MDPI

Article

Thyroid Disease Prediction using Statistical Machine Learning

BATTULA HARI KRISHNA - 2203A52006

[1] **Affiliation; 2203A52006@sru.edu.in**
* **Section - AIML AA**
† **SR UNIVERSITY**

Abstract: Thyroid disease prediction using statistical machine learning uses complex Machine learning algorithms to predict whether a person(he/her) have a thyroid disease. The complex, Machine Learning models i.e., SVM (Support Vector Machine) for classification, logistic regression, Perceptron Learning, and KNN (K nearest Neighbors), from blood tests to analyze the hormones levels responsible for thyroid disease. The above classification models perform different operations for predict the binary output. With this we can differentiate the diseased people and non-diseased people by utilizing their output with test results, out of all the models we choose the most effective model with highest accuracy to help the hospitals.

Keywords: Classification Models; Logistic Regression; Support Vector Machine; Perceptron learning; K-Nearest Neighbors;

1. Introduction

Thyroid is an important gland in the human body which releases thyroid hormones which are very vital in human body. If the gland fails to produce the hormones it causes imbalance in metabolism and effects the persons health.

In Thyroid disease is classified into two types one is where the gland doesn't produce necessary hormones levels where it is termed as hypothyroidism and the case where it produces more than enough which is termed as hyperthyroidism. The effects of hypothyroidism are fatigue, Weight gain, and dry skin and hair. Apart from this the effects hyperthyroidism are weight loss, muscle weakness etc. Even though it differs from person to person. Depending of the severeness Of the problem recommends the patients to take medicines, surgery, and Radioactive Iodine therapy.

Next, we discuss about how machine learning is used in Thyroid prediction, how the machine learning algorithms is efficient in predicting the output whether the person is diagnosed with thyroid or not. These advanced machine learning algorithms uses the data like gender, age, and the hormone levels. All these algorithms or models predicts whether the person is with thyroid or not and helps the doctors for early diagnose of the disease.

2. Literature Review

2.1. previous case studies

Here are few literature review on thyroid disease [1] [2] [3] [4] [5]

2.2. Challenges and Research Gaps

The probability of predicting the thyroid disease from the datasets collected from the websites Kaggle makes a tedious work like handling the complex the relationships between features, and selecting the appropriate model. Focusing on the difficulties make us choosing correct model for improved accuracy.

Data and Methodology

2.3. Data Description

Binary Class (positive or negative) This is the target column containing tags for the features.

1. In our thyroid disease prediction project, we're using 27 features, including age, sex, medical ion history, and blood test results goal is to classify patients as 0 (no thyroid issue) or 1 (thyroid issue).
2. These features provide essential information, such as thyroid hormone levels and patient history.
3. We'll employ various statistical models to make accurate predictions and assist in early thyroid problem detection.
4. By analyzing the dataset, we use appropriate model for predicting the disease, for improved health care.
5. The 'age' feature provides the data that which age group persons are more characterized by thyroid disease.
6. TSH measured is a critical thyroid function indicator.
7. TT4 represents is essential in diagnosing thyroid problems.

2.4. Data Analysis

The below given histogram shows the age vs binary class, helps understand how individual features contribute to classification problem. Features with distinct non-overlapping distributions for the given classes provides more informative for classification. By analyzing these type of histograms, we can make decisions about features selection, model choice, to improve the performance of the classification model.

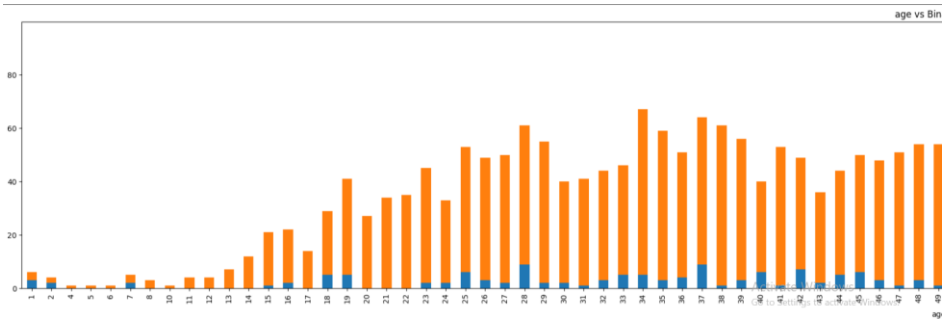


Figure 1. Age vs Binary Class

2.5. Data Preprocessing

Data preprocessing is a vital step in machine learning involves cleaning is a critical step in machine learning that involves cleaning, transforming, and placing raw data into a way suitable for model training. It plays an important role in ensuring a data of high quality and the machine learning model can learn insights from the data.

Below are the detailed steps in data preprocessing:

1. Data Cleaning: Data cleaning is an important part for data splitting, in which the string data is converted into numerical, and finding the Nan values in the data set and replacing with appropriate value i.e., mean, median, mode.
2. Data Transformation: Data Transformation is done with Normalization, or min max scaling, if not the high values in dataset effects the output, and the model can't predict it accurately.
3. Data Splitting: Train-Validation-Test Split is a method in which the dataset is divided into training, validation, and test sets. The training data is used for training the model, and test set for testing of the model for improved model performance.

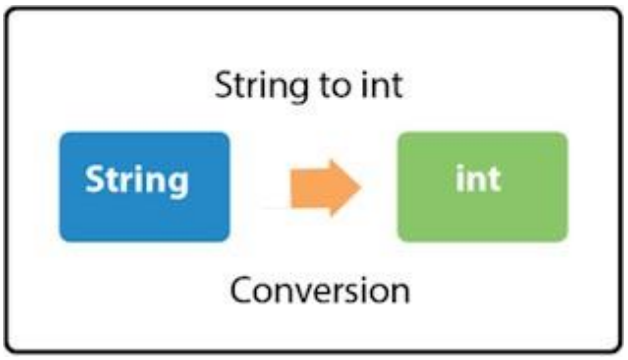


Figure 2. string to float conversion

4. Results

4.1. Logistic Regression

Accuracy: 94.17989417989418

Logistic regression is a popular statistical model used as binary classification problems.

In this method we may have one or more features for predicting the binary output.

In logistic regression the input is a linear function combines values between 0 and 1 representing the output as one binary class.

The logistic model is trained using likelihood function and parameters are estimated.

Unlike the linear regression here there are no continuous values, making perfect for binary classification problems

Logistic regression finds it applications in health care, business and etc.

The logistic regression model's efficiency is assessed by accuracy and precision.

This analysis helps us to identify the persons with thyroid (binary class 1) and persons without thyroid as (0 binary class)

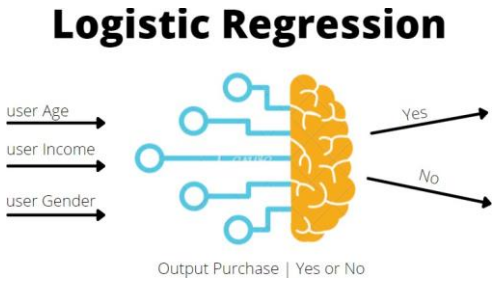


Figure 3. logistic*regression*

Results Section: Logistic Regression Model Performance

1.Accuracy: An accuracy score of 94.18 indicates that the model correctly predicted thyroid outcomes in nearly 94 out of every 100 cases.

This high accuracy shows that it is very efficient for predicting the persons with thyroid and persons without thyroid.

It considers many features like age, gender, and thyroid hormone levels.

It shows that the model is very well suited for thyroid disease prediction. Even though considering the performance metrics like f1-score, recall makes the model more effective

4.2. Support Vector Machine

To perform more advanced classification tasks, we use support vector machine (SVM) model for achieving better accuracy.

In a high-dimensional space, SVM find the better hyperplane that best separates class of data I.e., features.

The "support vectors" are the data points closest to the decision boundary which defines a maximized margin, the distance between support vector and data points.

SVM is used in extensively when dealing with non-linearly separable data.

It finds the best hyperplane the maximize the distance between the data points.

Accuracy: 94.44444444444444

The Support Vector Machine for classification finds that it has the best performance metrics among the logistic and SVM i.e., 94.5%

Accuracy: The Support Vector Machine (SVM) model having the accuracy 94.5% tells that it predicts the cases 95 out of 100 thyroid disease cases.

The confusion matrix, reveals that the model correctly identified 343 individuals without thyroid problems and 14 individuals with thyroid problems. However, it also produced 19 false negatives, incorrectly classifying 19 individuals with thyroid problems

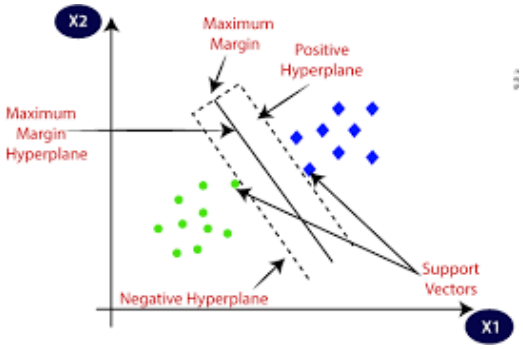


Figure 4. Support Vector Machine

as non-problems, and 2 false positives, mistaking two individuals without thyroid problems as having issues.

4.3. Perceptron Learning

Machine Learning foundations are made by perceptron learning.

It is used for binary Classifiers models.

The best example of single layer neural network is perceptron learning.

The first step in perceptron learning is after taking inputs is assigning weights to the input features and computes their sum and compares if the sum is above certain thrust hold it falls under one binary output if nit it falls under another binary output.

The process continues until convergence is achieved, where the model correctly classifies all training examples or a predefined number of iterations is reached.

The Perceptron learning model achieved an accuracy of 93.92, demonstrating its capability to make accurate predictions in thyroid diagnosis. ‘

The confusion matrix, revealing that the model correctly identified 344 individuals without thyroid problems and 11 individuals with thyroid problems.

However, it produced 22 false negatives, mistakenly categorizing 22 individuals with thyroid issues as non-problem cases, while generating only 1 false positive, incorrectly identifying a single individual without thyroid issues as having problems.

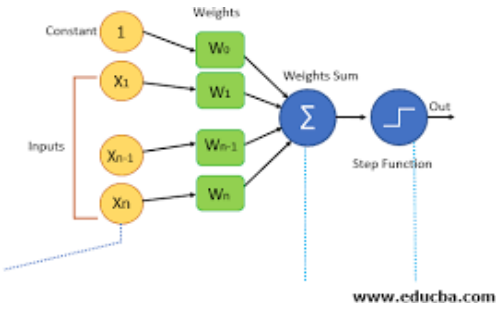


Figure 5. Perceptron Learning

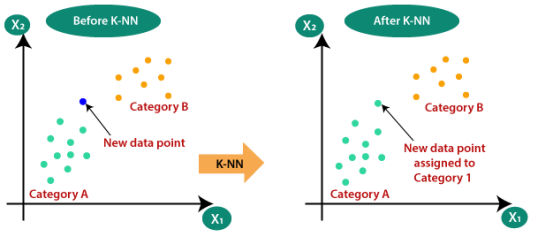
4.4. KNearestNeighbours

The simple machine learning algorithm used for both classification and regression problems are KNN.

It basically operates from the concept that similar data points belong to similar classes.

In KNN, predictions are made based on the majority class

KNN functions by calculating the distance between points i.e.,



between the selected point and all the remaining points in the training set.

The most common distance metric used is Euclidean distance, but other metrics can be applied based on the nature of the data.

K is a positive integer which says the number of nearest neighbors to compare.

Less K value is not suitable for complex or large dataset.

Accuracy:88.41

The k-Nearest Neighbors (KNN) classification model, with an accuracy of 88.41, tells us the model is a decent for predicting the output but not the optimal one as it have accuracy lesser than logistic, SVM, and perceptron models.

In KNN, when predicting the output it compares the K Nearest Neighbors and predicts this class output accordingly.

The model's effectiveness largely depends on selecting the right value of "k" and the distance metric used.

4.5. Bootstrap

Bootstrapping is a statistical machine learning model which continuously shuffles the data set. To create multiple new data sets, each of the same size as the old data set.

The main motto of boot strapping is to re-sample the data set of same size as the original data set. This uses the K nearest samples and predicts the output.

The accuracy majorly depends on k value if it is less the data set is very sensitive. Bootstrapping is less effective for large data sets.

4.5.1. Logistic Regression

Boot Strapping for logistic regression model helps in re sampling the data set to improve the performance of the logistic model. Parameter estimation is also very important part. This uses the K nearest samples and predicts the output. The accuracy majorly depends on k value if it is less the data set is very sensitive. Bootstrapping is less effective for large data sets.

4.5.2. Support Vector Machine

Boot strapping for Support vector Machine is best the way it finds the hyperplane with maximum margin. Also, it evaluates the performance metrics like f1 score, precision, recall, accuracy for re sampled data set and evaluates the performance and stability of the model.

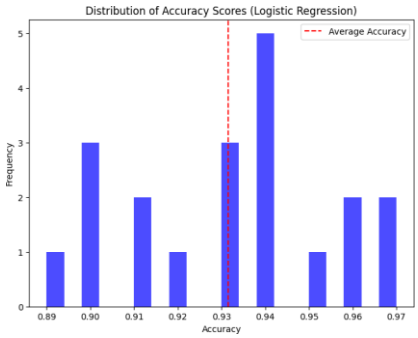


Figure 7. Bootstrap accuracy vs iterations

Understanding the stability and variability of the model's predictions through the bootstrap method.

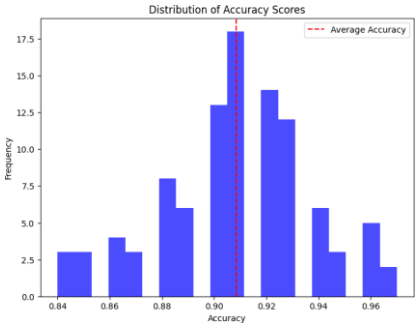


Figure 8. Bootstrap accuracy vs iterations

4.5.3. Perceptron learning

The unique part about boot strapping lies in the way how the training data impact the performance of the perceptron model. However perceptron learning performance is limited on bigger and complex data sets this is the limitation of perceptron learning.

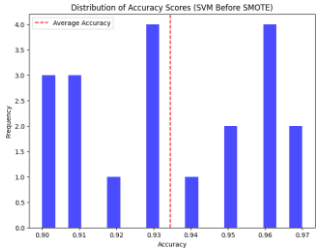


Figure 9. Bootstrap accuracy vs iteration

4.5.4. KNNeighbors

Boot strapping for K nearest neighbors referred as KNN is unique in its non parametric approach. It is a good algorithm for classification models. The performance of KNN lies in the parameter selection and the distance metric(k)..

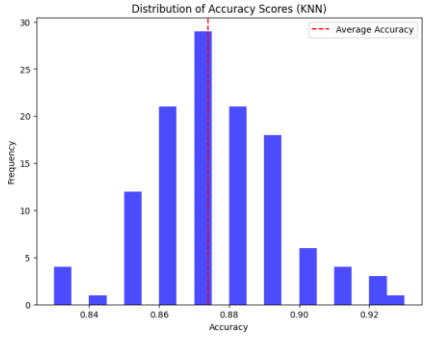


Figure 10. Bootstrap accuracy vs iterations

MODEL USED	ACCURACY
1)Logistic Regression	94%
2)Perceptron Learning	92%
3)Support Vector Machine	96%
4) KNN	84%

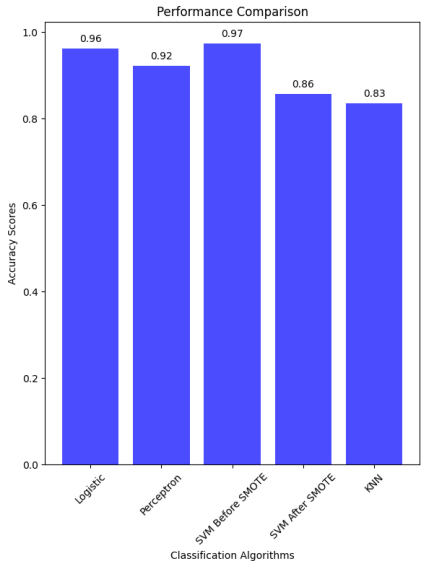


Figure 11. Accuracy vs Classification Models

5. Conclusion

Firstly, Logistic Regression model gets a accuracy of 96 percent, which shows it currently predicted 96 correctly out of 100 cases. Second Perceptron Learning achieves a 93 percent accuracy still demonstrates great predictive capability. Next Support Vector stands out with 97 percent accuracy for the data set, shows promising model for predicting the output. At last K-Nearest-Neighbours showing 83 percent accuracy tells that it is not suitable for predicting the output.

5.1. Summary

To conclude, Support Vector Machine have great accuracy for the data set mentions
that it is better Classification model for Thyroid disease prediction. Capstone project link [9]

6. References

1.case study 1: cross ref

2.Case study 2: cross ref

3.Case study 3: cross ref

4.Case study 4: cross ref

5.Case study 5: cross ref

6.Case study 6: cross ref

7.Case study 7: cross ref

8.data set : cross ref

9.Capstone project link: cross ref