Detailed Project report

Thyroid Disease Detection

**Introduction**

Thyroid disease a very common problem in India especially in female. Thyroid disorder can speed up or slow down the metabolism of the body. Hyperthyroidism and hypothyroidism are the most two common diseases caused by irregular functioning of thyroid gland.

Machine learning plays an important role in predicting the thyroid disease. Machine learning algorithms helps in early detection of the disease and to improve the quality of the life. Different classification algorithms can forecasts the presence of the disease such as Logistic regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine, XG Boost, KNN.

In the present work, the machine learning algorithms like logistic algorithm, Decision Trees, Random Forest, Support Vector Machine and XG Boost algorithm are used for classification of thyroid disease. Based on the obtained result Decision Trees algorithm is used to predict the thyroid disease. Web app is created to get data from users to predict the type of disease using flask. The user input and result is stored in Cassandra cloud database and the finally model is deployed on AWS cloud platform

**Objective**

The main goal of this project is to predict the risk of hyperthyroid and hypothyroid based on various factors of individuals. This project will play a decisive role in order to early detection, accurate identification of the disease and helps the doctors to make proper decisions and better treatment.

**Benefits**

* Detection of thyroid disease
* Eliminates human error
* Higher efficiency

**Architecture**

**Start**

**Data from UCI repository**

**Data pre-processing**

**Model building**

**Data from user**

**Application start**

**Pushing app to cloud**

**Cloud setup**

**Prediction**

**Save the user input and output at database**

**Display result**

**End**

**Work flow**

This is a workflow diagram for the thyroid detection

Data Collection

The dataset is collected from the UCI repository page. In this project patient’s details like age, sex, TSH, T3, FTI, TT4 and T4U levels are used to predict the result as shown below.

Age

Sex

FTI

TT4

T4U

T3

TSH

Data Exploration

Data exploration refers to the initial step in data analysis in which we use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

Data Pre-processing

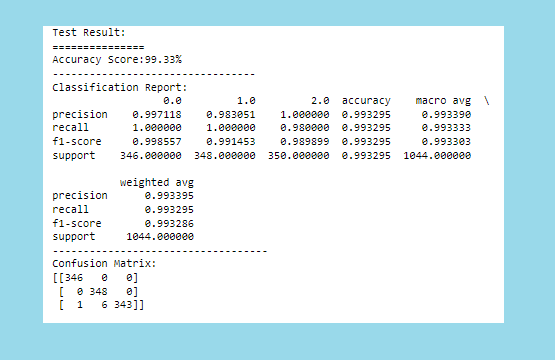
In this process, raw data is transformed into valid form for the machine learning models. This includes, data cleaning, imputing missing values using KNN imputers, removing outliers, drop unnecessary columns, imbalanced data set handling using random over sampler, encoding categorical columns etc.

Data Modelling

Various classification algorithms like Logistic Regression, Decision Tree, Support Vector Machine and XG Boost are used. Hyper parameter tuning was also performed. Decision Tree and XG Boost algorithms were given better results. Decision tree model is finalized to predict the result

Model evaluation

Model performance evaluated based on accuracy, confusion matrix, classification report. Multiple models were trained and evaluated based on accuracy score and recall. The best performing models were selected as final model. In our case, it is Decision tree and evaluation for the same is shown below



API Details or User Interface

Web app is created to get data from users to predict the type of disease using flask.

Database operation

The data from the user and corresponding result are inserted into the database. The database used is Cassandra

Deployment

We will be deploying the model to AWS.

**Q & A**

1) What’s the source of the data?

* The entire data is collected from the UCI machine learning repository.

2) What was the type of data?

* The data was a combination of numerical, categorical and null values.

3) How logs are managed?

* Once the code is executed a log file will be created in the same repository.

4) What techniques were used for data processing?

* Removing unwanted features/columns
* Visualizing relation of independent variable with each other through heat maps.
* Removing outliers
* Imputing null/missing values
* Encoding the categorical features
* Fixing the imbalanced data

5) How training was done and what models were used?

* After data processing, the training data was fitted into 4 different models. These included Logistic Regression, Support vector machine, XG Boost and Decision Tree.
* Finally based on performance metrics Decision tree was selected as final model

6) How prediction was done?

* Once the models were tuned, they were saved as a pickle file. The user input is given to pickle file and predicted is performed

7) What are the different stages of deployment?

* AWS account was created
* Flask was used to integrate the models and the simple HTML frontend.
* Requirement file was created which contained all the necessary libraries needed for the code to execute.
* Pushed to GitHub and then to AWS