Fever, Cough, Fatigue, Difficulty Breathing: Symptoms that may correlate with certain diseases.
- Age: The age of the patient (numerical).
- Gender: Patient gender, categorized as male or female.
- Blood Pressure: Categorized as low, normal, or high.
- Cholesterol Level: Categorized as low, normal, or high.
- Outcome Variable: Additional relevant health outcome data.

Upon initial examination, it is noted that most variables are categorical, with 'Age' being the sole numerical variable. This dataset will be subjected to various cleaning and preprocessing techniques to ensure accuracy and readiness for analysis.

Data Cleaning and Preprocessing

1. Filtering Diseases: After analyzing the 'Disease' column, we found a substantial number of unique diseases, many with only 1 to 5 samples, which are insufficient for reliable prediction. To enhance the robustness of our model, we filtered the dataset to retain only diseases with 10 or more samples, reducing the dataset's size to 83 entries with 10 features.
2. Handling Missing Values: A thorough examination of the dataset revealed no missing values, ensuring that the dataset is complete for analysis.
3. Removing Duplicates: We proceeded to eliminate any duplicate entries, resulting in a final dataset of 69 unique entries.
4. Encoding Categorical Variables: Categorical variables were transformed into numerical format using a defined mapping dictionary, converting the symptoms, gender, and other categorical data into integers for analytical compatibility.

Exploratory Data Analysis (EDA) Findings

- **Class Imbalance:** A pie chart analysis revealed significant class imbalance among diseases. Diseases such as Hypertension, Diabetes, and Migraine exhibited roughly 1.7 times fewer samples than Asthma, highlighting the necessity to address this imbalance for effective prediction modeling.
- **Univariate Analysis of Age: The analysis of the 'Age' variable revealed noteworthy trends:**

    - Individuals over 80 are more likely to experience strokes.
    - Hypertension and Osteoporosis increase with age, while Migraine and Hypertension are less prevalent in younger individuals (ages 20-30).
- **Significance of Other Variables:** A visual inspection of symptoms revealed that variables like cholesterol levels and fatigue exhibit significant variation across different diseases, indicating their potential as strong predictors in our model. For example, low blood pressure was correlated with a lack of stroke cases, further reinforcing its importance in disease prediction.

Insights Gained from the Data

The analysis highlighted that:

- Age is a Crucial Predictor: The distribution of diseases across age groups underscores the importance of age in predicting various conditions. However, caution must be taken due to limited samples in older age brackets.
- Impact of Symptoms: Symptoms such as fatigue, blood pressure, and cholesterol levels show substantial variation and significance across diseases, making them critical features in our predictive modeling efforts.
- Class Imbalance Must Be Addressed: The identified class imbalance necessitates strategies such as resampling or algorithmic adjustments to ensure model performance is not skewed towards more frequent classes.

Model Selection

- **Explanation of the Models Chosen**
- In this analysis, we have selected K-Nearest Neighbors (K-NN) and Support Vector Machines (SVM) as our machine learning algorithms. Both models are well-suited for classification tasks, particularly in the context of multi-class problems, which aligns with our objective of predicting various diseases based on a limited dataset.

Rationale for Selecting These Models

- **Suitability for the Task:**
- K-NN is a non-parametric algorithm that does not assume a specific distribution of the data. This flexibility makes it particularly useful for our dataset, which contains categorical variables and a limited number of samples.

- SVM, on the other hand, is effective in high-dimensional spaces and is robust against overfitting, especially in cases where the number of dimensions exceeds the number of samples. This characteristic is beneficial given our relatively small dataset.

Advantages:

- K-NN is intuitive and easy to implement, making it a good starting point for classification problems. It excels in situations where the decision boundary is irregular and can be visualized easily.
- SVM provides excellent performance in various classification tasks due to its ability to create complex decision boundaries. It also supports different kernel functions, allowing for adaptability to various types of data.

Brief Description of the Algorithms

K-Nearest Neighbors (K-NN):

- How It Works: For a given test instance, K-NN calculates the distance (typically using Euclidean distance) between the test instance and all other instances in the training dataset. It identifies the 'k' nearest neighbors and classifies the test instance based on the majority class among these neighbors. The choice of 'k' can significantly affect the model's performance, where smaller values may lead to noise sensitivity and larger values may oversmooth the decision boundary.

Support Vector Machines (SVM):

- How It Works: SVM transforms the feature space into a higher dimension using a kernel function (linear, polynomial, RBF, etc.) to find the optimal hyperplane that separates the classes. It focuses on maximizing the margin between the nearest data points of different classes (support vectors). The optimization problem is solved to find the best hyperplane, and predictions are made based on which side of the hyperplane the test instance lies.

In summary, both K-NN and SVM provide robust frameworks for tackling the multi-class disease prediction problem. Their strengths and adaptability to the dataset's characteristics will be thoroughly evaluated through model training and testing phases.

Fine-Tuning

Hyperparameter Tuning Process

The hyperparameter tuning process involved systematically adjusting the parameters of our chosen models (K-Nearest Neighbors and Support Vector Machines) to optimize their performance. For the SVM model, we focused on several key hyperparameters:

- **C**: This parameter controls the trade-off between achieving a low training error and a low testing error, which is crucial for managing overfitting.

- **Kernel**: The choice of kernel (linear, polynomial, radial basis function, etc.) influences the decision boundary and its flexibility.
- **Decision Function Shape**: This specifies how predictions are made in multi-class scenarios, particularly with One-vs-One (OVO) or One-vs-Rest (OVR) strategies.
- **Gamma**: This parameter defines the influence of a single training example, with low values meaning 'far' and high values meaning 'close.'
- **Shrinking**: This hyperparameter determines whether to use the shrinking heuristic, which can speed up the optimization process.

The tuning was performed using GridSearchCV, allowing us to explore various combinations of these parameters and evaluate their performance based on cross-validation.

Evaluation Metrics

For evaluation, we primarily focused on the **F1 score**, which provides a balance between precision and recall, particularly useful in scenarios with imbalanced class distributions. This metric is essential for understanding the model's effectiveness in correctly identifying positive cases while minimizing false positives.

Results of Tuning

After conducting hyperparameter tuning, the best score achieved was approximately **0.321**, with the following optimal parameters:

- **Best Params**:
  - 'svc__C': 1
  - 'svc__decision_function_shape': 'ovo'
  - 'svc__gamma': 'scale'
  - 'svc__kernel': 'rbf'
  - 'svc__shrinking': True

These results indicate that the combination of a radial basis function kernel and a value of **C** set to **1** significantly impacted the model's performance, highlighting the importance of these parameters in optimizing the SVM classifier.

We also analyzed the significance of other hyperparameters like the shrinking heuristic and the decision function shape. Interestingly, the presence or absence of the shrinking parameter did not affect the model's F1 score, suggesting that this hyperparameter may not be critical for our specific dataset. Similarly, using OVO versus OVR for the decision function shape produced no substantial difference in performance.

Challenges Faced

During the hyperparameter tuning phase, we encountered several challenges:

1. **Limited Data**: With only 69 samples available after filtering, achieving reliable results through cross-validation was challenging. The small sample size could lead to variability in model performance.
2. **Class Imbalance**: The imbalanced nature of our classes made it difficult to optimize the F1 score effectively, as some classes were significantly underrepresented.

3. **Computation Time**: Given the complexity of the SVM model and the number of hyperparameters, the tuning process required considerable computational resources and time.

Overall, while we successfully identified key hyperparameters influencing model performance, the limitations of our dataset necessitated cautious interpretation of the results. The next step is to apply the best model to the test data and analyze the predictions to understand the classification performance better.

Model Testing

Overview of Testing Methodology

The testing methodology for evaluating the models involved a combination of **cross-validation** and **train-test splitting**.

1. **Train-Test Split**: The dataset was divided into training and testing sets, with a typical split ratio of 80% for training and 20% for testing. This ensures that the model is trained on a substantial amount of data while retaining a separate set for unbiased evaluation.
2. **Cross-Validation**: To ensure robust model performance, cross-validation (typically k-fold cross-validation) was employed during hyperparameter tuning. This process involved dividing the training data into k subsets, training the model on k-1 subsets, and validating it on the remaining subset. This was repeated k times, allowing us to obtain average performance metrics and reduce overfitting.

Results from Testing Both Models

The best model achieved a test score of approximately **0.361**. The results were evaluated using multiple metrics:

- **Accuracy**: Measures the overall correctness of the model.
- **Precision**: Indicates the correctness of positive predictions.
- **Recall**: Measures the ability to identify all relevant instances.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.

Comparison of K-NN and SVM Performance

- **K-NN Performance**: K-Nearest Neighbors (K-NN) generally performed well with quick training times. However, its accuracy dropped significantly on the test set compared to the training set, indicating potential overfitting.
- **SVM Performance**: The Support Vector Machine (SVM) achieved a higher test score than K-NN, particularly excelling in predicting asthma cases. However, it struggled with other health conditions due to class imbalance.

While the SVM model demonstrated better overall performance, both models faced challenges in accurately predicting less frequent conditions, with SVM providing a more robust framework for differentiation among classes.

Discussion of Model Performance in Health Recommendations

The SVM model's performance, particularly its ability to predict asthma accurately, suggests its potential use as a secondary opinion tool in clinical settings. Despite the imbalance in the training dataset, the model can still aid healthcare providers in diagnosing asthma more effectively. However, the challenges in predicting other health conditions underscore the need for further enhancement.

To improve the model's generalization and performance across all classes, **data augmentation techniques** can be implemented. Methods such as:

- **Rotation and Scaling**: Modifying the existing data can create additional training examples.
- **Adding Noise**: Introducing slight variations to current data points can help the model learn more robust features.

These approaches may yield better results, allowing the model to predict less common conditions more accurately, thus enhancing its overall utility in health recommendations.

In conclusion, while the SVM model shows promise, especially in asthma detection, further work on data handling and model fine-tuning will be essential to achieve a more balanced performance across all targeted health conditions.

Conclusion

In conclusion, this project explored the prediction of diseases using basic medical information, focusing on developing machine learning models to enhance healthcare outcomes. The model achieved an F1 macro average score of **0.3611** across six disease classes, highlighting both its potential and the challenges inherent in disease prediction using limited data. This score indicates that while the model shows some capability, there is significant room for improvement in accurately predicting diseases based on basic medical information alone.

Key Findings

1. **Model Performance**: The model's performance was limited, particularly in predicting less frequent diseases, suggesting an imbalance in the training dataset.
2. **Algorithm Insights**: The comparison between K-NN and SVM revealed the strengths and weaknesses of each, with SVM performing better overall in this context, particularly with hyperparameter tuning.
3. **Data Challenges**: The accuracy of predictions was heavily influenced by the quality and diversity of the training data.

Evaluation of Model Effectiveness

The effectiveness of the current model is constrained by the available data and the feature set used. While it has shown some promise in identifying certain conditions, its predictive capabilities for other diseases are lacking. This indicates that simply relying on basic medical information may not be sufficient for accurate disease diagnosis.

Suggestions for Future Improvements

1. **Data Augmentation**: Implementing data augmentation techniques could help generate additional training samples, thereby improving the model's ability to learn from a more diverse dataset.
2. **Feature Engineering**: Exploring additional features beyond basic medical information, such as patient demographics, lifestyle factors, and historical health data, may enhance model accuracy.
3. **Alternative Algorithms**: Investigating other machine learning algorithms, such as ensemble methods (Random Forest, Gradient Boosting) or deep learning approaches, could yield better predictive performance.
4. **Hyperparameter Optimization**: Further fine-tuning of hyperparameters, especially using advanced techniques like grid search or randomized search, could help achieve better model performance.

Final Thoughts

Personalized health recommendations are crucial in enhancing patient care and improving health outcomes. By leveraging machine learning models that can accurately predict disease, healthcare providers can offer tailored recommendations, leading to proactive health management. This project demonstrates the importance of continuous improvement in predictive modeling and the need for collaboration between data scientists and healthcare professionals to achieve meaningful advancements in personalized healthcare solutions.

By addressing the challenges identified and implementing the suggested improvements, future work can lead to more reliable disease prediction models that contribute to better health outcomes for individuals.