A

PROJECT REPORT

ON

"PREDICTION OF THYROID DISEASE"

BY

Sandip Babasaheb Shelke
Sandip Bhimraj Atole

Submitted in Partial fulfillment of
Post Graduation Diploma in Data Science and AI
Savitribai Phule Pune University
For the Academic Year
2022-2023

UNDER THE GUIDANCE OF
Prof.Tausif Shaikh

Department of Technology,
Savitribai Phule Pune University,
Ganeshkhind, Pune-411007.

SAVITRIBAI PHULE PUNE UNIVERSITY

A

PROJECT REPORT

ON

"PREDICTION OF THYROID DISEASE"

BY

Sandip Babasaheb Shelke
Sandip Bhimraj Atole

Submitted in Partial fulfillment of
Post Graduation Diploma in Data Science and AI
Savitribai Phule Pune University
For the Academic Year
2022-2023

UNDER THE GUIDANCE OF

Prof.Tausif Shaikh

Department of Technology,
Savitribai Phule Pune University,
Ganeshkhind, Pune-411007

DEPARTMENT OF TECHNOLOGY

CERTIFICATE

This is to certify that **Sandip Babasaheb Shelke, Sandip Bhimraj Atole**

has successfully completed his project on

"Thyroid Disease Prediction" in partial fulfillment of

3rd Semester work for his Post Graduation Diploma in Data Science and AI

under Savitribai Phule Pune University, for the academic year 2022-2023.

Prof.Tausif Shaikh	**Dr. Manisha Bharati**	**Dr.Aditya Abhyankar**
(Project Guide)	**(Course Coordinator)**	**(H.O.D)**

Signed By
(External Examiner)

Place : PUNE

Date : / /

Student's Declaration

we undersigned student's of Department of Technology, Pune Univeristy 3rd semester, declare that project titled **Prediction of Thyroid Disease using Machine Learning** is a result of our own work and our indebtedness to other work publications, references, if any, have been duly acknowledged. If we are found guilty of copying any other report or published information and showing as our original work. I understand that we shall be liable and punishable by Institute or University, which may include Fail in examination, Repeat study and resubmission of the report or any other punishment that Institute or University may decide.

Name of Student:
Sandip Shelke
Sandip Atole

Enrollment Number:PGD22DS52 & PGD22DS51

Signature:

Date:

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deepest gratitude and Indebtedness to those who contributed through their valuable time and guidance in helping us to achieve success in our project. For the development and production of this we feel indebted: To Prof.Tausif Shaikh of DOT, Pune University and the staff of the Department of Technology for their guidance, suggestion and the encouragement in thick and thin. We also thank our parents for supporting our project work and for financing our efforts. We thank our friends and colleagues for their helping hands, kind advice and their worthy support which made an encouraging environment to work within. Finally, we would like to thank all the people directly or indirectly concerned

Sandip Shelke (PGD22DS52)
Sandip Atole (PGD22DS51)

Abstract

The Thyroid gland is a vascular gland and one of the most important organs of a human body. This gland secretes two hormones which help in controlling the metabolism of the body. The two types of Thyroid disorders are Hyperthyroidism and Hypothyroidism. When this disorder occurs in the body, they release certain type of hormones into the body which imbalances the bodyâs metabolism. Thyroid related Blood test is used to detect this disease but it is often blurred and noise will be present. Data cleansing methods were used to make the data primitive enough for the analytics to show the risk of patients getting this disease. Machine Learning plays a very deciding role in the disease prediction. Machine Learning algorithms, SVM - support vector machine, decision tree, logistic regression, KNN - K-nearest neighbours, ANN Artificial Neural Network are used to predict the patientâs risk of getting thyroid disease. Web app is created to get data from users to predict the type of disease.

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Introduction:

Machine learning (ML) is an application of artificial intelligence (AI) that gives systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer models that can independently access data and use it in learning. Learning begins with observations or data such as examples, first-hand experiences, or instructions to look for data patterns and make better decisions in the future based on the examples we provide. The main goal is to make computers learn and act like humans and to improve their learning over time in an independent manner by providing them with data and information in the form of observations and interactions in the real world. Deep learning is a type of machine learning and artificial intelligence (AI) that mimics the way humans acquire certain types of knowledge. Deep learning is an important part of data science, which includes statistics and predictive modeling. It is extremely useful for data scientists who need to collect, analyze, and interpret large amounts of data. Deep learning makes this process faster and easier. The evolvement computational biology is used in healthcare industry. It allows collection of stored patient data for the prediction of the disease. There are prediction algorithms which are available for the diagnosis of the disease at early stages. The medical information systems are rich of datasets but there are only few intelligent systems which can easily analysis the disease. Over a period of time, the machine learning algorithms have started playing a crucial role in resolving the complex and non-linear problems in the developing model. In any disease prediction models are used to override the features that can be selected from different datasets which can be used in classification in healthy patient as accurate as possible. If this is not done, misclassification can lead to a healthy patient getting unnecessary treatment.

The Thyroid gland is an endocrine gland present in the human neck beneath the Adamâs apple which help in secretion of thyroid hormone that influence the rate of metabolism

and protein synthesis. The thyroid hormones are useful in counting how briskly the heart beats and how fast we burn calories. The thyroid secretes two types of active hormones called levothyroxine (T4) and triiodothyronine (T3). These hormones help in regulating the body temperature. These also aid in energy-bearing and transmission in every part of the body and decisive in protein management. Iodine is considered as the main building block of the thyroid gland. Itâs prostrated in few specific problems. Undersupply of these hormones can lead to hyperthyroidism. There are many originations related to hyperthyroidism and underactive thyroids. There are various kinds of medications like thyroid surgery is liable to ionizing radiation, continual tenderness of the thyroid, deficiency of iodine and lack of enzyme to make thyroid hormones

1.2 Problem Statement

According to statistics, thyroid disorders are on the rise in India. Approximately 1 in 10 Indian adults suffer from thyroid problem. It has been estimated that around 42 million peoples suffer from thyroid disease. Predicting thyroid disorder by doctor is a tedious process which might lead to negative prediction, only experienced doctor can examine the case properly. To assist doctors machine learning can help them in diagnosis of disease and reduces their burden.

Chapter 2

Objective and Scope of Project

The main goal of this project is to predict the risk of thyroid disease based on various factors of individuals. Thyroid disease is a common cause of medical diagnosis and prediction, with an onset that is difficult to forecast in medical research. It will play a decisive role in order to early detection, accurate identification of the disease and helps the doctors to make proper decisions and better treatment. This software system will be a Web application This system will be designed to detect the thyroid disease for better disease management, improved interventions, and more efficient health-care resource allocation. More specifically, Early detection of any preventable diseases is important for better disease management. This system is designed to predict the thyroid disease from patient information such as age, disease history, lab results, procedures and medications.

Chapter 3

Research Methodology

3.1 Data Source

Thyroid disease is a common cause of medical diagnosis and prediction, with an onset that is difficult to forecast in medical research. It will play a decisive role in order to early detection, accurate identification of the disease and helps the doctors to make proper decisions and better treatment. The collected dataset consisted of medical history of 3771 different patient with different age.This dataset provides us with all needed data i.e, the medical features such as Age ,on thyroxine,sick,pregnant,Lithium,Goiter,Tumor,Hypopituitary,TSH,T3,TT4,T4U etc. of the patient that help us in detecting the patient that is diagnosed with any thyroid disease or not.

```
•[2]: path = open("ThyroidDataset.csv")#here we are defining the path of our dataset file
      dataframe = pd.read_csv(path)#here we are using the read_csv function to read the csv file from the path
      dataframe #here we are rpinting the dataframe
```

[2]:

	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	TI
0	41	F	f	f	f	f	f	f	f	f	...	t	125	t	1.14	t	109	f	
1	23	F	f	f	f	f	f	f	f	f	...	t	102	f	?	f	?	f	
2	46	M	f	f	f	f	f	f	f	f	...	t	109	t	0.91	t	120	f	
3	70	F	t	f	f	f	f	f	f	f	...	t	175	f	?	f	?	f	
4	70	F	f	f	f	f	f	f	f	f	...	t	61	t	0.87	t	70	f	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3767	30	F	f	f	f	f	f	f	f	f	...	f	?	f	?	f	?	f	
3768	68	F	f	f	f	f	f	f	f	f	...	t	124	t	1.08	t	114	f	
3769	74	F	f	f	f	f	f	f	f	f	...	t	112	t	1.07	t	105	f	
3770	72	M	f	f	f	f	f	f	f	f	...	t	82	t	0.94	t	87	f	
3771	64	F	f	f	f	f	f	f	f	f	...	t	99	t	1.07	t	92	f	

Figure 3.1: Data Source

3.2 Understanding about Dataset

Feature Name	Type of Feature
Age	Age of the person
Sex	Male,female
on thyroxine	True/False
Query on thyroxine	True/False
On antithyroid medication	True/False
Sick	True/False
Pregnant	True/False
Thyroid surgery	True/False
I131 treatment	True/False
Query hypothyroid	True/False
Query hyperthyroid	True/False
Lithium	True/False
Goiter	True/False
Tumor	True/False
Hypopituitary	True/False
Psych	True/False
TSH measured	True/False
TSH	Thyroid stimulating Harmone floating value
T3 measured	True/False
T3	Triiodothyronine value
TT4 measured	True/False
TT4	Thyroxine value
T4U measured	True/False
T4U	Numerical value
FTI measured	True/False
FTI	Free Thyroxine Index
TBG measured	True/False
TBG	Thyroid-Binding Globulin value

3.3 Visualization:

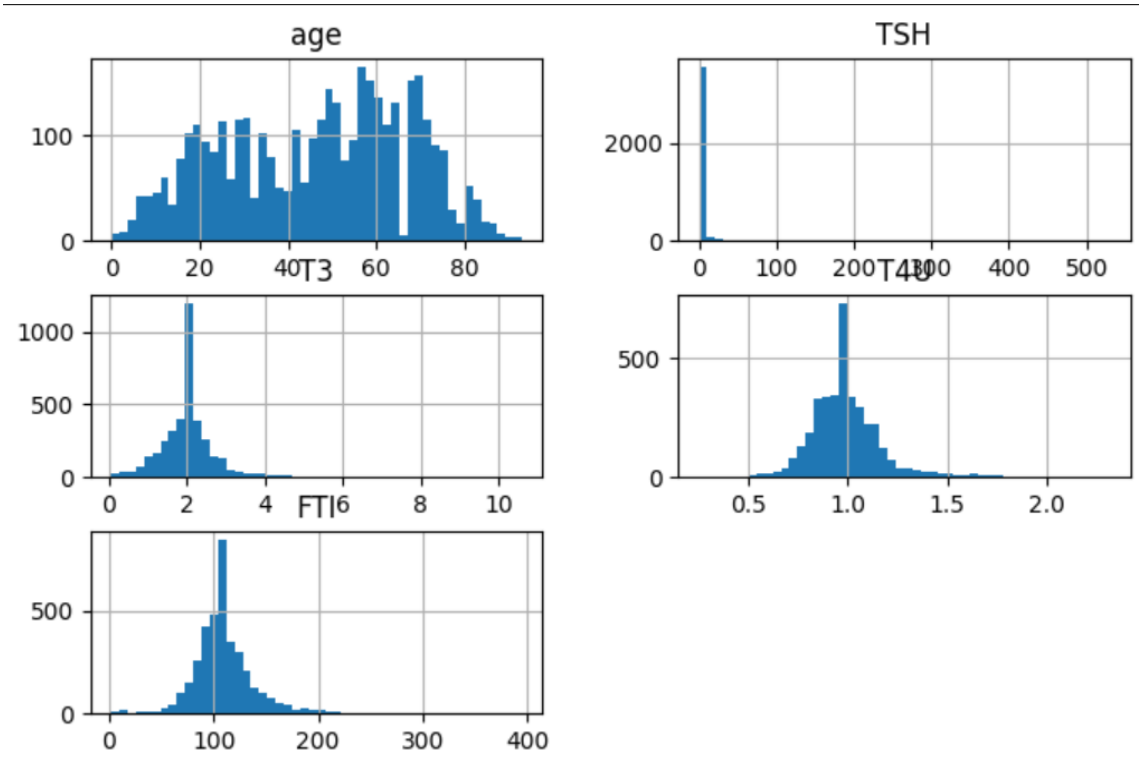

Figure 3.2: Histograms of Dynamic features

3.4 Rational for the study:

Studying thyroid disease detection through machine learning offers several compelling reasons:

1. **Early Detection:**
    - Machine learning models can assist in early identification of thyroid disorders, enabling timely interventions and improved patient outcomes.
2. **Improving Accuracy:**
    - ML models can potentially enhance diagnostic accuracy by analyzing complex patterns in data that might not be easily discernible through traditional methods.
3. **Efficiency:**
    - Automated systems can streamline the diagnostic process, reducing the time and resources required for initial screenings and assessments..
4. **Personalized Medicine:**
    - ML algorithms can aid in tailoring treatment plans by analyzing individual patient data, leading to more personalized and effective therapies.
5. **Handling Complexity:**
    - Thyroid diseases often involve intricate interplays of various factors. Machine learning can handle and analyze multifaceted data, including genetic, environmental, and clinical information, to provide a comprehensive understanding of these conditions.
6. **Research Advancements:**
    - Conducting such studies contributes to the development and validation of innovative technologies in healthcare, fostering advancements in medical research and technology.

    Overall, employing machine learning for thyroid disease detection holds promise in revolutionizing diagnostic capabilities, enhancing patient care, and furthering our understanding of theseÂ conditions.

3.5 Significance of the study:

The significance of using machine learning for thyroid disease detection lies in its potential to improve accuracy, speed, and efficiency in diagnosing thyroid conditions. ML models can analyze large amounts of data, including patient records, lab results, to identify patterns that might not be easily noticeable by human experts. This can lead to earlier detection, personalized treatment plans, and better overall patient outcomes.

3.6 Limitations of the Project:

Detecting thyroid diseases using machine learning has its challenges and limitations:

1. **Data Quality:** Limited or poor-quality data can affect the model's performance. Incomplete or biased datasets might lead to inaccurate predictions.

2. **Class Imbalance:**Thyroid diseases might be relatively rare compared to the overall dataset, causing class imbalances. This imbalance can affect the model's ability to correctly predict these conditions.

3. **Interpretability**: Many machine learning models, especially complex ones like neural networks, can lack interpretability. Understanding why a model makes a particular prediction might be challenging, especially in critical medical applications where interpretability is crucial.

4. **Generalization:**Models trained on specific datasets might not generalize well to new or diverse populations. Variations in demographics, genetics, or environmental factors can affect the model's performance.

5. **Clinical Validation:** Using sensitive health data raises ethical issues, including data privacy, consent, and potential biases in the data or algorithmic decisions.Clinical Validation: Model predictions need to be validated and tested rigorously in clinical settings to ensure their accuracy and reliability before real-world deployment.

   Addressing these limitations requires careful data curation, model development, rigorous testing, and collaboration with healthcare professionals to ensure the accuracy, interpretability, and ethical use of machine learning in thyroid diseaseÂ detection.

3.7 Encoding-Decoding

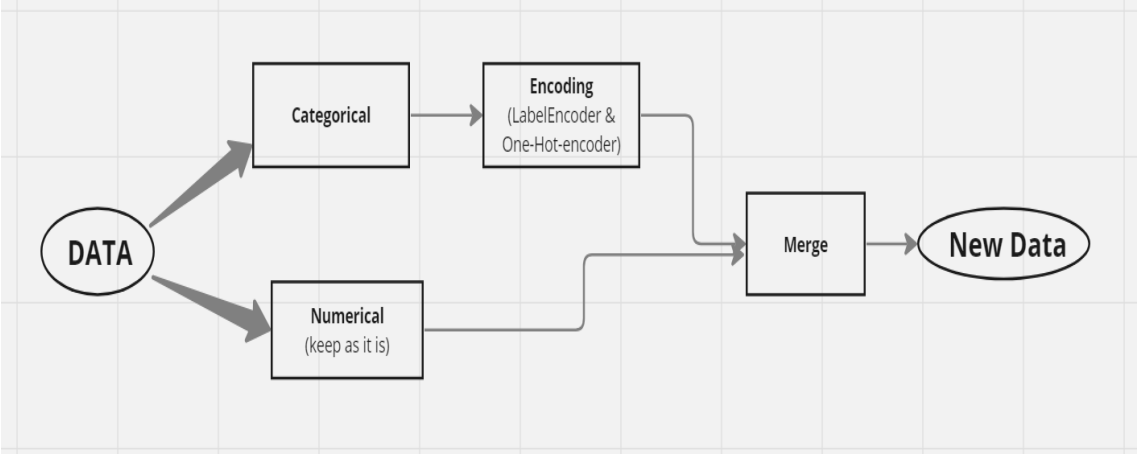

Figure 3.3: Stratergy of Encoding Data

(a) **What is categorical variable encoding and why do we need it?** Categorical variables usually have strings for their values. Many machine learning algorithms do not support string values for the input variables. Therefore, we need to replace these string values with numbers. This process is called categorical variable encoding.

(b) **Categorical Variable Conversion (LabelEncoder and OneHotEncoder):**
LabelEncoder: Converts each category into a unique integer. It is primarily used when the categorical variable has ordinal relationship between categories. It doesn't create additional columns, so it's not suitable for nominal variables with more than two categories.
OneHotEncoder: Creates binary columns (dummy variables) for each category within a categorical variable. It's used for nominal variables, and it is effective when there is no inherent order among the categories.

```
[29]: from sklearn.preprocessing import LabelEncoder
      le = LabelEncoder()

[30]: cols = dataframe.select_dtypes(include=['object'])

[31]: for i in cols.columns:
          try:
              dataframe[i] = le.fit_transform(dataframe[[i]])
          except:
              continue
```

Figure 3.4: Operation on Data

Performed Label Encoding on following features

age,sex,on thyroxine,query on thyroxine,on antithyroid medication,sick,pregnant,thyroid surgery,I131 treatment,query hypothyroid,query hyperthyroid,lithium,goitre,tumor,hypopituitary,psych,TSH,T3,TT4,T4U,FTI,binaryClass

3.8 KNNImputer:

The KNNImputer is a handy tool for handling missing values by imputing them based on the values of their nearest neighbors. Here's a brief overview of how you can use it in a machine learning project:

1. **Data Preprocessing:** Import the necessary libraries (pandas, sklearn), load your dataset, and identify missing values.

2. **Implement KNNImputer:**Use KNNImputer from the sklearn.impute module to fill in missing values. Set the number of nearest neighbors (n neighbors) and other relevant parameters.

3. **Fit and Transform**:Apply the imputer on your dataset using the fit transform method to fill in the missing values.

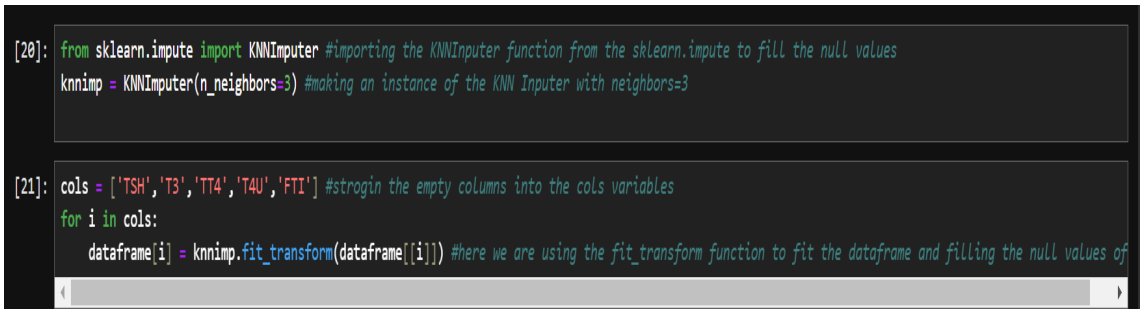

Figure 3.5: Snapshot of imputing missing values using KNNImputer

3.9 Feature-Selection:

Feature selection is crucial for machine learning models in healthcare projects like thyroid disease detection. Some common methods for feature selection include:

1. **Correlation Analysis:**Identifying features highly correlated with the target variable.

2. **Feature Importance:**Techniques like Random Forest or Gradient Boosting can rank features based on their contribution to model accuracy.

3. **Recursive Feature Elimination(RFE)**: Iteratively removing the least significant features and training the model until the best subset is found.

4. **L1 Regularization (Lasso):**Encourages sparsity in coefficients, effectively selecting important features.

5. **Univariate Selection:** SelectKBest or SelectPercentile methods using statistical tests to choose the most relevant features.

   For thyroid disease detection, relevant features might include thyroid hormone levels, TSH, T3, T4, thyroid antibodies, age, gender, symptoms, etc. The choice of features can greatly impact model performance and interpretability.

Chapter 4

Modelling

4.1 Overview of the Data

The thyroid dataset consists of a table with 3771 records and 30 features. Features are distributed as 7 continuous features and 23 categorical features. There are a total 3017 patients in the training set and 705 patients in the test set.

```
[2]: path = open("ThyroidDataset.csv")#here we are defining the path of our dataset file
     dataframe = pd.read_csv(path)#here we are using the read_csv function to read the csv file from the path
     dataframe #here we are rpinting the dataframe
```

[2]:

	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	TI
0	41	F	f	f	f	f	f	f	f	f	...	t	125	t	1.14	t	109	f	
1	23	F	f	f	f	f	f	f	f	f	...	t	102	f	?	f	?	f	
2	46	M	f	f	f	f	f	f	f	f	...	t	109	t	0.91	t	120	f	
3	70	F	t	f	f	f	f	f	f	f	...	t	175	f	?	f	?	f	
4	70	F	f	f	f	f	f	f	f	f	...	t	61	t	0.87	t	70	f	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3767	30	F	f	f	f	f	f	f	f	f	...	f	?	f	?	f	?	f	
3768	68	F	f	f	f	f	f	f	f	f	...	t	124	t	1.08	t	114	f	
3769	74	F	f	f	f	f	f	f	f	f	...	t	112	t	1.07	t	105	f	
3770	72	M	f	f	f	f	f	f	f	f	...	t	82	t	0.94	t	87	f	
3771	64	F	f	f	f	f	f	f	f	f	...	t	99	t	1.07	t	92	f	

Figure 4.1: Data Snapshot

4.2 Separating feature and target variables and doing train test split

After finding doing data preprocessing and getting the required data, now we separate the target and feature variables into two sets. These sets will be used for the train test split where we put 80% training and 20% for testing.

```
from sklearn.model_selection import train_test_split #importing the train test split function from model selection of skelarn
X_train,X_test,y_train,y_test = train_test_split(df2,y,test_size=0.20,random_state=42) #dividing the dataset into training and testing dataset3#


from sklearn.metrics import accuracy_score
```

Figure 4.2: train test split

4.3 Data Analysis Steps

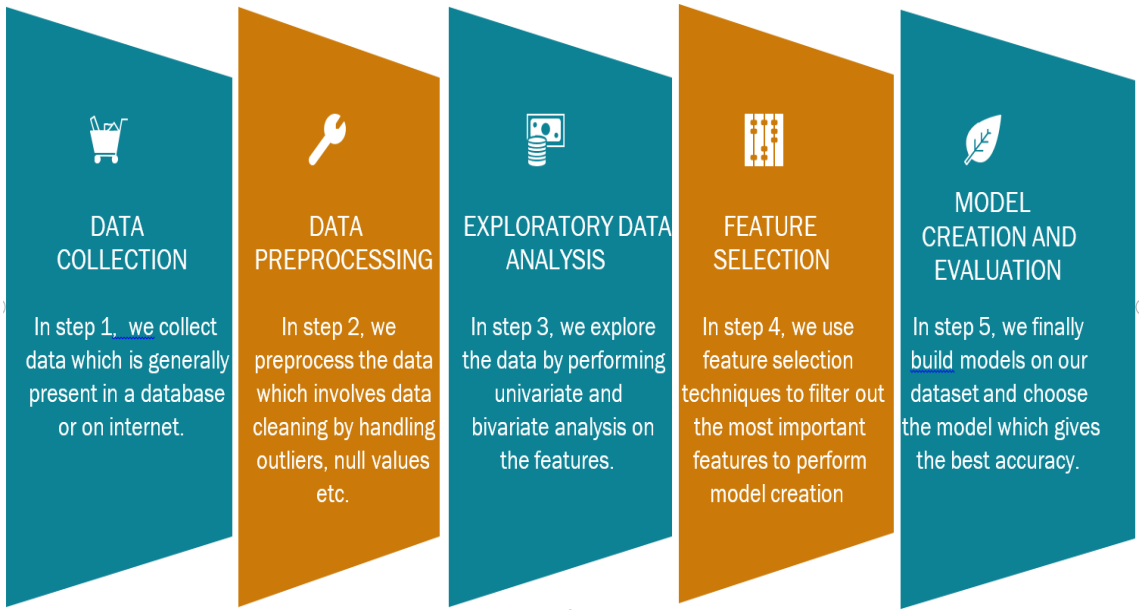

Figure 4.3: Data Analysis Steps

4.4 Models to be applied

To evaluate the thyroid disease prediction, multiple ML classifiers can be used:Decision Trees, SVM with RBF kernel,K-Nearest Neighbors(KNN), Gaussian Naive Bayes, logistic regression, XGBoost, and random forest classifier.

Decision-Tree-Classifier:

A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It creates a tree-like model of decisions and their consequences. The algorithm splits the dataset into subsets based on the most significant attribute, making decisions at each node until a leaf node is reached. Each branch represents a decision or outcome.

Strengths: Easy to understand and interpret, handles both categorical and numerical data, and can capture non-linear relationships.

Weaknesses: Prone to overfitting if the tree is too deep, sensitive to small variations in the data.

Random-Forest-Classifier:

A random forest algorithm is a supervised machine learning algorithm that is extremely popular and used for classification and regression problems in machine learning. We know that a forest is made up of numerous trees and the more trees there are, the more robust it is. The larger the number of trees in a random forest algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains multiple decision trees for different subsets of the given data set and averages them to improve the prediction accuracy of that data set. It is based on the concept of ensemble learning, which combines multiple classifiers to solve a complex problem and improve the performance of the model.

Logistic-Regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc., but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving classification problems.

4.5 Performance

Model	Accuracy
Logistic Regression 1	96.02 %
Random Forest Classifier	95.36 %
Decision Tree Classifier	99.02 %

4.6 Tools used:

Figure 4.4: logos

Chapter 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

A thyroid disease detection model was developed using three ML classification modeling techniques. This project detects people with thyroid diseases by extracting the patient's medical history leading to fatal thyroid disease from a dataset containing the patient's medical history such as age, blood test, medical health, etc. This thyroid disease detection system helps a patient based on his/her clinical information that they have been diagnosed with a previous thyroid disease. The algorithms used in building the given model are DecisionTreeClassifier, random forest classifier and ANN. The accuracy of our model is 99%. By using these computer-aided techniques, we can predict the patient quickly and better and the costs can be reduced significantly. In summary, this project helps us detect the patients who has diagnosed with thyroid disease by cleaning the dataset and applying DecisionTreeClassifier and ANN to achieve an average of 99% accuracy for our model, which is better is than the previous models.

5.2 Future Enhancement

Enhancing a thyroid disease detection project using machine learning can involve various strategies:

1. **Data Augmentation:** Augment the existing dataset by collecting more diverse samples to account for different demographics, ethnicities, and regions. This can help improve the model's generalization.

2. **Advanced Feature Engineering:** Explore additional relevant features or transformations that might better capture the nuances of thyroid health, such as ratios of hormone levels, temporal patterns, or interactions between variables.

3. **Ensemble Methods:** Experiment with ensemble techniques like Random Forests, Gradient Boosting, or stacking multiple models to combine their predictive power and enhance performance.

4. **Deep Learning Architectures:** Consider using deep learning models like neural networks or convolutional neural networks (CNNs) to automatically learn hierarchical representations from raw data, potentially capturing complex patterns for improved detection.

5. **Hyperparameter Tuning:**Optimize hyperparameters of the chosen models through techniques like grid search, random search, or Bayesian optimization to further enhance performance.

6. **Handling Imbalance:** If the dataset suffers from class imbalance, employ techniques like oversampling, undersampling, or using algorithms that handle imbalanced datasets effectively.

7. **Interpretability:** Focus on making the model more interpretable by using techniques like SHAP values, LIME, or model-specific feature importance to understand the model's predictions and increase trust.

8. **External Data Sources:** Incorporate additional external data sources (e.g., research studies, medical records) to enrich the dataset and improve the model's predictive capabilities.

9. **Continuous Monitoring and Updating:** Implement a system for continuous monitoring of the model's performance in real-world scenarios and update it periodically as new data becomes available.

Chapter 6

REFERENCES

1. Ankita Tyagi and Ritika Mehra. (2018).Interactive Thyroid Disease Prediction System using Machine Learning Techniques published on ResearchGate.

2. International Journal of Engineering Applied Sciences and Technology, 2021 Vol. 5, Issue 9, ISSN No. 2455-2143, Pages 173-177 Published Online January 2021 in IJEAST (http://www.ijeast.com)

3. Thyroid Disease: Causes, Symptoms, Risk Factors, Testing Treatment (clevelandclinic.org)

4. Proceedings of the International Conference on Machine Learning (ICML), International Conference on Healthcare Informatics (ICHI).

5. Thyroid Disease Dataset from Kaggle: Look for datasets specific to thyroid disease that could be used for training and testing machine learning models.