**UNIVERSITY OF NAIROBI**

SCHOOL OF COMPUTING AND INFORMATICS

227 - 2nd Year Project (Retake)

**HYPOTHYROID DISEASE DETECTION SYSTEM**

BY:

MBURU RYAN LEEROY

P15/130645/2018

[ryanleeroy@students.uonbi.ac.ke](mailto:ryanleeroy@students.uonbi.ac.ke)

Supervisor :

Mr Eric Ayienga

Github :

<https://github.com/mbururyan/Hypothyroid-Disease-Detection-System.git>

# DEDICATION

I would like to dedicate this project to my loving family who have provided financial support throughout the completion of this project.

Their financial support is what has guided and motivated me throughout the duration of the project as I did not present it the first time.

I would also like the faculty for their motivation and allowance for a retake for the undertaking of this task.

# ACKNOWLEDGEMENT

I greatly thank my supervisor, Mr Eric Ayienga for his guidance, supported advice as I undertook this project.

His wisdom and expertise was a good guiding compass for the success of the project.

# TABLE OF CONTENTS

[***DEDICATION***](#_mqwx7qn0b0s) ***2***

[***ACKNOWLEDGEMENT***](#_3r5gsdusgmr) ***3***

[***TABLE OF CONTENTS***](#_jxyf19fp7pts) ***4***

[***INTRODUCTION / BUSINESS UNDERSTANDING***](#_9o9zkxn0xjq3) ***5***

[Background](#_cis7x6cbpvhy) 5

[Problem Statement](#_66fvsmpmux40) 6

[Proposed Solution](#_v1wjodbjkf6f) 7

[Other required tests](#_r31bmzvw7fhn) 7

[Objectives](#_hzy5rjhz94gr) 8

[Data Collection](#_1ehrmo052z6) 9

[Project Schedule (Gantt Chart )](#_9gl45i2am5b4) 9

[**DATA UNDERSTANDING / SYSTEM ANALYSIS**](#_lflf281nxxht) **11**

[Introduction](#_2vtjvv27uf9y) 11

[Importance](#_aei5bnkmmuc9) 11

[Methodology](#_owuqepih0pth) 11

[Significance](#_5vt9b9x8l996) 11

[Stages :](#_i13g27jn6niq) 11

# INTRODUCTION / BUSINESS UNDERSTANDING

This is the stage in a data mining project where the project idea at hand is expounded and getting to know the business reasons as to why this project at hand is important to be completed.

## Background

Hypothyroid disease, or better known as thyroid disease, is a health complication that arises around the neck area when the patient’s body fails to produce enough thyroid, or produces too much thyroid.

Thyroid disease can affect anyone, symptoms being prevalent among men, women and children alike. It is more common among women, as they have a 5 times probability of contracting the disease as opposed to men or children.

People with a family history in thyroid disease and the elderly are also more likely to contact the ailment.

*Thyroid gland*

The thyroid gland is a small organ located at the front of your neck, and wraps around the trachea. The said gland produces vital hormones that help for normal functioning of someone’s body. The hormones also control the body’s metabolism.

When your thyroid doesn’t work properly, it can impact your entire body.

If your body makes too much thyroid hormone, you can develop a condition called [hyperthyroidism](https://my.clevelandclinic.org/health/diseases/14129-hyperthyroidism).

If your body makes too little thyroid hormone, it’s called [hypothyroidism](https://my.clevelandclinic.org/health/diseases/12120-hypothyroidism).

Both conditions are serious and need to be treated by your healthcare provider.



*Syns and Symptoms*

Symptoms of hypothyroidism include:

1. Fatigue
2. Gaining weight
3. Forgetfulness
4. Frequent menstrual periods (among women)
5. Dry hair
6. Course voice
7. Intolerance to cold temperatures

Early detection of the disease is important as it leads to early treatment to the thyroid gland.

## Problem Statement

Thyroid disease can be examined in a variety of ways such as imaging using ultrasound, blood tests and physical tests. Unfortunately, most healthcare centers in Kenya **dont have the resources for imaging** using ultrasound save for the major Hospitals in Kenya. This is due to financial constraints. This system will help health care systems save on costs and as well diagnose their patients appropriately.

Thyroid Disease, more predominantly in 3rd world countries, is usually **misdiagnosed** for other ailments such as diabetes or tonsillitis due to swelling around the neck. This system will aim to help eradicate the problem of misdiagnosis around health care facilities around our country, as it will thrive to diagnose patients, on high accuracy.

## Proposed Solution

A system specifically tailored for healthcare facilities will be built and tested. The system will attempt to bring accurate diagnosis of the disease to even the most obscure clinics in Kenya so as to raise awareness on the disease and diagnose patients early and cheaply.

This will be done by diagnosing the patients using a website that will take in details of the patient and measurements of a patient's blood done by the doctor/physiotherapist.

The website/system will run the information through a Deep learning model and an output/Diagnosis will be provided.

*Required tests:*

1. TSH Measurement (Thyroid stimulating Hormone)
2. T4 : Thyroxine
3. T3 : Triiodothyronine
4. TT4

### Other required tests

1. Age of patient
2. Sex
3. Pregnancy status (If Female)
4. Medication query
5. Goiter condition

Using Python and R, the data will be cleaned and Exploratory Data analysis performed on it so as to gather as much information about the data in hand and the disease.

*Data Cleaning techniques to be implemented will include :*

1. Checking for Null values
2. Duplicates
3. Data Types
4. Column Names
5. Uniformity in the data
6. Outliers and Anomaly Detection
7. Feature Selection using random forests

*EDA techniques include :*

1. Univariate analysis
2. Bivariate analysis
3. Multivariate analysis
4. Correlation

*Machine Learning*

The data will be separated into training data for building the model and testing data for model evaluation.

I will build multiple ML models and compare them.

A model with 90% or more accuracy will be deemed sufficient and used in deployment

*Implementation*

Using web scripting languages such as HTML/CSS and python, a website will be built that a healthcare personnel can input details of the patient as stated above and the model will predict if the patient has the disease or not

The aim of the project is to save lives of Kenyans, using Machine Learning.

## Objectives

*Main Objectives*

1. Develop a Deep Learning model that predicts whether a patient has hypothyroid disease or not
2. Develop a user friendly website/User Interface that users can interact with and get results on whether a patient has hypothyroid disease, or not.
3. Raise awareness about the disease in Kenya

*Other Objectives*

1. Outline the model’s accuracy in prediction after thorough hyperparameter tuning.
2. Using data analysis, gather insights on which gender is most affected by hypothyroid disease.
3. Which age group is most affected by hypothyroid disease.
4. Are pregnant women more vulnerable to getting the disease?
5. Are the elderly (people over 60 years) more vulnerable to the disease?

## Data Collection

### Data Sourcing

The data was sourced by myself, and the final data was found from Kaggle. It was not clean nor was there any solution to the problem.

1. Dataset - > <https://www.kaggle.com/yasserhessein/thyroid-disease-detection-using-deep-learning/data>
2. Epidemiology and insight on the ailment ->

<https://my.clevelandclinic.org/health/diseases/8541-thyroid-disease#:~:text=The%20two%20main%20types%20of,swelling>

The project will be stored in GitHub, where the code can be accessed by anyone for academic purposes

### Written Interview

To have better insight of the project at hand as I am not a health expert, I interviewed my friend who is a 4th year Medicine student at UoN on the disease at hand and how Machine Learning comes to play in their field of work.

This is how it went :

Data Collection Technique: **Written Interview**

Interview Questions:

**Name and field of study.**

Muli Felix Musembi

Bachelor of Dental Surgery

**Years of experience in the medical field (If student, what year)**

Student at University of Nairobi, Department of Dental Sciences

Year IV

**Any hands-on experience in the medical field?**

Yes. The Department of Dental Sciences features a dental hospital run by the university students under supervision by experienced lecturers and the university in general. Here, I have been stationed at a variety of specialized clinics that are separated depending on the services they offer. They include; Oral Diagnosis clinic, Radiology, Prosthetic Clinic, Orthodontic Clinic, Periodontal Clinic etc.

**What field of medicine do you major in?**

Dentistry

**What do you know about hypothyroid disease?**

Hypothyroidism is a disease of the endocrine system whereby the thyroid gland does not produce enough thyroid hormone to meet the body’s requirements.

Depending on the cause of the condition, hypothyroidism can be classified into either primary, secondary or tertiary hypothyroidism. In primary hypothyroidism, the thyroid gland is directly affected hence it is unable to produce sufficient thyroid hormone. This is mainly caused by deficiency of iodide which is an important component required to synthesize thyroid hormone. Other causes include; trauma to the thyroid gland and autoimmune diseases that damage the thyroid gland e.g. Hashimoto’s disease.

Secondary hypothyroidism is caused by reduced secretion of Thyroid Stimulating Hormone (TSH) by the pituitary gland. TSH is an enzyme that stimulates the thyroid gland to produce thyroid hormone. Reduced concentration of TSH in blood circulation results in reduced levels of thyroid hormone resulting in secondary hypothyroidism.

Tertiary hypothyroidism is caused by reduced secretion of Thyroid Releasing Hormone (TRH) by the hypothalamus. The function of TRH is to stimulate the pituitary gland to secrete Thyroid Stimulating Hormone (TSH) which in turn stimulates the thyroid gland to secrete thyroid hormone. Reduced levels of TRH results in low circulating levels of thyroid hormone leading to tertiary hypothyroidism.

The common presenting symptoms of hypothyroidism include; weight gain, cold intolerance, fatigue, weakness, depression, myalgia, constipation and bradycardia. Further complications may develop e.g., goitre, cretinism and myxedema.

The treatment of hypothyroidism is usually to treat the underlying cause of the condition. Management can be done by giving thyroid hormone supplements e.g., levothyroxine which is an analogue of thyroid hormone thyroxine (T4).

**Is it widely known by Kenyan citizens?**

I do not think so.

Goiter as a manifestation of hypothyroidism may be a well-known symptom as it involves a relatively large swelling of the neck region around the thyroid gland area. However, this symptom is associated more with hyperthyroidism than hypothyroidism.

**If not, what can be done to spread awareness?**

Community health education. Campaigns can be carried out to educate the public on hypothyroidism, the causes and common presenting symptoms. They can also be educated on preventive mechanisms e.g. adequate intake of iodine and screening to check for autoimmune diseases.

**What do you know about Machine Learning / Artificial Intelligence?**

Machine learning is the ability of a system to analyze data, recognize patterns in the data and therefore adapt and improve through complex algorithms. Artificial intelligence is the ability of a computer system to carry out complex functions that mimic human behavior and abilities.

**Is AI/ML being used in the medical field in any way?**

Yes.

**If so, give a vivid example and how it has helped the medical sector.**

Artificial Intelligence has simplified the process of taking and interpreting cephalograms in the radiology sector. Complex algorithms enable the computer system to use a digital sensor to detect X-rays and display the radiograph on screen. From there the computer can then denote important landmarks on the radiograph and draw relations between observable anatomical features. The system subsequently interprets the acquired information in order to come up with a diagnosis. This has helped especially in the orthodontic sector where the relationship between various anatomical landmarks can then be used to diagnose different dental malformations and plan for orthodontic and orthognathic surgery.

In addition, artificial intelligence has also been used in carrying out various surgical operations. This is especially true for procedures that require high accuracy and fine control e.g. coronary artery bypass graft (CABG) and hysterectomy (surgical removal of the uterus). Artificial systems are able to carry out these procedures with the required level of accuracy and effectiveness hence reducing the risk of human error. This increases the positive outcome of such procedures.

**In your opinion, can ML/AI assist medical experts in their day-to-day work?**

Yes. ML/AI has shown that it can carry out complex procedures with high accuracy and efficiency hence reducing the workload for medical experts while also improving the accuracy and efficiency of their work.

**What do you like/dislike about health/medical websites?**

Like; Medical websites contain a vast amount of knowledge regarding specific topics of interest to a researcher, a medical practitioner or even a student. They provide deep, concise information that can be used academically. Some websites even provide modern trends in the medical field hence enabling the practitioner to keep up with the ever-improving methods of clinical procedures.

Dislike: Some medical websites contain out-of-date information that can provide wrong information or information that has been improved for higher efficiency. In addition, some websites have a very complex user-interface that makes it difficult to actually navigate through different pages or sections while looking for complete information. Some websites are also not very pleasing to look at since they may have a rigid structure.

**What constructive criticism can you give to improve users' experience with the said websites?**

Improvement of user-interfaces – Simplification of user interfaces and clear instructions on navigation makes it easier for the user to actually navigate through different web pages and increase user satisfaction. Grouping of related information can also improve the user experience while reading through the information especially if the information provided is presented in an aesthetic manner.

Websites also need to ensure that they constantly update their information to keep up with modern medical practices.

**Do doctors/medical practitioners use websites in their day-to-day work? If so, how do they help?**

Yes.

Medical websites serve a variety of functions;

* Medical students can acquire academic information from educative websites
* Medical researchers use websites to search for important articles, journals and related research papers in order to further build their research proposals.
* Medical practitioners can acquire information on medical procedures that are changing with time.
* Medical practitioners can refer to websites in order to refresh their memory on clinical information they may have forgotten.

Websites make it easy to access the relevant information through the internet without having to refer to hardcopy books that may or may not be available.

**In your opinion, can Artificial Intelligence be used to diagnose a patient? If so, give an example and how it helps over traditional diagnosis**

Yes.

In radiography, Cone Beam Computed Tomography (CBCT) utilizes artificial intelligence to generate 3-dimensional images taken through exposure to X-rays. The resultant image displayed on the screen is of high quality and can be manipulated to be viewed in different sections. This 3D image is then used to make a more accurate diagnosis e.g. in checking for direction of displacement of bone in mandibular fractures or pattern of fracture of the zygoma bone (cheek bone).

This is a lot more efficient since traditional diagnosis would have utilized two-dimensional (2D) radiographs that are more difficult to interpret. They also required multiple exposure of the patient to X-rays in order to generate views from different angles. This predisposes the patient to cancer among other conditions.

Thank you.

## Methodology

As this project is a Data Science project, I went with a bit of a hipster approach and will use the CRISP-DM methodology. (**Cross-Industry Standard Procedure for Data Mining.)**

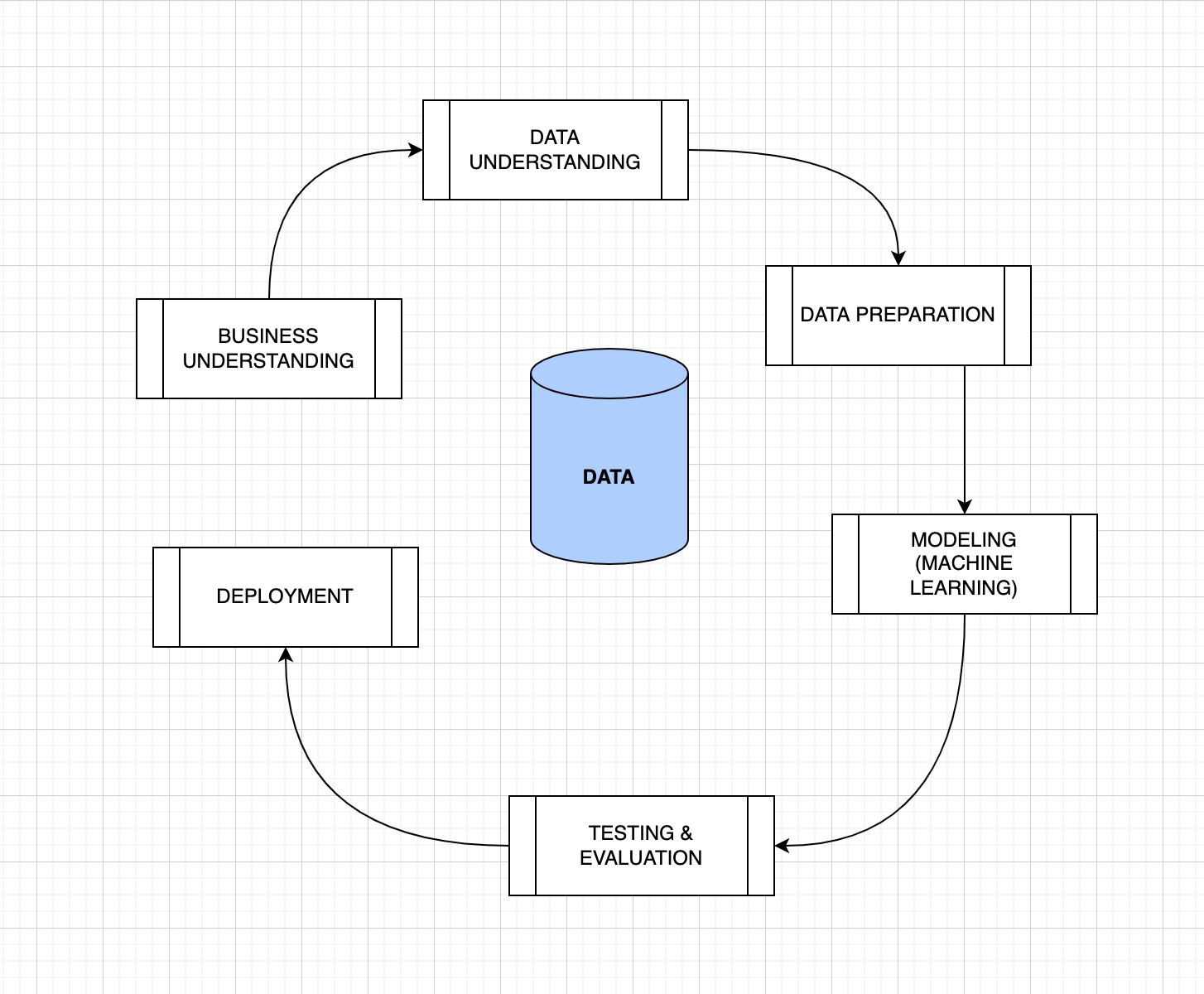
### Significance

This is a methodology that is widely used and accepted in the Data Science field that incorporates the most important stages of a data science project in an organized and structured manner.

The methodology was created by IBM experts

### Stages :

The said approach has 6 stages, that will be graphically illustrated below



1. **Business Understanding**

This is the initial stage, where the data is collected and business prospects derived from the data. This is the most crucial stage, as it shows whether the project is feasible to be done or not and just a waste of the programmer’s and stakeholders time and/or resources.

1. **Data Understanding**

This is the stage where the collected data is scrutinized from face value, the size, complexity and other identifications specified before any manipulation.

This stage ensures the data is of the highest quality to be cleaned and analyzed.

1. **Data Preparation / Data Cleaning**

This stage is usually the most time-consuming in the whole project, as the data at hand is meant to be cleaned without losing a lot of data. This requires the data scientist to take risks and good decision making .

Fun fact : This usually takes about 50% of the data scientist’s time

1. **Modeling**

This is the stage where the cleaned data is implemented into a machine Learning model of the developer’s choice and predictions are done. This is the most rewarding process of the cycle.

1. **Testing & Evaluation**

The developed model, before deployment, has to be tested so as to ensure it performs greatly, with good accuracy, without overfitting.

1. **Deployment**

The system, so as to ‘come to life’, has to be implemented in some sort of dashboard, website or mobile application.

Users at this point should be able to interact with your website and have the develop model do some cool predictions.

1. **Feedback (Not a must)**

This is the final stage where feedback from the users is sought after so as to improve the overall performance and user quality of the system at hand.

## Project Schedule

Will use a gantt chart to visually display activities to be accomplished against time.

|  | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Planning |  |  |  |  |  |  |  |
| Data Analysis |  |  |  |  |  |  |  |
| Machine  Learning |  |  |  |  |  |  |  |
| User Interface  Design |  |  |  |  |  |  |  |

## Feasibility Study

This is an assessment of the operational, technical and economic merits of a proposed project.

The feasibility study is a preliminary review of the facts to see if it is worthy to commence with the project.

### Operational Feasibility

It is a measure of how well the system solves the problem at hand.

In our case, the system accurately predicts whether the patient has hypothyroid disease, with an accuracy of 98 %.

It solves its sole purpose, with high accuracy at that.

### Economic Feasibility

This is a measure of how well the project used its allocated resources, and whether the financial money assigned was used well or more was needed.

In this case, the project used little to no money.

If an actual hospital buys the project, some money may be needed to buy an unique IP.

### Schedule Feasibility

This is a measure of how reasonable the allocated time was.

In this case, 8 weeks to complete the project was more than enough time needed.

## Functional & Non-Functional Requirements

### Functional Requirements

This is what our system is necessarily expected to do.

1. Give accurate prediction on whether a patient has hypothyroid disease or not
2. A user friendly website
3. Information and intractable graphs (e.g) bar charts and pie charts that give information about the disease
4. Users should have instant access to the website
5. Website should be deployed and accessible at any given machine
6. Deployed website should be accessible at any given time
7. Website should be operational

### Non-Functional Requirements

This are usually non-mandatory, but are necessary for assessing the product’s quality

1. ML model should be scalable and light so as to make the website fluid
2. Website should give prediction in less than 5 seconds
3. Intractable informatory graphs should work seamlessly

## Use Case Diagram

This is where the persons who will be interacting with the website are illustrated according to what tasks they will be able to perform.

### Actors :

1. Doctor
2. System Dev / Data Scientist

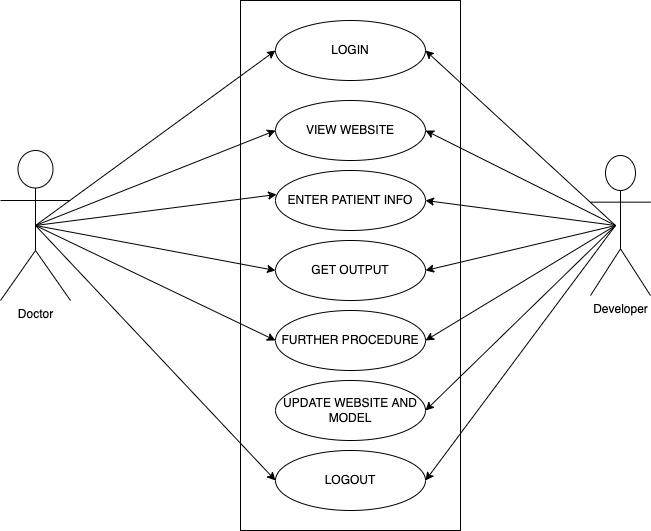
### Doctor :

1. Login
2. View Website
3. Enter Patient’s info
4. Get Output / Prediction
5. Further procedure on disease if positive
6. Logout

### Developer :

Can Perform all the above and as well added privileges such as

1. Improving website layout
2. Improving/Changing ML model



# DATA UNDERSTANDING / SYSTEM ANALYSIS

## Introduction

This is the second stage of the project, where the data at hand is scrutinized and understood and the system to be built is analyzed and well defined.

The methodology being used in the project will be well discussed as well.

### Importance

The significance of this stage is that it improves the project’s quality in its lifecycle and ensures that all components within the project work in unison to accomplish its purpose.

## Source of Data

The data was collected from Kaggle, a popular Data Science platform. It is in csv format.

Here is the link to the dataset :

<https://github.com/mbururyan/Hypothyroid-Disease-Detection-System/blob/main/hypothyroid.csv>

## Data Description

The dataset has 3, 770 rows of data and 30 Columns.

Some of the most important columns include :

1. *Age* - age of the patient
2. *Sex* - gender of the patient
3. *On thyroxine* - if the patient is taking the said medication
4. *Sick* - if the patient is sick
5. *Thyroid surgery* - if the patient has had any previous surgery on the thyroid gland
6. *Goiter* - if the patient has had any complications with his/her goiter
7. *TSH* - medical measurement
8. *T4* - medical measurement
9. *TT4* - medical measurement
10. *TTU* - medical measurement

The data is mostly of object data type, which is a mix of integers and strings.

The data is mostly nominally encoded, with values such as :

1. Sex column : F - Female and M - Male
2. Measurements column : f : false and t - true

## Data Quality

From visually inspecting the data, it is observed to have a lot of missing values, and inconsistency in the values.

This will be handled in the data preparation phase of the project.

# DATA PREPARATION / CLEANING

This is one of the most important and time consuming stages in a Data Science project. It usually takes up 40 - 60 % of the total amount of time of a project’s time and effort.

It is an extremely important phase as it prepares the data adequately for data analysis.

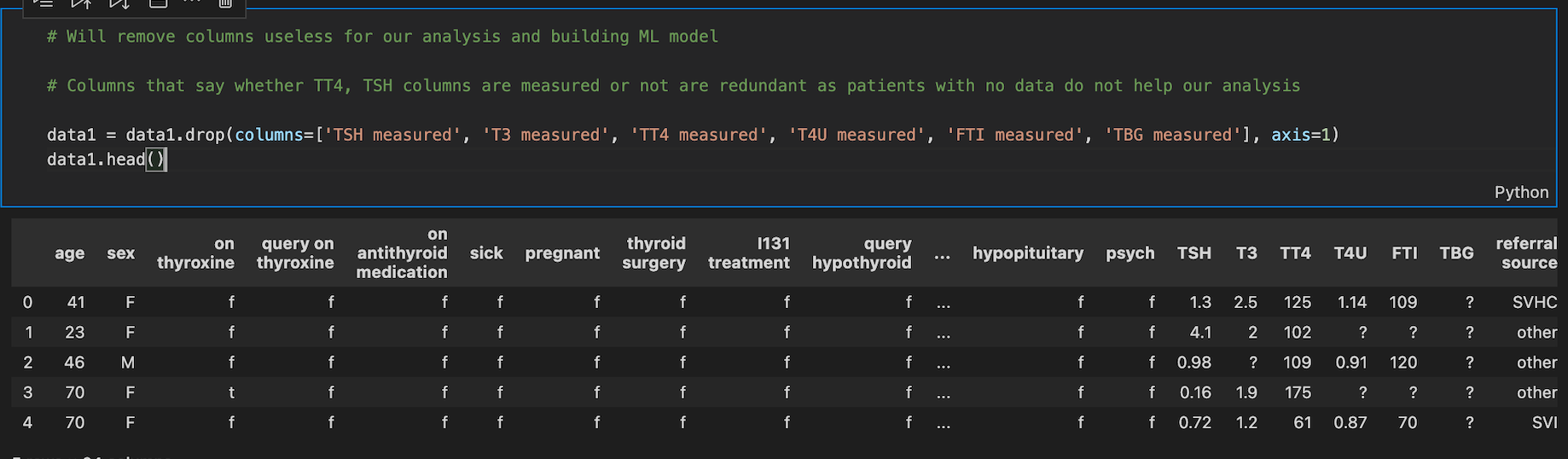
The following tasks usually happen in this stage :

* Data Merging (When there are two or more datasets)
* Data Selection
* Records aggregation
* Deriving new attributes ( Encoding )
* Handling null values
* Handling duplicate values
* Data Uniformity
* Handling outliers in the data

## Data Selection

This is where columns that are not necessary in modeling are usually dropped. These columns are usually known as **features** when handling a Machine Learning project.

In our dataset, we decided to drop the columns that had information on when a value e.g TT4 had been measured. This was important so as to avoid redundancy.



The columns indicated in the python code were dropped from our dataset.

## Handling missing values (Null Values)

Missing values / null values are a common occurrence in data related projects as humans usually are prone to mistakes.

Missing data often leads to a lot of bias in the data or plainly incorrect analysis.

When handling missing data in this project, we used two techniques :

* Imputation (replacing the missing data)
* Dropping

Here are the features that had missing values in our data :

Age : 1

Sex : 150

TSH : 369

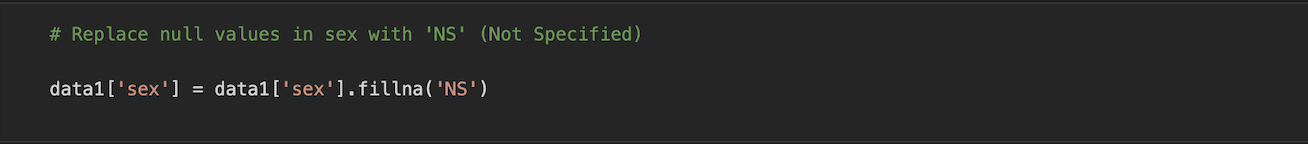
T3 : 400

FTI : 385

#### ***I. Imputation***

The dataset had a gender column named ‘sex’ which had some missing values. To avoid losing data, we went with the imputation method, where we replaced NaN values with ‘NS’ standing for Not Specified.

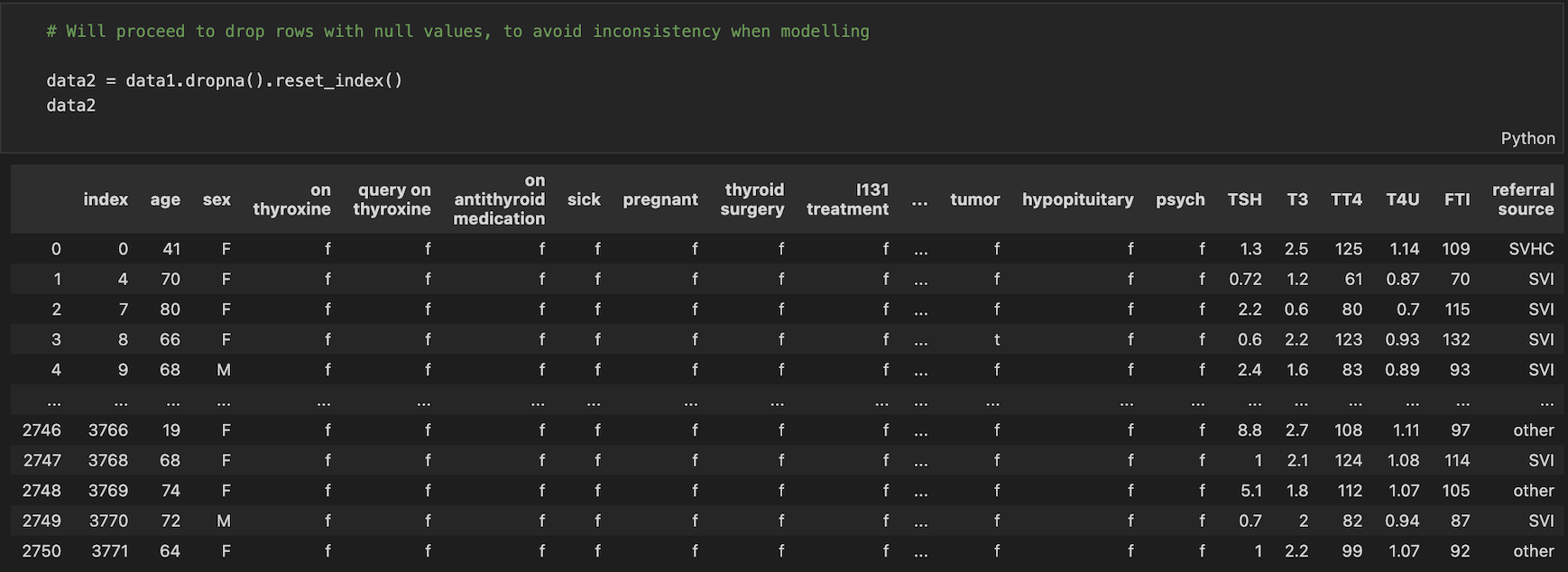
My reasoning was that some people are not comfortable with stating their gender. It's the 21st century, there are even non binary people hehe



#### ***II. Dropping missing values***

As with the rest of the missing values, most of them were the technical inputs on their blood pressure etc such as FTI, therefore felt as replacing them with maybe the mean would lead to incorrect values.

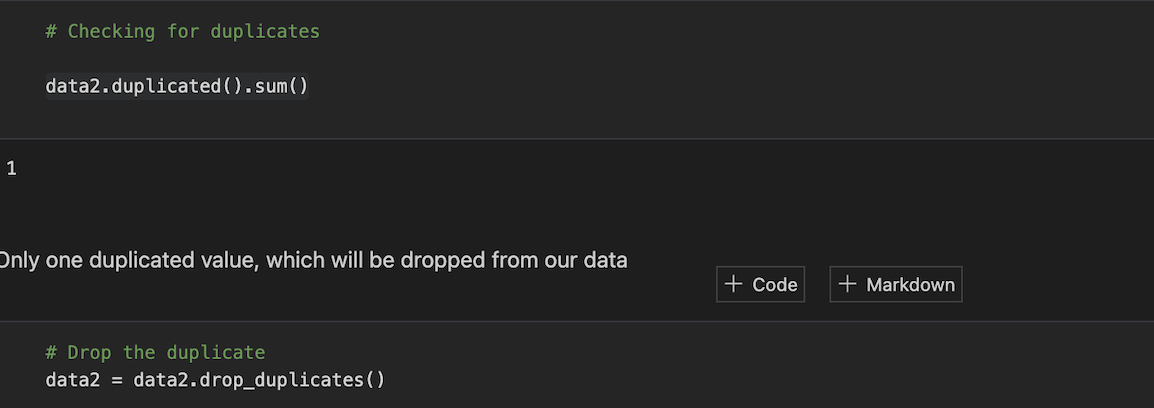
Decided to drop them altogether.



About 400 records of data were lost, but were left with more than 2000 accurate records of data to work with when modeling.

## Handling Duplicates

Duplicate values just cause redundancy in the data, therefore after inspection they were immediately dropped.



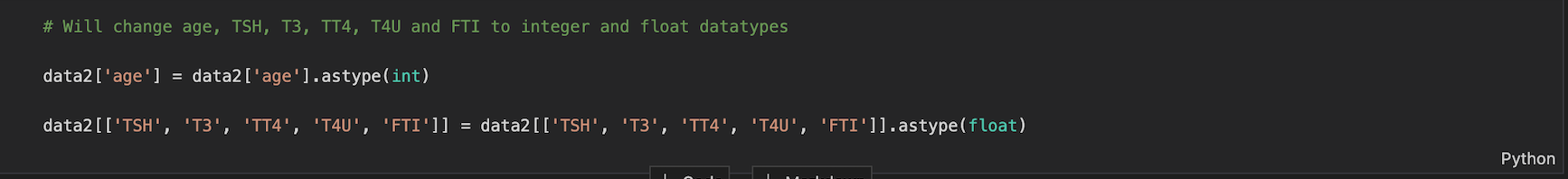
As seen, the data had only a single duplicated record, which was dropped.

## Data Uniformity

Here is where data types are checked and corrected if necessary.

In our data, most of the data was in ‘object’ type data type, which was to cause problems ahead as there were alot of integers in the data.

‘Age’ feature was changed to integer and 'TSH', 'T3', 'TT4', 'T4U', 'FTI' were changed to float datatype, as shown below.

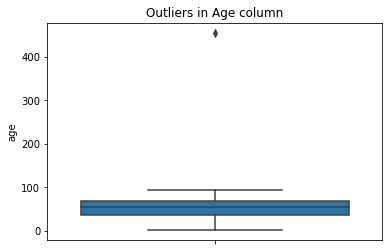


## Handling Outliers

When coding, I initially stopped at data uniformity and proceeded with working on the project. However, when I was checking the distribution of ages of patients, I noticed an abnormal pattern, with the age hitting past 300.

This was an indication of outliers existing in the data, so revisited data preparation and handled the anomaly.

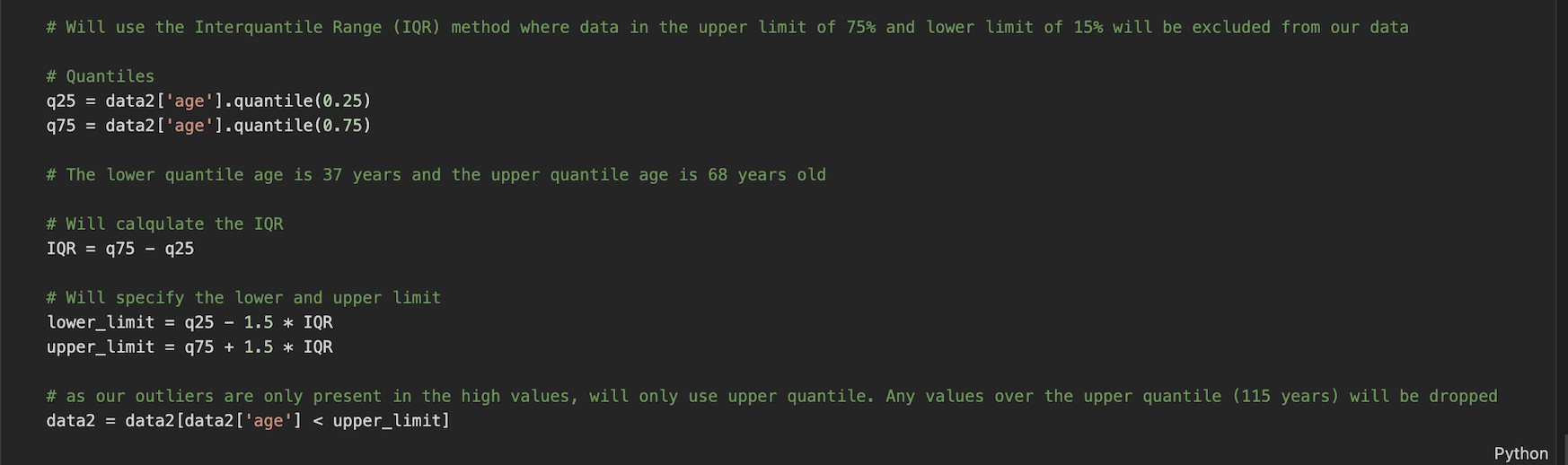
Used boxplots from the seaborn library to plot outliers present in age column



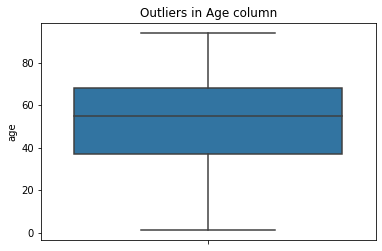
As seen by that ‘dot’ above the boxplot, the data claims that there is a patient over 400 years old, which is impossible. This anomaly had to be handled.

Used a technique called IQR method (Interquartile Range ).

* Separated the data into two quartiles, under 25% and over 75%.
* Calculate the IQR, which is the upper quartile minus the lower quartile
* Specified the lower and upper limits. Lower limit was specified by **subtracting** the lower quartile by 1.5 of the IQR. THe upper limit was the opposite i.e adding the upper quartile to the value of 1.5 times the IQR
* As the outlier was a lot, only specified to every value above the upper limit to be dropped.
* The data is now clean from outliers.



Plotted another boxplot, this time after dropping the outliers, and the result is outstanding.



The data is now clean from any outliers.

The data after preparation, is now ready for analysis and modeling.

# EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis or commonly referred to as EDA, is the critical process of performing investigation on our data so as to extract meaningful insights and notice any patterns/trends within the data. This is with the use of statistical and graphical methods.

There are three stages in EDA :

1. Univariate analysis - Analysis of one feature
2. Bivariate analysis - Relationship between two variables
3. Multivariate analysis - Interaction between multiple fields

## Univariate Analysis

This is where a single variable, or in our case, a single feature’s distribution and statistical behavior are identified so as to gather better business driven insights from our data.

### **Patients’ Ages**

A statistical summary was done on the patient’s ages and here were the results :

*Total number of patients* : 2749

*Average age* : 52

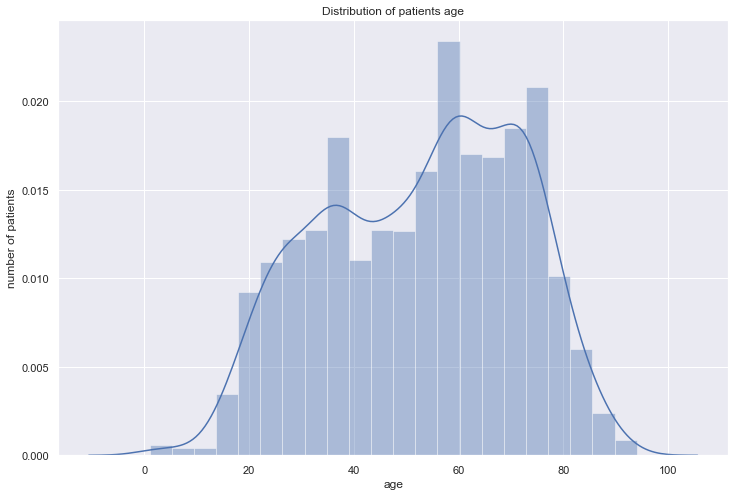
*Standard dev of ages* : 18.8

*Minimum age* : 1 year old

*Maximum age* : 94 years old

### **Distribution of Patients’ Ages**

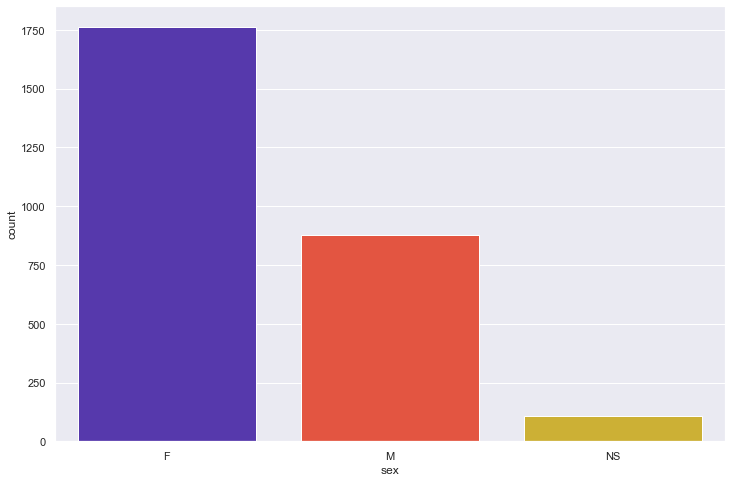
Plotted the patients ages on a histogram, the ages against the number of patients and here was the result.



As seen, the distribution of the patients’ ages is slightly skewed to the right. This indicates that a large portion of the patients of hypothyroid are of the older generation.

### **Gender of Patients**

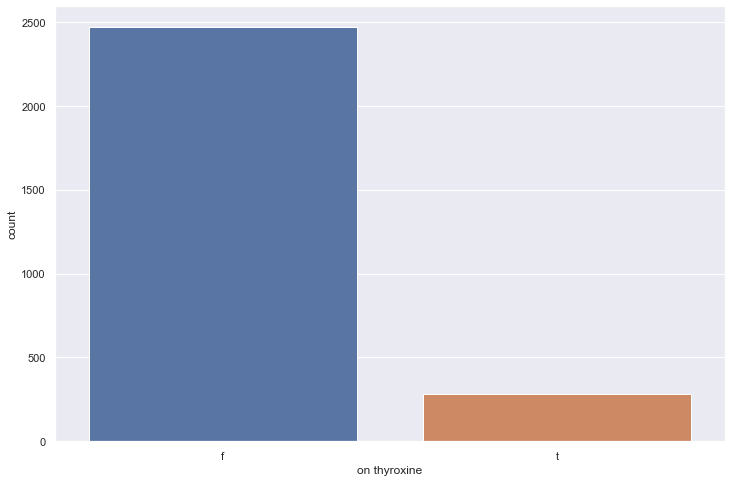
Here is a visual representation of the count of patients per gender



As seen, most of the patients are female.

### Patients on thyroxine medication

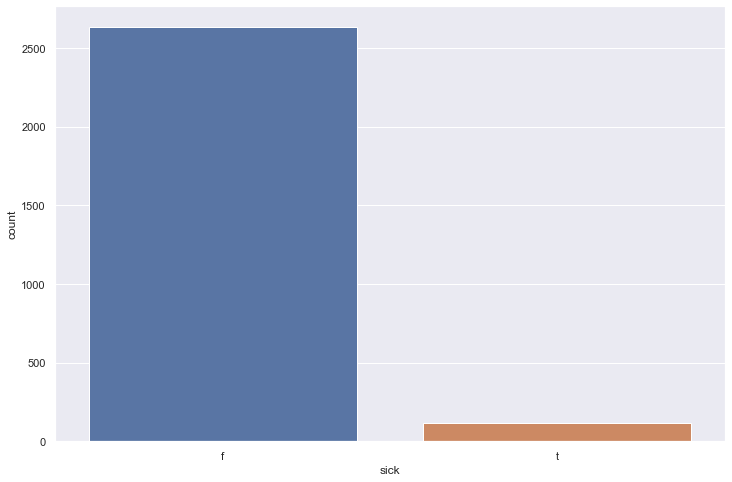
Here is a visual representation of the count of patients taking thyroxine medication



As visualized, most of the patients are not on any thyroxine medication.

### **Sickly Patients**

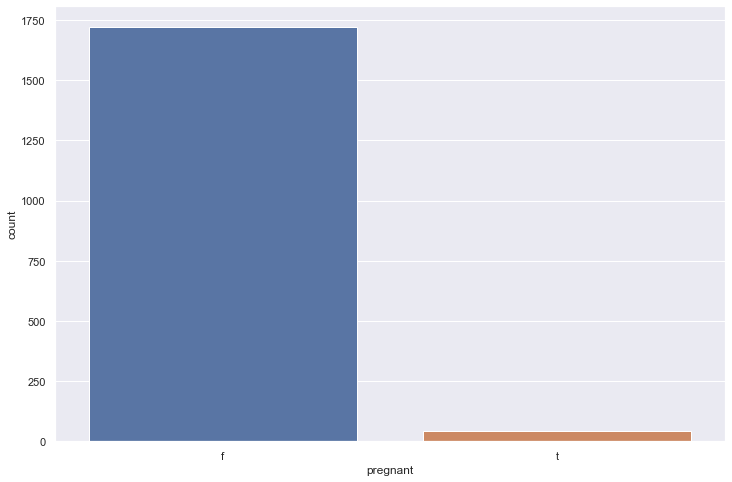
Let's see if most patients had any symptoms or not.



Most patients were deemed healthy when taking the test.

### **Pregnant patients**

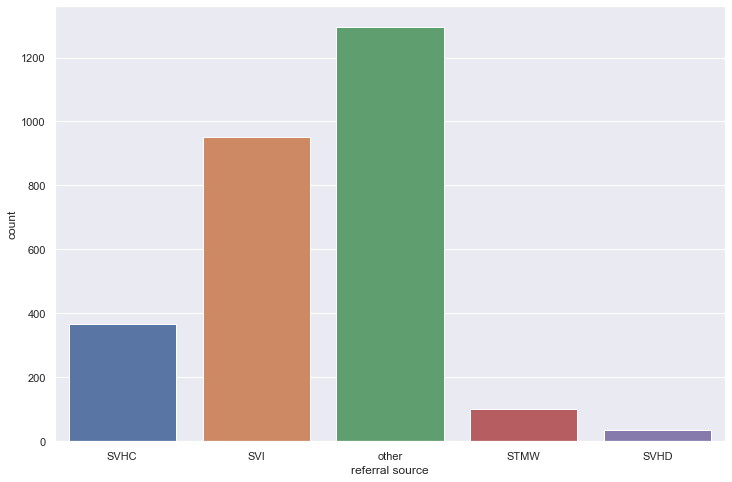
Among the female patients, will plot if most of them were pregnant or not



Most female patients were not pregnant during the test.

### **Source of Referral**

Will plot which referral source was the most common.



Unidentified referral sources were in fact the most common.

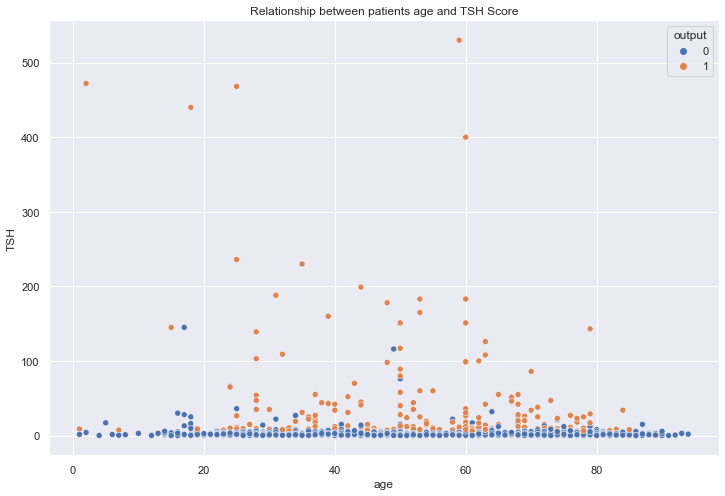
## Bivariate & Multivariate Analysis

This is where two or more features are compared with each other and relationships drawn from them, if any exists

Will use scatter plots and a heatmap to visualize the features’ correlation with each other.

### **Age vs TSH score of sick and healthy patients**

Will plot ages of patients vs their TSH score and note any trend if it exists. The data points are segregated into sick and healthy patients

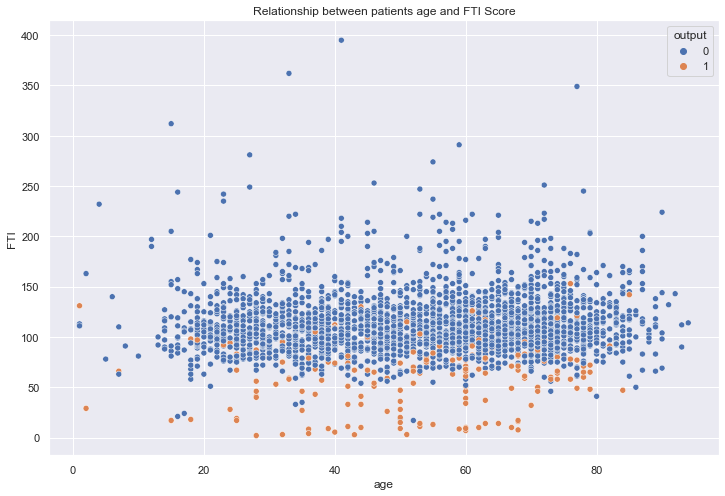


There is no relationship between the patients’ ages and the TSH score.

However, it is noted from the scatter plot that patients with higher TSH scores were most likely to have hypothyroid disease over patients with low TSH scores.

### **Age vs FTI score of sick and healthy patients**

Will plot ages of patients vs their FTI score and note any trend if it exists. The data points are segregated into sick and healthy patients as well.



From this scatterplot, the first observation gathered is that there is no relationship whatsoever between the FTI scores and the ages of patients.

The second observation is that there is a relationship between FTI scores and outcome of the disease, where patients with low FTI scores had higher chances of having hypothyroid disease ove patients with higher FTI scores.

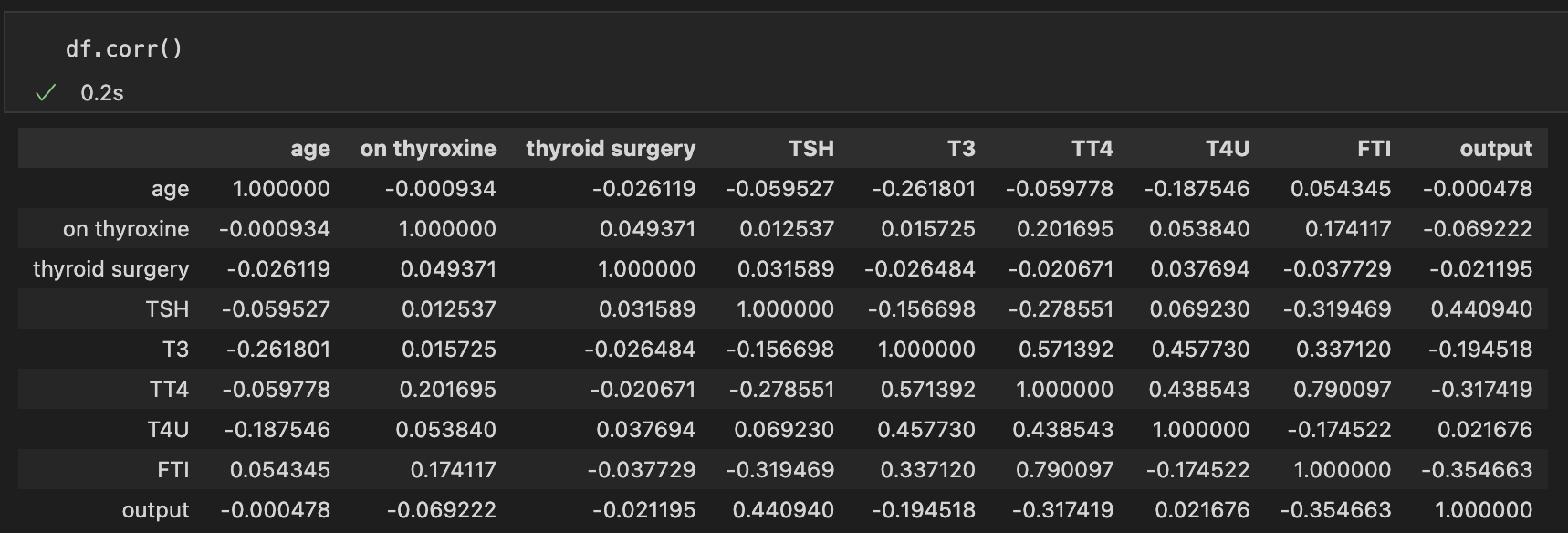
### **Correlation**

Correlation is the statistical evaluation of the strength of a relationship between two variables and how one attribute affects the other.

There are three types of correlation :

1. **Positive correlation** - A positive correlation happens when two variables move in the same direction. This means that when one attribute’s value increases, the other increases as well.
2. **Negative correlation -** A negative correlation is where two variables move in opposite directions. This means that when one attribute’s value increases, the other decreases and vice versa
3. **Weak / zero correlation -** This happens when two variables have no correlation existing between them, therefore one attribute’s value does not affect the other

Here are the correlations of our features, and the most important insights will be indicated at the bottom.



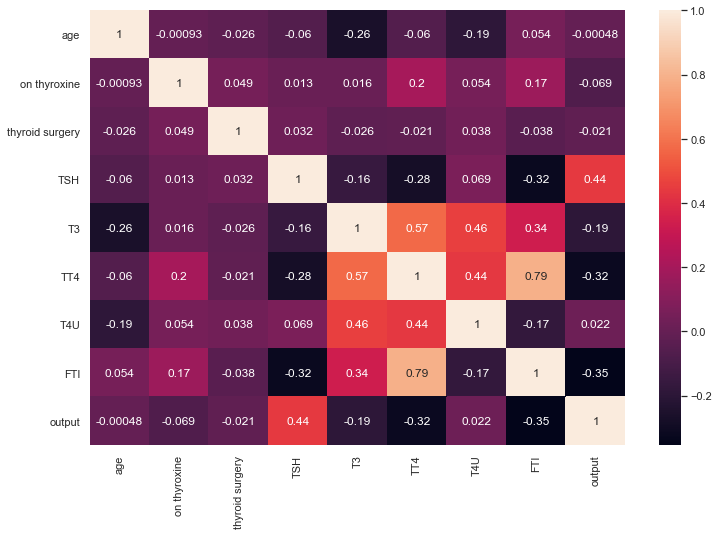
Most of the features are negatively correlated with each other. This indicates that when one attribute value increases e.g. T3, the age of the patient is lower and vice versa as the two features have a correlation of **-0.262.**

An instance of positive correlation in our dataset is TT4 and FTI, which has a correlation of **0.79**, which is a strong positive correlation. This indicates that when TT4’s value increases, so will FTI's value and vice versa.

### **Correlation Heatmap**

So as to have a better representation of correlations, heatmaps are used. They offer a more attractive option of viewing features’ correlations over crunching raw numbers.

Here is the data’s heatmap.



Here is a visual representation of the attributes’ correlation, where lighter tones indicate high correlation and darker tones represent low to zero correlation.

# MODELING

Before modeling commences, the data has to be pre-processed so as to ensure its in the right format to be used in the desired Machine Learning Model(s).

## Data Pre-Processing

This is the stage during modeling, where the data is prepared to be fed into a particular Machine Learning model.

This usually happens after analysis of the previously cleaned data, so here it is assumed the data is free from null values and outliers.

During this stage, the data scientist (in this case myself) is already aware of the Machine Learning models he/she will implement, and will curate the data specifically to cater for the model’s needs.

This is so as algorithms such as random forests do not need the data to be normalized/standardized etc.

The following tasks took place in this stage :

* Feature encoding
* Feature selection
* Normalization / standardization
* Data splitting

### Feature Encoding

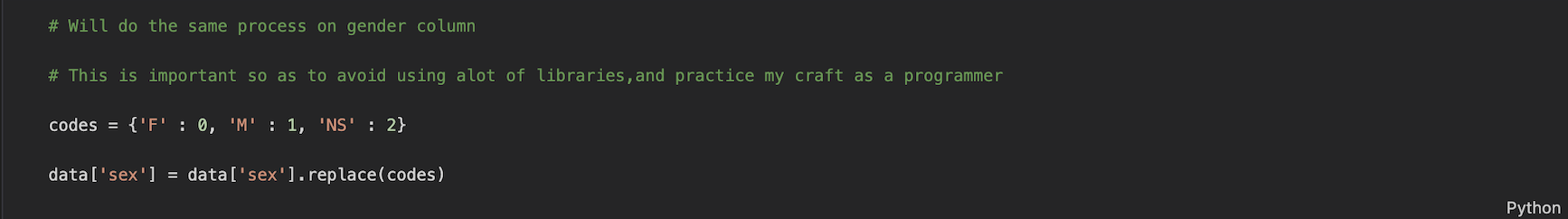
This is where categorical data that is in string form (Yes/No, M/F) is transformed to numerical data (1/0)

This is crucial as Machine Learning models can only be fed numeric data.

In the Data Science community, there exists numerous libraries that help in encoding. But so as to avoid using libraries, we went ahead and initialized codes per category using python dictionaries. To avoid redundancy, initialized the dictionary in a python function that runs the codes per column name.

The codes were then mapped onto the data, where any occurrence of **t(true)** was replaced with **1** and any occurence of **f(false)** was replaced with **0.**

The same was done in the gender column, where **M(male)** was replaced with **1**, **F(Female)** with **0** and **NS(Not Specified**) replaced with **2**

****

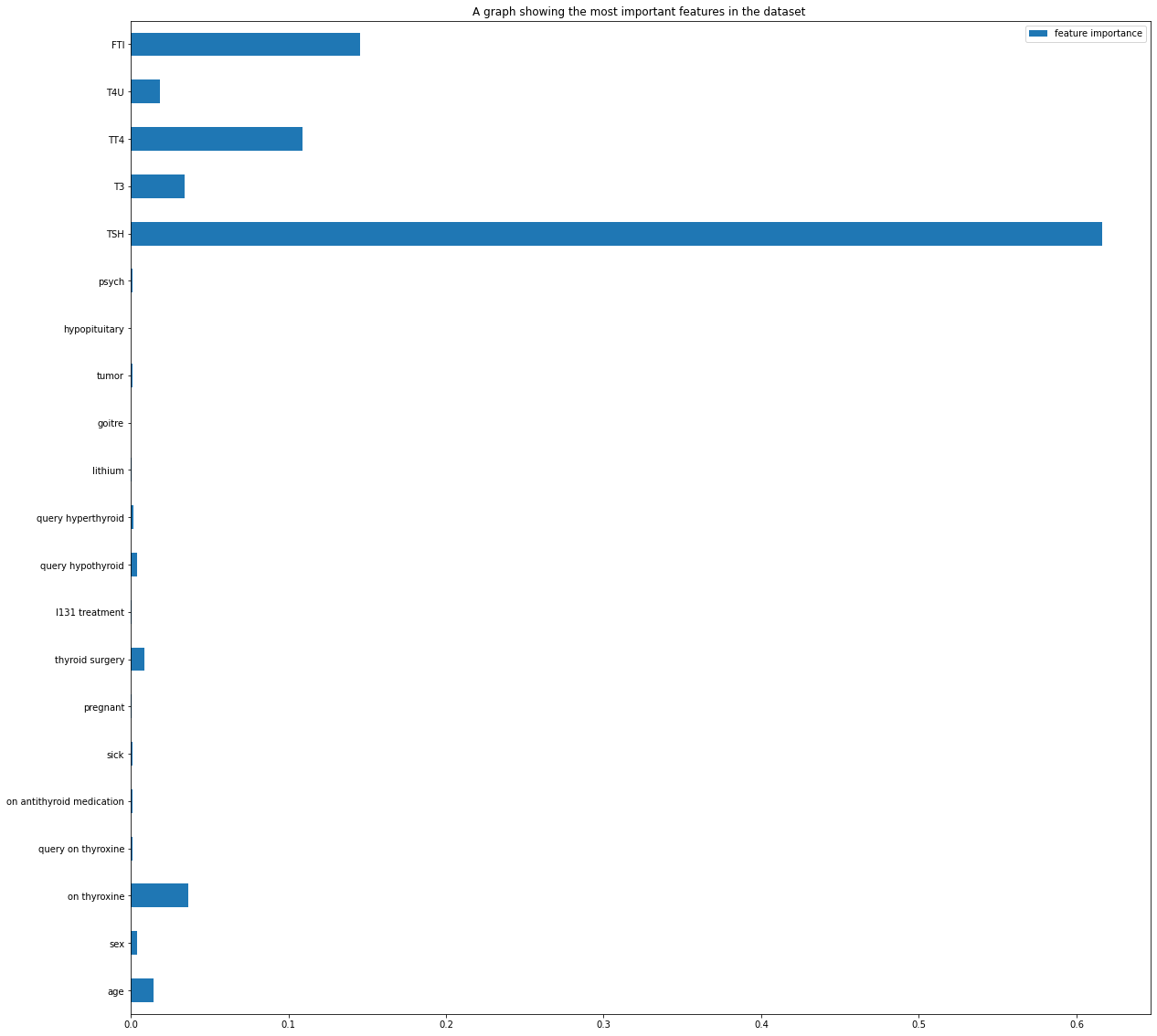
### Feature Reduction

Feature selection is the pre-processing technique where the most important features are chosen to be input into the model. These ‘important features’ are the columns that contribute the most in prediction. This is so as to :

* avoid unnecessary features being input into the model.
* Save time in deployment as the end user can put less inputs and get a similar result.

As it is difficult and almost impossible to predict which features contribute the most in prediction, what I did is build a quick random forest model. The Random Forest model in sklearn has a neat feature that allows the developer to see the most important features, and they can even be plotted.

This is exactly what I did and here was the outcome :



As seen, the blue bars represent how ‘important’ the column is to prediction. The top 5 or 6 features were picked for modeling and they were :

1. TSH (Most important)
2. T3
3. TT4
4. T4U
5. FTI
6. On thyroxine
7. Age (Least important)

### Feature Scaling / Standardization

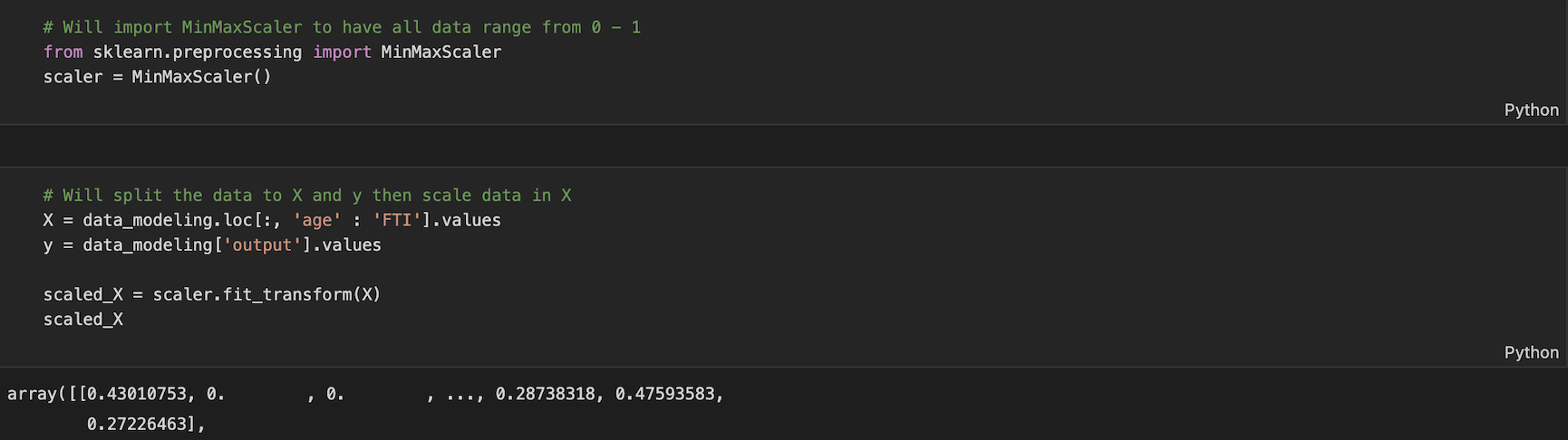
Data is usually in different scales and measurements such as meters and kilograms, age and weight etc. Raw data fed into a Machine Learning in this form often leads to the model performing extremely poorly.

A technique that happens to solve this is the data being scaled/standardized into a fixed range so as to handle the highly varying magnitude of data e.g. in our data **age** and **FTI.**

The technique implemented in our data was the Min-Max Scaling. This is where all the data is fixed into a range of 0 - 1. This is the mathematical formula.



IN our code, we imported the MinMaxScaler model from preprocessing and after initialization and splitting the data into X(features) and y(label), fitted it with our model and transformed X using fit\_tranform.



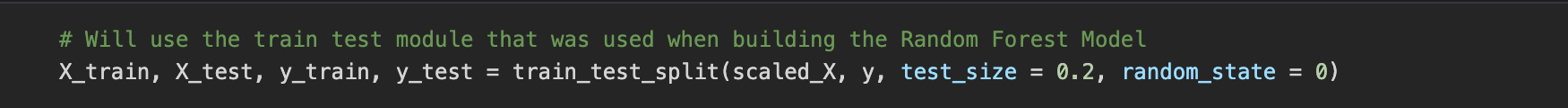
As seen by the first array output, all the data is in a range of 0 - 1.

### Data Splitting

This is usually the very last process before starting to train the model where the data is split into training data and testing data.

The training data (usually 80% of the data) is used to train the model and the testing data (20% of the data ) is used to test the model when evaluating the model’s accuracy, RMSE score etc.

Used a random state of 0 so as to allow the splitting to be random and to avoid bias when splitting.



The data now in the form of X\_train, X\_test, y\_train and y\_test are now ready to be used in the Machine Learning model of our liking.

## Machine Learning

Here is where Machine Learning comes to play. Machine Learning models are implemented and the best performing one chosen for deployment. It is said by Data Scientists that this is where all the hard work finally pays off.

**CHOOSING WHICH MODEL TO USE**

In our case, the data does not have too many features, which eliminates the need of a neural networking model.

The data is labeled, so supervised learning models will be necessary.

The models to be used include :

1. Naive Bayes
2. Random Forest

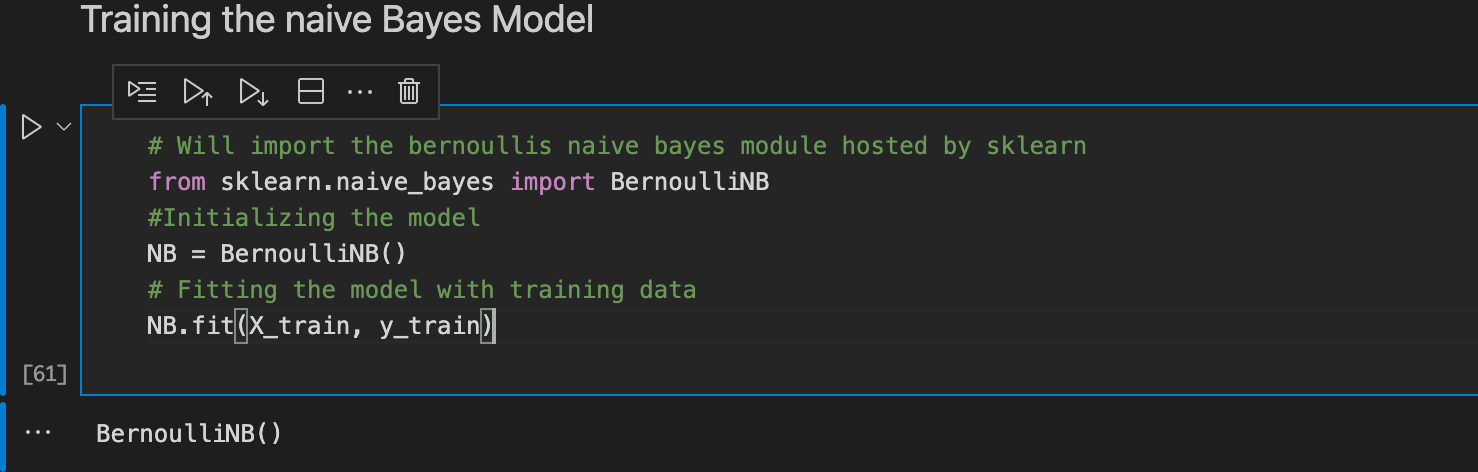
### Naive Bayes

This is a predominantly classification algorithm that aims to calculate the probability of on escenario (in this case if a patient has hypothyroid disease) based on its relationship with another scenario(their ages, TSH score etc)

As our output is in bernoulli (0/1), we will use Bernoulli NB module.

#### Training

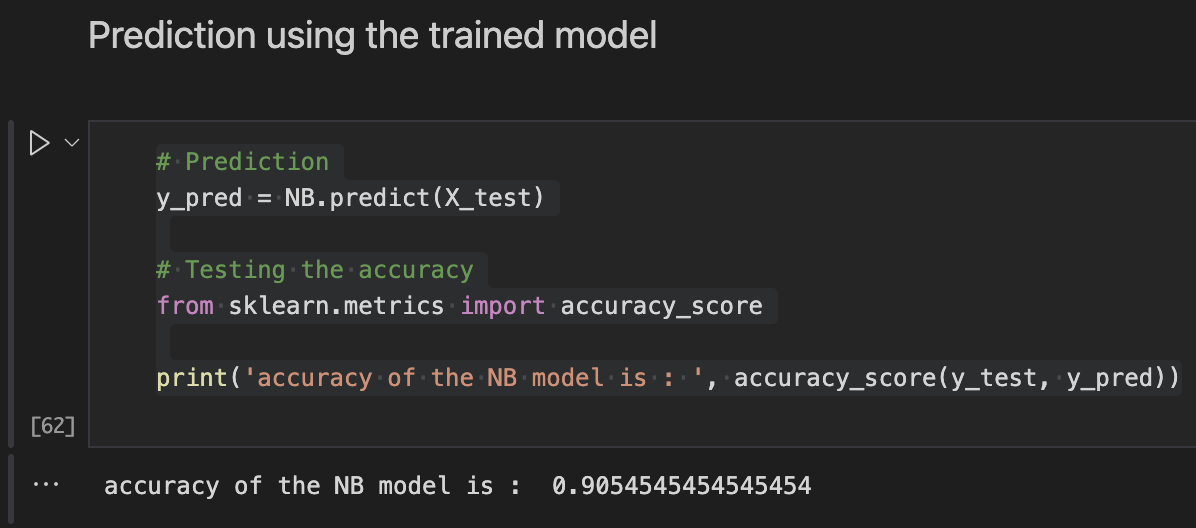
The model was called from SKLearn and the data was fed into the model.



#### Testing

After the NB model is trained, a y\_pred variable was declared where the model will attempt to predict the outcome of the disease using the test set data.

The y\_pred answers were compared with the test data and an accuracy of 90.5% was found, using accuracy score.



However, after vigorous analysis, I found out that the outcomes were all 0, which is inaccurate, despite the accuracy scores.

### Random Forest

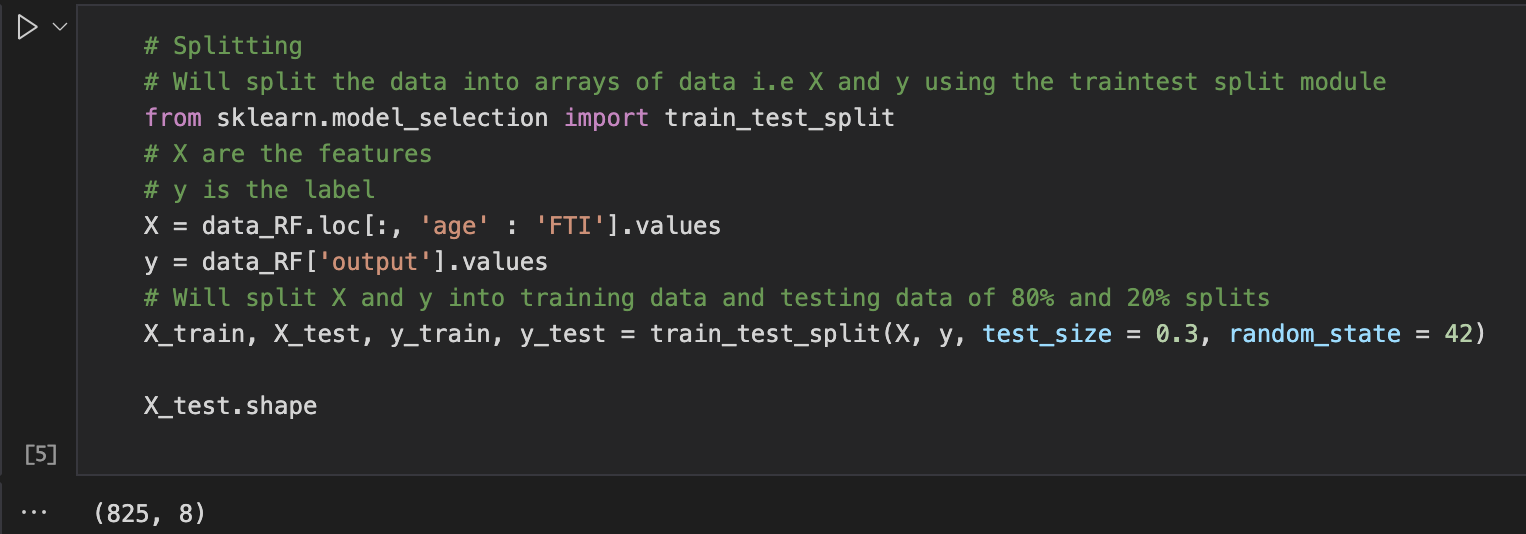
Random Forest is a popular Supervised ML algorithm that I used and was confident in.

How it works is by building multiple decision trees and taking the majority vote as the output.

Advantages of RF over Decision Trees

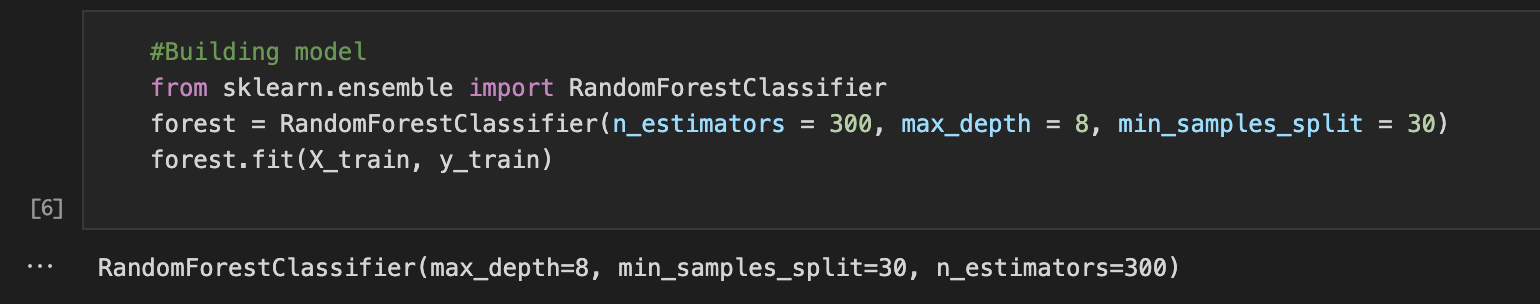
1. Decision Trees are prone to overfitting, unlike Random Forest which takes the most popular voting of multiple sub trees, therefore eliminating the possibility of overfitting.
2. Splitting

The data is split into training data and testing data



#### Training

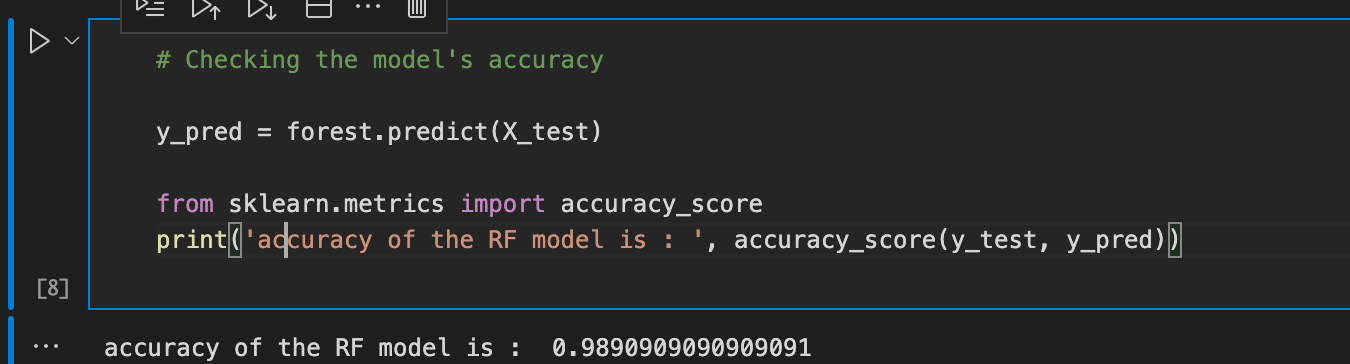
The model is called from ensemble in Sklearn, and the training data is used to fit the model.



#### Testing

Using the testing data, the model is used to predict and the predicted values are compared with the Original output.

Accuracy score is used to get the percentage of the accuracy.



The model has a 98% accuracy, which is fantastic news.

The model was then exported and used in the website, to predict outcomes of the disease.

# DEPLOYMENT

Here is the stage where all the hard work, sweat and tears usually pays off. The built model should be able to take external/user defined data and provide an output.

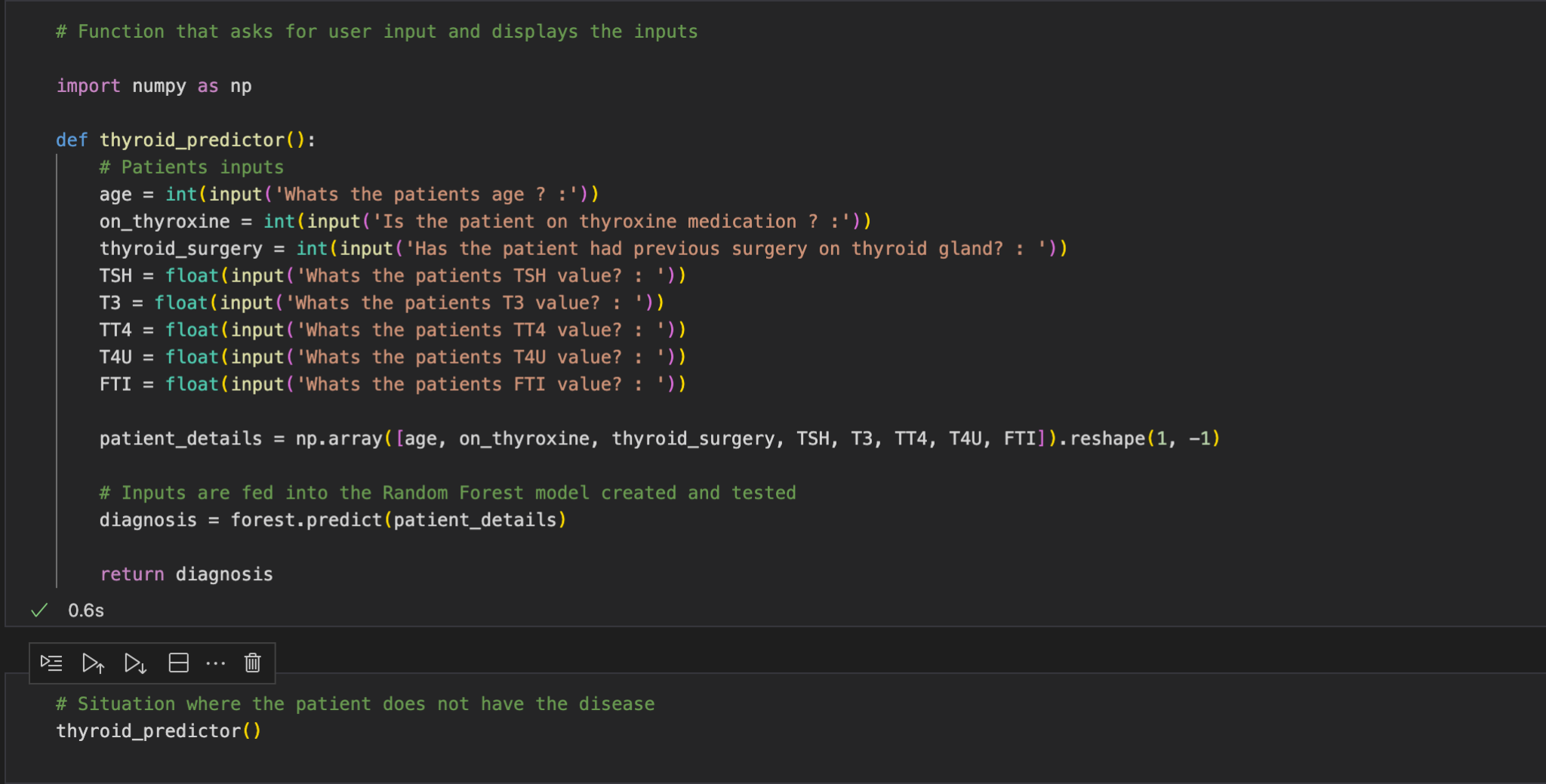
This is so provided the number of inputs are exactly the same as the number of inputs used to fit the model. The data types should be compatible as well and the programmer should have code that preprocesses the foreign data if necessary.

## Draft Deployment

This was done on the python notebook, where immediately after building the model and checking its accuracy, I wrote a python function that asks the user for raw data, (can guess if need be) and the data input by the user is formatted as an array and fed into the model.

The data should have 8 inputs, all numerical values.

Here was the outcome :



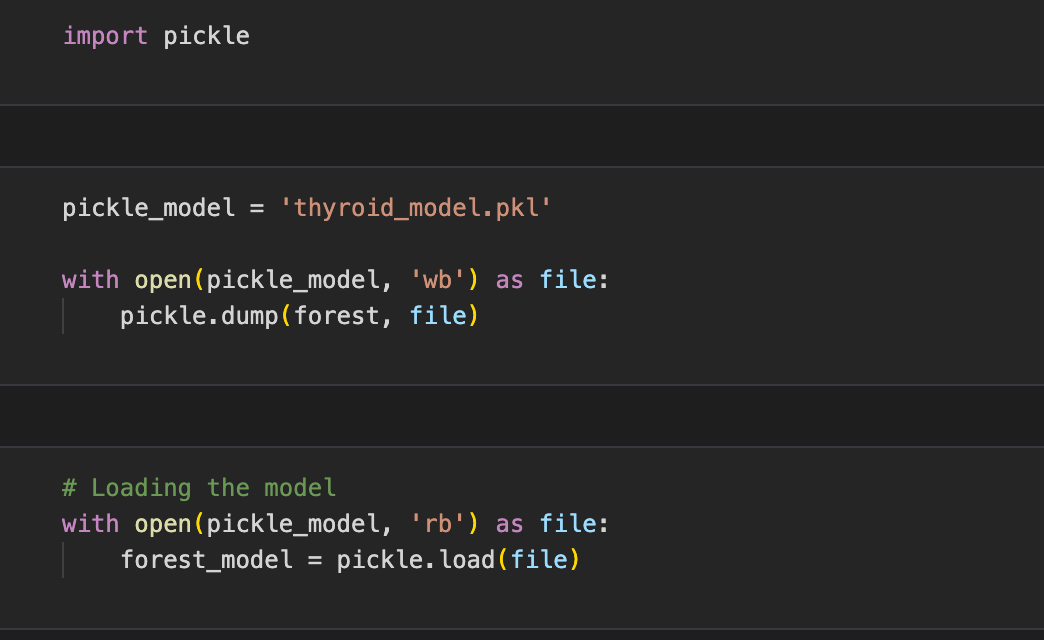
First 3 inputs are integers, and the rest are in float format(can use decimals)

## Exporting Model

So as to ensure the model can work anywhere at any time, it was required to be exported from the notebook so as to be a separate entity.

The model was exported using pickle, which is a popular python library used for model exportation.

The final model was a portable 13kb file, which can do wonders despite its mere size.



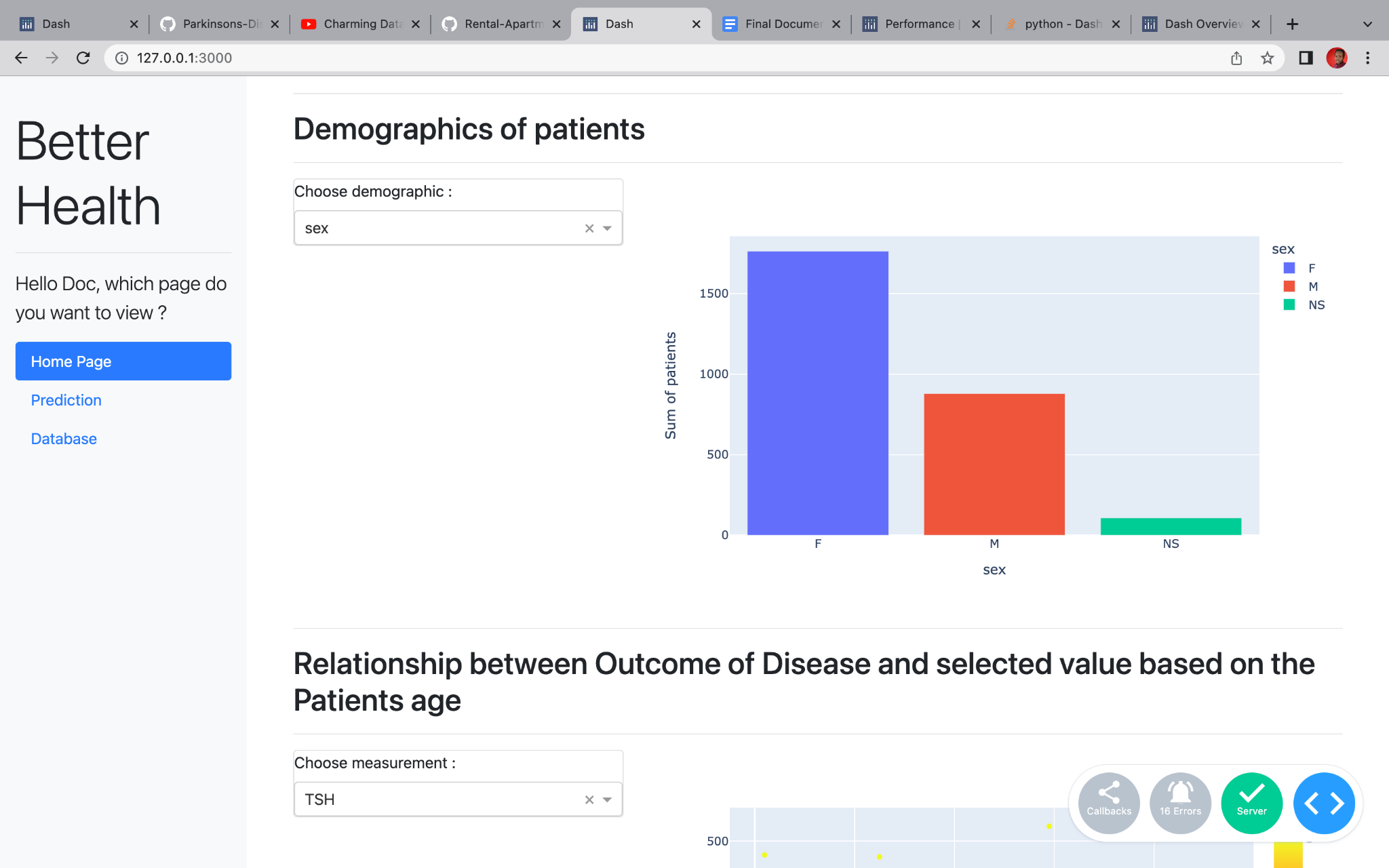
## Dashboarding

To ensure data scientists can show all the analysis done in a presentable manner, a platform called Plotly dash was developed to ensure they can quickly develop websites and portray their visualizations in a presentable manner.

It's the platform I used to deploy my project as well, making use of intractable bar charts, scatter plots and disease prediction in a whole entire page, on the same website.

### Analysis Dashboard

Here the Exploratory Data Analysis charts explained in the previous section are deployed in a presentable manner, where a drop down was implemented to allow toggling between different features and gather information, all within a singular graph.

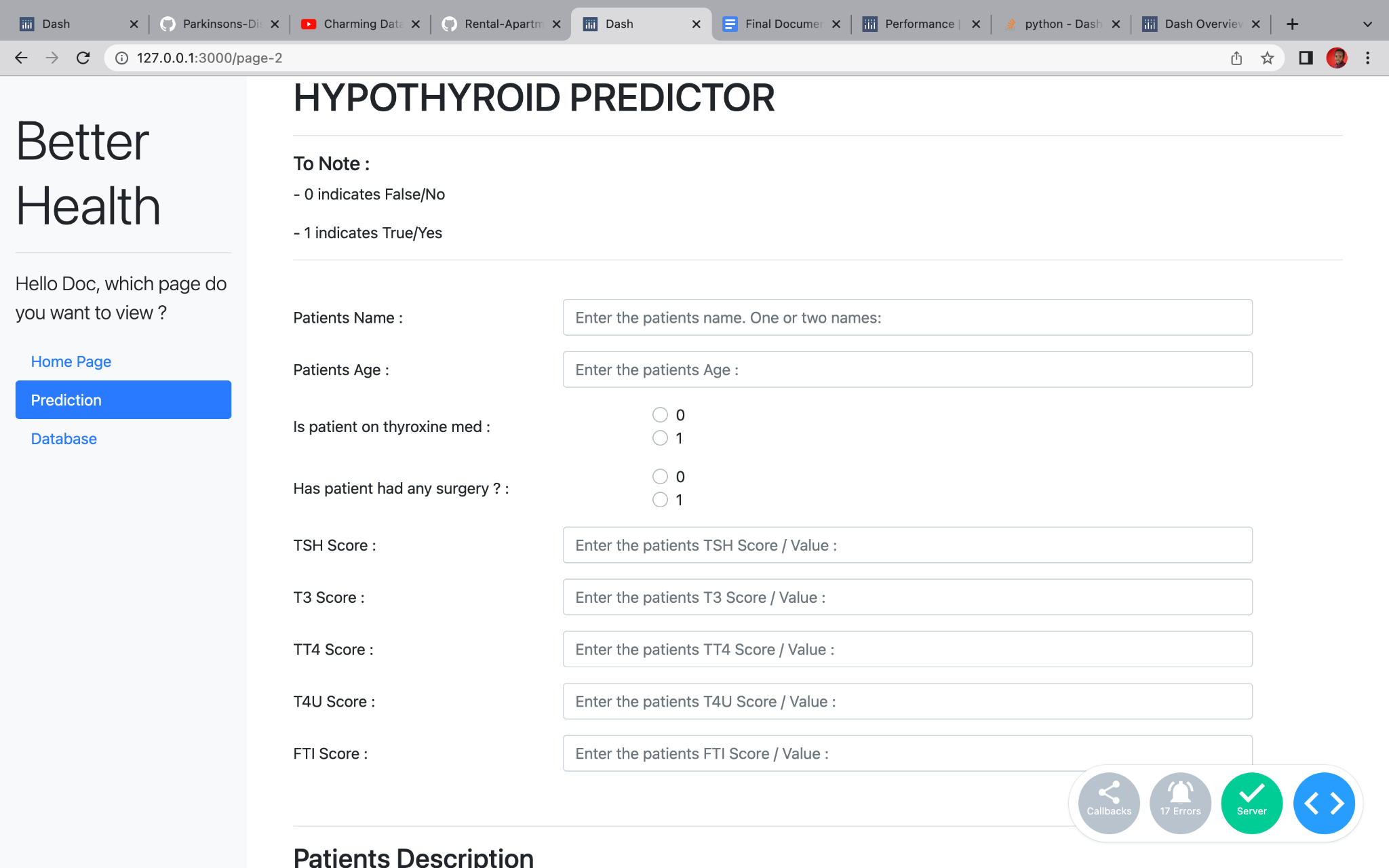


### Prediction Dashboard

Here is where the Machine Learning model developed is practically used, where a user, whether a doctor or anyone can input information to his/her desire.

The info is fed into the ML model in the backend, a prediction is made and presented for the end user to view.

Very interesting.



# CONCLUSION

This was primarily a data science project, combined with some web development for my 2nd year project retake assessment.

The objectives were met on the stipulated dates and the model was built successfully and deployed on a dashboard.

More practice on Django would mean that a better and modern website would be developed that can be used in medical institutions, and is an area I will work on as a student.

Thank you.

## 