Machine learning algorithm for detecting thyroid-related disorders to increase thyroid illness diagnosis accuracy

Advance Machine Learning / Spring 2024 / Prof. Ashok Kumar Patel

By Group – 11

Ajeet Singh – 48
Manas Vishal – 58
Rohith Kumar – 13

Final Project Proposal

Problem Statement:

Machine learning techniques have been increasingly used in the medical field to improve the accuracy of diagnoses and treatment plans. The use of an ML algorithm for detecting thyroid-related disorders to increase thyroid illness diagnosis accuracy. This approach involves using data-based decisions for diagnosing thyroid dysfunction, which is a classification problem that can be solved using Machine Learning techniques.

About the Datasets:

The dataset contains records of thyroid diagnoses from the Garvan Institute, collected between 1984 and early 1987. Each record consists of 29 attribute values, followed by a diagnosis code. The attributes include various medical conditions, patient characteristics, and laboratory measurements related to thyroid function. The diagnosis codes are strings of letters indicating diagnosed conditions, with "-" indicating no condition requiring comment. And a diagnosis of form "X|Y" indicates a diagnosis consistent with X but more likely Y.

The dataset is structured to facilitate the development of decision trees for classifying thyroid conditions into several categories, including hyperthyroid conditions, hypothyroid conditions, binding protein status, general health, replacement therapy, and discordant results. This classification is crucial for understanding the nature of thyroid diseases and for developing effective treatment plans.

Source:

Quinlan, Ross. (1987). Thyroid Disease.
UCI Machine Learning Repository.
https://doi.org/10.24432/C5D010.

Solution Approach:

Given the complexity and the nature of the thyroid disease dataset, which includes a mix of continuous and categorical variables, and considering the need for accurate prediction of multiple classes of thyroid conditions, a combination of machine learning algorithms would be beneficial. The choice of algorithms should be based on their ability to handle both types of data and their performance in classification tasks, especially in the context of imbalanced classes.

1. Random Forest (RF): Random Forest is a powerful ensemble learning method that can handle both continuous and categorical data. It is known for its ability to avoid overfitting and its robustness to outliers. The study mentioned 0.99 accuracy in predicting ten thyroid diseases using Random Forest, indicating its effectiveness in this context.
2. Gradient Boosting Machine (GBM): GBM is another ensemble method that builds a series of weak learners (typically decision trees) to form a strong predictive model. It is particularly effective in handling imbalanced datasets and can capture complex patterns in the data.
3. Support Vector Machine (SVM): SVMs are effective in high-dimensional spaces and are versatile as different Kernel functions can be specified for the decision function. They are particularly useful when the dataset is not linearly separable.
4. AdaBoost: AdaBoost is an ensemble method that combines multiple weak learners to create a strong learner. It is effective in handling noisy data and can improve the accuracy of the model.
5. Deep Learning Models: For a more complex and nuanced understanding of the data, deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks can be used. These models can capture complex patterns and dependencies in the data, which might be beneficial for predicting thyroid conditions.

It would be beneficial to experiment with these algorithms, possibly in combination, and use cross-validation to evaluate their performance. The choice of the best algorithm or combination of algorithms would depend on the specific characteristics of the dataset and the performance metrics (accuracy, precision, recall, etc.) that are most relevant to the problem at hand.