**Detailed Project Report**

**Thyroid Disease Prediction**

|  |  |
| --- | --- |
| Written By | Kollipati Vamshi Krishna |
| Version | 1.0 |
| Date | 25-06-2022 |

**Document Change Control Record**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Reviews

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Reviewer** | **Comments** |
|  |  |  |  |

### Approval Status:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Review date** | **Reviewed By** | **Approved By** | **Comments** |
|  |  |  |  |  |

# **Introduction**

Thyroid disease a very common problem in India, more than one crore people are suffering with the disease every year. Especially it is more common in female. Hyperthyroidism and hypothyroidism are the most two common diseases caused by irregular function of thyroid gland. Thyroid disorder can speed up or slow down the metabolism of the body. In the world of rising new technology and innovation, healthcare industry is advancing with the role of Artificial Intelligence. Machine learning algorithms can help to early detection of the disease and to improve the quality of the life. This study demonstrates the how different classification algorithms can forecasts the presence of the disease. Different classification algorithms such as Logistic regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine have been tested and compared to predict the better outcome of the model.

1. **Objective**

The main goal of this project is to predict the risk of hyperthyroid and hypothyroid based on various factors of individuals. Thyroid disease is a common cause of medical diagnosis and prediction, with an onset that is difficult to forecast in medical research. It 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

**3. Architecture**

Following workflow was followed during the entire project.

**Dataset**

**THYROID**

**DISEASE**

**DETECTION**

SEX

PSYCHOLOGICAL SYMPTOMS

AGE

ON THYROXINE MEDICATION

GOITRE

TOTAL THYROXINE(TT4)

TSH LEVEL

ON ANTI

-

THYROID MEDICATION

FREE THYROXINE INDEX

HYPOPITUITARY

**Data Analysis Steps**

DATA

COLLECTION

DATA

PREPROCESSING

EXPLORATORY DATA

ANALYSIS

FEATURE

SELECTION

MODEL

CREATION AND

EVALUATION

In step 1, we collect

data which is generally

present in a database

or on internet.

In step 2, we

preprocess the data

which involves data

cleaning by handling

outliers, null values

etc.

In step 3, we explore

the data by performing

univariate and

bivariate analysis on

the features.

In step 4, we use

feature selection

techniques to filter out

the most important

features to perform

model creation

In step 5, we finally

build models on our

dataset and choose

the model which gives

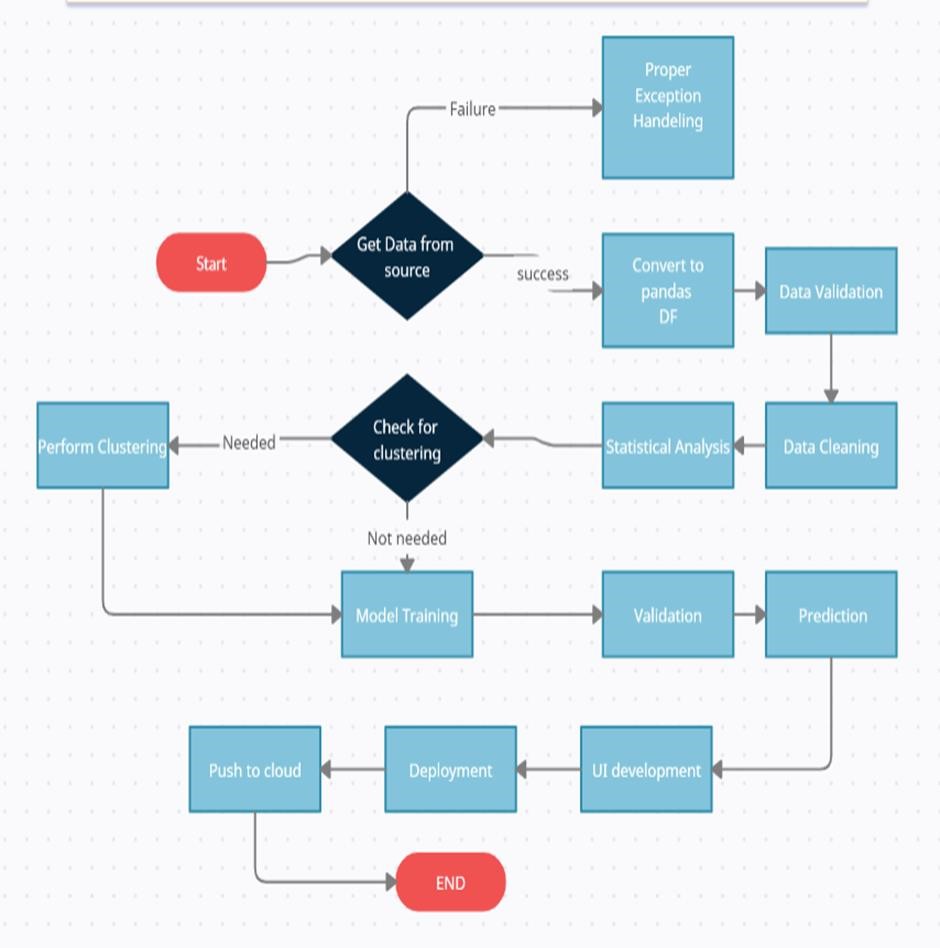
the best accuracy.

**Random Forest Model**

**INTRODUCTION:**

* The random forest classifier is a supervised learning algorithm which we can use for regression and classification problems. It is among the most popular machine learning algorithms due to its high flexibility and ease of implementation.
* It is called Random Forest because it consists of multiple decision trees just as a forest has many trees. On top of that, it uses randomness to enhance its accuracy and combat overfitting, which can be a huge issue for such a sophisticated algorithm. These algorithms make decision trees based on a random selection of data samples and get predictions from every tree. After that, they select the best viable solution through votes.
* Random Forest Classifier being ensembled algorithm tends to give more accurate result. This is because it works on the principle i.e., number of weak estimators when combined forms strong estimator. Even if one or few decision trees are prone to noise, overall results would tend to be correct. Even with small number of estimators (=30), it gives us high accuracy as 97%.

**MODEL TRAINING AND VALIDATION WORKFLOW**



**MODEL TRAINING AND VALIDATION WORKFLOW**

**Data Collection**

* Thyroid Disease Data Set from UCI Machine Learning Repository
* For Data Set: <https://archive.ics.uci.edu/ml/datasets/thyroid+disease>

**Data Pre-processing**

* Missing values handling by Simple imputation (median strategy)
* Outliers' detection and removal by boxplot and percentile methods
* Categorical features handling by ordinal encoding and label encoding
* Feature scaling done by Standard Scalar method
* Imbalanced dataset handled by SMOTE - Over sampling
* Feature selection done by forward feature selection

**MODEL TRAINING AND VALIDATION WORKFLOW**

**Model Creation and Evaluation**

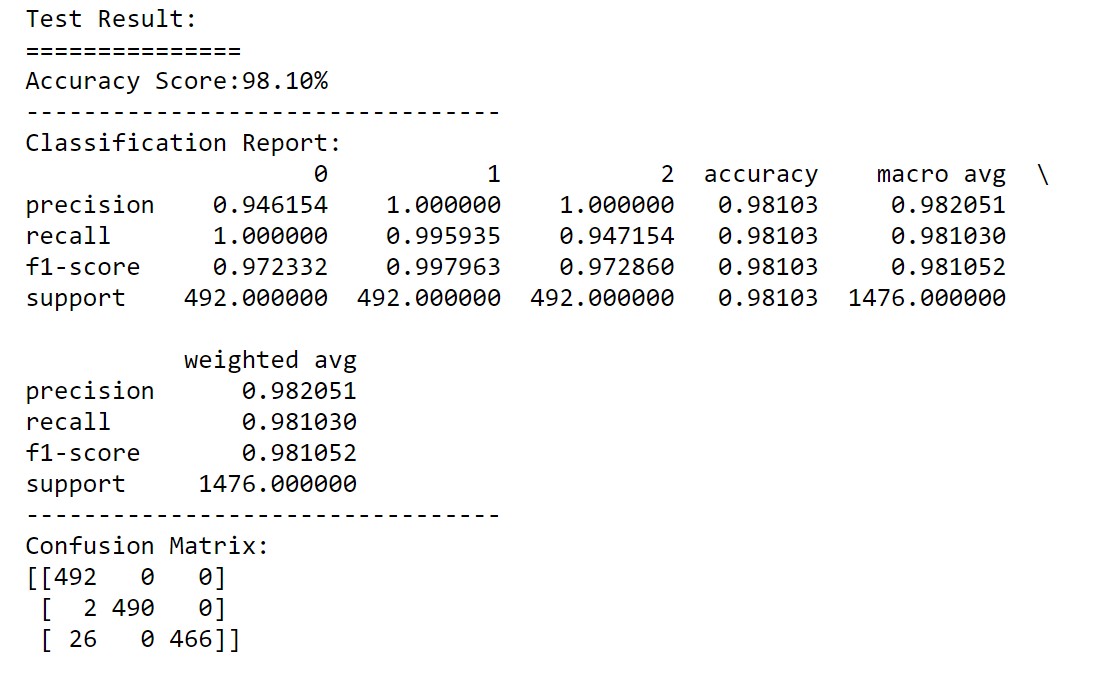
❑ Various classification algorithms like Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine tested.

❑ Random Forest, Decision Tree and Logistic regression were given better results. Random Forest was chosen for the final model training and testing.

❑ Hyper parameter tuning was performed.

❑ Model performance evaluated based on accuracy, confusion matrix, classification report.

**Model Prediction Results**

**on Test Dataset**

**Model Deployment**

**Model Deployment**

*  The final model is deployed on Heroku using Flask framework.

**Frequently Asked Questions**

Q1) What is the source of data?

The data for training is obtained from famous machine learning repository. UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/thyroid+disease>

Q2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q3) What’s the complete flow you followed in this Project?

Refer slide 7th, 8th and 9th for better understanding.

Q4) After the File validation what you do with incompatible file or files which didn’t pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q5) How logs are managed?

We are using different logs as per the steps that we follow in training and prediction like model training log and prediction log etc. And then sub log are inside those folder

Q 6) What techniques were you using for data pre-processing?

• Removing unwanted attributes

• Visualizing relation of independent variables with each other and output variables

• Checking and changing Distribution of continuous values

• Removing outliers

• Cleaning data and imputing if null values are present.

• Converting categorical data into numeric values.

Q 7) How training was done or what models were used?

• First Data validation done on raw data.

• Then Data preprocessing done on final CSV file.

• We did clustering over the data to divide it on desired cluster based on elbow method.

• Various model such as Decision Tree, Random Forest and XGBoost models are trained on all clusters and based on performance, for each cluster different model is saved.

Q 8) How Prediction was done?

• The testing files are shared by the client. We Perform the same life cycle till the data is clustered.

• Then on the basis of cluster number model is loaded and perform prediction. In the end we get the accumulated data of predictions.

Q 9) What are the different stages of deployment?

• After model training and finalizing all models. We created required files for deployment.

• Finally deployed our model on Heroku.

Q 10) How is the User Interface present for this project?

• Created form for user input prediction.

• UI is very user friendly and easy to use

**Thank You**