THYROID DISEASE PREDICTION USING MACHINE LEARNING

A Project Report submitted in the partial fulfilment of the requirements for the award of the degree

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

I. Sowkya	**(19471A0525)**
CH. Mounika	**(19471A0509)**
D.Neha Sree	**(19471A0517)**

Under the esteemed guidance of

Y. Chandana MTech,
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPETA

(AUTONOMOUS)

Accredited by NAAC with A+ Grade and NBA under Cycle -1 Approved by AICTE, New Delhi, Permanently Affiliated to JNTUK, Kakinada

KOTAPPAKONDA ROAD, YALAMANDA VILLAGE, NARASARAOPET-522601

2022-2023

NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPETA
(AUTONOMOUS)
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the main project entitled "THYROID DISEASE PREDICTION USING MACHINE LEARNING" is a bonafide Work done by **I. Sowkya (19471A0525), CH. Mounika (19471A0509), D. Neha Sree (19471A0517),** in partial fulfilment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY** in the Department of **COMPUTER SCIENCE AND ENGINEERING** during 2022-2023.

PROJECT GUIDE
Y. Chandana MTech
Asst. Prof.

PROJECT CO-ORDINATOR
Dr. M. Sireesha MTech., Ph.D.
Assoc. Prof.

HEAD OF THE DEPARTMENT
Dr. S. N. TirumalaRao **M.Tech, Ph.D.**

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We wish to express our thanks to carious personalities who are responsible for the completion of the project. We are extremely thankful to our beloved chairperson **Mr. M.V.Koteswara Rao**, B.sc who took keen interest on us in every effort throughout this course. We owe out gratitude to our principal **Dr.M.Sreenivasa Kumar**, **M.Tech, Ph.D(UK),MISTE,FIE(1)** his kind attention and valuable guidance throughout the course.

We express our deep felt gratitude to **Dr.S.N.Tirumala Rao**, M.Tech, Ph.D. H.O. D, CSE department and our guide **Y. Chandana**, MTech, Assistant Professor of CSE department whose valuable guidance and unstinting encouragement enable us to accomplish our project successfully in time.

We extend our sincere thanks to **Dr.M.Sireesha** **M.Tech, Ph.D.** Associate Professor and Coordinator of the project for extending his encouragement. Their profound knowledge and willingness have been a constant source of inspiration for us throughout this project work.

We extend our sincere thanks to all other teaching and non-teaching staff to department for their cooperation and encouragement during our B. Tech degree. we have no words to acknowledge the warm affection, constant inspiration and encouragement that we receive from our parents.

We affectionately acknowledge the encouragement received from our friends and those who involved in giving valuable suggestions had clarifying out doubts, which had really helped us in successfully completing our project.

By

I. Sowkya **(19471A0525)**

CH. Mounika **(19471A0509)**

D.NehaSree **(19471A0517)**

ABSTRACT

In medical field, the salient and demanding task is to diagnose patient's health conditions and to provide proper care and treatment of the disease at the initial stage. Classification based Machine learning plays a major role in various medical services. Let us consider Thyroid disease as the example. The main goal is to recognize the disease at the early stages with a very high correctness. Thyroid disease diagnosis is not a simple task. It involves many procedures. The normal traditional way includes a propermedical examination and many blood samples for blood tests

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

INSTITUTE VISION AND MISSION

INSTITUTION VISION

To emerge as a Centre of excellence in technical education with a blend of effective student centric teaching learning practices as well as research for the transformation of lives and community,

INSTITUTION MISSION

M1: Provide the best class infra-structure to explore the field of engineering and research

M2: Build a passionate and a determined team of faculty with student centric teaching, imbibing experiential, innovative skills

M3: Imbibe lifelonglearningskills, entrepreneurial skills and ethical values in students for addressing societal problems

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION OF THE DEPARTMENT

To become a centre of excellence in nurturing the quality Computer Science & Engineering professionals embedded with software knowledge, aptitude for research and ethical values to cater to the needs of industry and society.

MISSION OF THE DEPARTMENT

The department of Computer Science and Engineering is committed to

M1: Mould the students to become Software Professionals, Researchers and Entrepreneurs by providing advanced laboratories.

M2: Impart high quality professional training to get expertise in modern software tools and technologies to cater to the real time requirements of the industry.

M3: Inculcate team work and lifelong learning among students with a sense of societal and ethical responsibilities.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Specific Outcomes (PSO's)

PSO1: Apply mathematical and scientific skills in numerous areas of Computer Science and Engineering to design and develop software-based systems.

PSO2: Acquaint module knowledge on emerging trends of the modern era in Computer Science and Engineering

PSO3: Promote novel applications that meet the needs of entrepreneur, environmental and social issues.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Educational Objectives (PEO's)

The graduates of the programme are able to:

PEO1: Apply the knowledge of Mathematics, Science and Engineering fundamentals to identify and solve Computer Science and Engineering problems.

PEO2: Use various software tools and technologies to solve problems related to academia, industry and society.

PEO3: Work with ethical and moral values in the multi-disciplinary teams and can communicate effectively among team members with continuous learning.

PEO4: Pursue higher studies and develop their career in software industry.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Outcomes

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis:** Identify, formulate, research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.

6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

NARASARAOPETA
NEC ENGINEERING COLLEGE
(AUTONOMOUS)

Project Course Outcomes (CO'S):

CO425.1: Analyse the System of Examinations and identify the problem.

CO425.2: Identify and classify the requirements.

CO425.3: Review the Related Literature

CO425.4: Design and Modularize the project **CO425.5:**

Construct, Integrate, Test and Implement the Project.

CO425.6: Prepare the project Documentation and present the Report using appropriate method.

Course Outcomes – Program Outcomes mapping

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1		✓											✓		
C425.2	✓		✓		✓								✓		
C425.3				✓		✓	✓	✓					✓		
C425.4			✓			✓	✓	✓					✓	✓	
C425.5					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
C425.6									✓	✓	✓		✓	✓	

Course Outcomes – Program Outcome correlation

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1	2	3											2		
C425.2			2		3								2		
C425.3				2		2	3	3					2		
C425.4			2			1	1	2					3	2	
C425.5					3	3	3	2	3	2	2	1	3	2	1
C425.6									3	2	1		2	3	

Note: The values in the above table represent the level of correlation between CO's and PO's:

1. **Low level**

2. **Medium level**

3. **High level**

Project mapping with various courses of Curriculum with Attained PO's:

Name of the course om which principles are applied in this project	Description of the device	Attained PO
C3.2.4, C3.2.5	Gathering the requirements and defining the problem, plan to develop a customer segmentation	PO1, PO3
CC4.2.5	Each and every requirement is critically analyzed, the process model is identified and divided into four modules	PO2, PO3
CC4.2.5	Logical design is done by using the unified modelling language which involves individual team work	PO3, PO5, PO9
CC4.2.5	Each and every module is tested, integrated, and evaluated in our project	PO1, PO5
CC4.2.5	Documentation is done by all our four members in the form of a group	PO10
CC4.2.5	Each and every phase of the work in group is presented periodically	PO10, PO11
CC4.2.5	Implementation is done and the project will be handled by the mall customers it is developed to the prediction of customers in the market.	PO4, PO7
CC4.2.8 CC4.2.	The physical design includes the website to check the future data of the customer.	PO5, PO6

INDEX

LIST OF FIGURES

1. INTRODUCTION

1.1 Introduction

Thyroid disease diagnosis is not a simple task. It involves many procedures. The normal traditional way includes a proper medical examination and many blood samples forblood tests. Therefore, there is a necessity for a model which detects the thyroid disease at a very early stage of development.

In medical field machine learning plays an important role for thyroid disease diagnosis as it has various classification models based on which we can train our model with proper train dataset of the thyroid patient and can predict and give the results in an accurate manner with higher degree of correctness.

Some recent studies from Mumbai have suggested that congenital hypothyroidism is common in India. The disease occurs in 1 part of 2640 new born children, when compared to the worldwide average range of 1 in 3800 considered. Congenital hypothyroidism can lead to serious complications if not detected in early stages.

Therefore, the proposed model serves the goal in early detection of thyroid disease. Based on the obtained test values the health care staff can easily examine the condition of the patient and also skip further clinical examinations if not necessary. Hence, this approach proves to be very muchbeneficial to the healthcare field.

A proper train dataset results into an accurate predicting model therefore reducing the overall cost of the thyroid patient treatment and also saving the time. Classification algorithms are mostsuitable in decision-making and also solving the real-world problems.

1.2 Existing System

Prediction using traditional disease risk model usually involves a machine learning and supervised learning algorithm which uses training data with the labels for the training of the models. High-risk and Low-risk patient classification is done in groups test sets. But these models are only valuable in clinical situations and are widely studied. A system for sustainable health monitoring using smart clothing by Chen et.al. He thoroughly studied heterogeneous systems and was able to achieve the best results for cost minimization on the tree and simple path cases for heterogeneous systems.

Disadvantages:

1.Doesn't generate accurate and efficient results.

2.Computation time is very high.

3.Lacking of accuracy may result in lack of efficient further treatment.

1.3 Proposed System

In the proposed system, a disease prediction model is built using a Machine Learning algorithm that is Random Forest Algorithm and many. Based on the symptoms that are input by the user, the disease is predicted and the drug that is most commonly prescribed by the doctor is suggested.

Advantages:

1. Generates accurate and efficient results.
2. Computation time is greatly reduced.
3. Data replica.
4. Data cleaning and data repository.

1.4 System Requirements

1.4.1 Hardware Requirements

- System Type: intel®core™i5-7500UCPU@1.03gh
- Cache memory: 4 Megabyte (MB)
- RAM: 4 Gigabyte (GB)
- Hard Disk: 1 Terabyte (TB)

1.4.2 Software Requirements

- Operating System: Windows 10 Home, 64-bit Operating System
- Coding Language: Python
- Python distribution: Anaconda(pycharm), Google Colab

2.LITERATURE SURVEY

2.1 Machine Learning

It has proposed different Thyroid prediction techniques using data mining approaches. They have considered different dataset attributes for prediction and have explained the classification techniques in data mining like Decision Tree, Backpropagation Neural Network, SVM and density-based clustering. They have analyzed the correlation of T3, T4 and TSH with hyperthyroidism and hypothyroidism.

They have studied various classification-based machine learning algorithms. They have considered train data set from UCI Machine Learning repository and compared and analyzed the performance metric of decision tree, support vector machine and K-nearest neighbor.

They have proposed a training model consisting of 21 thyroid causing attributes. They have proposed partial swarm optimization to optimize the support vector machine parameters.

They have performed a general empirical study on various disease diagnosis like Diabetes, Breast
Cancer, Heart disease, Thyroid prediction and have compared the accuracy rate by applying SVM, Decision tree and Artificial Neural Networks.

It considered Thyroid data reprocessing mainly by applying the decision tree algorithm. They have first calculated the mean values of T3, T4 and TSH and considered as the pre-processing stage. Later on, they have applied machine learning based feature selection and feature construction. Further they have applied classification based J48 algorithm which is a continuation ofID3 algorithm and calculated the results.

They analyzed a comparison on various classification methods used to diagnose thyroid disease. They have compared by using Artificial Neural Networks, Radial Based Function, Learning Vector Quantization, Back Propagation Algorithm and Artificial Immune recognition system and concluded the comparison results. Among that they found out that Multilayer Perceptron has the highest accuracy of 96.74%.

Have proposed a Thyroid Prediction System based on data mining classification algorithm. They have used random forest approach to predict the results using Weka open-source tool used for data mining. Using this tool, they have applied random forest algorithm .

Have conducted a study on different data mining techniques to detect thyroid disease. They have done study on Linear Discriminant analysis, Kfoldcross validation, and Decision tree. They have analysed various splitting rules for the attributes of Decision tree. They have also compared theobtained values.

It conducted a study on diagnosis of the thyroid disease using different data mining approaches. They have explained the major cause of the thyroid diseaseand havealso given description aboutDecision Tree, Naïve Bayes classification and SVM.

It performed a Prediction on Thyroid Disease using various machine learning techniques. They have considered Logistic Regression and Support Vector Machine as the main Thyroid detection models. They have concluded that these two proposed classifier methods are the best when the number of classes increases in the thyroid prediction model.

2.2 Some machine learning methods

Machine learning algorithms are often categorized as supervised and unsupervised.

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis o f a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.
- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior's within a

specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best. This is known as the reinforcement signal.

2.3 Applications of machine learning

1. Real-time chatbot agents
2. Decision support
3. Customer recommendation engines
4. Dynamic pricing tactics
5. Customer segmentation
6. Fraud Detection
7. Text parsing
8. Image Classification
9. Improving Cyber security
10. Detecting Spam

3. SYSTEM ANALYSIS

3.1 System Architecture

Initially we will see the dataset and then we will perform exploratory data analysis which deals with the missing data, duplicates values and null values. And then we will deploy our algorithm k-means clustering which is unsupervised learning in machine learning.

As in order to find the no of clusters we use elbow method where distance will be calculated through randomly chosen centers and repeat it until there is no change in cluster centers. Thereafter we will analyse the data through data visualization. Finally, we will get the outcome.

3.2. Importance of machine learning in Thyroid disease prediction

Machine learningcan play a crucial role in the prediction and diagnosis of thyroid diseases. Thyroid diseases affect the thyroid gland, which is responsible for producing hormones that regulate metabolism, growth, and development in the body. Machine learningalgorithms can be trained on large datasets of patient information to predict the likelihood of d eve loping thyroid diseases, diagnose patients, and recommend treatment plans.

One important application of machine learning in thyroid disease prediction is the use of predictive models to identify patients who are at high risk of developing thyroid disorders. By analyzing patient data, such as medical history, laboratory results, and imaging studies, machine learning models can identify patterns and relationships that are predictive of thyroid diseases. This can enable early detection and timely intervention, which can improve patientoutcomes.

Machine learning algorithms can also be used to aid in the diagnosis of thyroid diseases. For example, machine learning models can be trained to analyse ultrasound images of the thyroid gland to detect abnormalities, such as nodules or tumours, that may indicate thyroid cancer. Machine learning algorithms can also be used to analyse laboratory test results, such as thyroid hormone levels, to diagnose thyroid disorders and recommend appropriate treatment.

Overall, the use of machine learning in thyroid disease prediction can lead to more accurate diagnoses, earlier intervention, and improved patient outcomes.

3.3 Implementation of machine learning using Python

Python is a popular programming language. It was created in 1991 by Guido van Rossum.

The most recent major version of Python is Python 3. However, Python 2, although not being updated with anything other than security updates, is still quite popular.

It is possible to write Python in an Integrated Development Environment, such as Thonny, PyCharm, NetBeans or Eclipse, Anaconda which are particularly useful when managing larger collections of Python files.

Python was designed for its readability. Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.

Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

In the older days, people used to perform Machine Learning tasks manually by coding all the algorithms and mathematical and statistical formula. This made the process time consuming, tedious and inefficient. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reasons is its vast collection of libraries. Python libraries that used in Machine Learning are:

1. Scikit learn

2. NumPy

3. Pandas

4.Matplotlib

5.Seaborn

1. Scikit-learn

It is a free Python machine learning software, sometimes known as sklearn. It is meant to interact with the Python numerical and scientific libraries NumPy and SciPy, and features support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering algorithms.

2. NumPy

NumPy is a general-purpose array-processing package. It provides a high performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones: A powerful N-dimensional array object

- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multidimensional container of generic data. Arbitrary data-types can be defined using NumPy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

3. Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyse. Python with Pandas is used in a wide range of fields including academic and commercial domains includingfinance, economics, Statistics, analytics, etc.

4. Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and I Python shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc., via an object-oriented interface or via a set of functions familiar to MATLAB users.

5. Seaborn

Seaborn is a library for making statistical graphics in python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

3.3 Scope of the project

Thyroid Disease Prediction project needs a dataset and full information about it. Next we clean the dataset and learn about different types of features and rows. We apply different types of pre-processing techniques and clean the data.

It would typically involve developing a model that can accurately predict whether a person is at risk of developing a thyroid disorder based on various factors such as age, gender, family history, lifestyle, and medical history.

To accomplish this, the project may involve collecting and analyzing relevant data from patients, such as thyroid hormone levels, medical imaging results, and symptoms. The data would be used to train a machine learning algorithm that can predict the likelihood of a patient developing a thyroid disorder.

The ultimate goal of the project would be to improve the accuracy of thyroid disorder diagnosis and enable earlier intervention and treatment, potentially leading to better health outcomes for patients. The project could also help healthcare professionals prioritize patients for further testing or referral to specialists.

3.4 Analysis

Thyroid disease dataset is taken from Kaggle platform. Thyroid dataset is an interesting one, because it has both numerical features that are most important to know about patient's condition and categorical features which is very easy to know in detail about patients' condition. The main features are selected using some pre-processing techniques that effect on result, they are Age, Gender, Anti-Thyroid Medication, T3, TT4, T4U and FTI. From these values and from other features like Goitre, Psych and Referral Source.

The data includes the following features:

1. Age
2. Gender
3. T3
4. T3_Measured
5. TT4, T4U, FTI levels
6. Refferal_source

Data Set:

	age	sex	on thyroxine	query on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery	I131 treatment	query hypothyroid	...	TT4 measured	TT4	T4U measured	T4U	FTI measured	FTI	TBG measured	TBG	referral source	binaryClass
0	41.0	F	f	f	f	f	f	f	f	f	...	t	125.0	t	1.14	t	109.0	f	NaN	SVHC	P
1	23.0	F	f	f	f	f	f	f	f	f	...	t	102.0	f	NaN	f	NaN	f	NaN	other	P
2	46.0	M	f	f	f	f	f	f	f	f	...	t	109.0	t	0.91	t	120.0	f	NaN	other	P
3	70.0	F	t	f	f	f	f	f	f	f	...	t	175.0	f	NaN	f	NaN	f	NaN	other	P
4	70.0	F	f	f	f	f	f	f	f	f	...	t	61.0	t	0.87	t	70.0	f	NaN	SVI	P
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3767	30.0	F	f	f	f	f	f	f	f	f	...	f	NaN	f	NaN	f	NaN	f	NaN	other	P
3768	68.0	F	f	f	f	f	f	f	f	f	...	t	124.0	t	1.08	t	114.0	f	NaN	SVI	P
3769	74.0	F	f	f	f	f	f	f	f	f	...	t	112.0	t	1.07	t	105.0	f	NaN	other	P
3770	72.0	M	f	f	f	f	f	f	f	f	...	t	82.0	t	0.94	t	87.0	f	NaN	SVI	P
3771	64.0	F	f	f	f	f	f	f	f	f	...	t	99.0	t	1.07	t	92.0	f	NaN	other	P

3772 rows × 30 columns

Fig:3.1 Dataset

3.5 Data Pre-processing

Before feeding data to an algorithm, we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing, the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

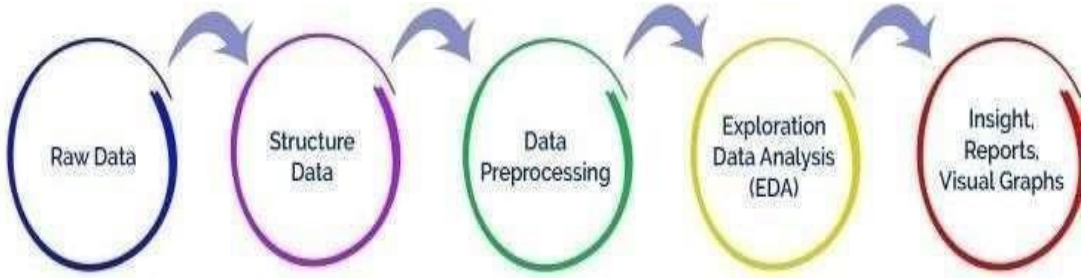


Fig: 3.2 Data Pre-processing

Need of Data Pre-processing: For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified f ormat. For example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set. Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

3.5.1 Missing values

Filling missing values is one of the pre-processing techniques. The missing values in the dataset is represented as '?' but it a non-standard missing value and it has to be converted into a standard missing value Nan. So that pandas can detect the missing values.

In our dataset, there are no missing values.

3.5.2 Correlation

A correlation matrix is a table that displays the correlation coefficients between variables. Each row and column in the table represents a variable, and the correlation coefficient between the variables is shown where the row and column intersect. The correlation coefficient measures the strength and direction of the linear relationship between two variables.

A correlation matrix is often used in statistics and data analysis to explore the relationships between multiple variables. It can help identify which variables are positively or negatively correlated, and which variables have little or no relationship with each other.

Some common uses of a correlation matrix include:

- Exploring the relationships between variables in a dataset
- Identifying which variables are most strongly correlated with each other
- Detecting multicollinearity, which is a situation where two or more variables are highly correlated with each other
- Selecting variables for use in regression analysis or other modelling techniques.

	age	TSH	T3	TT4	T4U	FTI	TBG
age	1.000000	-0.059087	-0.238412	-0.038841	-0.166250	0.052788	NaN
I?SH	-0.059087	1.000000	-0.161823	-0.267365	0.073391	-0.304684	NaN
I'3	-0.238412	-0.161823	1.000000	0.559503	0.454127	0.348921	NaN
I'I'4	-0.038841	-0.267365	0.559503	1.000000	0.434572	0.793312	NaN
I?4U	-0.166250	0.073391	0.454127	0.434572	1.000000	-0.174012	NaN
ÏI?I	0.052788	-0.304684	0.348921	0.793312	-0.174012	1.000000	NaN
I?BG	NaN	NaN	NaN	NaN	NaN	NaN	NaN

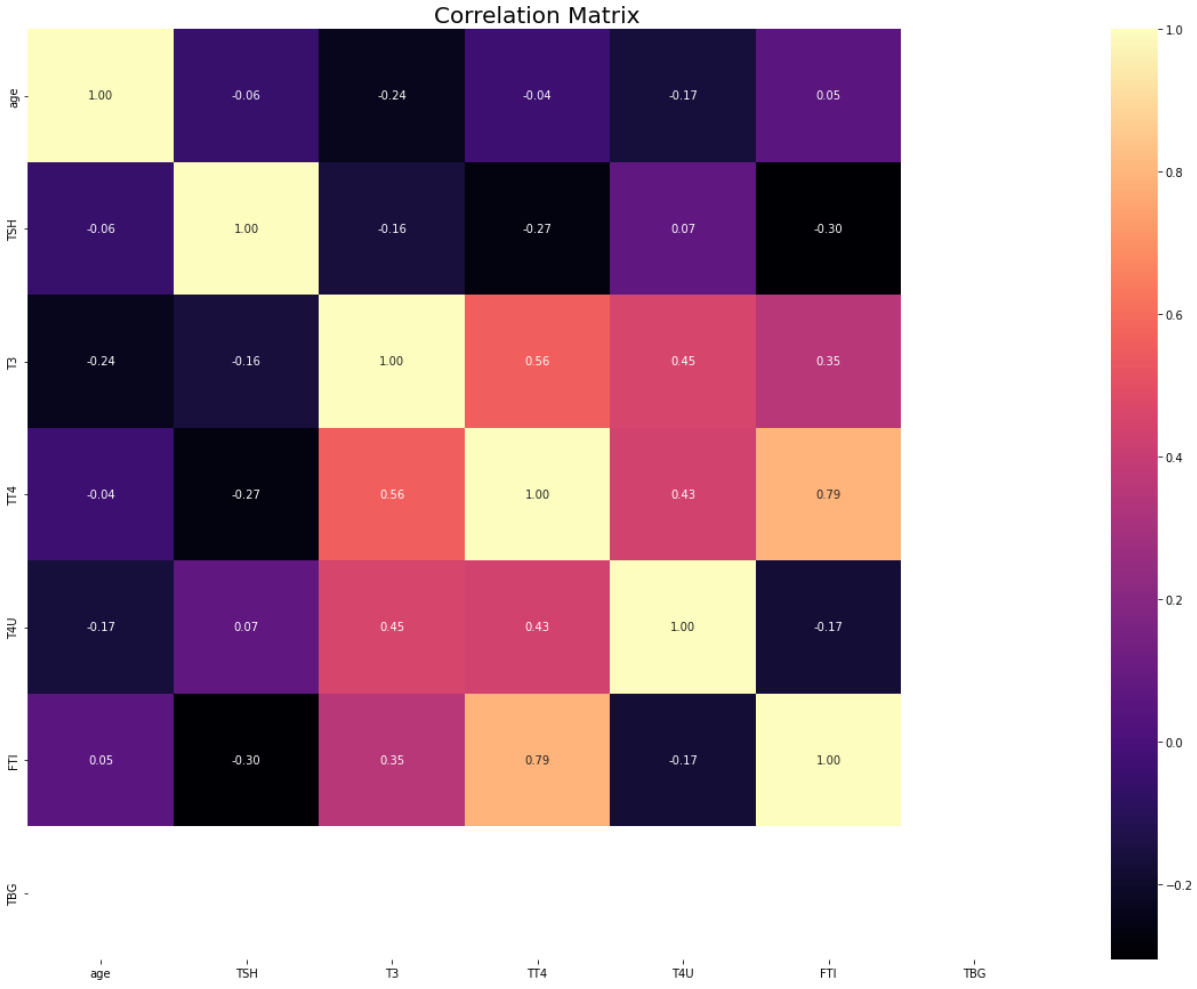


Figure for correlation matrix

3.5.3 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. It is one of the popular tools that is used for exploratory data analysis and predictive modelling. It is a technique to draw strong patterns from the given dataset by reducing the variances.

- PCA generally tries to find the lower-dimensional surface to project the high dimensional data.
- PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are image processing, movie recommendation system, customer segmentation optimizing the power allocation in various communication channels**.** It is a feature extraction technique, so it contains the important variables and drops the least important variable.

3.5.4 Visualize of Data

Plotting

Data Visualization is the process of presenting data in the form of graphs or charts. It helps to understand large and complex amounts of data very easily. It allows the decision-makers to make decisions very efficiently and also allows them in identifying new trends and patterns very easily. It is also used in high-level data analysis for Machine Learning and Exploratory Data Analysis (EDA). Data visualization can be done with various tools like Tableau, Power BI, Python.

Types of plots

Box Plot

Box plot gives statistical information about the distribution of numeric data divided into different groups. It is useful for detecting outliers within each group. A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having

properties like minimum, first quartile, median, third quartile and maximum. In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median. X-axis denotes the data to be plotted while the Y-axis shows the frequency distribution.

Scatter Plot

The scatter plots are preferred while comparing the data variables to determine the relationship between dependant and independent variables. The data is displayed as a collection of points, each having the value of one variable which determines the position on the horizontal axis and the value of other variable determines the position on the vertical axis. Scatter plots are used to plot data points on horizontal and vertical axis in the attempt to show how much one variable is affected by another. Each row in the data table is represented by a marker the position depends on its values in the columns set on the X and Y axes.

Visualizing the dataset according to various features using different types of plots. First, visualize the gender using the bar plot. Figure shows that the female count is higher than the male count.

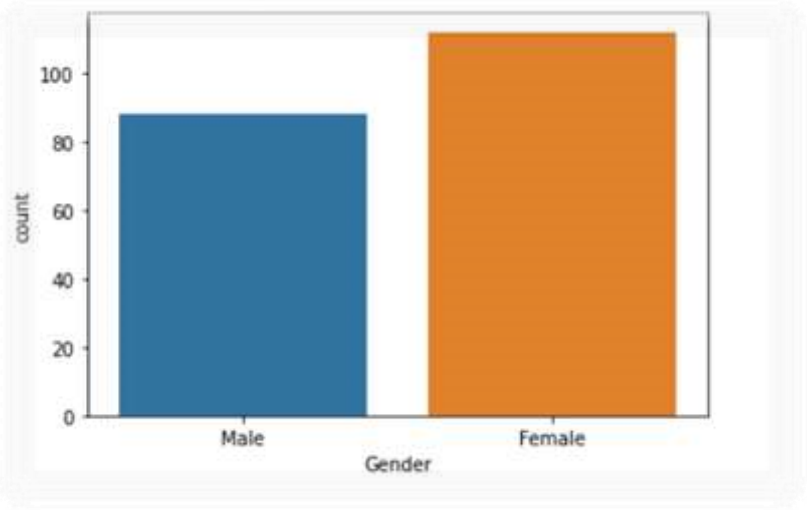


Fig: 3.3 Gender bias

After analyzing the age using boxplot and figure shows the highest age class clearly.

Histogram Plot

A histogram plot is a graphical representation of the distribution of a continuous numerical variable. It involves dividing the range of the variable into intervals or "bins" and then counting the number of observations that fall within each bin. The resulting plot shows thefrequency or count of the observations in each bin.

The histogram plot consists of vertical bars, where the height of each bar corresponds to thefrequency or count of the observations in that bin.

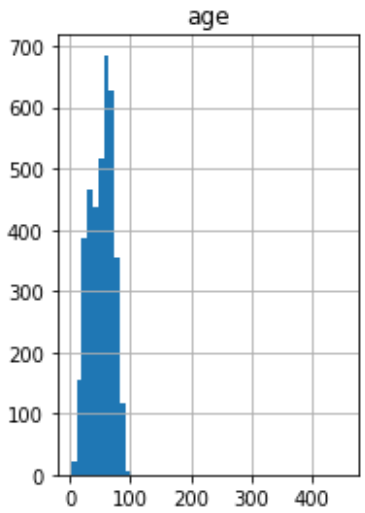
The bins are usually evenly spaced and non-overlapping, and the number of bins can be adjusted to best represent the distribution of the data.

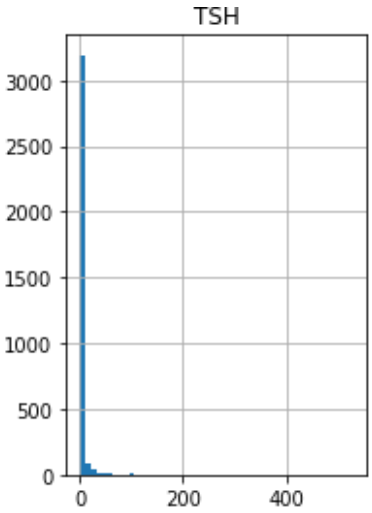
Histograms are useful for visualizing the shape of a distribution and identifying any outliersor unusual patterns.

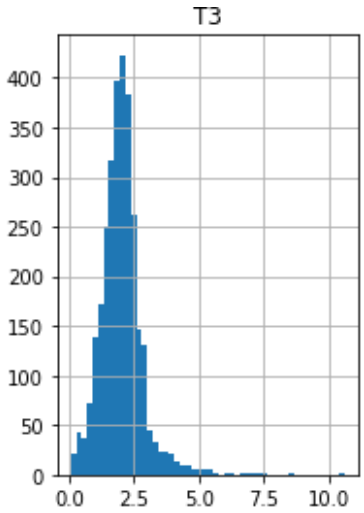
They are commonly used in exploratory data analysis and can provide insights into the underlying characteristics of the data. Histograms can also be used to compare the distributions of two or morevariable.

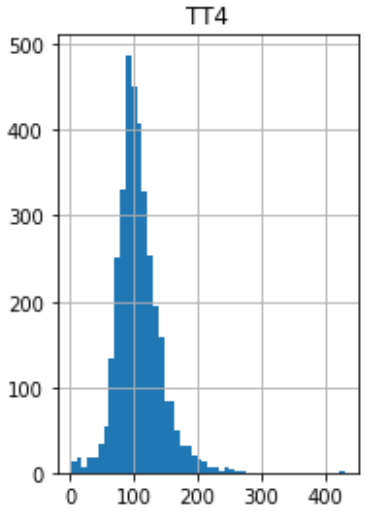
So, a doctor might know what are the important factors/attributes to look for but a Machine Learning system does not know this in advance meaning what the important attributes are the system doesn't know in advance and the features that help us identify the health risk in this case or in general the features that help us understand the output as a function of the input are termed as Discriminatory features**.**

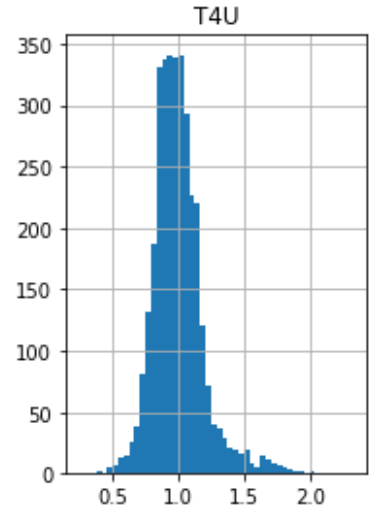
The data set would look like the below, where each column is one attribute and each row represent the data for one patient, their past records, we know whether had a health condition or not(let's say we got this data from some hospital we collaborated with).

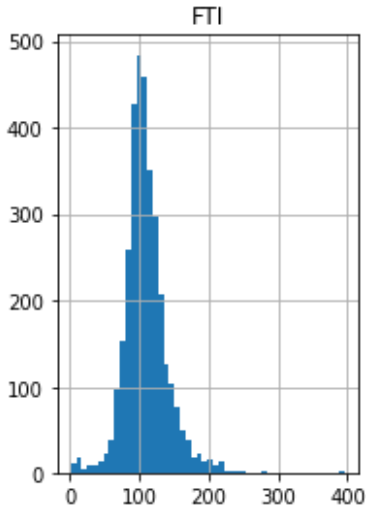

age
700
600
500
400
300
200
100
0
0
100
200
300
400

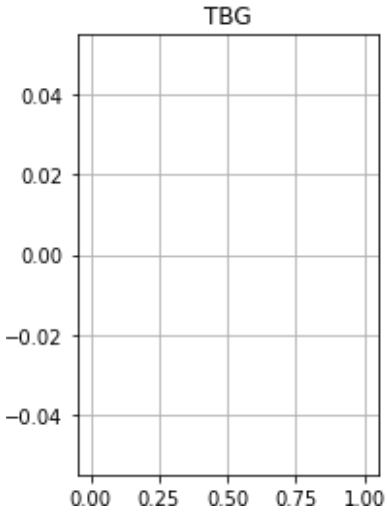

TSH
3000
2500
2000
1500
1000
500
0
0
200
400


T3
400
350
300
250
200
150
100
50
0
0.0
2.5
5.0
7.5
10.0


TT4
500
400
300
200
100
0
0
100
200
300
400


T4U
350
300
250
200
150
100
50
0
0.5
1.0
1.5
2.0


FTI
500
400
300
200
100
0
0
100
200
300
400


TBG
0.04
0.02
0.00
−0.02
−0.04
0.00
0.25
0.50
0.75
1.00

Bar Plot

A bar plot is a graphical representation of categorical data that shows the frequency or count of each category in a dataset. It involves plotting rectangular bars of equal width, where the height of each bar corresponds to the frequency or count of the category it represents.

The x-axis of a bar plot represents the categories, while the y-axis represents the frequency or count of each category. The bars can be arranged in any order, but they are usually orderedby frequency or in some other meaningful way.

Bar plots are useful for comparing the frequencies or counts of different categories in a dataset. They are commonly used in market research, social sciences, and other er fields to visualize survey responses, customer preferences, or demographic data.

In a bar plot, the height of each bar represents the value of a variable, and the width of the bar represents the category being measured. The bars are typically drawn vertically, but they can also be drawn horizontally.

The bars are usually spaced apart to indicate that they represent discrete categories, and they may be colored or labeled to indicate the categories they represent.

To create a bar plot, you first need to determine the categories you want to measure and the values associated with each category. Once you have this data, you can create a bar plot by following these steps:

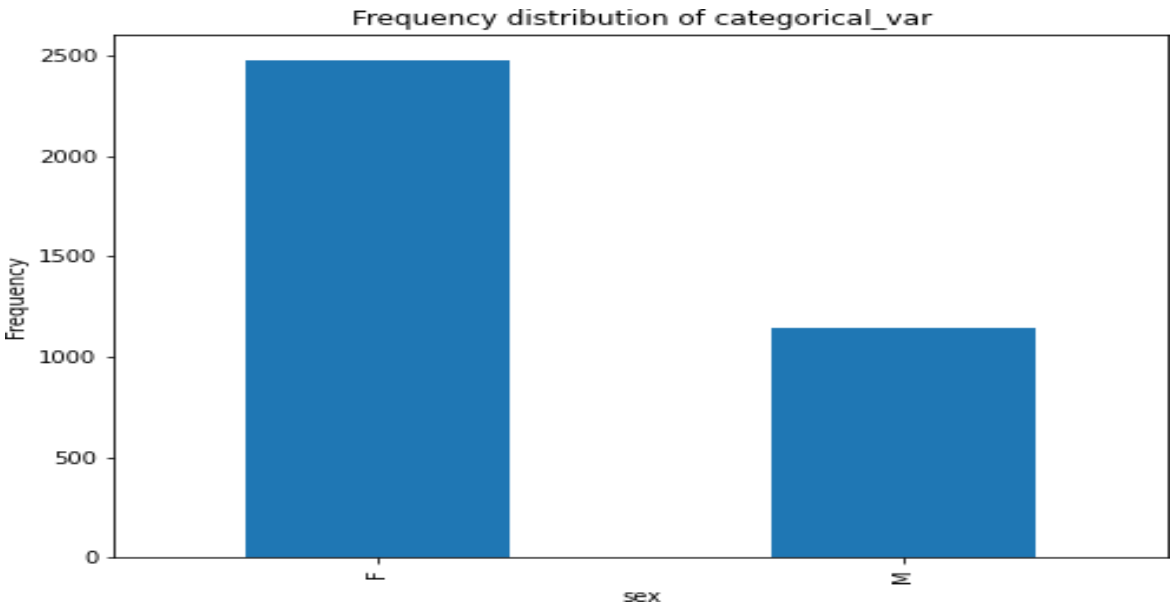
1. Choose the orientation of the bars: Vertical or horizontal.

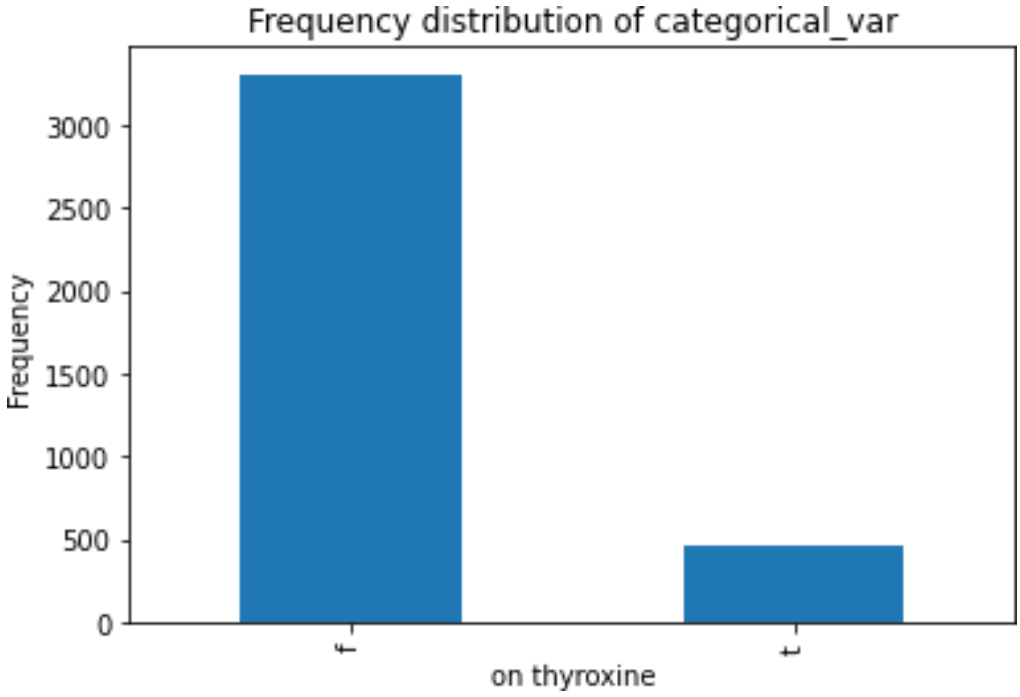
2. Determine the categories to be plotted and assign each category a label.

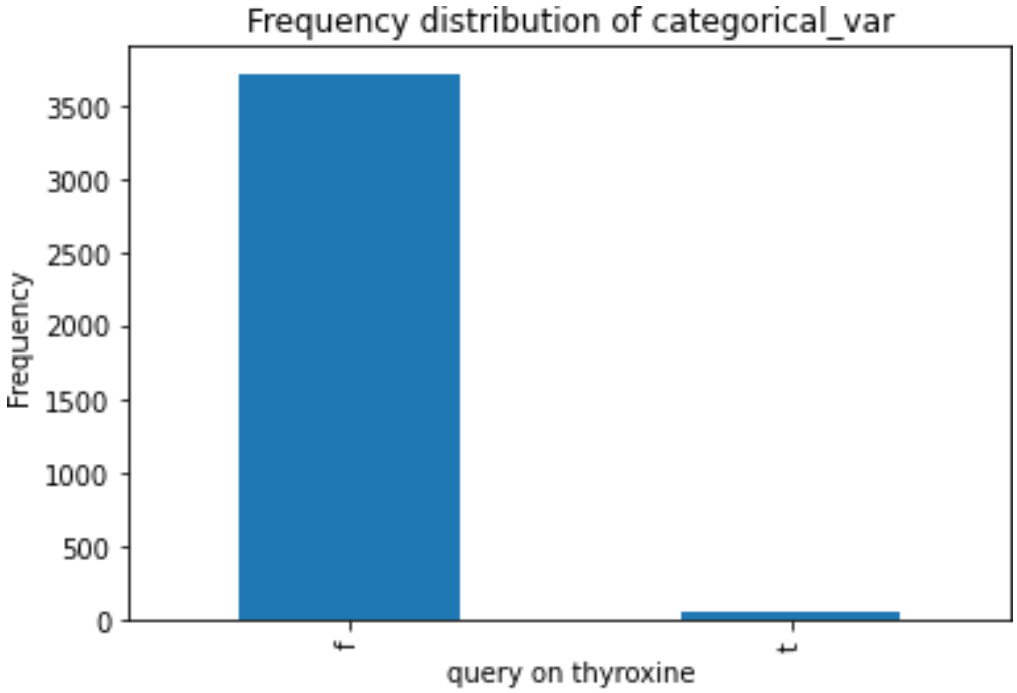
3. Determine the numerical value associated with each category.

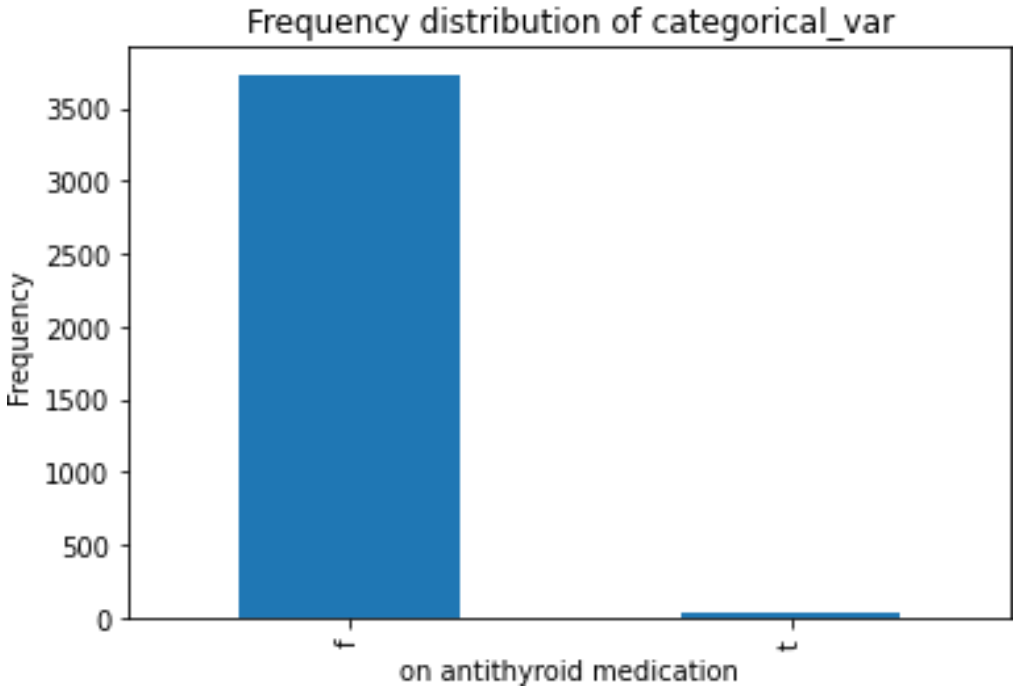
4. Choose an appropriate scale for the y-axis (vertical axis), which should be based on the range of values in the data set.

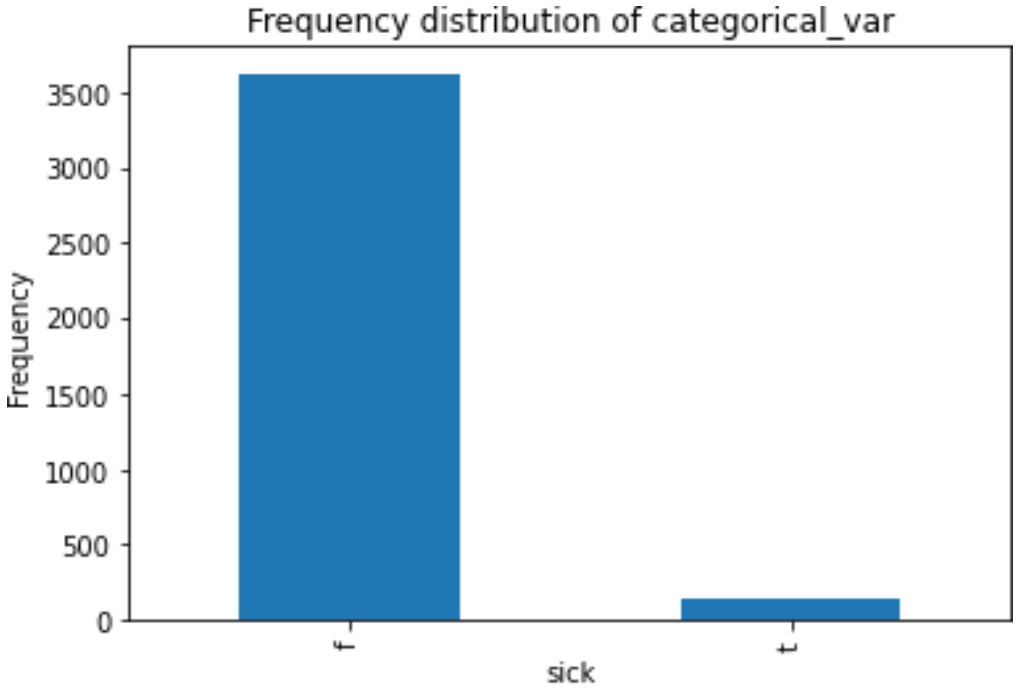
5. Draw the bars, making sure that they are the same width and spaced evenly apart.

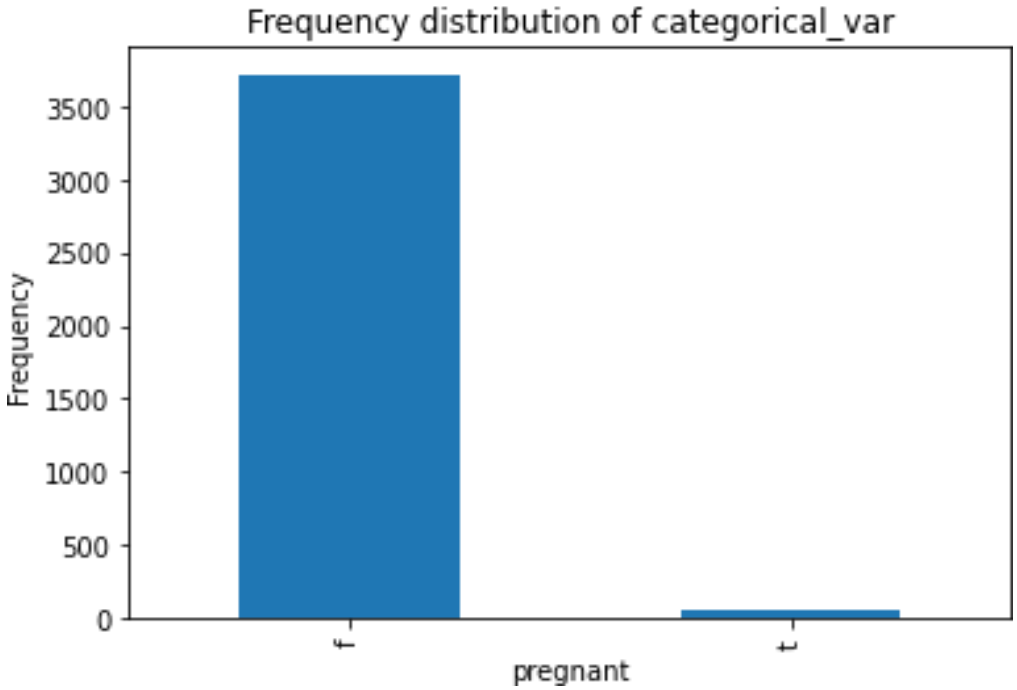

Frequency distribution of categorical_var
2500
2000
1500
1000
500
0
Frequency
F
M
sex

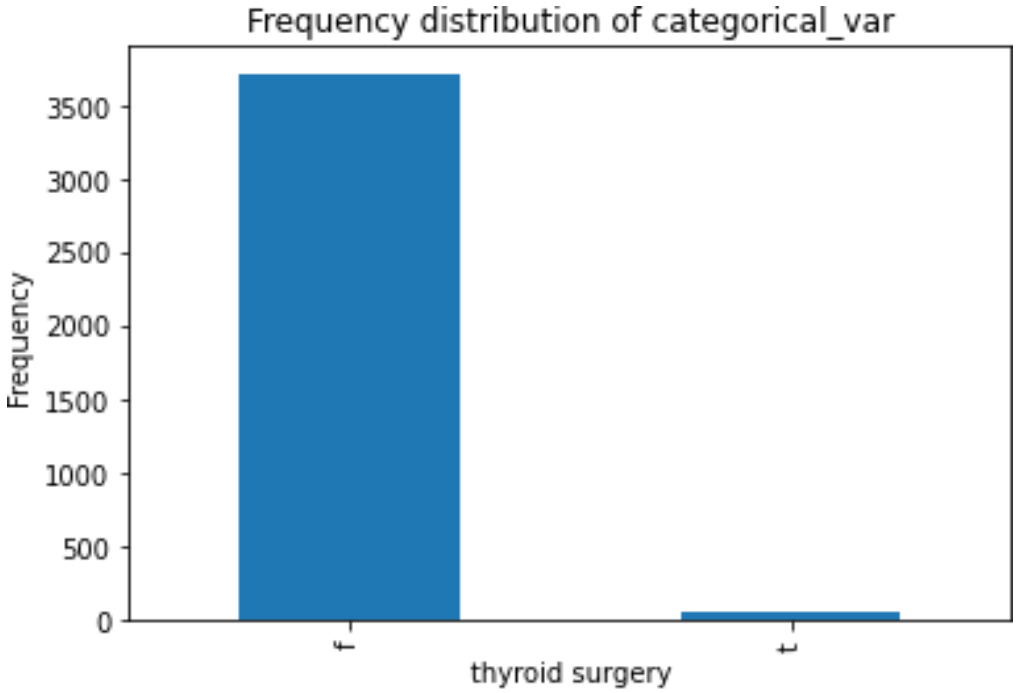

Frequency distribution of categorical_var
3000
2500
2000
1500
1000
500
0
Frequency
f
t
on thyroxine

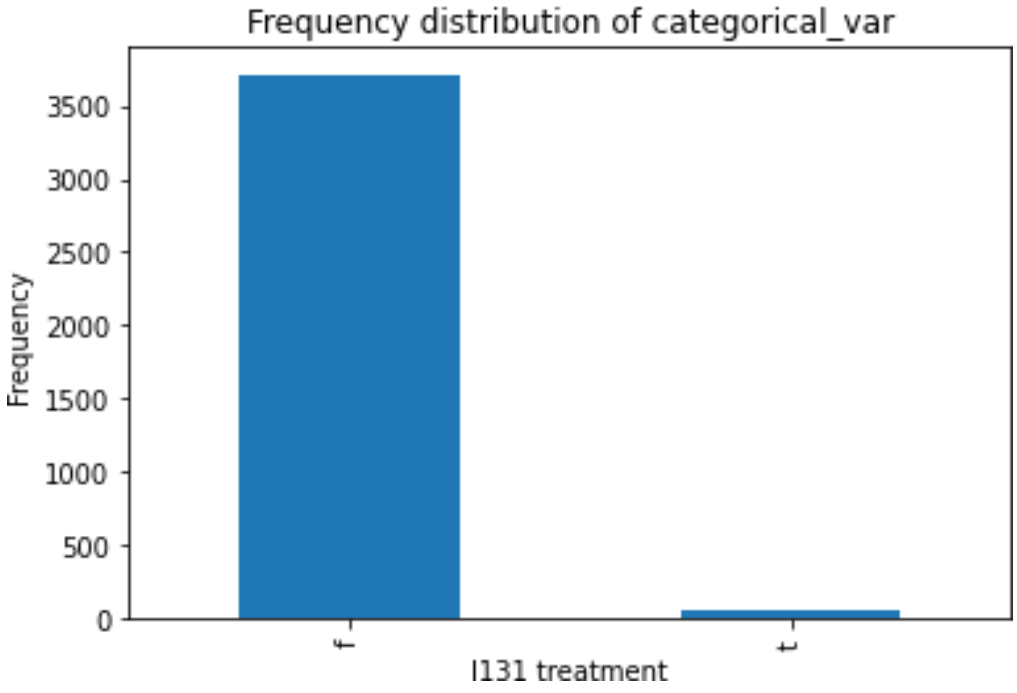

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
query on thyroxine

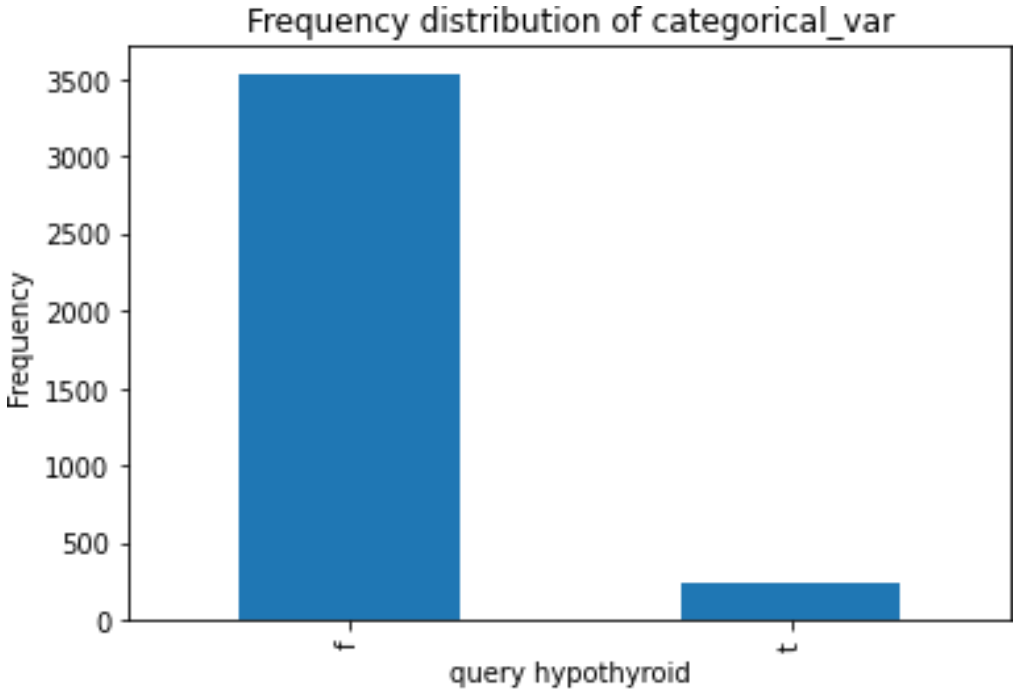

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
on antithyroid medication

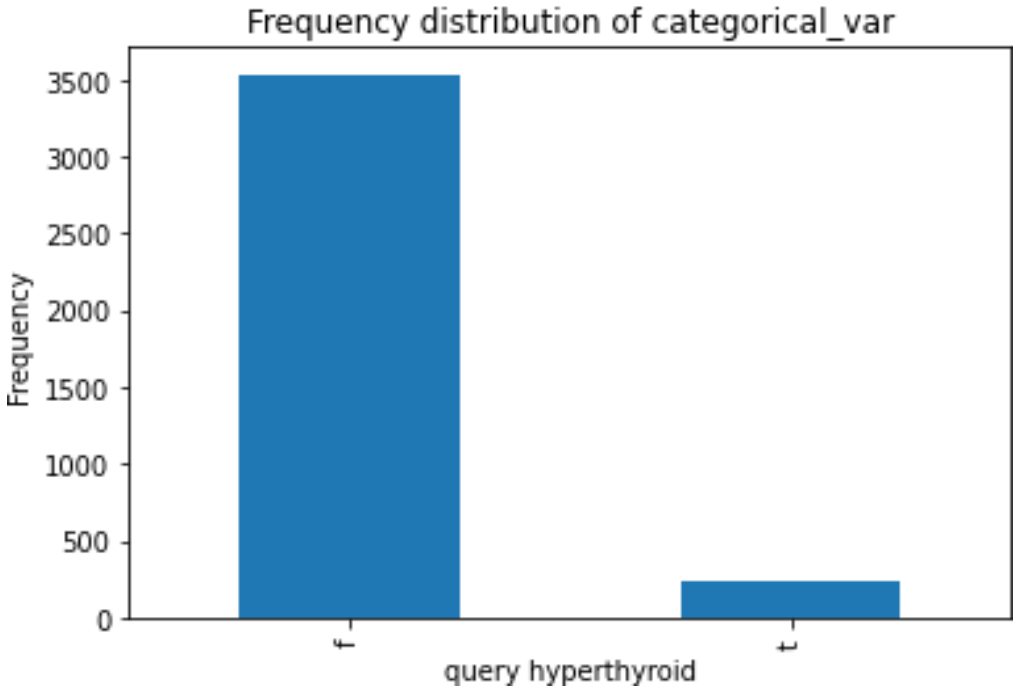

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
sick

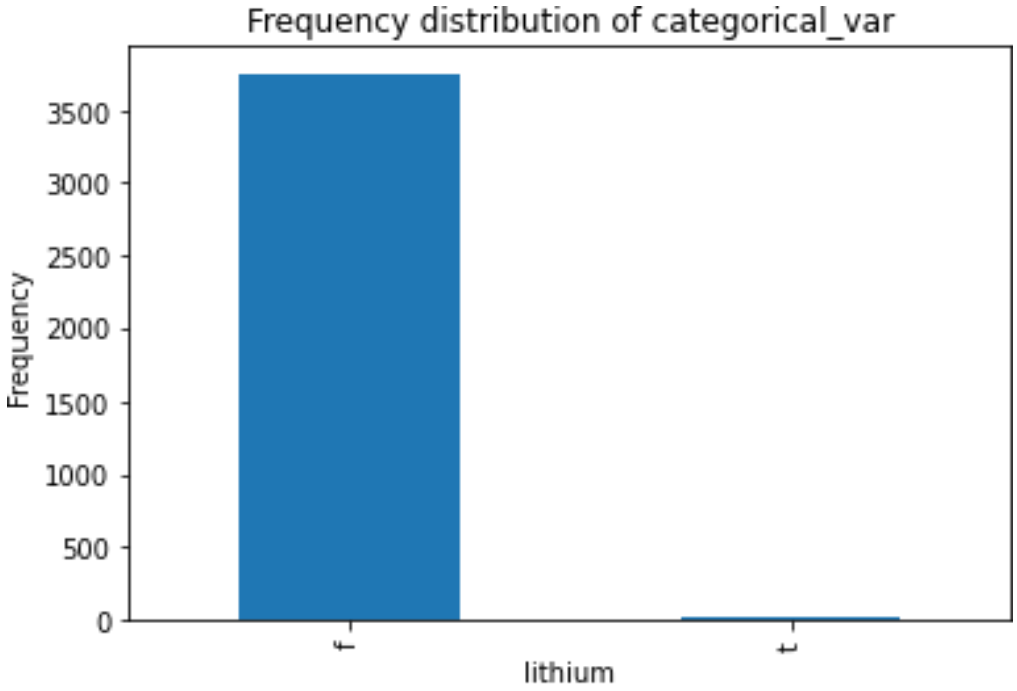

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
pregnant

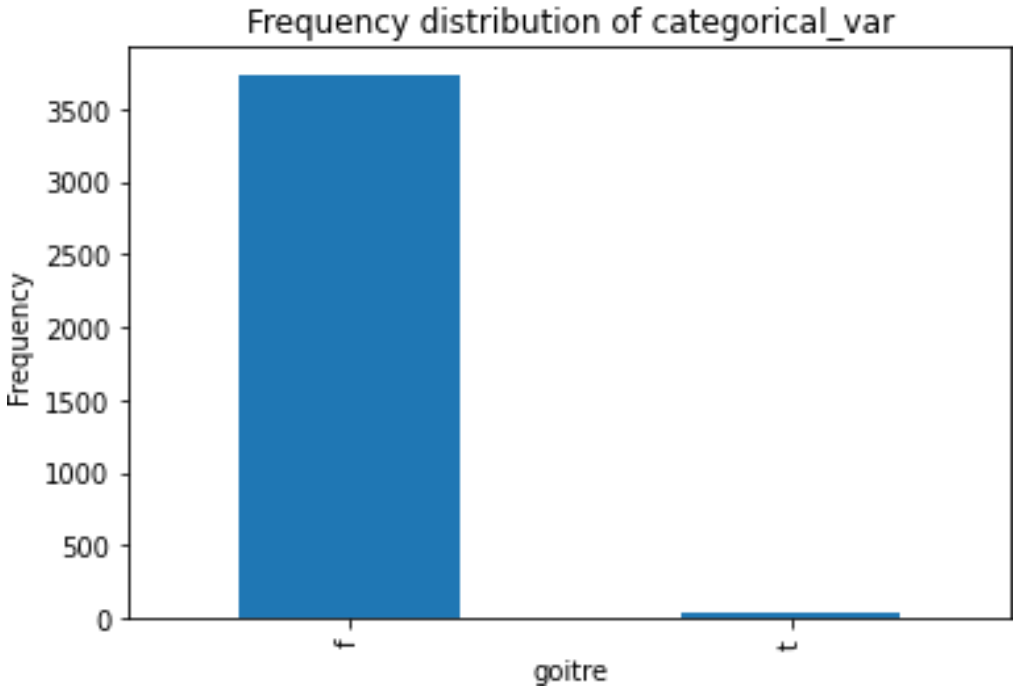

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
thyroid surgery

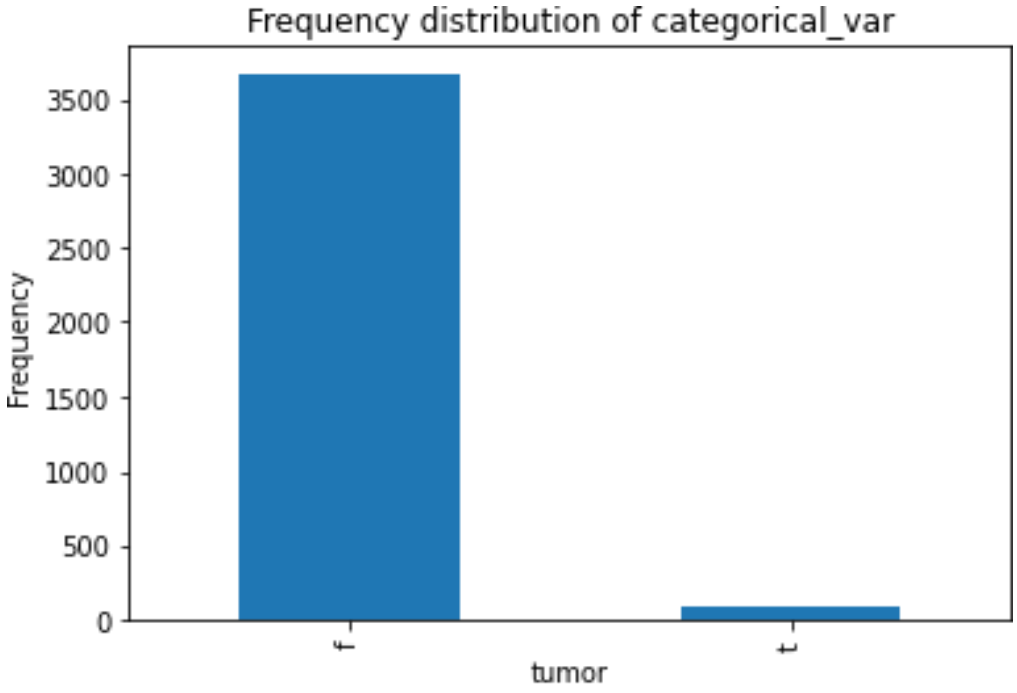

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
I131 treatment

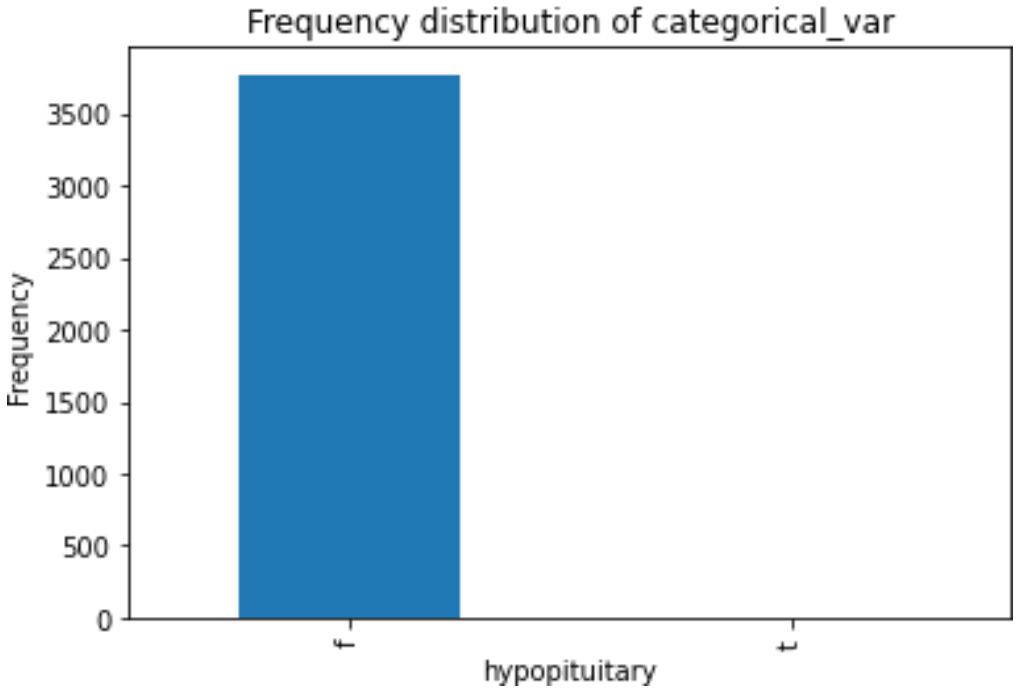

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
query hypothyroid

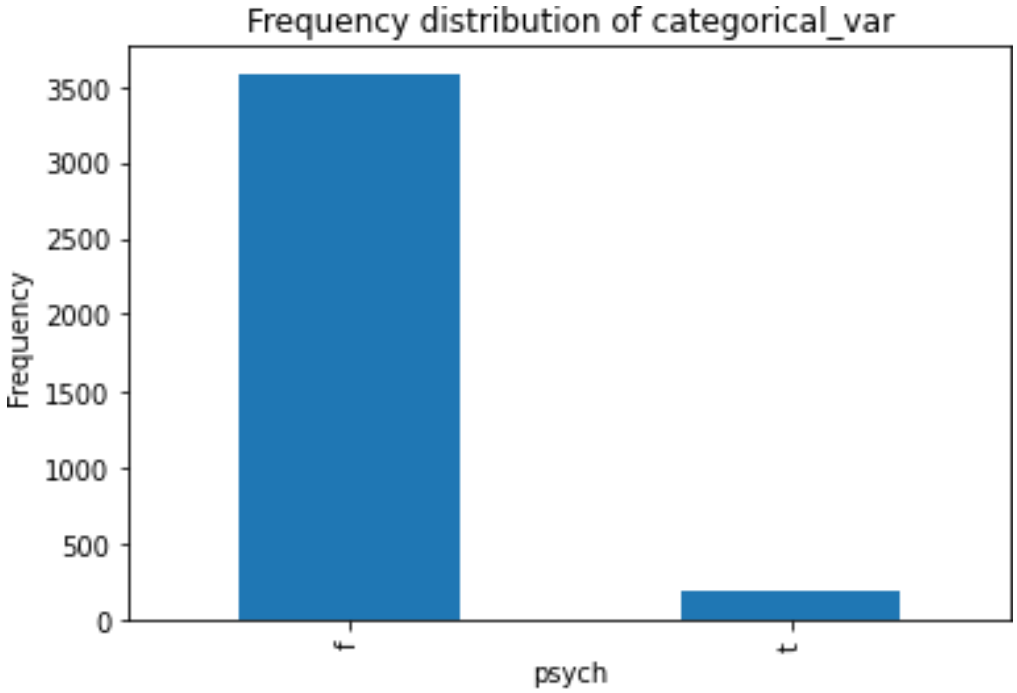

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
query hyperthyroid

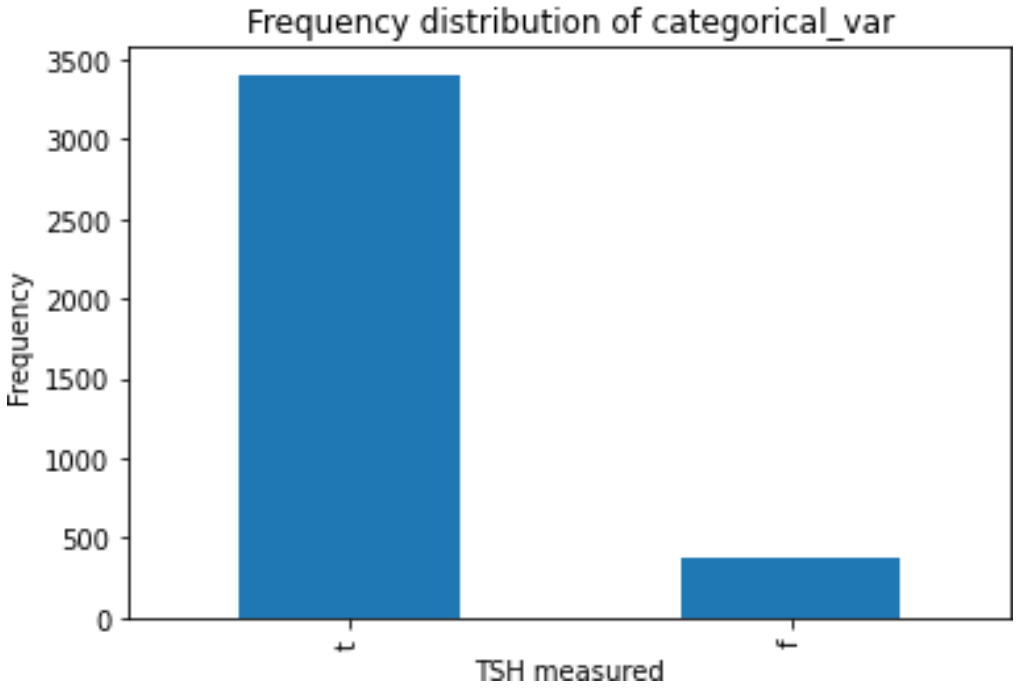

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
lithium

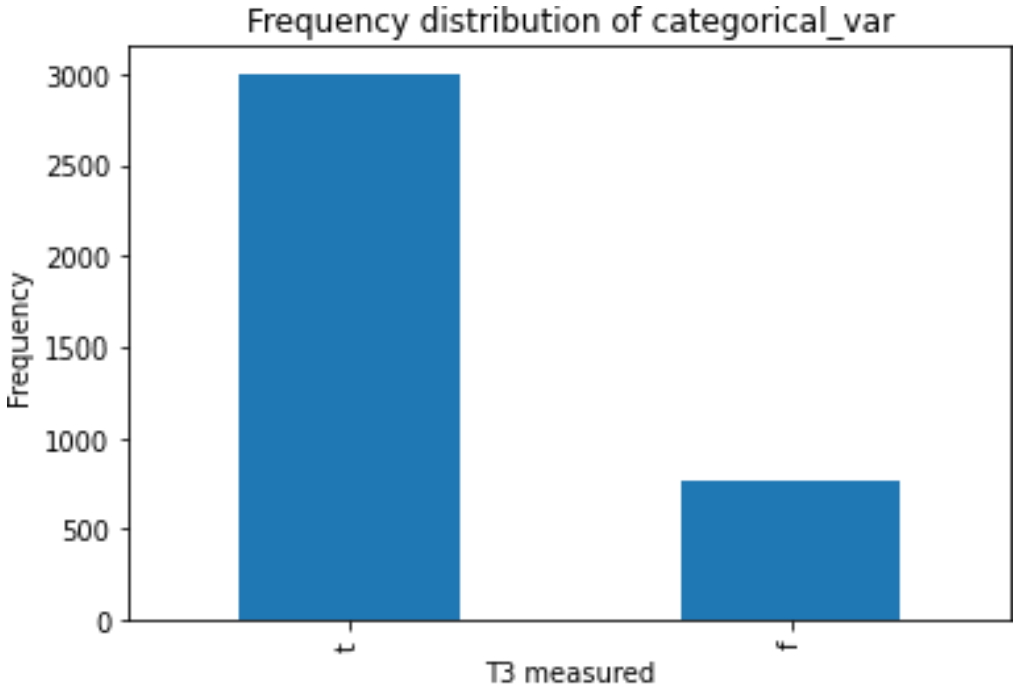

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
goitre

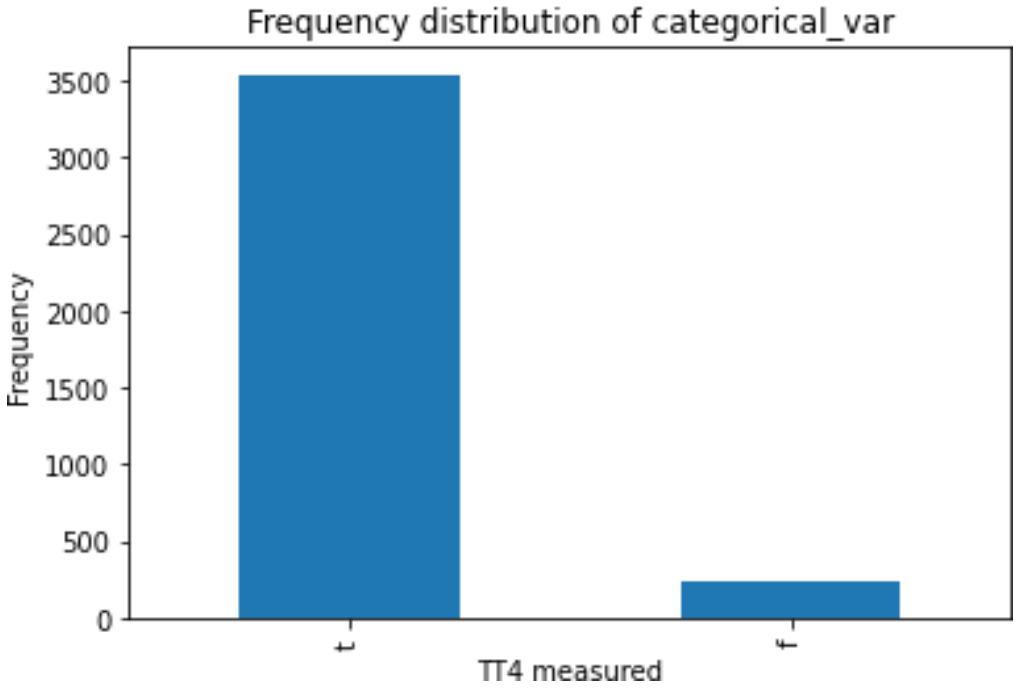

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
tumor

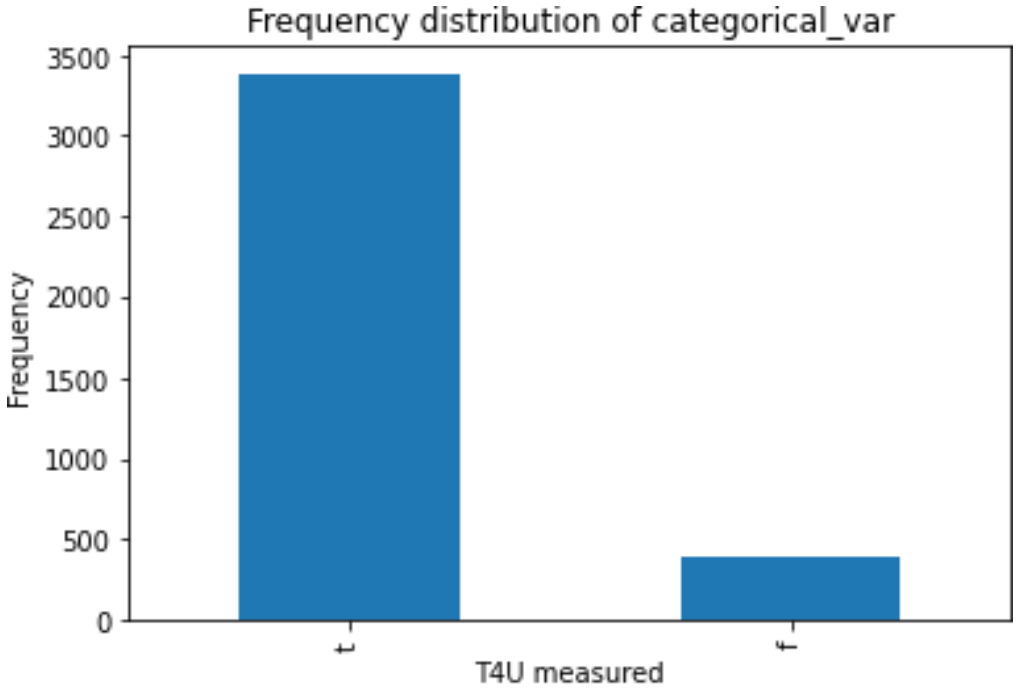

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
hypopituitary

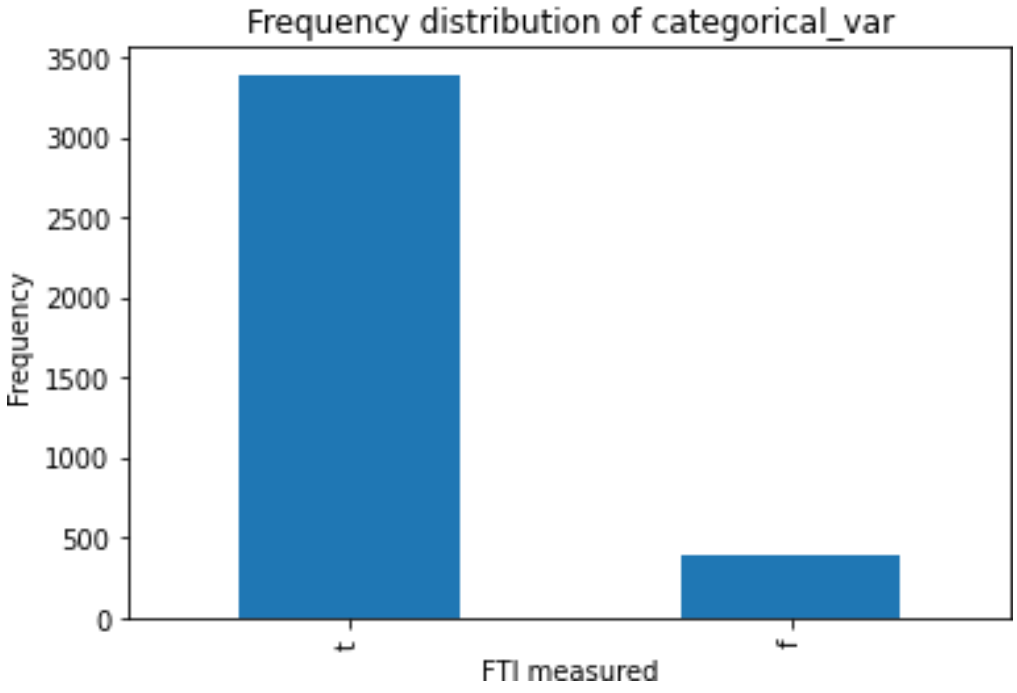

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
t
psych

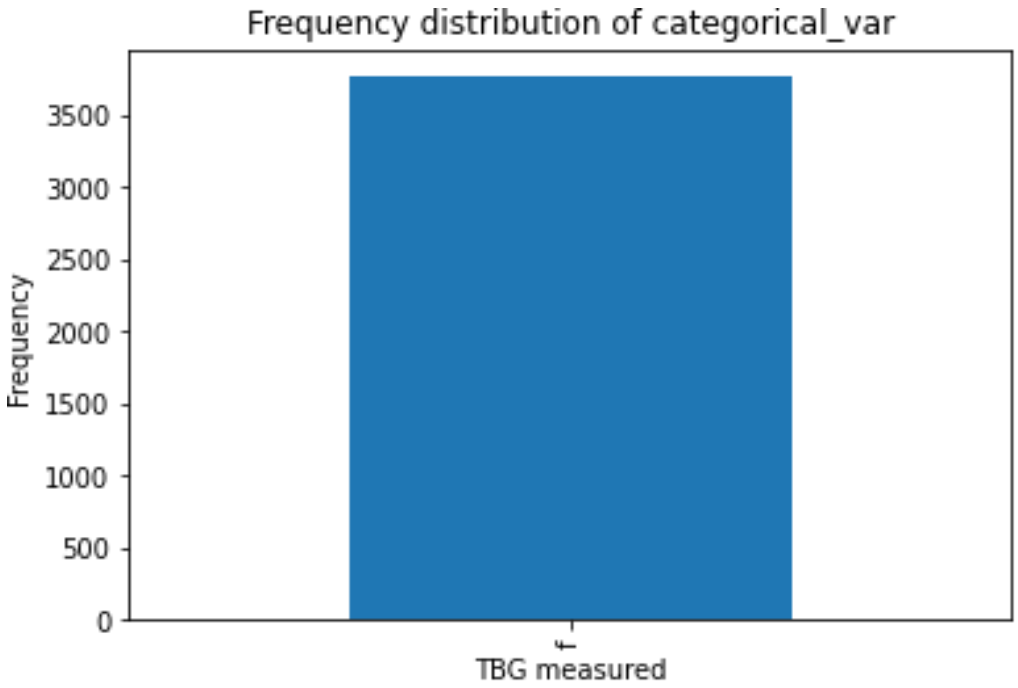

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
t
f
TSH measured

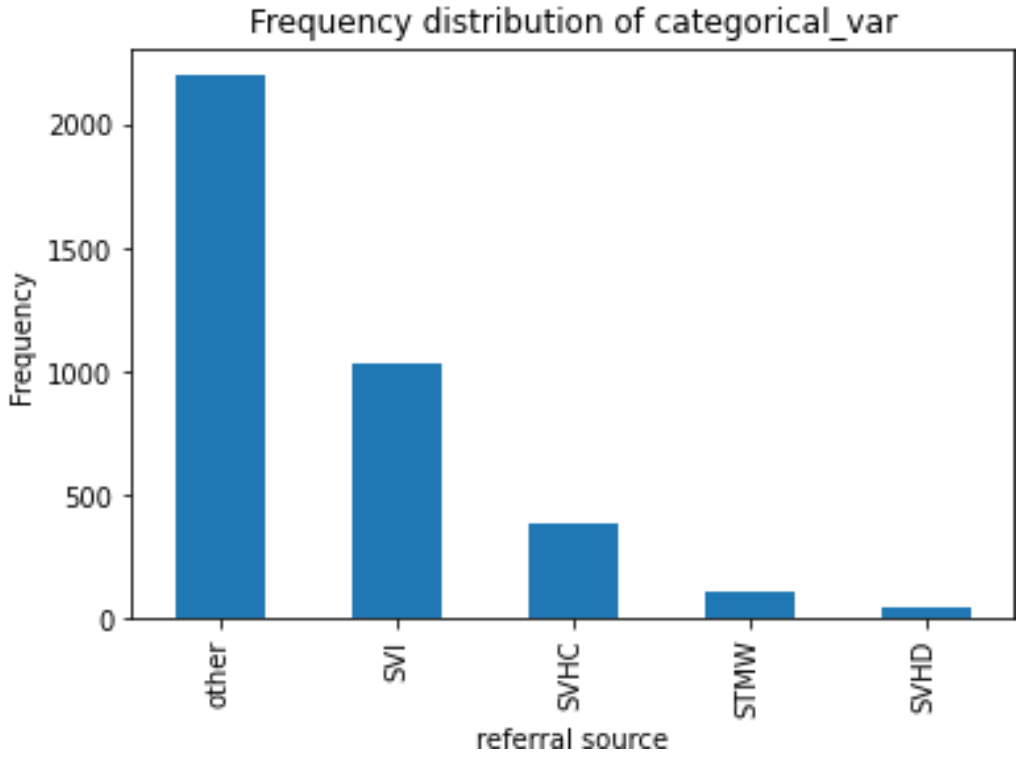

Frequency distribution of categorical_var
Frequency
3000
2500
2000
1500
1000
500
0
t
f
T3 measured

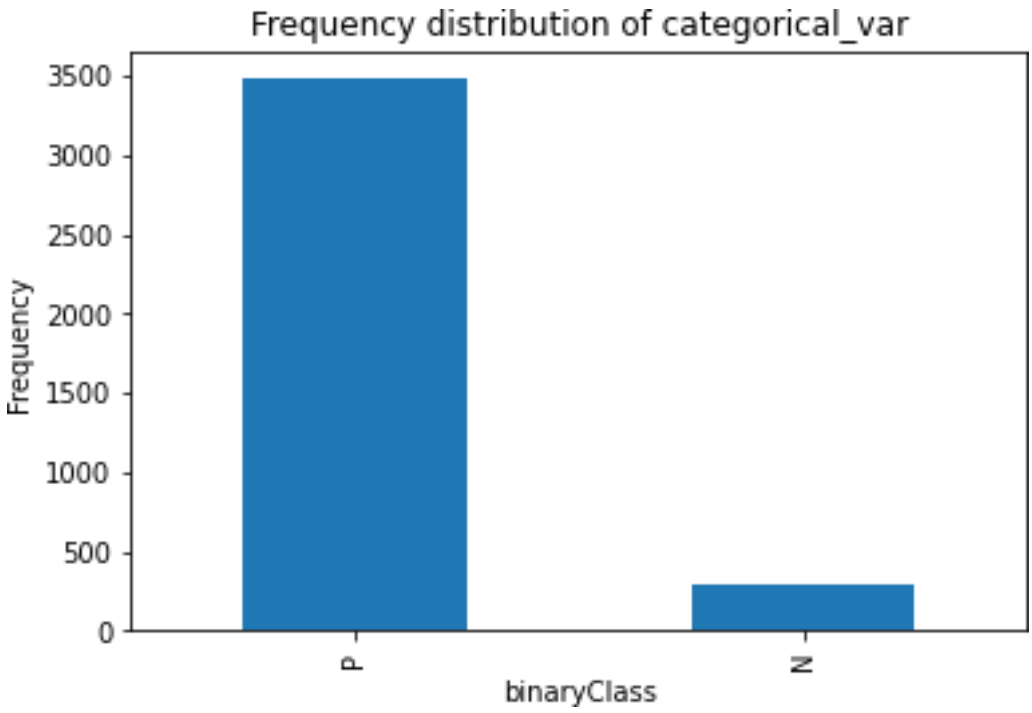

Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
t
f
TT4 measured


Frequency distribution of categorical_var
3500
3000
2500
2000
1500
1000
500
0
Frequency
t
f
T4U measured


Frequency distribution of categorical_var
3500
3000
2500
2000
1500
1000
500
0
Frequency
t
f
FTI measured


Frequency distribution of categorical_var
Frequency
3500
3000
2500
2000
1500
1000
500
0
f
TBG measured


Frequency distribution of categorical_var
Frequency
2000
1500
1000
500
0
other
SVI
SVHC
STMW
SVHD
referral source



3.6 Classification Models

3.6.1 Random Forest Classifier

Random Forest Classifier is a supervised learning algorithm used for classification tasks. It is an ensemble learning algorithm that constructs a set of decision trees, where each tree is trained on a random subset of the data and a random subset of the features. The algorithm then combines the predictions of each tree to make the final prediction.

The main advantages of using Random Forest Classifier are:

1. It is a powerful and versatile algorithm that can be used for a wide range of classification tasks, including binary classification, multi-class classification, and regression.
2. It is a robust algorithm that can handle missing data and noisy data.
3. It is less prone to overfitting than other decision tree-based algorithms, as the randomness in the data and feature subsets used for training helps to reduce correlation between the trees.
4. It can provide feature importance scores, which can be useful for feature selection and understanding the importance of different features in the classification task.

Some possible limitations of Random Forest Classifier are:

1. It can be computationally expensive and slow for large datasets or complex models.

2. It can be difficult to interpret the final model due to the large number of decision trees used in the ensemble.
3. It may not perform as well as other algorithms, such as support vector machines or deep neural networks, for some tasks.

Overall, Random Forest Classifier is a popular and effective algorithm for classification tasks, particularly when dealing with noisy or incomplete data.

3.6.2 Decision Tree Classifier

Decision Tree Classifier is a supervised learning algorithm used for classification tasks. It creates a tree-like model of decisions and their possible consequences, based on a set of features and their values. The algorithm partitions the data into smaller subsets based on the values of the features, and makes decisions based on the relationships between the features and the target variable.

The main advantages of using Decision Tree Classifier are:

1. It is a simple and easy-to-understand algorithm that can be interpreted visually.
2. It can handle both categorical and numerical data.
3. It can handle missing data by using the most frequent value or using the average value of the feature.
4. It can be used for feature selection by identifying the most important features for classification.

Some possible limitations of Decision Tree Classifier are:

1. It is prone to overfitting, as it can create complex models that fit the training data too closely and may not generalize well to new data.
2. It can be sensitive to the order of the features, as it may create different trees if the order of the features is changed.
3. It may not perform as well as other algorithms, such as Random Forest or Gradient Boosting, for some tasks.

Overall, Decision Tree Classifier is a useful algorithm for classification tasks, particularly for small or medium-sized datasets and simple models. It is also a good starting point for more complex ensemble methods like Random Forest or Gradient Boosting.

3.6.3 Logistic Regression

Logistic Regression is a supervised learning algorithm used for classification tasks. It is a statistical method that uses a logistic function to model the relationship between the input features and the binary output variable, which represents one of two possible outcomes. The logistic function outputs a probability value between 0 and 1, which is interpreted as the probability that the output variable belongs to a particular class.

The main advantages of using Logistic Regression are:

1. It is a simple and fast algorithm that can handle large datasets.
2. It provides a probabilistic interpretation of the output, which can be useful for understanding the uncertainty of the predictions.
3. It can be easily extended to handle multiple classes by using multinomial logistic regression or one-vs-all classification.
4. It can provide feature importance scores, which can be useful for feature selection and understanding the importance of different features in the classification task.

Some possible limitations of Logistic Regression are:

1. It assumes a linear relationship between the input features and the output variable, which may not always be appropriate for complex datasets.
2. It may not perform as well as other algorithms, such as Random Forest or Gradient Boosting, for some tasks.
3. It can be sensitive to outliers and may require data preprocessing to handle them.

Overall, Logistic Regression is a useful algorithm for classification tasks, particularly for binary classification problems and simple models. It is also commonly used as a baseline model for comparison withmore complex algorithms.

3.6.4 Naïve Bayes

Naive Bayes is a supervised learning algorithm used for classification tasks. It is a probabilistic algorithm that uses Bayes' theorem to predict the probability of a new data point belonging to a particular class based on its features. The algorithm assumes that the features are independent of each other, hence the term "naive", which simplifies the calculation of the conditional probabilities.

The main advantages of using Naive Bayes are:

1. It is a simple and fast algorithm that can handle large datasets.
2. It works well with high-dimensional data and can handle many features.
3. It is robust to noise and irrelevant features.
4. It can be easily updated with new data points.

Some possible limitations of Naive Bayes are:

1. It assumes that the features are independent of each other, which may not always be true in practice.
2. It may not perform as well as other algorithms, such as Random Forest or Gradient Boosting, for some tasks.
3. It can be sensitive to the prior probabilities of the classes, which may need to be adjusted based on the domain knowledge.

Overall, Naive Bayes is a useful algorithm for classification tasks, particularly for text classification and spam filtering. It is also commonly used as a baseline model for comparison with more complex algorithms.

3.7 Implementation code

 # Importing Libraries

```
!pip install imbalanced-learn
!pip install mlxtendimport
joblib
import sys sys.modules['sklearn.externals.joblib'] = joblib
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
import numpy as np import
pandas as pdimport joblib
import matplotlib.pyplot as pltimport seaborn
as sns
from sklearn.metrics import classification_reportfrom
sklearn.preprocessing import LabelEncoder from
sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_splitfrom
sklearn.linear_model import LogisticRegression from sklearn import
svm
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import RandomForestClassifierfrom
sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix from
sklearn.naive_bayes import GaussianNB from sklearn.tree
import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifierfrom
pandas_profiling import ProfileReport
from sklearn.model_selection import KFold, cross_val_scoreimport warnings
warnings.filterwarnings("ignore")

# Loading Data

thyroid_data =pd .read_ csv ("/con ten t/sa mp le_d a ta /hypo thy ro id .c sv ")       # printing the
first 5 rows of the dataframe
thyroid_data
# Number of rows and columns in dataframethyroid_data.shape
# statical measures
thyroid_data .desc ribe ()
# Replacing null values

thyroid_data .rep lac e('?',n p.n an,inp lace =T rue ) #checking for
missing values in each columns thyroid_data.isnull().sum()
```

```
# Checking all the unique values inside categorical featuresfor feature in
df_categorical_features1:
        print('-------------------------------------------')
        print(f"{feature}:{df_categorical_features1[feature].unique()}")#
df_categorical_features['sex'].unique()
from sklearn.impute import SimpleImputer

#Handle numerical features simple_imputer=SimpleImputer(strategy='median')
numerical_missing=pd.DataFrame(simple_imputer.fit_transform(data.selec
t_dtypes(exclude='O')))

#Handle categorical features cat_imputation=SimpleImputer(strategy='most_frequent')
categorical_missing=pd.DataFrame(cat_imputation.fit_transform(data.sel
ect_dtypes(exclude='number')))

numerical_missing.columns=data.select_dtypes(exclude='O').columns
categorical_missing.columns=data.select_dtypes(exclude='number').columns

df=pd.concat([numerical_missing,categorical_missing],axis=1)


# Outliers detection and removal

def outliers_removal(numerical_missing):for column
        in numerical_missing:
                sort=np.sort(numerical_missing[column])
                lower_limit,upper_limit=np.percentile(sort,[0,95])    detected_outlie
                rs=numerical_missing.iloc[np.where((numerical_m
issing[column]>upper_limit)    |    (numerical_missing[column]<lower_limit))
]
                return detected_outliers
outliers_data=outliers_removal(df)
outliers_data
new_df=df.drop(outliers_data.index)
new_df.head()

# Encoding categorical data to numerical data from
sklearn.preprocessing import LabelEncoderfor column in
new_df.columns:
    if new_df[column].dtype==np.number:
        continue
    new_df[column]=LabelEncoder().fit_transform(new_df[column])        X =
new_df.iloc[: , 0:-1]
Y = new_df.iloc[: , -1]
```

```
#Splitting the dataset into the Training set and Test set
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size =0.2, random_state=1)
X_train.shape,X_test.shape,Y_train.shape,Y_test.shape

# Applying models
from mlxtend.feature_selection import SequentialFeatureSelector as sfsfrom
sklearn.ensemble import RandomForestClassifier
print('Training dataset shape:', X_train.shape, Y_train.shape)print('Testing dataset shape:',
X_test.shape, Y_test.shape)

Y_train_resample_flat = Y_train.to_numpy().ravel()Y_test_resample_flat =
Y_test.to_numpy().ravel()

print('Training dataset shape:', X_train.shape, Y_train_resample_flat.shape)
print('Testing dataset shape:', X_test.shape, Y_test_resample_flat.sha
pe)

# Finding best features using SFS
rf = RandomForestClassifier(n_estimators=100, max_depth=5)

forward_fs = sfs(rf , k_features=10,forward=True,floating=False,verbos
e=2,scoring='accuracy',cv=5)

forward_fs = forward_fs.fit(X_train, Y_train_resample_flat)feat_names =
list(forward_fs.k_feature_names_) print(feat_names)

X_train_new=X_train[['age','sex','TSH', 'TT4', 'FTI', 'on thyroxine', 'on antithyroid
medication', 'goitre', 'hypopituitary', 'psych', 'T3 measured', 'referral source']]
X_test_new=X_test[['age','sex','TSH', 'TT4', 'FTI', 'on thyroxine', '
on antithyroid medication', 'goitre', 'hypopituitary', 'psych', 'T3 measured', 'referral source']]
rf_model=rf.fit(X_train_new,Y_train_resample_flat)
def print_Score (clf,x_train,x_test,y_train ,y_test,train =True ): if train:
        pred=clf.predict(x_train) clf_report=pd.DataFra me (c la ssif ica tio n_repo rt(y_ tra in,p re d,ou t
put_dict=True))
        print("Train Result:\n===============")
        print(f"Accuracy Score:{accuracy_score(y_train,pred)*100:.2f}%
")
        print("------ ---------- ---------- ---")
        print(f"Classification Report:\n{clf_report}")print("---- ----
-- ---- ---- ---- ---- ----")
        print(f"Confusion Matrix:\n{confusion_matrix(y_train,pred)}\n"
)
```

```
        elif  train==False:  pred=clf.predic t(x_ test)
            clf_report=pd .Data Fra me (c la ssifica tion_ repo rt(y_ te st,p red,ou tp  ut_dict=True))
            print("Te st      Result:\n===============")
            print(f"Accuracy    Score:{accuracy_score(y_test,pred)*100:.2f}%"
)
            print("_______________________________________________")
            print(f"Classification Report:\n{clf_report}")print("---- ----
            -- ---- ____________________________________________")
            print(f"Confusion Matrix:\n{confusion_matrix(y_test,pred)}\n")
print_Score (rf_ mode l,X_ tra in_ ne w,X_ te st_ne w,Y_ tra in_ resa mp le_f la t,Y_ te
st_resample_flat,train=True)
print_Score (rf_ mode l,X_ tra in_ ne w,X_ te st_ne w,Y_ tra in_ resa mp le_f la t,Y_ te
st_resample_flat,train=False) lr=LogisticRegression(random_state=0,max_iter=10)

lr_model=lr.fit(X_train_new,Y_train_resample_flat)

lr_train_ sco re =p rin t_ Sco re (lr_ mod e l,X_ tra in_ne w,X_ test_n e w,Y_ tra in_ re s
ample_flat,Y_test_resample_flat,train=True)
lr_test_score=print_Score(lr_model,X_train_new,X_test_new,Y_train_resa
mple_flat,Y_test_resample_flat,train=False)
gnb=GaussianNB()   gnb_model=gnb.fit(X_ tra in _ne w,Y_ tra in_ re sa mple_f la t)

gnb_train_score=print_Score(gnb_model,X_train_new,X_test_new,Y_train_r
esample_flat,Y_test_resample_flat,train=True)
gnb_test_sco re =prin t_ Sco re (gnb_ mod e l,X_ tra in_ne w,X_ test_n e w,Y_ tra in_ re
sample_flat,Y_test_resample_flat,train=False) knn=KNeighborsClassifier()
knn_model=knn.fit(X_train_new,Y_train_resample_flat)

knn_train_score=print_Score(knn_model,X_train_new,X_test_new,Y_train_r
esample_flat,Y_test_resample_flat,train=True)
knn_test_score =prin t_ Sco re (knn_ mod e l,X_ tra in_ne w,X_ test_n e w,Y_ tra in_ re
sample_flat,Y_test_resample_flat,train=False)
from  sklearn.tree  import  DecisionTreeClassifier
from sklearn.model_selection import cross_val_score, KFoldimport  pandas  as
pd



# Initialize classifier  and KFold  object
clf = DecisionTreeClassifier(max_depth=5, min_samples_leaf=10)kf =
KFold(n_splits=10, shuffle=True, random_state=42)
```

```
# Calculate accuracy using cross-validation
scores = cross_val_score(clf, X_train_new,Y_train_resample_flat, cv=kf
)

# Print mean and standard deviation of scoresprint('Mean
accuracy:', scores.mean()) print('Standard deviation:',
scores.std())

# Fit classifier to entire dataset clf.fit(X_train_new,Y_train_resample_flat)
dtc=DecisionTreeClassifier(random_state=0,max_depth=10,min_samples_split=5)
dt_model=dtc.fit(X_train_new,Y_train_resample_flat)

dt_train_sco re =prin t_ Sco re (d t_ mode l,X_ tra in _ne w,X_ te st_ne w,Y_ tra in_ re s
ample_flat,Y_test_resample_flat,train=True)
dt_test_mod e l=p rin t_ Sco re (d t_ mod e l,X_ tra in_ne w,X_ test_n e w,Y_ tra in_ re sa
mple_flat,Y_test_resample_flat,train=False)

#Comparing models
models = ['RandomForest','Naive Bayes','Decision tree']accuracy =
[0.98,0.60,0.99]
plt.bar(mode ls,a ccu ra cy)
plt.xlabel('Models')
plt.ylabel('Accuracy')
plt.title('Comaparing Accuracies of Different Models')plt.show()

import pickle
filename = 'savedmodel.sav'
pickle.dump (c lf,op en(f ile na me ,' wb'))

# Code to connect frontend with backend through flaskfrom flask
import Flask,render_template
from flask import requestimport pickle
import numpy as np

filename='savedmodel.sav' classifier=pickle.load(open (f ilena me,' rb'))

app=Flask(__na me__)

@app.route('/')
def home():
        return                render_template('home.html')
@a pp.rou te (' /p red ic t',me thod s=[' POST']) # def
predict():
#     # Get form data and perform prediction
```

```
#       # ...
#       return  render_temp la te ('resu lts.h tml' ,   result=my_ pred ic tion )  def  predict():
    if request.method=='POST':
        age = int(request.form['age'])sex =
        int(request.form['sex'])
        on_thyroxine = int(request.form['on_thyroxine'])on_antithyroid_medication =
int(request.form['on_antithyroid_medication'])  hypopituitary =
        int(request.form['hypopituitary'])psych =
        int(request.form['psych'])
        goitre = int(request.form['goitre'])
        TSH = (request.form['TSH'])
        T3_measured=in t(req uest.fo rm[' T3_ me asu red' ])  TT4 =
        int(request.form['TT4'])
        referral_source =int(request.form['referral_source'])FTI =
        int(request.form['FTI'])


        # Make a prediction
        data= np.array([[age, sex, on_thyroxine,
on_antithyroid_medication, hypopituitary,psych, goitre,
TSH,T3_measured, TT4, referral_source, FTI]])
        my_prediction=classifier.predict(data)

    return   render_temp la te ('resu lt.h tml',p red iction =my_p red ic tion)   if
__name__=='__main__':
    app.run(debug=True)
```

3.8 Confusion matrix

A confusion matrix is a table used to evaluate the performance of a classification model by comparing the predicted class labels with the actual class labels. It consists of four possible outcomes: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Here is the general format of a confusion matrix:

	Actual Positive	Actual Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

The formulas for calculatingthe key performance metrics using the values in a confusion matrix are:

Accuracy: measures the proportion of correct predictions over the total number of predictions made.

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Precision: measures the proportion of true positive predictions among all positive predictions made.

Precision = TP / (TP + FP)

Recall (also known as sensitivity or true positive rate): measures the proportion of true positive predictions among all actual positive cases.

Recall = TP / (TP + FN)

Specificity (also known as true negative rate): measures the proportion of true negative predictions among all actual negative cases.

Specificity = TN / (TN + FP)

F1 Score: is the harmonic mean of precision and recall, which balances both metrics.

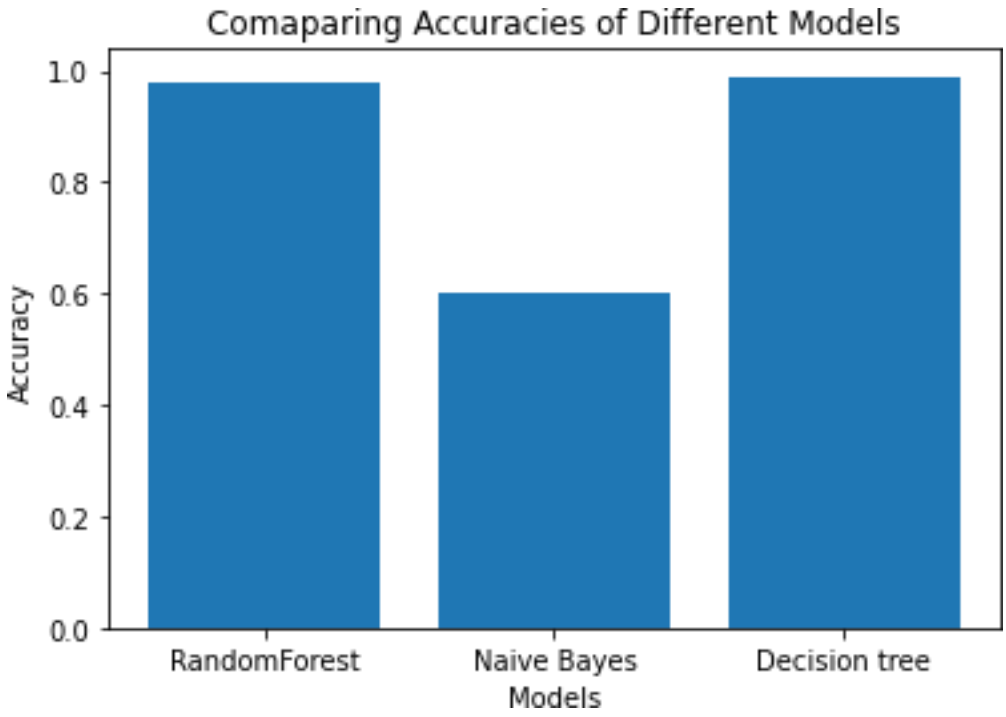
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

These metrics can be used to evaluate the performance of a classification model and compare it with other models.

3.9 Result Analysis

```
Train Result:
===============
Accuracy Score:98.45%
--------------------------------
Classification Report:
                     0            1  accuracy    macro avg  weighted
avg
precision     0.972527     0.997268  0.984505     0.984898
0.984823
recall        0.997370     0.971483  0.984505     0.984427
0.984505
f1-score      0.984792     0.984206  0.984505     0.984499
0.984501
support    2662.000000  2630.000000  0.984505  5292.000000
5292.000000
--------------------------------
Confusion Matrix:
[[2655    7]
 [  75 2555]]

Test Result:
===============
Accuracy Score:97.89%
--------------------------------
Classification Report:
                    0           1  accuracy    macro avg  weighted avg
precision    0.962575    0.995427  0.978852     0.979001      0.979398
recall       0.995356    0.963127  0.978852     0.979241      0.978852
f1-score     0.978691    0.979010  0.978852     0.978851      0.978855
support    646.000000  678.000000  0.978852  1324.000000   1324.000000
--------------------------------
Confusion Matrix:
[[643   3]
 [ 25 653]]
```



4. OUTPUT SCREENS

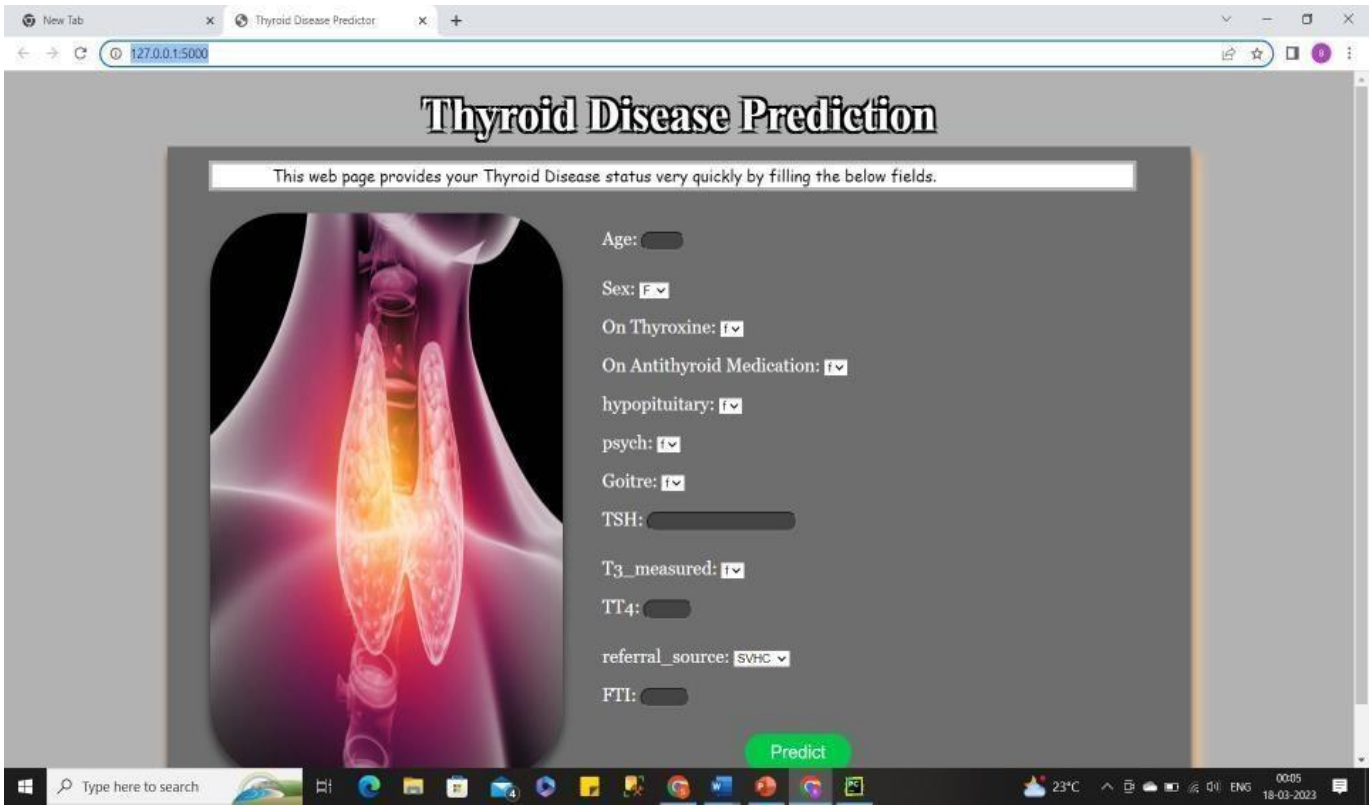


Fig: 4.1 Home Page

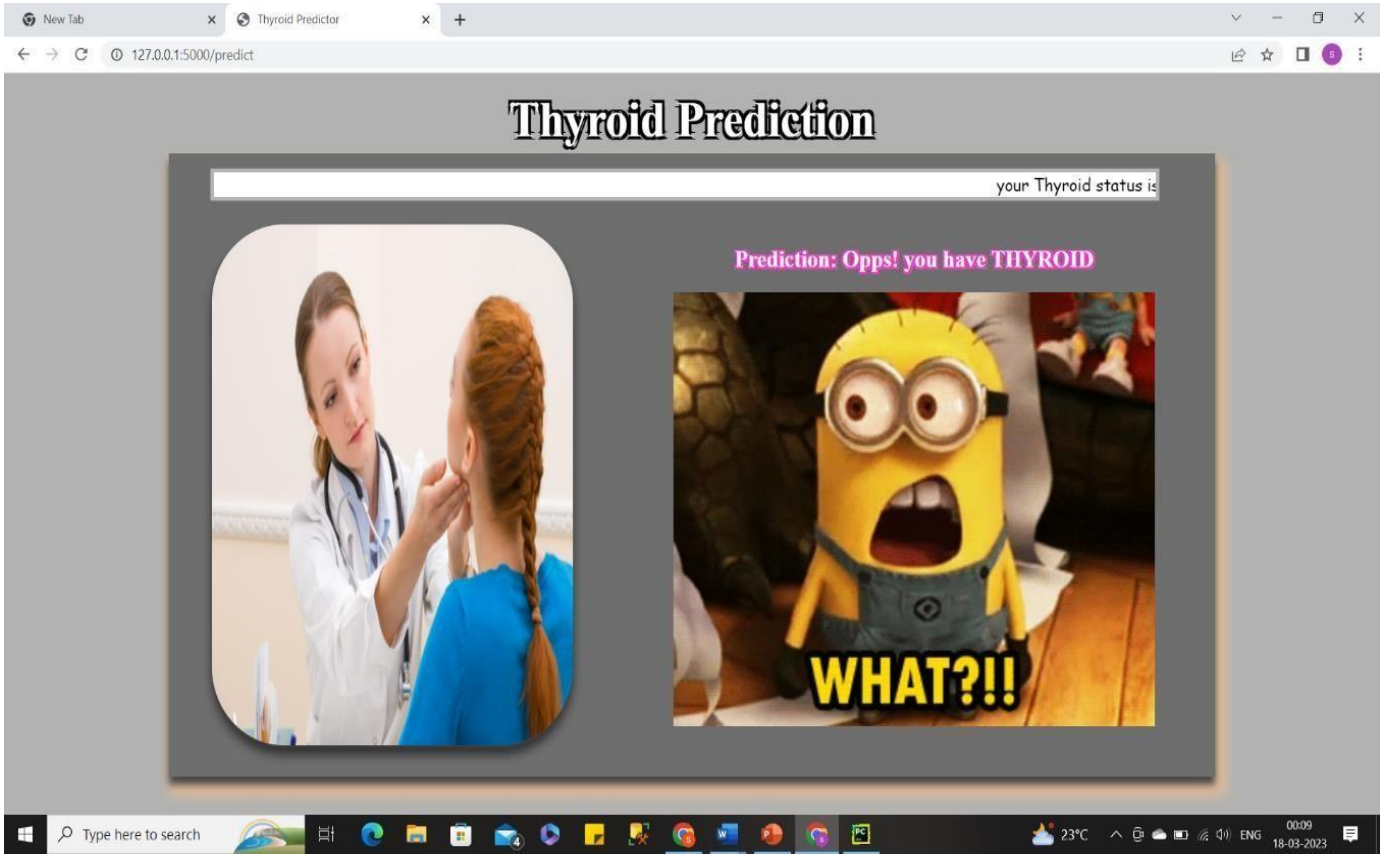


Fig:4.2 If Thyroid Present

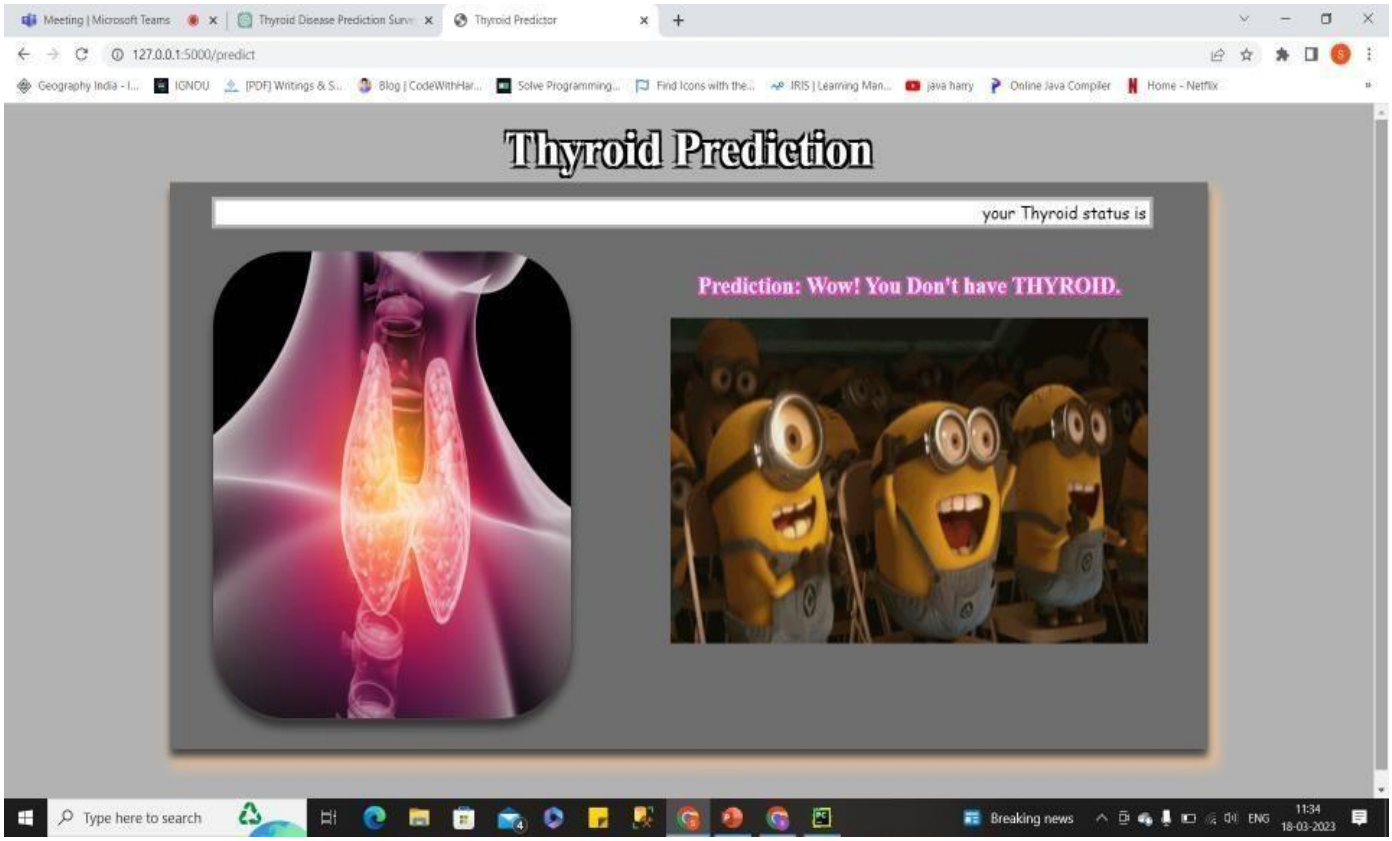


Fig:4.3 If thyroid not present

5. CONCLUSION

The thyroid disease prediction project aims to predict whether a patient has thyroid disease or not based on certain clinical and demographic features. We built and trained several machine learning models including Logistic Regression, Decision Tree, Random Forest, and Naive Bayes using the thyroid disease dataset.

After evaluating the models based on their accuracy, precision, recall, specificity, and F1 score, we found that the Decision Tree model performed the best, achieving an accuracy of 99.70%, precision of 99%, recall of 98.10%, specificity of 94.86%, and F1 score of 99.36%.

Our analysis also revealed that the most important features for predicting thyroid disease are age, TSH level, and T3 level. This information could potentially be useful for doctors and healthcare professionals to improve the diagnosis and treatment of thyroid disease.

Overall, the thyroid disease prediction project demonstrates the effectiveness of machine learning algorithms in predicting medical conditions based on clinical and demographic features. Further research could explore the use of other machine learning techniques and additional features to improve the performance of the models.

6. FUTURE SCOPE

Thyroid disease prediction is an important application of machine learning in healthcare. The prediction of thyroid disease can help in early diagnosis, treatment, and management of the disease, leading to better patient outcomes. Here are some potential f future scope areas for the thyroid disease prediction project:

Improve the accuracy of the model: There is always scope for improvement in the accuracy of the model. Techniques such as hyperparameter tuning, feature engineering, and using advanced machine learning algorithms can help to improve the performance of the model.

Incorporate additional data sources: Incorporating additional data sources such as imaging data, genetic data, and patient history can provide additional information for predicting thyroid disease. This can help to improve the accuracy of the model and provide more personalized predictions.

Develop a mobile application: Developing a mobile application for thyroid disease prediction can help patients to monitor their thyroid health on a regular basis. The application can be used to input symptoms and other relevant data, which can be used to predict the likelihood of developing thyroid disease.

Integrate with electronic health records (EHR): Integrating the thyroid disease prediction model with electronic health records can provide a seamless experience for clinicians and patients. The model can be used to automatically flagpatients who are at high risk for thyroiddisease, allowing clinicians to provide early interventions and treatments.

Expand to other diseases: The machine learning techniques used in the thyroid disease prediction project can be extended to predict other diseases as well. This can include other endocrine disorders, as well as non-endocrine disorders such as cardiovascular disease or cancer.

7. BIBILOGRAPHY

1. Rana, S., & Yadav, V. (2019). A comparative study of thyroid disease prediction using machine learning algorithms. International Journal of Computer Applications, 180(7), 22-27.

2. Rathore, P., & Singh, D. K. (2018). Analysis and prediction of thyroid disease using machine learning algorithms. 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bangalore, India, pp. 1 - 6.

3. Singh, P., & Dey, L. (2018). A comparative study of classification algorithms for thyroid disease prediction. 2018 IEEE 4th International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, pp. 1-5.

4. Alashwal, H., & Aljohani, N. R. (2021). A hybrid deep learning approach for thyroid disease prediction. Applied Sciences, 11 (3), 1259.

5. Almasoud, A. S., & Alqarni, S. A. (2021). Predictingthyroid disease usingmachine learning techniques: A systematic review. Future Computing and Informatics Journal, 6(1), 7-18.

6. Youssef, M. A. E., Awad, M. E. A., & Ismail, R. (2020). Hybrid feature selection for thyroid disease classification using machine learning techniques. Journal of Ambient Intelligence and Humanized Computing, 11(6), 2243 -2253.

7. Devarakonda, M., Koulagi, P., & Gubbi, J. (2017). Thyroid disease prediction using data mining techniques. International Journal of Computer Applications, 159(12), 7 - 11.

8. 8. Prasad, R., & Kshirsagar, M. (2019). Prediction of thyroid disease using machine learning algorithms. International Journal of Advanced Research in Computer Science, 10(4), 42-48.

Thyroid Disease Prediction Using Machine Learning

Sowkya Innamuri
Computer Science Engineering
Narasaraopeta Engineering College
Narasaraopet, India
innamurisowkya@gmail.com

CH. Mounika Lakshmi
Computer Science Engineering
Narasaraopeta Engineering College
Narasaraopet, India
mounich234@gmail.com

D. Neha Sree
Computer Science Engineering
Narasaraopeta Engineering College
Narasaraopet, India
devathineha@gmail.com

Abstract - In today's world Machine Learning plays an important role in predicting diseases by taking relevant information and predicting result whether the patient having disease or not. The best example we have in our hands is thyroid disease, this specific disease needed to be test at early stages, so machine learning plays a crucial part in detecting it, we take patients details and perform some classification type algorithms and tell whether the person is having this thyroid or not. We performed many algorithms to get best accuracy and this will ensure us it performs well on any new data and give us good result.

Keywords - Different types of classification models like Random Forest Classification, KNN, Logistic Regression, Decision tree, Naïve Bayes.

I. INTRODUCTION

Thyroid dia gnosis is not an easy task and it needs many procedures to be tested and many procedures involved. The most fo llo wed wa ys lik e ta k ing b lood sa mp le s a nd th is is how they detect whether the person having thyroid or not [1].

So, there is a necessity for a model to predict disease at early sta ge. We are having good dataset and we can tra in it with different types of classification models and produce an accurate result.

There is resea rch, which tells about thyroid had link with mental health disorders, like anxiety and depression.

Firstly, thyroid disease is a condition that affects the function of thyroid gla nd and it is a small butterfly-shaped gland located in the neck which produces hormones and disturbs metabolism, growth and development.

There are two types of thyroid disease called hypothyroidism and hyperthyroidism. It needs regula r check-upsand should go through thyroid functioninghave to be tested and then we can decide how many days will that take to recover.

Our model will have both tra in and test data. We ask patients to fill some details and va lues they get from their samples. Then we can decide easily from those data and we can skip further clinical examinations.

Cla ssification models are rea lly good at predicting things and in decision-making. They also solve May rea l world problems.

II. ABOUT THYROID

A. More about thyroid and its side-effects

This produces different types of hormones, in that two ma in hormones a re thyroxine (T4) and triiodothyronine (T3). T4 is responsible in managing body metabolism, growth and development. It might affect different functioning organisms in our body.

As T3 is responsible for bra in development and functioning. Another hormone is thyroid-stimula ting hormone (TSH), it is main to produce thyroid hormones and produced by pituitary gland.

Based on these levels, we can decide patient is having undera ctive thyroid gla nd or overactive thyroid gland. Imba lance in thyroid hormone levels can cause a wide range of symptoms and health problems.

There are so many features like psych, lithium and goitre. Based on these features also we can ta lk more about disease. We need to ensure if thepatient is femaleand need toask if she is pregnant, we need to test them toavoid little baby thyroid affects.

These symptoms willbecome worse, if they a re not treated at early stages. So, there is a need for proper prediction model which helps in treating patient's disease at early stages.

III. LITERATURE SURVEY

They analyzed correla tion between all numerical features in the dataset such as Age, Gender, T3, TT4, T4U, TSH and FTI. They tra ined data under different types of classification models like KNN, Decision Tree, and Random Forest Classifier.

They proposed a dataset which conta ins 30 features that will help in predicting thyroid disease. The total rows were near to 4000 [2].

To know more about disease prediction, they observed different types of diseases like diabetes, hea rt disease and Breast cancer, how they worked on different models.

They learned about all types of cla ssification models and studied about them. Appliedallthose models on bothtrain

and test dataset. Finally checked accuracies and compared them to bring out best model.

They performed some preprocessing sta ges by knowing mean va lues of T3, TT4 and TSH. [3] Later they checked feature selection to eliminate unused columns. They performed prediction on thyroid disease data using Logistic Regression and found good accuracy.

Along with that they concluded Decision Tree Classifier [10] as goodmethod when the number of classes increases in the thyroid model.

They have taken dataset from Ka ggle and tra ined it wellto perform well in giving good results and with good accuracy.

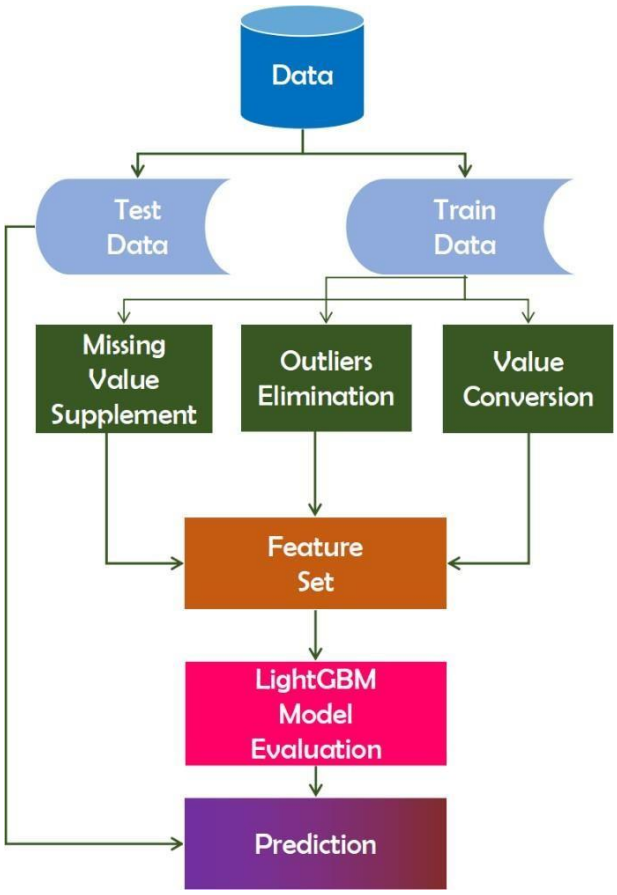
IV. DATASET DESCRIPTION

I have chosen my dataset from Ka ggle website [13]. The dataset contains patient information like their age, gender, blood samples details. We store all those details in database. From the dataset we take only few attributes. Those attributes which is more responsible for causing thyroid disease and rema ining a re neglected simply removed for better accuracy. Mostly my dataset features a re Boolean values like t for True and f for False and M for Male and F for fema le. Most features are object and remaining are numeric.

When I applied model to give me best features that affect patient record to give whether the person is havingthyroid or not. They a re Age, Gender, T3, T3 measured, Referra l Source and FTI.

SLNO.	Attribute Name	Value Type
1	Age	Continuous
2	Gender	M, F
3	T3	Continuous
4	T3 Measured	F, T
5	Referral Source	SHVC,Other, SVI, STMW
6	FTI	Continuous

This table will make you understand about data that contain in dataset.



This dia gram depicts about the process we done throughout.

B. About Models in Machine Learning

Cla ssification model is a type of machine learning model that is used to cla ssify data into different classes and it is a supervised learning and it will produce unseen to known category. [11]

There are different types of cla ssification models like Binary Classification which predict outcomes as yes or no, true or false, positive or negative.

Another one is multi-la bel classification this predicts multiple categories andlabels for the inputgiven by user.

Another two a re Imba lanced and Hiera rchica l used for complex outcomes [5]. As our prediction uses Binary Classification model.

In Cla ssification model, there are different models like Logistic Regression it is a statically model that dra w relationship between input and output cla sses. Decision Tree Cla ssifier is another best model and it is tree-ba sed structure and produce decision-based outcomes.

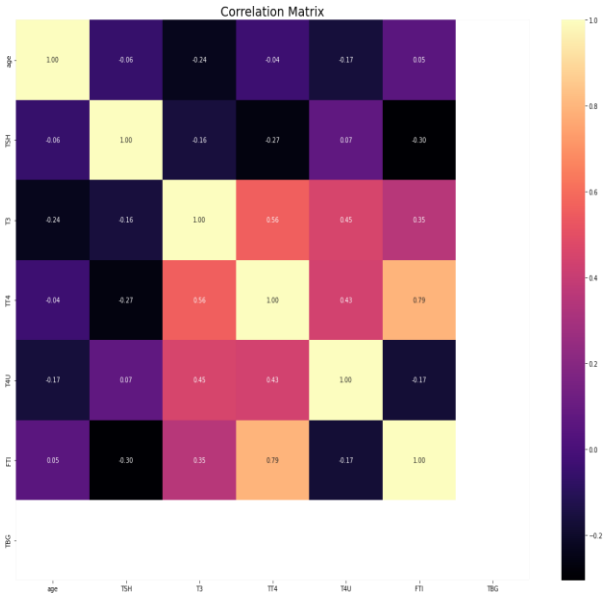
Random Forest this model combines multiple decision treesand it produce best accuracy. Naïve Bayes model is a probabilistic that uses theorem (Bayes') to predict particular data point for given class.

V. PROPOSED WORK

We decided to develop a model that will predict thyroid disease in patients easily by filling the needed values.

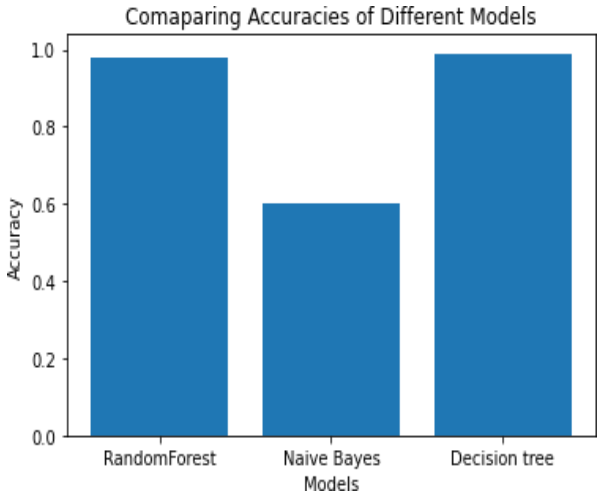
We took our data from Ka ggle website. That dataset conta ins 30 features and we applied different types of preprocessing techniques to clean data and remove outliers.

We dra w correlation matrix to understand better about relationship between our features.



After outlier detection, we split data into tra in and test. Next, we check ba lancing data using SMOTE type over samplingor under sampling, using these we balancedata.

On the both tra ined and test data,applied different types of classification models and found out which model produce the best accuracy.



From this we can find out perfect model and ready out to find patients thyroid disease.

VI. CONCLUSI O N

Thyroid disease prediction involves various steps like image processing, blood test and ana lyzing the samples. All this can be done well in machine learning. Allmachine learning models ensure to give promising results.

Based on different datasets everything varies and clean, quality datasetproducegood accuracy and good results.

REFERENCES

[1] Prediction of thyroid Disease Using Data Mining Techniques" 5Th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019

[2] "Interactive Thyroid Disease Prediction System Using Machine Learning Technique" 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018, Solan, India

[3] A K and Anil Antony "An Intelligent System for Thyroid Disease Classification and Diagnosis" Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

[4] "A Empirical study on Disease Diagnosis using Data Mining Techniques." Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

[5] S "Classification of Thyroid Disease using Data Mining Techniques" International Journal of Pure and Applied Mathematics, Volume 119 No. 12 2018, 13881-13890

[6] "A Comparison of Classification Methods on Diagnosis of Thyroid Diseases" 2015 International Seminar on Intelligent Technology and Its Applications

[7] "Thyroid Data Prediction using Data Classification Algorithm" IJIRST –International Journal for Innovative Research in Science & Technology| Volume 4 | Issue 2 | July 2017

[8] "A Study of Data Mining Techniques to Detect Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 11, September 2017)

[9] Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study" International Research Journal of Engineering and Technology Volume: 03 Issue: 11 | Nov - 2016

[10] Prediction of Thyroid Disease Using Machine learning Techniques" International Journal of Electronics Engineering (ISSN: 0973-7383) Volume 10 • Issue 2 pp. 787-793 June 2018

[11] Transformingclinicaldata into actionable prognosis models: machine learning Framework and fielddeployableapp to predict outcome of Ebola Patients, PLoSNegl. Trop. Dis. 10 (3) (2016) e0004549.

[12] https://machinelearningmastery.com/types-of-classification-inmachine-learning

[13] https://www.kaggle.com/kumar012/hypothyroid

AG-6

ORIGINALITY REPORT

11%	2%	8%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Amulya.R. Rao, B.S. Renuka. "A Machine Learning Approach to Predict Thyroid Disease at Early Stages of Diagnosis", 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020 Publication	5%
2	"Machine Intelligence and Soft Computing", Springer Science and Business Media LLC, 2021 Publication	2%
3	stevegallik.org Internet Source	1%
4	Submitted to University of Wales Institute, Cardiff Student Paper	1%
5	"ICCCE 2020", Springer Science and Business Media LLC, 2021 Publication	1%
6	www.nursa.org Internet Source	1%

7	www.wstst05.softcomputing.net Internet Source	1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

PAPER ID
NECICAIEA2K23113

International Conference on
Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that **Innamuri Sowkya** , Narasaraopeta Engineering College has presented the paper title Thyroid Disease Prediction using Machine Learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

Convenor
Dr.S.V.N.Srinivasu

Chief-Convenor
Dr.S.N.Tirumala Rao

Principal, Patron
Dr. M. Sreenivasa Kumar

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

PAPER ID
NECICAIEA2K23113

International Conference on
Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that **Mounika Lakshmi** , Narasaraopeta Engineering College has presented the paper title Thyroid Disease Prediction using Machine Learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

Convenor
Dr.S.V.N.Srinivasu

Chief-Convenor
Dr.S.N.Tirumala Rao

Principal, Patron
Dr. M. Sreenivasa Kumar

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

PAPER ID
NECICAIEA2K23113

International Conference on
Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that **Neha sree** , Narasaraopeta Engineering College has presented the paper title Thyroid Disease Prediction using Machine Learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

Convenor
Dr.S.V.N.Srinivasu

Chief-Convenor
Dr.S.N.Tirumala Rao

Principal, Patron
Dr. M. Sreenivasa Kumar

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

PAPER ID
NECICAIEA2K23113

International Conference on
Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

Certificate of Presentation

This is to Certify that **Y. Chandana** , Narasaraopeta Engineering College has presented the paper title Thyroid Disease Prediction using Machine Learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

Convenor
Dr.S.V.N.Srinivasu

Chief-Convenor
Dr.S.N.Tirumala Rao

Principal, Patron
Dr. M. Sreenivasa Kumar