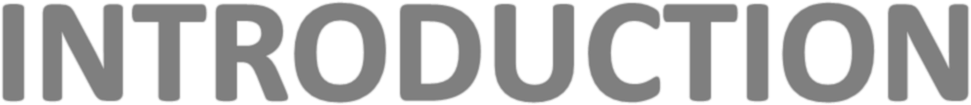
THYROID DISEASE DETECTION

Detailed Project Report

By- Vinayak Verma



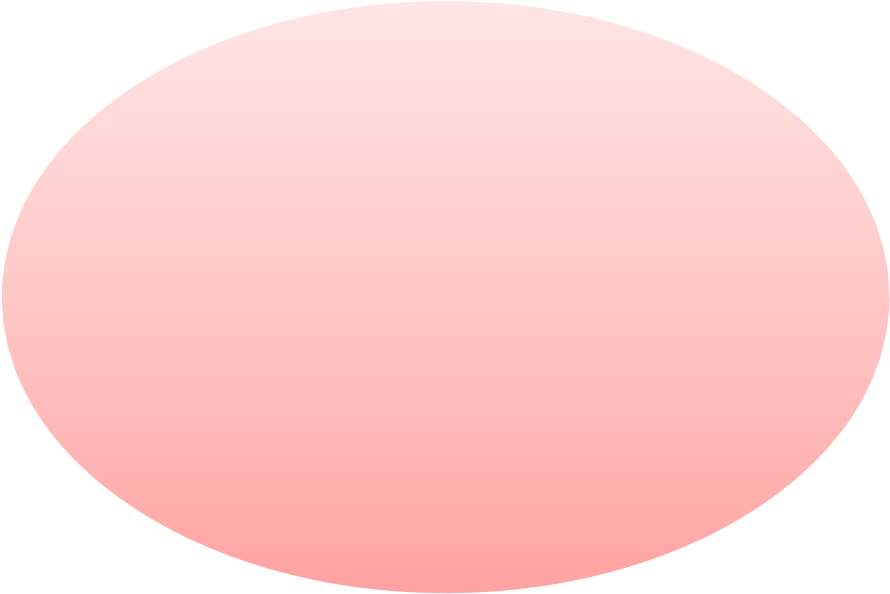
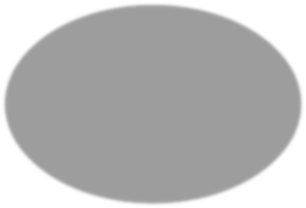
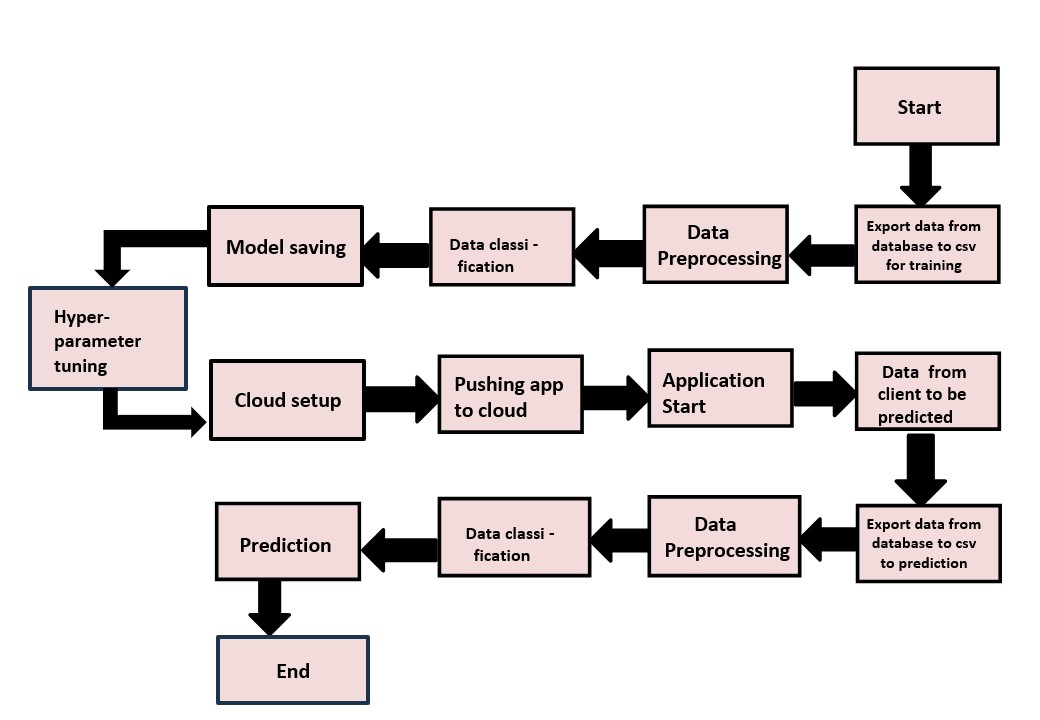
In India, thyroid disease affects approximately one out of every ten individuals, with women between the ages of 17 and 54 being the most commonly affected. Severe thyroid disorders can result in cardiovascular issues, increased blood pressure, high cholesterol levels, depression, and decreased fertility. The thyroid gland produces two essential hormones, total serum thyroxin (T4) and total serum triiodothyronine (T3), which are crucial for regulating metabolism, energy levels, body temperature, and protein synthesis, thereby ensuring proper cell and organ function.

Thyroid disease primarily manifests as either Hyperthyroidism or Hypothyroidism, conditions where the thyroid gland's dysfunction either speeds up or slows down the body's metabolism. With the advancement of technology, the healthcare sector is increasingly utilizing Artificial Intelligence to enhance patient care. Machine learning algorithms, in particular, hold significant potential for early disease detection and improving overall quality of life.

This study explores the use of classification algorithms, specifically Random Forest, to predict the presence of thyroid disease. By evaluating these algorithms, the research aims to identify the most effective model for early disease detection, thereby contributing to improved healthcare outcomes.



The main goal of this project is to accurately predict the risk of hyperthyroidism and hypothyroidism in individuals by considering various factors. Thyroid disease is a common and difficult medical condition to predict in research. Our efforts in this study are vital as they aim to facilitate early detection and precise identification of the disease, leading to informed decisions and better treatment outcomes for patients. By leveraging predictive modeling, we strive to provide doctors with valuable insights, enhancing their ability to deliver timely and effective healthcare to those suffering from thyroid disorders.



Age

Sex

TSH

Level

Total

Thyroxine(TT4)

Free

Thyroxine

Index

Thyroxine

Medication

AntiThyroid

Medication

Goitre

Test

Hypopituitary

Psychological

Symptoms

T3

Level

Measure

T4

Level

Measure

Iodine

Test

Lithium

Test

Tumour

Test

TBG

Blood

Test

**Thyroid**

**Disease**

**Detection**



Data

Collection

Data

P

r

ep

r

ocess

ing

Exploratory

D

a

t

a

A

nal

y

s

is

Feature

S

e

lect

i

on

Model

C

r

e

a

ti

o

n

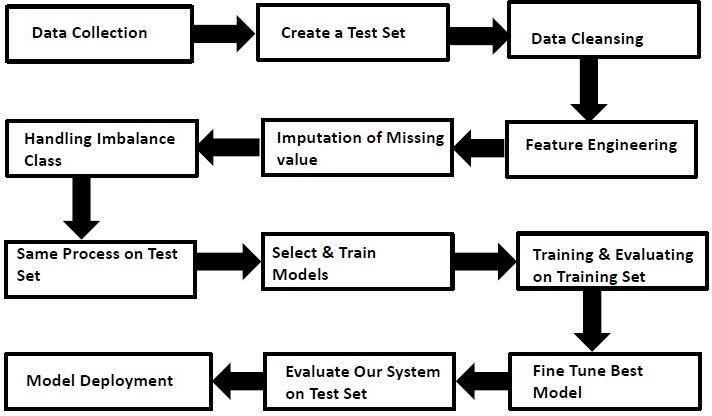
&

Evaluation

**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](https://archive.ics.uci.edu/ml/datasets/thyroid%2Bdisease)

Data Pre-Processing

* Categorical features handling by ordinal encoding and label encoding
* Feature scaling done by Standard Scalar method
* Imbalanced dataset handled by Random Over sampling
* Drop unnecessary columns

# MODEL TRAINING AND VALIDATION WORKFLOW

Model Creation and Evaluation

* 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.

# Random Forest Classifier Model INTRODUCTION

The random forest model is an ensemble learning method frequently used in machine learning for classification and regression. It works by creating multiple decision trees during the training phase and merging their predictions for the final result. Each tree is trained using a random subset of the data, and at each node, a random subset of features is selected. This randomness promotes diversity among the trees and helps to mitigate overfitting.

For prediction, the random forest compiles the outputs from individual trees, employing majority voting for classification tasks or averaging for regression tasks. This ensemble method improves the model’s accuracy, stability, and capacity to manage large datasets with high-dimensional feature spaces.

The random forest is particularly suited for thyroid disease prediction due to its effectiveness in handling high-dimensional medical data while minimizing the risk of overfitting. Its ensemble approach boosts accuracy and robustness, which is essential for reliable diagnosis. Additionally, the model's feature importance analysis helps identify significant indicators, offering valuable insights for medical professionals. Its capability to manage missing data and noisy features makes the random forest a powerful and adaptable algorithm, contributing to enhanced and efficient thyroid disease detection.

Reasons to Use Random Forest Classifier Model:

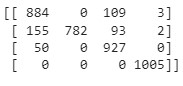
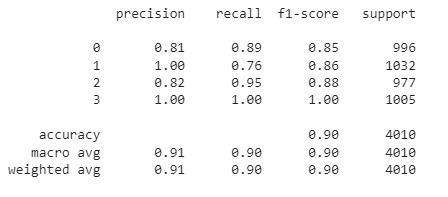
- High execution speed.

- Improved model performance.

**MODEL PREDICTION RESULTS**

## ON TEST DATASET

**ClassificationReport ConfusionMatrix**



**Accuracy**

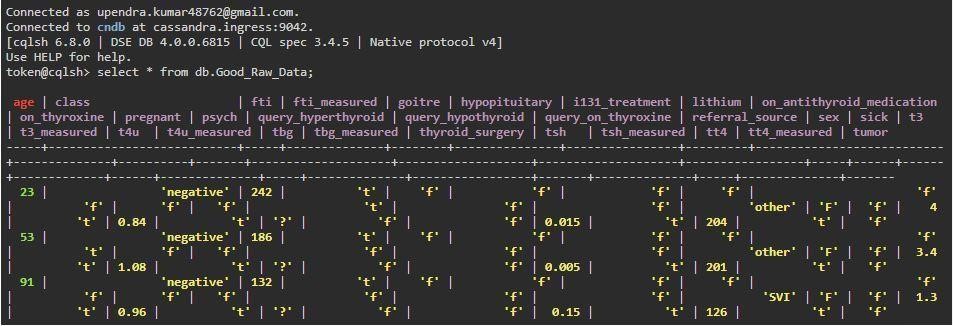


**DATABASE CONNECTION &**

**DEPLOYMENT**

**Database Connection**

* Cassandra Database used for this project.



**Model Deployment**

* The final model is deployed on Streamlit .