THYROID DISEASE DETECTION

Detailed Project Report

By- Arkoprovo Ghosh

INTRODUCTION

In India, thyroid disease affects at least one out of ten individuals. This condition predominantly occurs in women aged 17 to 54. In severe cases, thyroid disorders can lead to cardiovascular complications, increased blood pressure, elevated cholesterol levels, depression, and reduced fertility. The thyroid gland produces two crucial hormones, total serum thyroxin (T4) and total serum triiodothyronine (), which play a vital role in regulating the body's metabolism, energy levels, temperature, and protein synthesis for proper cell and organ functioning.

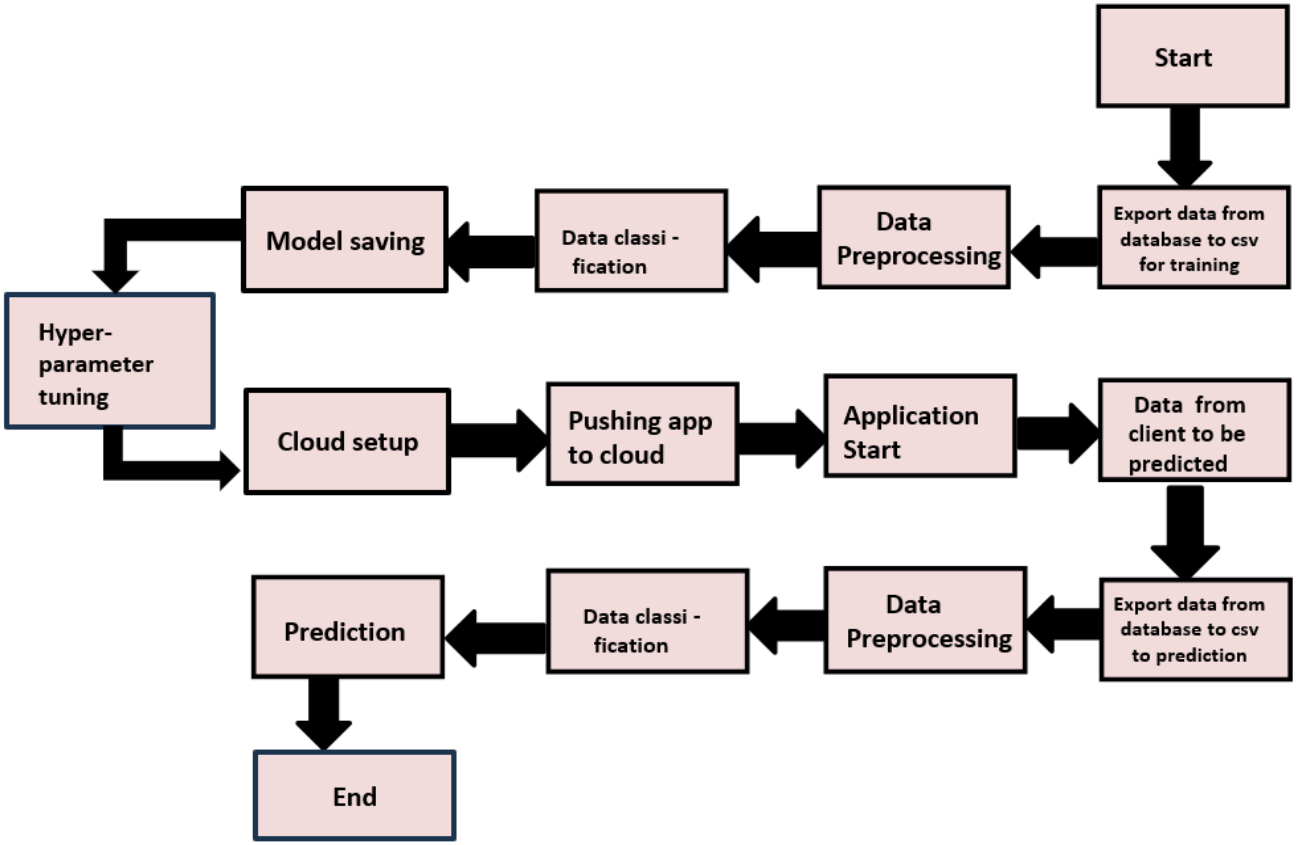
Thyroid disease manifests mainly as Hyperthyroidism and Hypothyroidism, where the thyroid gland's irregular function either accelerates or decelerates the body's metabolism. In the modern era of technological advancements, the healthcare industry is leveraging Artificial Intelligence to improve patient care. Machine learning algorithms offer potential benefits for early disease detection and enhancing the overall quality of life.

This study focuses on employing classification algorithms-Random Forest to predict the presence of thyroid disease. By comparing these algorithms, the research aims to identify the most effective model for forecasting the disease, aiding in earlier detection and better healthcare outcomes.

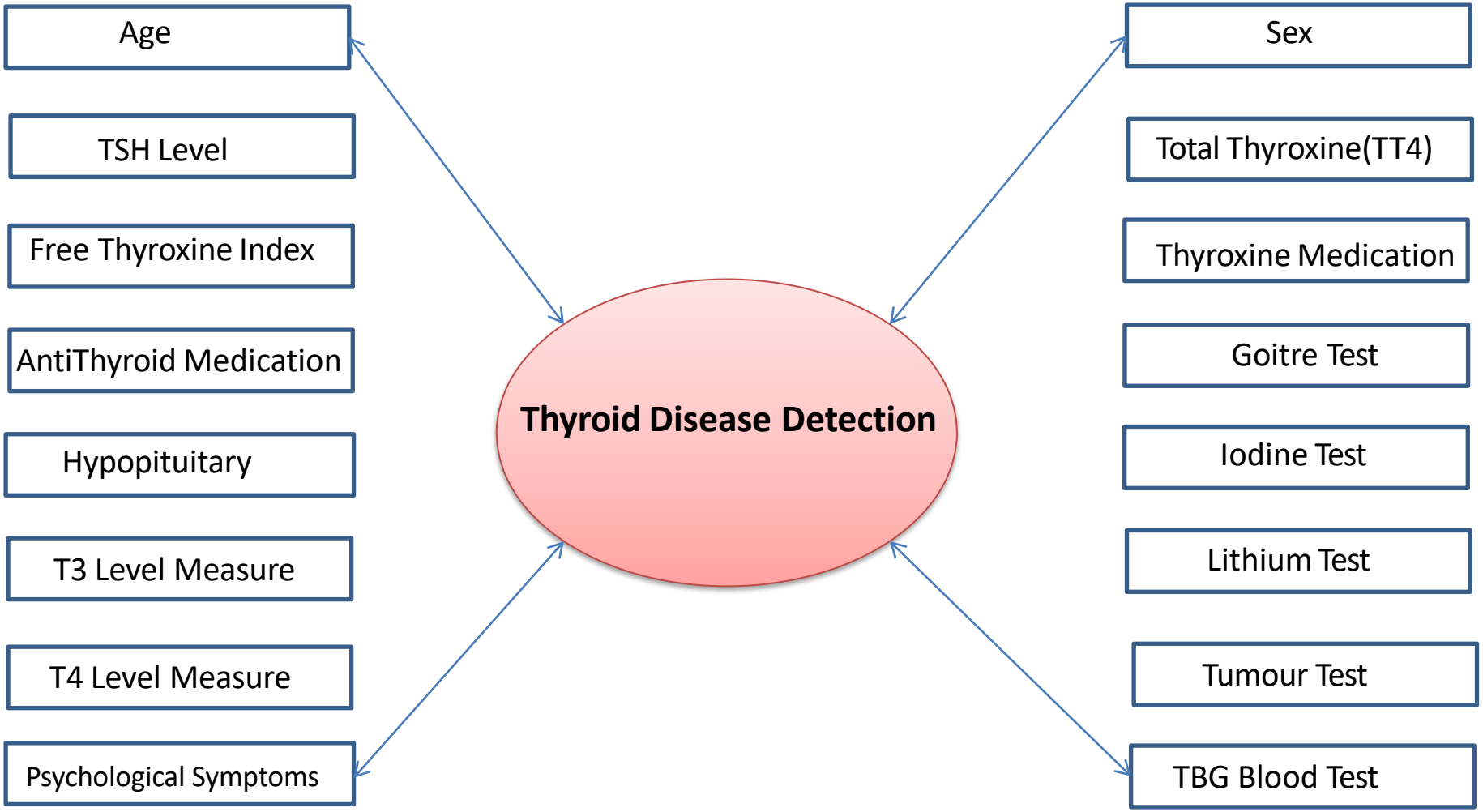
OBJECTIVE

The primary objective of this project is to accurately predict the risk of hyperthyroidism and hypothyroidism in individuals based on various factors. Thyroid disease is a prevalent and challenging medical condition to forecast in research. Our efforts in this study are crucial as they aim to enable early detection and precise identification of the disease, leading to informed decisions and improved treatment outcomes for patients. By harnessing the power of predictive modeling, we seek to empower doctors with valuable insights, enhancing their ability to provide timely and effective healthcare to those affected by thyroid disorders.

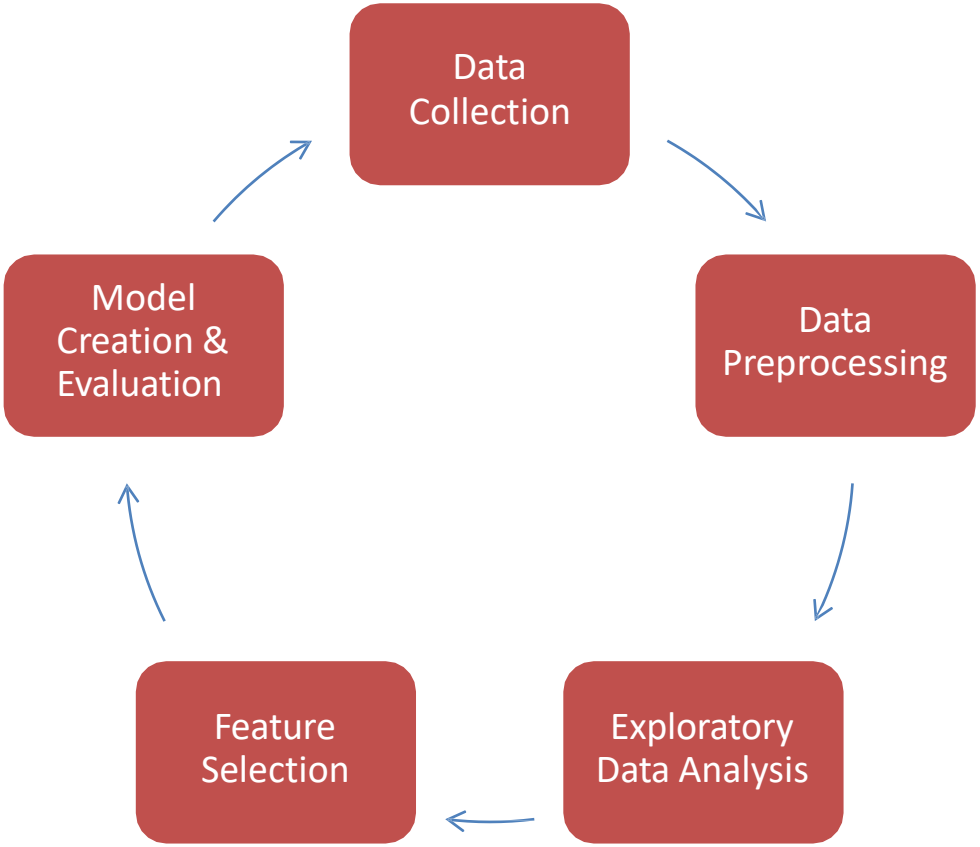
ARCHITECTURE



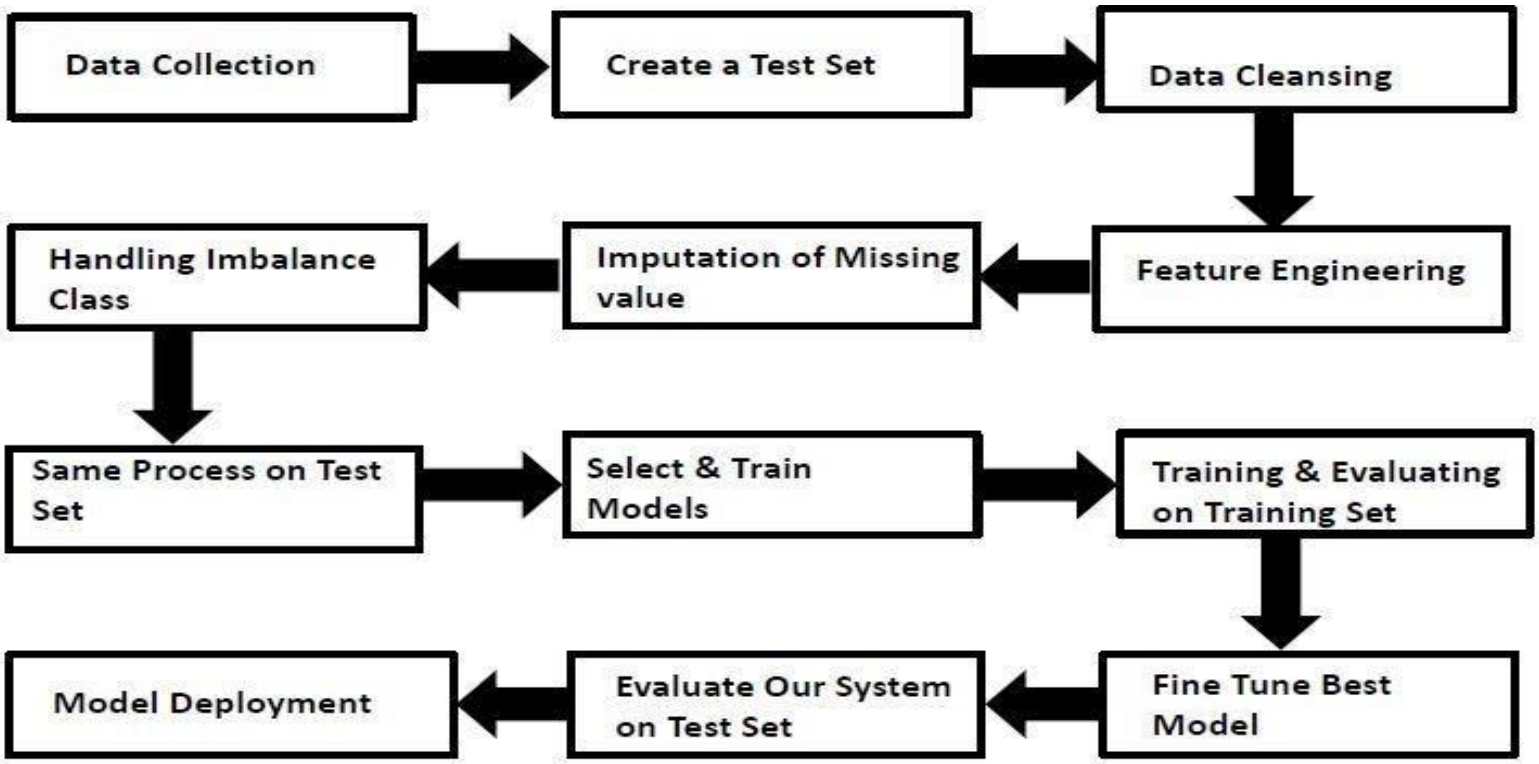
DATASET


Age
TSH Level
Free Thyroxine Index
AntiThyroid Medication
Hypopituitary
T3 Level Measure
T4 Level Measure
Psychological Symptoms
Thyroid Disease Detection
Sex
Total Thyroxine(TT4)
Thyroxine Medication
Goitre Test
Iodine Test
Lithium Test
Tumour Test
TBG Blood Test

Data Analysis Steps



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

- 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 technique widely used in machine learning for classification and regression tasks. It operates by constructing multiple decision trees during training and combining their predictions to produce a final output. Each tree is trained on a random subset of the data, and at each node, a random subset of features is considered. This randomness ensures diversity among the trees and reduces overfitting.

During prediction, the random forest aggregates the outputs from individual trees, either through majority voting for classification tasks or averaging for regression tasks. This ensemble approach enhances the model's accuracy, stability, and ability to handle large datasets with high-dimensional feature spaces.

Random forest is chosen for model prediction in thyroid detection because it can effectively handle high-dimensional medical data, reducing the risk of overfitting. Its ensemble approach enhances accuracy and robustness, crucial for reliable diagnosis. Moreover, the model's feature importance analysis aids in identifying significant indicators, providing valuable insights for medical practitioners. With the ability to handle missing data and noisy features, random forest proves to be a powerful and versatile algorithm, contributing to improved and efficient thyroid detection.

Reason to use Random Forest Classifier model:

- It has high execution speed.
- It gives better model performance.

MODEL PREDICTION RESULTS ON TEST DATASET

Classification Report

	precision	recall	f1-score	support
0	0.81	0.89	0.85	996
1	1.00	0.76	0.86	1032
2	0.82	0.95	0.88	977
3	1.00	1.00	1.00	1005
accuracy			0.90	4010
macro avg	0.91	0.90	0.90	4010
weighted avg	0.91	0.90	0.90	4010

Confusion Matrix

```
[[ 884    0  109    3]
 [ 155  782   93    2]
 [  50    0  927    0]
 [   0    0    0 1005]]
```

Accuracy

```
Accuracy: 0.8950124688279302
```

DATABASE CONNECTION & DEPLOYMENT

Database Connection

- Cassandra Database used for this project.

```
Connected as upendra.kumar48762@gmail.com.
Connected to cndb at cassandra.ingress:9042.
[cqlsh 6.8.0 | DSE DB 4.0.0.6815 | CQL spec 3.4.5 | Native protocol v4]
Use HELP for help.
token@cqlsh> select * from db.Good_Raw_Data;

 age | class                    | fti | fti_measured | goitre | hypopituitary | i131_treatment | lithium | on_antithyroid_medication
| on_thyroxine | pregnant | psych | query_hyperthyroid | query_hypothyroid | query_on_thyroxine | referral_source | sex | sick | t3
| t3_measured | t4u   | t4u_measured | tbg | tbg_measured | thyroid_surgery | tsh    | tsh_measured | tt4 | tt4_measured | tumor
-----+--------------------------+-----+--------------+--------+---------------+----------------+---------+---------------------------
+--------------+----------+-------+--------------------+-------------------+--------------------+-----------------+-----+------+------
+--------------+------+-------------+-----+-------------+-----------------+------+-------------+-----+-------------+-------
  23 |                'negative' | 242 |          't' |   'f' |           'f' |          'f' |    'f' |                          'f'
|          'f' |      'f' |   'f' |                't' |             'f' |                'f' |         'other' | 'F' |  'f' |    4
|         't' | 0.84 |          't' | '?' |           'f' |           'f' | 0.015 |          't' | 204 |          't' |  'f'
  53 |                'negative' | 186 |          't' |   'f' |           'f' |          'f' |    'f' |                          'f'
|          't' |      'f' |   'f' |                'f' |             'f' |                'f' |         'other' | 'F' |  'f' |  3.4
|         't' | 1.08 |          't' | '?' |           'f' |           'f' | 0.005 |          't' | 201 |          't' |  'f'
  91 |                'negative' | 132 |          't' |   'f' |           'f' |          'f' |    'f' |                          'f'
|          'f' |      'f' |   'f' |                'f' |             'f' |                'f' |           'SVI' | 'F' |  'f' |  1.3
|         't' | 0.96 |          't' | '?' |           'f' |           'f' |  0.15 |          't' | 126 |          't' |  'f'
```

Model Deployment

- The final model is deployed on Streamlit .