**THYROID DETECtion**

**DETAILED PROJECT REPORT (DPR)**

CREATOR : LALATENDU DALAI

TABLE OF CONTENTS

● Problem Statement

● Objectives

● Steps Followed

● Visualisations

Problem Statement

Thyroid disease is a common cause of medical diagnosis and prediction, with an onset

that is difficult to forecast in medical research. The thyroid gland is one of our body's

most vital organs. Thyroid hormone releases are responsible for metabolic regulation.

Hyperthyroidism and hypothyroidism are one of the two common diseases of the thyroid

that releases thyroid hormones in regulating the rate of body's metabolism.

The main goal is to predict the estimated risk on a patient's chance of obtaining thyroid

disease or not.

Objective

**Objective :**

* The dataset contains the features and characteristics of thyroid disease

and the target variable estimating the risk on a patient's chance of obtaining

thyroid disease.

* The aim of this project is to use the given data and create a machine learning

model which will predict the estimated risk on a patient's chance of obtaining

thyroid disease or not.

Dataset Information

From Garavan Institute # Documentation: as given by Ross Quinlan.

6 databases from the Garavan Institute in Sydney, Australia

Approximately the following for each database:

\*\* 2800 training (data) instances and 972 test instances

\*\* Plenty of missing data

\*\* 29 or so attributes, either Boolean or continuously-valued

additional databases, also from Ross Quinlan, are also here

\*\* Hypothyroid.data and sick-euthyroid.data

\*\* Quinlan believes that these databases have been corrupted

\*\* Their format is highly similar to the other databases

A Thyroid database suited for training ANNs

\*\* 3 classes

\*\* 3772 training instances, 3428 testing instances

Steps followed

**Data Extraction**: This step involves extracting the data from different sources relevant to the problem statement or obtaining data from the client.

**Data Preprocessing**: In data preprocessing step, we check if there missing data, duplicate values, and datatypes of each feature. In our dataset, there are some null and duplicate values.

**Exploratory Data Analysis:** This step includes bivariate and univariate analysis of features. Checking outliers using boxplots, and outlier treatment is carried out as well. Distribution of the features are plotted to see to what extent our data is skewed.

**Feature Engineering:** In this part, the datatypes of the features were checked whether it belongs same datatypes or different datatypes. Outliers were checked using boxplot but there is no such major outliers in the dataset.

**Model Implementation:** After train and test splitting, pipeline containing Standard Scaler and Label Encoder was fitted to several models such as DecisionTree Classifier , RandomForest Classifier,XGB Classifier, KNeighbors Classifier, etc. Their f1 score were obtained and it was determined that DecisionTree Classifier performs better than other models.

**Model Evaluation:** Test dataset is used to evaluate the model 20% of dataset was separated for testing. Predicted results of the model are compared with the actual data to check the amount of error. As there was no considerable change after , it helped us to overcome overfitting and perform better on new data.

**Designing UI:** For this project, a user interface is built using Apache Air Flow.It is a open-source platform for developing, scheduling, and monitoring batch-oriented workflows.

**Designing a server:** A server should be created to run the UI application continuously. Amazon Web Service EC2 instance is used to create a virtual server for the application.Elastic Compute Cloud (EC2) is a virtual server in AWS for running applications on the AWS infrastructure.

**Deployment:** The codes for this machine learning model should be deployed to the cloud, so that when data is entered into the application, our code runs, and a user gets the result online. In this stage, we containerized the code using Docker and deploy the model to AWS.

Visualisation

Numerical features

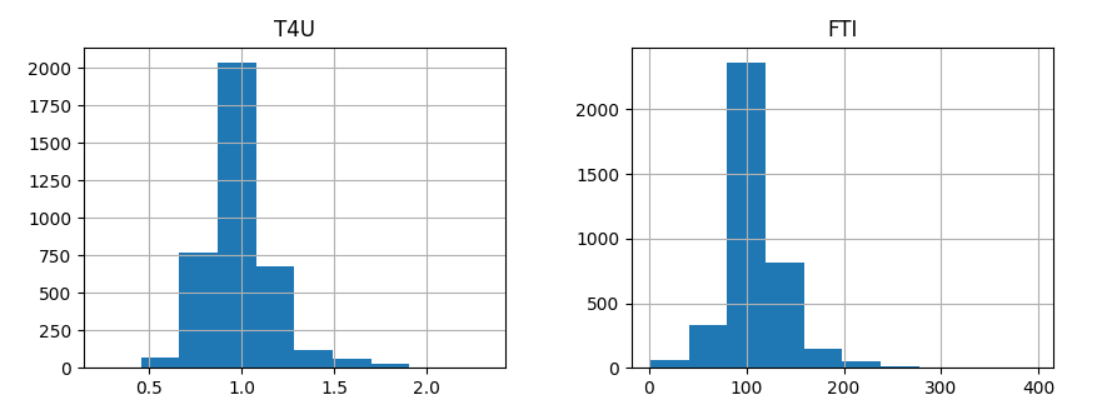
A graph of a bar and a bar of a graph

Description automatically generated with medium confidence

A graph of a bar and a bar graph

Description automatically generated with medium confidence

# 



Distributions

A graph with blue lines

Description automatically generated

A graph of a number of blue lines

Description automatically generated with medium confidence

A graph with a line

Description automatically generated

A graph of a graph

Description automatically generated

Checking Outliers

A comparison of a graph

Description automatically generated with medium confidence

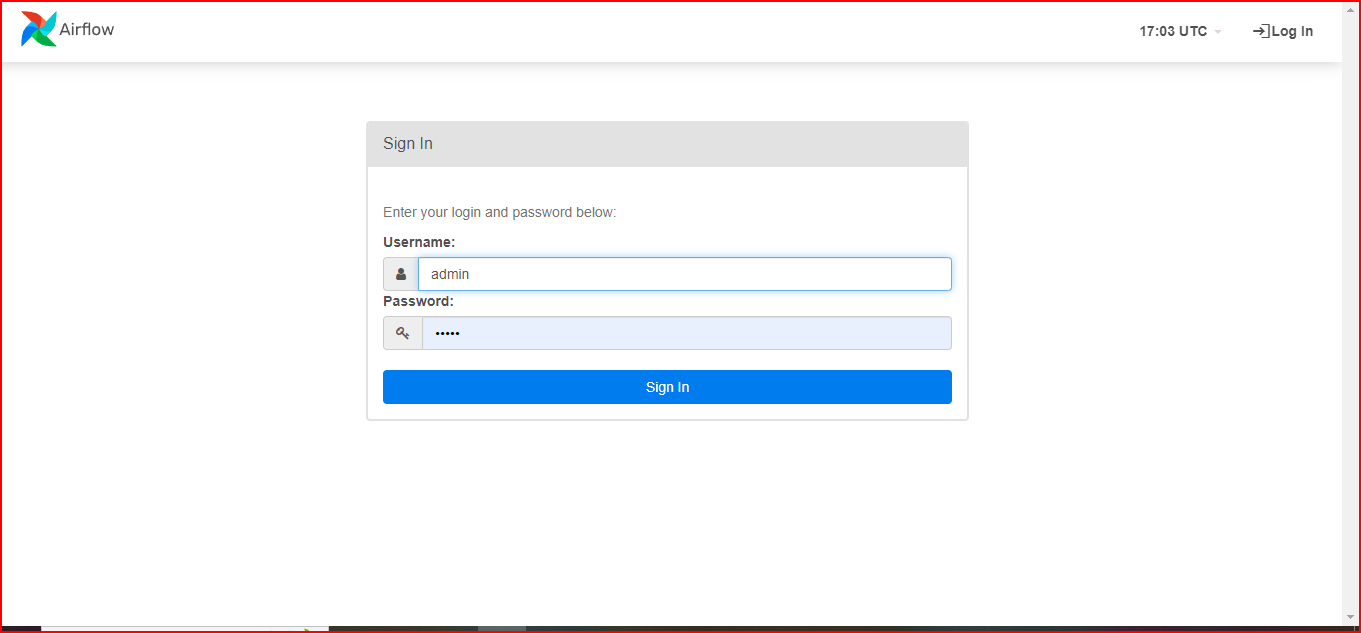
A comparison of a graph

Description automatically generated with medium confidence

A comparison of a graph

Description automatically generated

Web interface



A screenshot of a computer

Description automatically generated

**App link**: http://ec2-35-154-207-161.ap-south-1.compute.amazonaws.com:8080/

Summary

* The target column has 4 class type one is ‘compensated hypothyroid’ which has 194 counts, second is ‘negative’which 3481 counts,third is ‘primary hypothyroid’ and the forth is ‘secondary hypothyroid’ so we have unequal counts for classes in our data.
* The dataset contains the data for all 3 genders. i.e.Male, female and trans.
* The dataset contain some missing values and all the columns have same data types.
* The dataset does not contain any major outlier.
* The DecisionTree Classifier model has the best accuracy on both train and test data.

Q & A

* What is the type of the data ?

The type of the data is categorical.

* Which cloud platform is used for the deployment?

Amazon AWS is used for the deployment.

* What is the source of the data?

The dataset is taken from UCI repository.

* How logs are managed?

We are using different logs as per the steps that we follow in training and prediction like model training logs and prediction logs ,etc.

* Is your model 100% sure about the predictions?

Looking at the results, yes we are sure. However it is recommended that you also take help from someone who is expert.

**Thank You**