THYROID DISEASE DETECTION

Detailed Project Report

INTRODUCTION

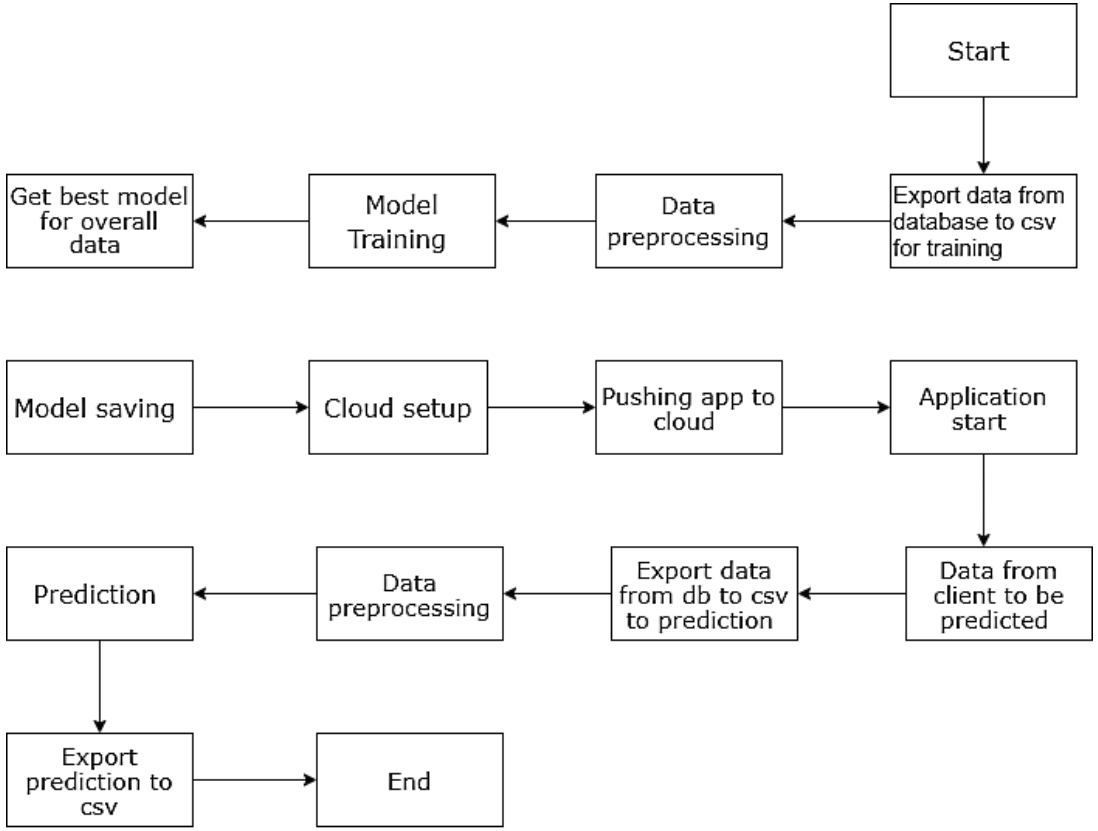
The thyroid gland is a vital organ in the human body, secreting two hormones—total serum thyroxin (T4) and total serum triiodothyronine (T3)—that regulate metabolism. Common thyroid disorders include Hyperthyroidism, associated with excessive thyroid hormone circulation, and Hypothyroidism, often due to a lack of thyroid hormone generation and poor alternate therapy.

Machine learning, a powerful tool in disease prediction, is used to analyse and classify thyroid diseases based on a dataset from the UCI machine learning repository. Various algorithms such as SVM (support vector machine), decision tree, logistic regression, KNN (K-nearest neighbours), decision tree, random forest, and xgboost are employed to predict the risk of thyroid disease in patients.

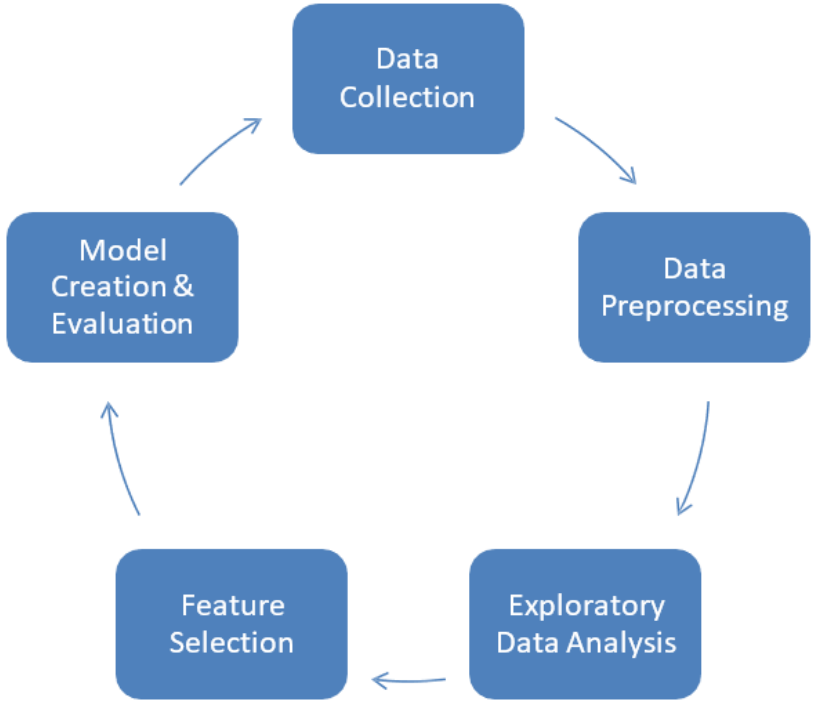
OBJECTIVE

The main goal is to predict the estimated risk on a patient's chance of obtaining thyroid disease or not.

ARCHITECTURE



MACHINE LEARNING PROJECT STEP



1.DATA COLLECTION

Data is obtained from the following link

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

2.DATA PREPROCESSING

- Missing values handling by Simple imputation
- Outliers' detection and removal by boxplot and percentile methods
- Categorical features handling by one hot encoding and label encoding
- Feature scaling done by Standard Scalar method
- Imbalanced dataset handled by SMOTE -Over sampling
- Drop unnecessary columns

3.MODEL CREATION AND EVALUATION

- Various classification algorithms like Random Forest, XG Boost, Logistic Regression, Decision tree etc. tested.
- Random Forest, Xgboost and decision tree all were given better results. Random Forest was chosen for the final model training and testing, with highest accuracy of 99.4%
- Model performance evaluated based on accuracy

```
================================================================================

Best Model Found , Model Name : Random Forest , Accuracy Score : 0.9946977730646872

================================================================================
```

RANDOM FOREST CLASSIFIER MODEL

Random Forest is an ensemble learning algorithm widely used for both classification and regression tasks. It leverages a collection of decision trees during training, combining their predictions for enhanced accuracy and robustness. Random Forest introduces randomness in the construction of individual trees, creating diversity and reducing overfitting.

Reasons to Use Random Forest:

- High Predictive Accuracy: Random Forest often provides high accuracy by combining the strengths of multiple decision trees.
- Resistance to Overfitting: Random Forest is less prone to overfitting due to the variability introduced in the tree-building process.
- Feature Importance: It offers insights into feature importance, helping users understand the contributions of different features.

4.WEBPAGE DEVELOPMENT AND DEPLOYMENT

- Webpage is developed using FLASK framework
- Deployment of the model using AWS (AMAZON WEBSERVICE)

THANK YOU