**THROAT CANCER DETECTION**

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

in

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

by

**K.SAIRAM (2103A51519)**

**S.NAGASAI VINAY (2103A51290)**

**R.JAYANTH REDDY (2103A51489)**

Under the guidance of

**Mr. S.NARESH KUMAR**

Assistant Professor, Department of CSE.

**

**Department of Computer Science and Artificial Intelligence**

**

**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that project entitled **“THROAT CANCER DETECTION**" is the bonafied work carried out by K.**SAIRAM , S.NAGASAI VINAY, R.JAYANTH REDDY** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

**Mr. S Naresh Kumar Dr. M.Sheshikala**

Asst. Professor, Assoc. Prof. & HOD (CSE),

S R University, S RUniversity,

Ananthasagar, Warangal. Ananthasagar, Warangal.

**ACKNOWLEDGEMENT**

We express our thanks to Course co-coordinator **Mr.** **S.Naresh Kumar, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude toHead of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**Table of Contents**

**Chapter No. Throat Cancer Detection Page No.**

1. **Introduction**
   1. Overview 5
   2. Problem Statement 6
   3. Existing system 7
   4. Proposed system 7
   5. Objectives 7
   6. Architecture 8
2. **Literature survey**
   * 1. Document the survey done by you 10
3. **Data pre-processing** 
   1. Dataset description 11
   2. Data cleaning 17

3.3 Data Visualization 21

1. **Methodology**
   1. Procedure to solve the given problem 35
   2. Model architecture 44
   3. Software description 45
2. **Results and discussion** 46
3. **Conclusion and future scope** 46………
4. **References**  47

**CHAPTER-1.INTRODUCTION**

* 1. **OVERVIEW**

Throat cancer detection is the process of identifying the presence of cancerous cells in the throat. The goal of throat cancer detection is to identify cancer at an early stage when it is most treatable.There are different methods for detecting throat cancer, including physical exams, imaging tests, and biopsies. Physical exams involve a doctor examining the throat and looking for any abnormal growths or lumps. Imaging tests, such as X-rays, CT scans, or MRIs, create images of the throat that can reveal any abnormalities or signs of cancer. A biopsy involves taking a sample of tissue from the throat and examining it under a microscope to determine if it contains cancerous cells.

Artificial intelligence and machine learning techniques can be used to help improve the accuracy and efficiency of throat cancer detection. By analyzing large amounts of medical data and images, AI algorithms can identify patterns and markers that are associated with throat cancer, helping doctors make more accurate diagnoses. AI-powered tools can also assist in identifying cancerous lesions in images, which can be difficult for human eyes to detect.

Overall, throat cancer detection is an important area of research that can help improve outcomes for patients with this disease. By combining traditional medical techniques with AI and machine learning approaches, we can improve our ability to detect and treat throat cancer.

* 1. **PROBLEM STATEMENT**

Develop a machine learning model that can accurately detect the presence of throat cancer in patients based on the analysis of their medical records and symptoms. The model should be able to analyze a set of input variables, including age, sex, smoking habits, alcohol consumption, throat pain, hoarseness, difficulty swallowing, and other relevant medical history, and provide a binary output indicating the likelihood of throat cancer. The dataset used for training and testing the model should be large enough to provide robust results and should be sourced from reliable medical sources. The model's performance should be evaluated using appropriate metrics such as accuracy, sensitivity, specificity, and F1-score. The ultimate goal of this project is to provide a tool that can aid doctors in making more informed diagnoses and treatment decisions for patients with throat cancer.

* 1. **EXISTING SYSYEMS**

Image-based systems: These systems use images of the throat, taken using an endoscope or other imaging tools, to detect signs of cancer. The images are analyzed using machine learning algorithms that can identify abnormal cells and tissues.

Voice analysis systems: These systems analyze the patient's voice to detect signs of cancer. They use machine learning algorithms to analyze changes in pitch, tone, and other vocal characteristics that may indicate the presence of cancer.

Electronic health record (EHR) systems: These systems use patient data, such as medical history, lab results, and imaging studies, to detect signs of cancer. Machine learning algorithms are used to analyze this data and identify patterns that may indicate the presence of cancer.

Symptom-based systems: These systems use patient-reported symptoms, such as sore throat, difficulty swallowing, or hoarseness, to detect signs of cancer. Machine learning algorithms are used to analyze the symptoms and determine the likelihood of cancer.

Overall, these systems aim to improve the accuracy and efficiency of throat cancer detection, allowing for earlier diagnosis and better outcomes for patients

* 1. **PROPOSED SYSTEM**

The proposed system will focus on analyzing and detecting the number of confirmed throat cancer cases based on input data. The system will detect the throat cancer by analyzing several quantities such as thyroxine levels and other information collected from diagnosis and will be used to accurately detect the throat cancer.

* 1. **DEFINE OBJECTIVES**

The objective of an AIML project for throat cancer detection would be to develop an algorithm or system that can accurately identify the presence of throat cancer in patients using medical imaging data, such as CT scans or MRIs.

The system would need to be able to analyze the images and detect any signs of abnormal growths or tumors in the throat area. The goal would be to create a tool that could assist medical professionals in making faster and more accurate diagnoses of throat cancer, potentially leading to earlier treatment and better patient outcomes.

* 1. **OVERALL ARCHITECTURE**

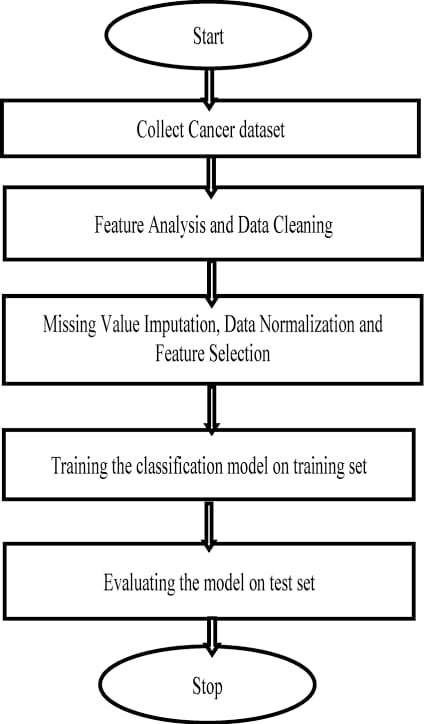
1. Data collection: Gathering a large and diverse dataset of throat cancer patients and healthy individuals, including their medical records, diagnostic reports, imaging studies, and other relevant information.

2. Preprocessing and feature extraction: Cleaning and preparing the dataset for analysis, including removing any noise or irrelevant data, selecting the relevant features, and transforming the data into a suitable format for ML algorithms.

3. Model selection and training: Selecting the appropriate ML algorithm(s) based on the data and the problem, and training them on the preprocessed data to build predictive models that can accurately distinguish between throat cancer patients and healthy individuals.

4. Model evaluation and validation: Evaluating the performance of the trained models using various metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC, and validating them using independent datasets or cross-validation techniques.

5. Deployment and monitoring: Deploying the best-performing model(s) in a real-world clinical setting and monitoring their performance and accuracy over time, including updating the model as needed to reflect changes in the data or the disease.



**Chapter -2: LITERATURE SURVEY**

* 1. **SURVEY:**

Reddy, D. K. N., “Prior Prediction and Impediment of Cancer using Machine Learning Process”, International Journal of Psychosocial Rehabilitation, 24(4), 2020

* Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D.,

Almangush, A.,“Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool”, VirchowsArchiv, 475(4), 489-497, 2019.

* Liu, Huicong, Wei Dong, Yunfei Li, Fanqi Li, JiangjunGeng, Minglu Zhu, Tao Chen, Hongmiao Zhang, Lining Sun, Chengkuo Lee, "An epidermal sEMG tattoo-like patch as a new human–machine interface for patients with loss of voice", Microsystems

&Nanoengineering 6(1), 1-13, 2020.

* Latchoumi, T. P., &Sunitha, R. Multi agent systems in distributed datawarehousing. In 2010 International Conference on Computer and Communication Technology (ICCCT) (pp. 442447). IEEE, 2010.
* Loganathan, J., Janakiraman, S., &Latchoumi, T. P. (2017). A Novel Architecture for Next Generation Cellular Network Using Opportunistic Spectrum Access Scheme. Journal of Advanced Research in Dynamical and Control Systems,(12), 1388-1400.
* Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W., Ho, R. C., “Characterizing Artificial Intelligence Applications in Cancer Research: A Latent Dirichlet Allocation Analysis”, JMIR Medical Informatics, 7(4), e14401, 2019.
* Korach, Z. T., Cato, K. D., Collins, S. A., Kang, M. J., Knaplund, C., Dykes, P. C., Chang, F.,

“Unsupervised Machine Learning of Topics Documented by Nurses about Hospitalized Patients Prior to a Rapid-Response Event”, Applied Clinical Informatics, 10(05), 952-963, 2019 .

* G. Castellano, L. Bonilha, L.M. Li, F. Cendes,“Texture analysis of medical images”, Clinical Radiology, 59, 1061–1069, 2004.
* Pereda, M., Estrada, E., “Visualization and machine learning analysis of complex networks in hyperspherical space”, Pattern Recognition, 86, 320-331, 2019.
* Feltes, B. C., Chandelier, E. B., Grisci, B. I., Dorn, M., “CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in

Cancer Research”, Journal of Computational Biology, 26(4), 376-386, 2019.

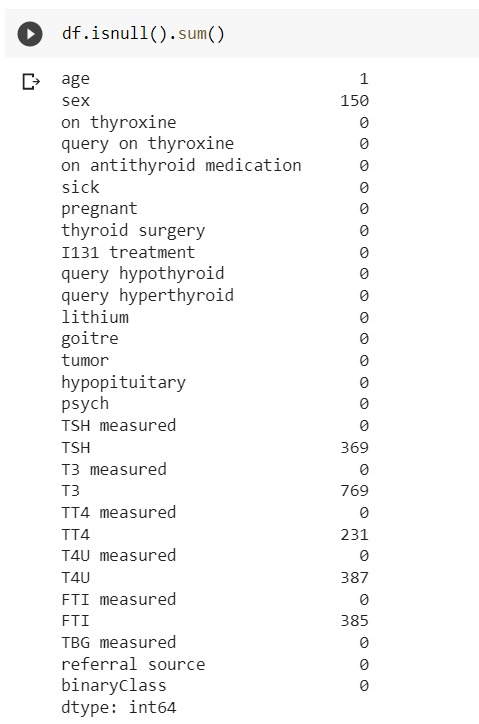
**CHAPTER-3 :DATA PREPROCESSING**

* 1. **DATA SET DESCRIPTION:**
* This data set contains 30 attributes , which were then divided into 972 test instances
* This dataset contains 30 attributes
* The entire dataset has the statistical data.
* The attributes are:

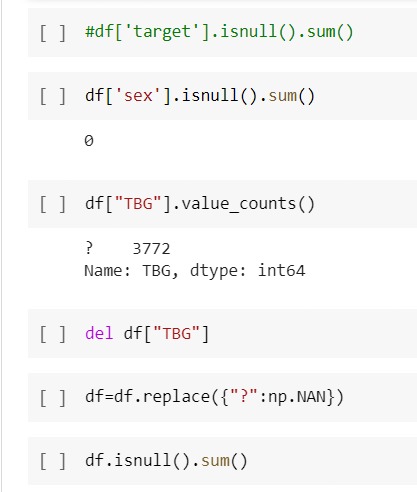
1. AGE
2. Sex
3. on thyroxine
4. query on thyroxine
5. on antithyroid medication
6. sick
7. Pregnant
8. Thyroid surgery
9. I131
10. query hypothyroid
11. query hypothyroid
12. lithium
13. goitre
14. Tumour
15. Hypopituitary.
16. Psych
17. TSH measured
18. TSH
19. T3 measured
20. T3
21. TT4 MEASURED
22. TT4
23. T4U measured
24. T4U
25. FTI measured
26. FTI
27. TBG measured
28. BINARY CLASS
29. dtype: object
    1. **DATA CLEANING:**

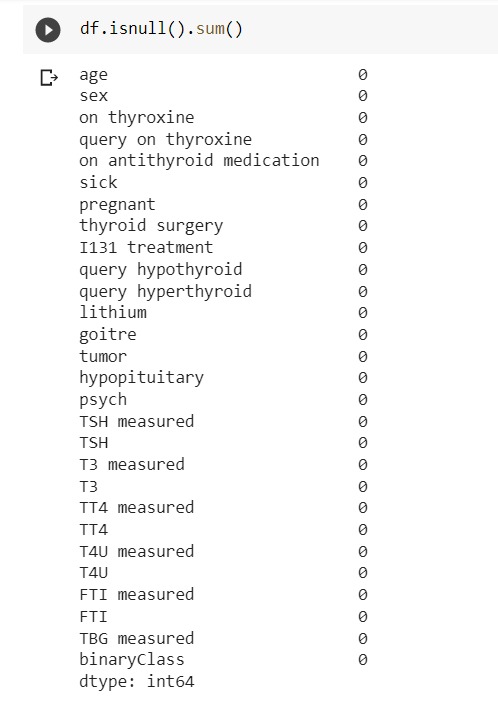
Data cleaning is an important step in any machine learning task. It involves identifying and correcting errors and inconsistencies in the dataset to ensure that the data is accurate, complete, and ready for analysis.

Here are some common data cleaning steps that can be applied to sculpture detection datasets:

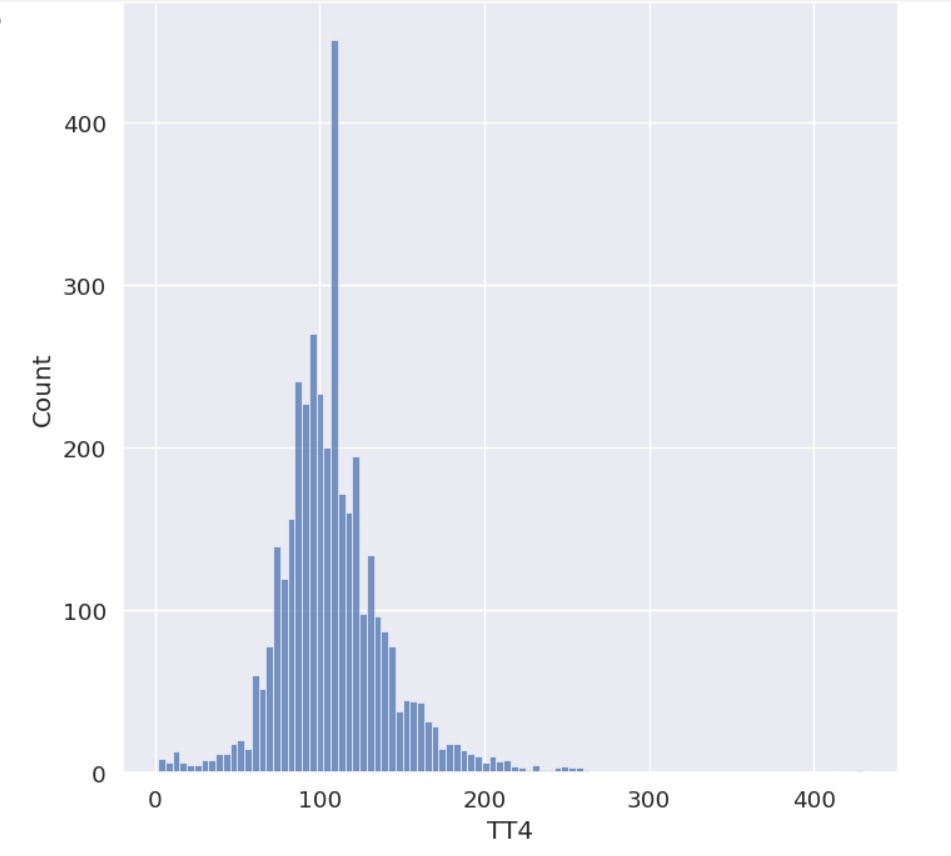
1. Handling Missing Data: One of the most common issues in statistical data is missing values. You can either remove the rows or columns with missing values, impute them with the mean, median, or mode of the column, or use advanced techniques like interpolation or machine learning algorithms to predict the missing values.
2. Handling Categorical Data: Statistical data can have categorical variables, such as gender or occupation. You can convert categorical data into numerical data using techniques like one-hot encoding or label encoding, which convert categorical variables into a format that can be used in statistical analysis.
3. Data Standardization: Statistical data may have variables with different units or scales, which can affect the analysis. You can standardize the data by scaling the variables to have the same mean and variance.
4. Data Normalization: Normalizing the data involves transforming it to a standard distribution, such as the normal distribution. This can help improve the performance of statistical models that assume a normal distribution of the data.
5. Handling Duplicates: Duplicate records can occur in statistical data, which can bias the analysis. You can remove duplicates using techniques like dropping rows or columns with duplicate values, or merging duplicate records into a single record.
6. 

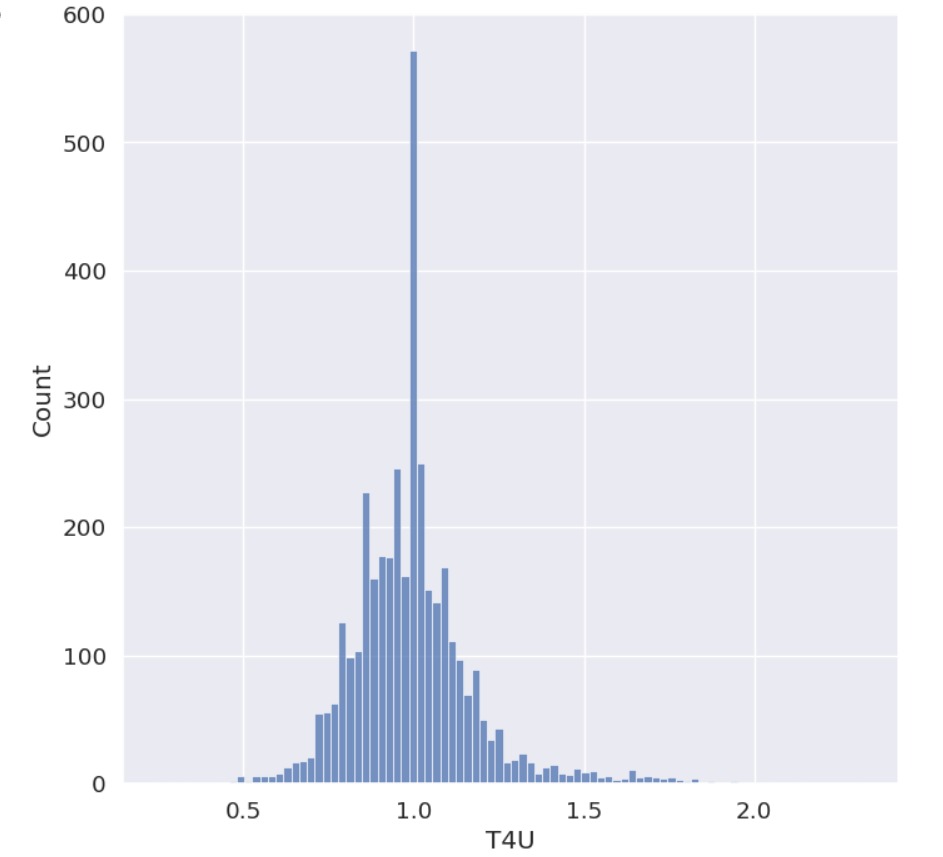


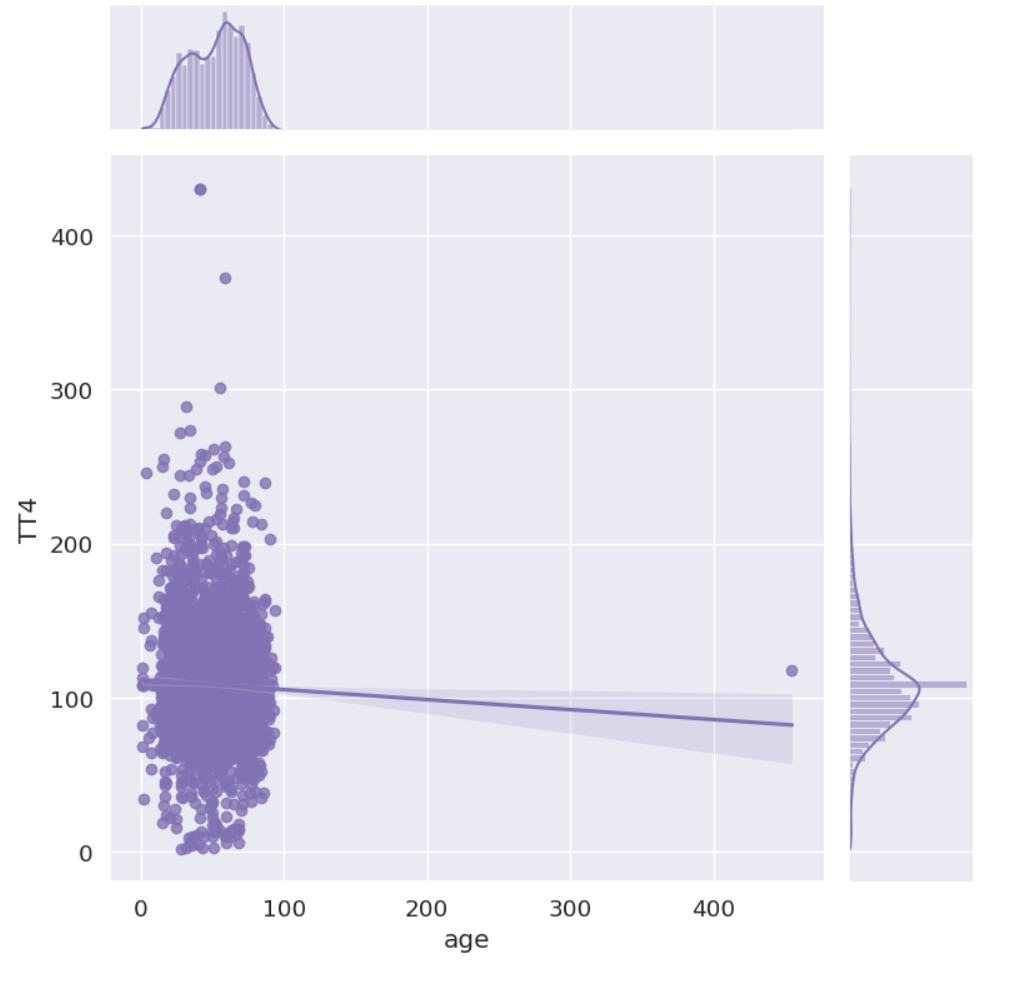
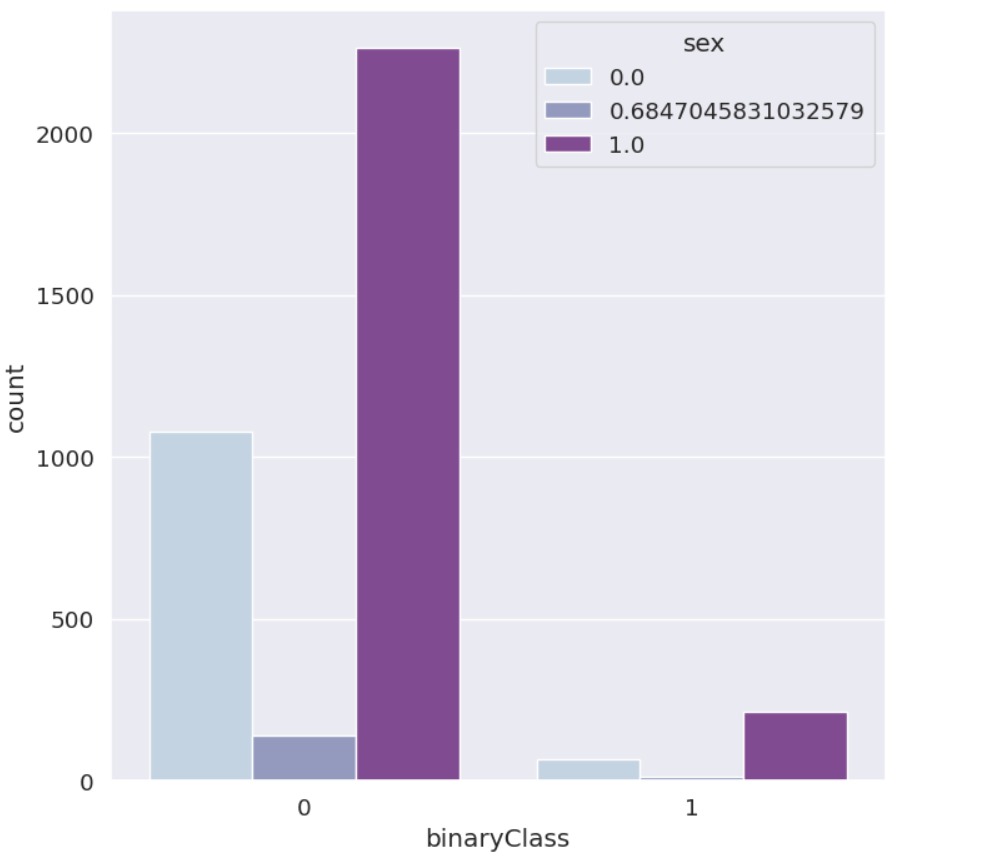


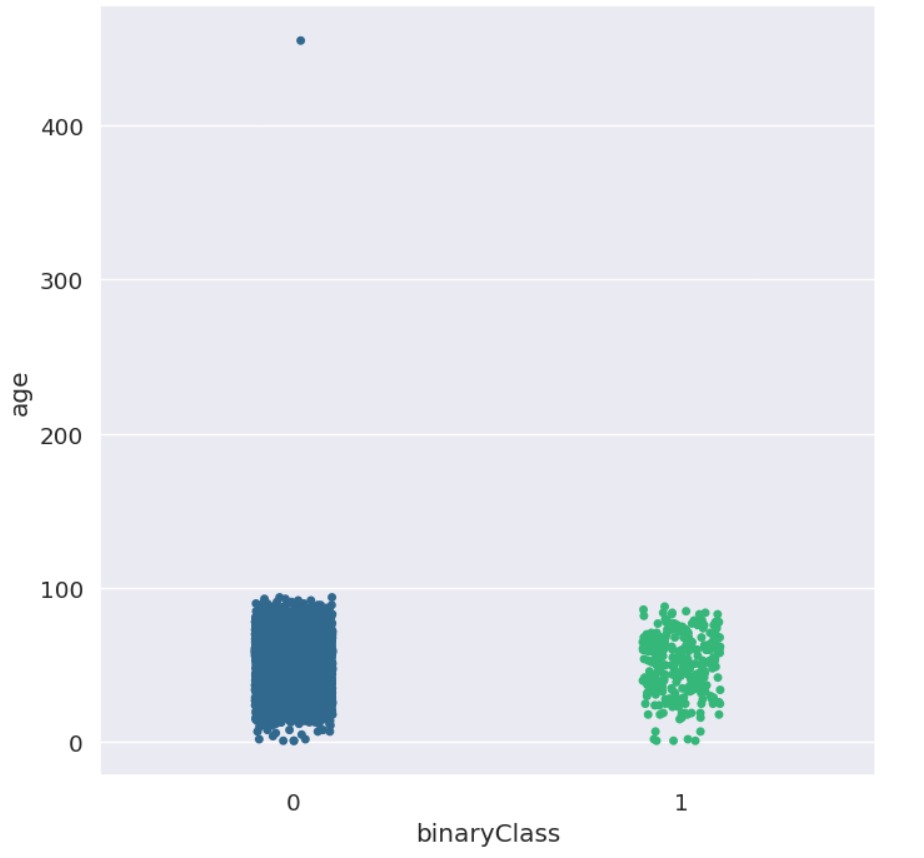


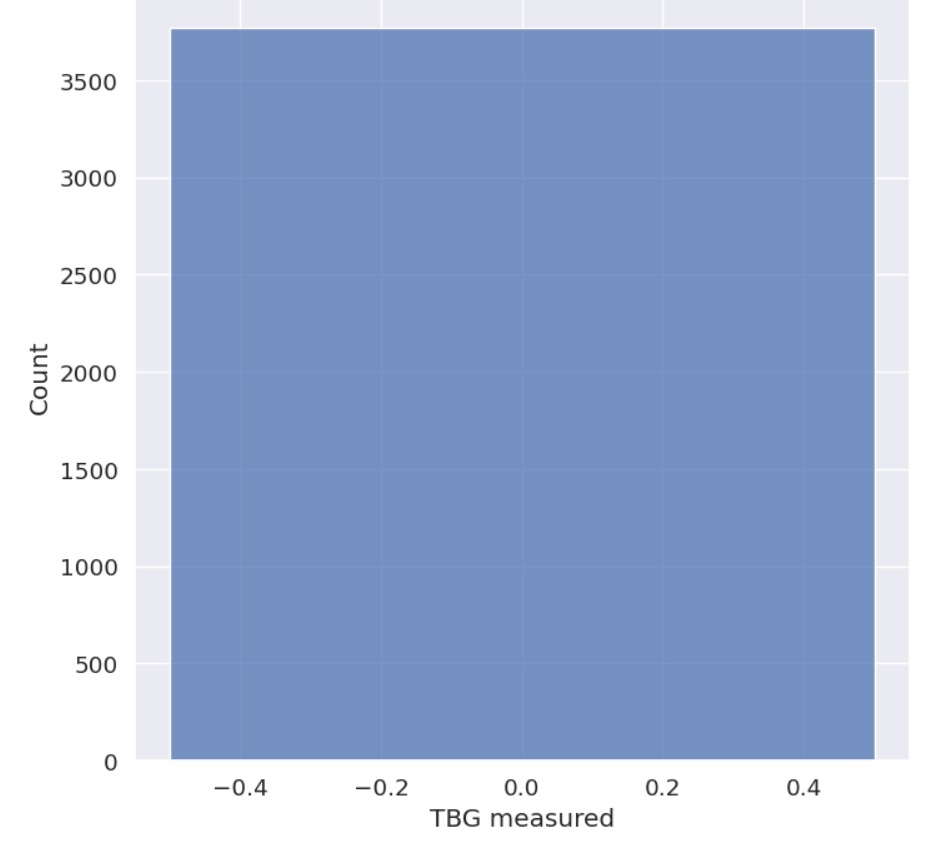
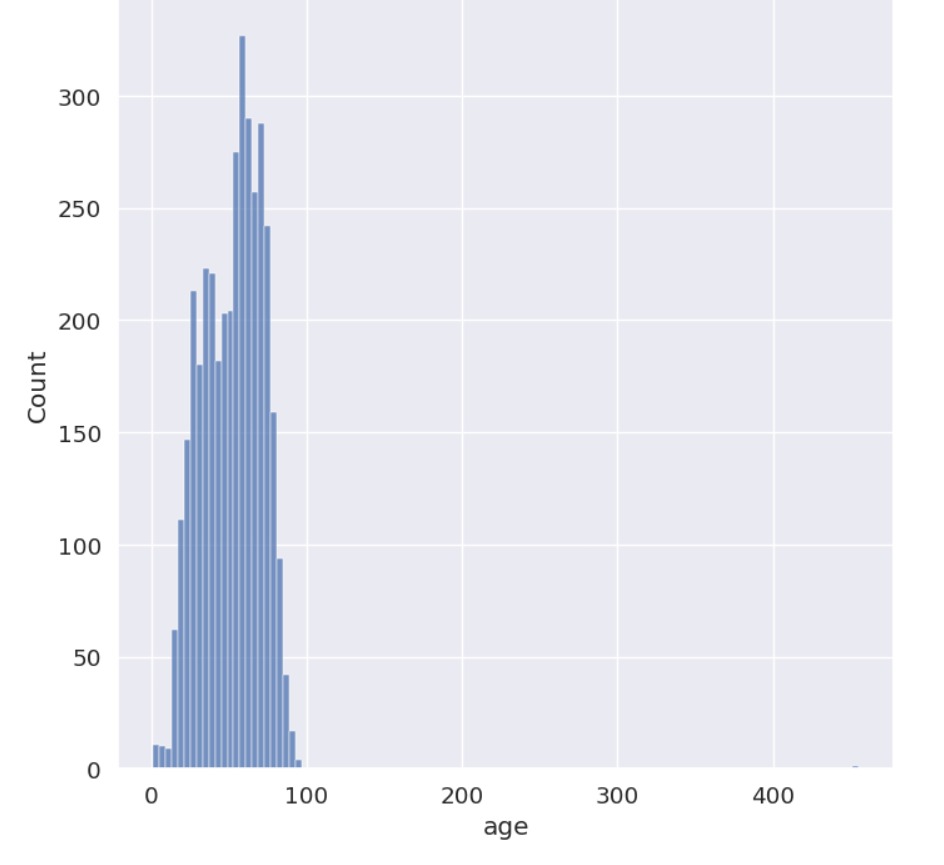
* 1. **DATA VISUALISTION:**

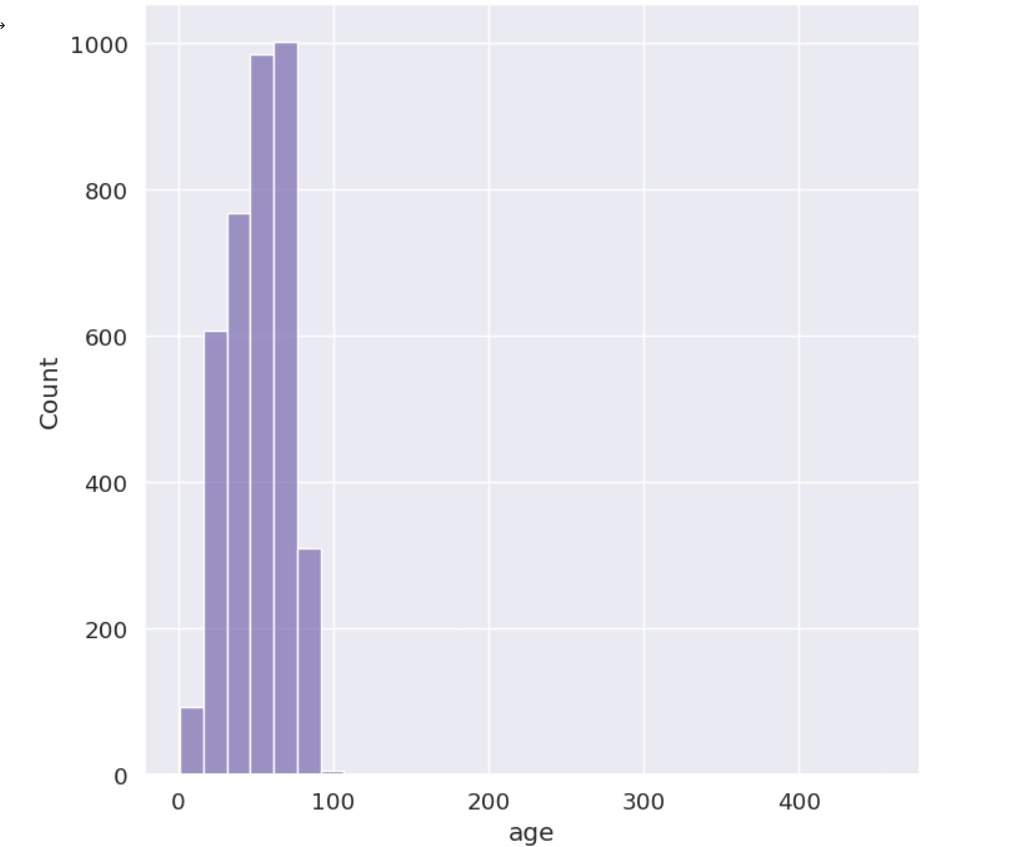


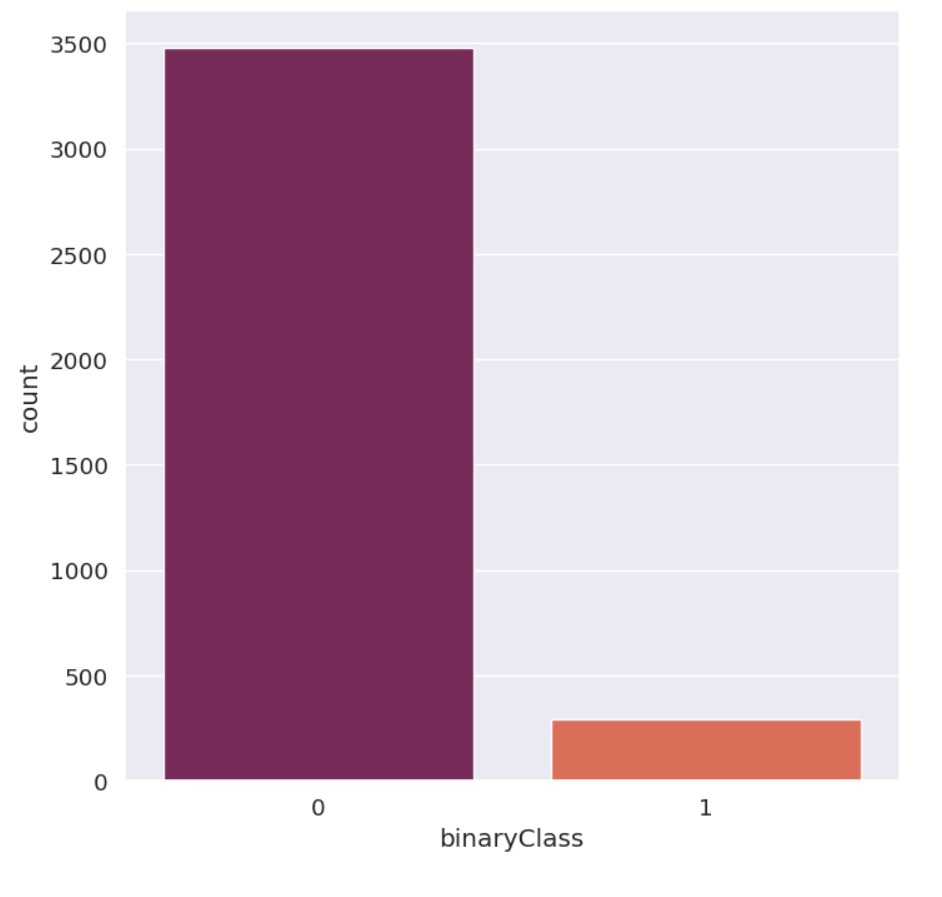


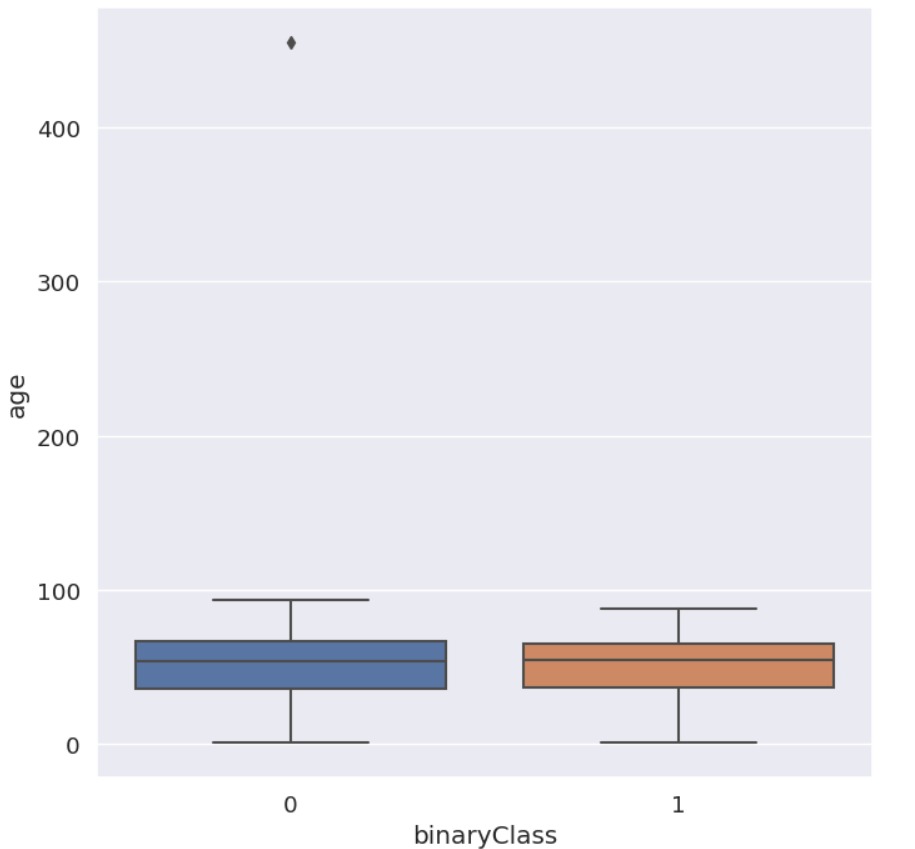


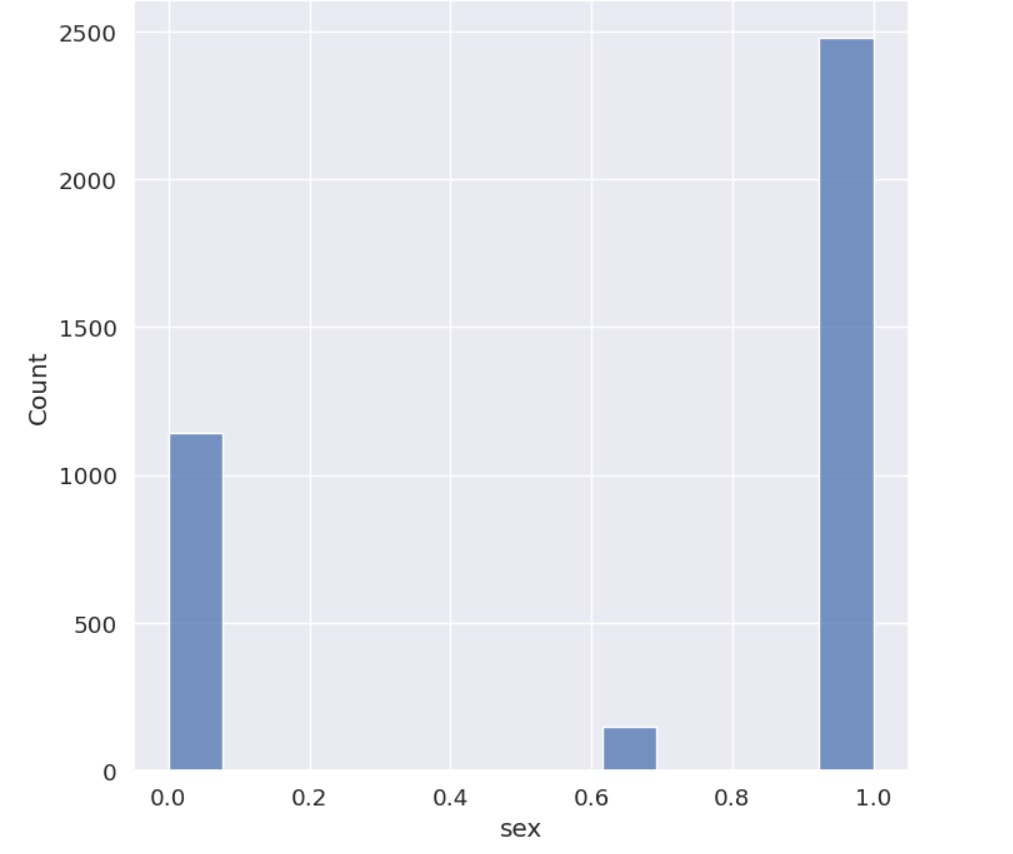
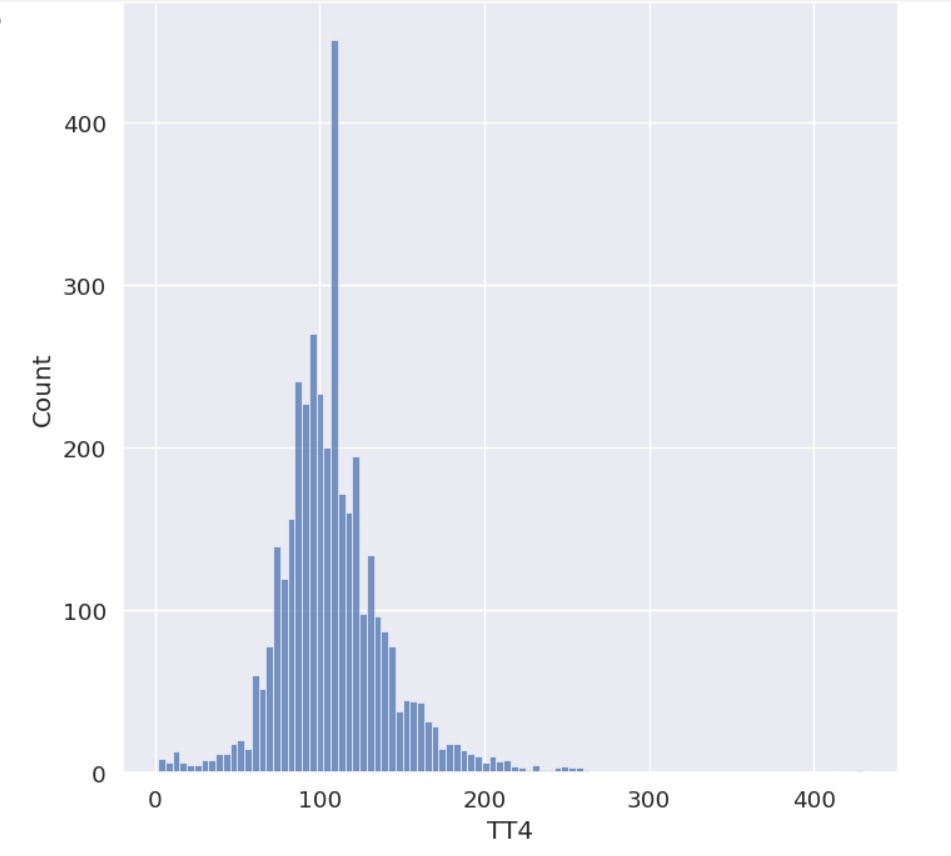


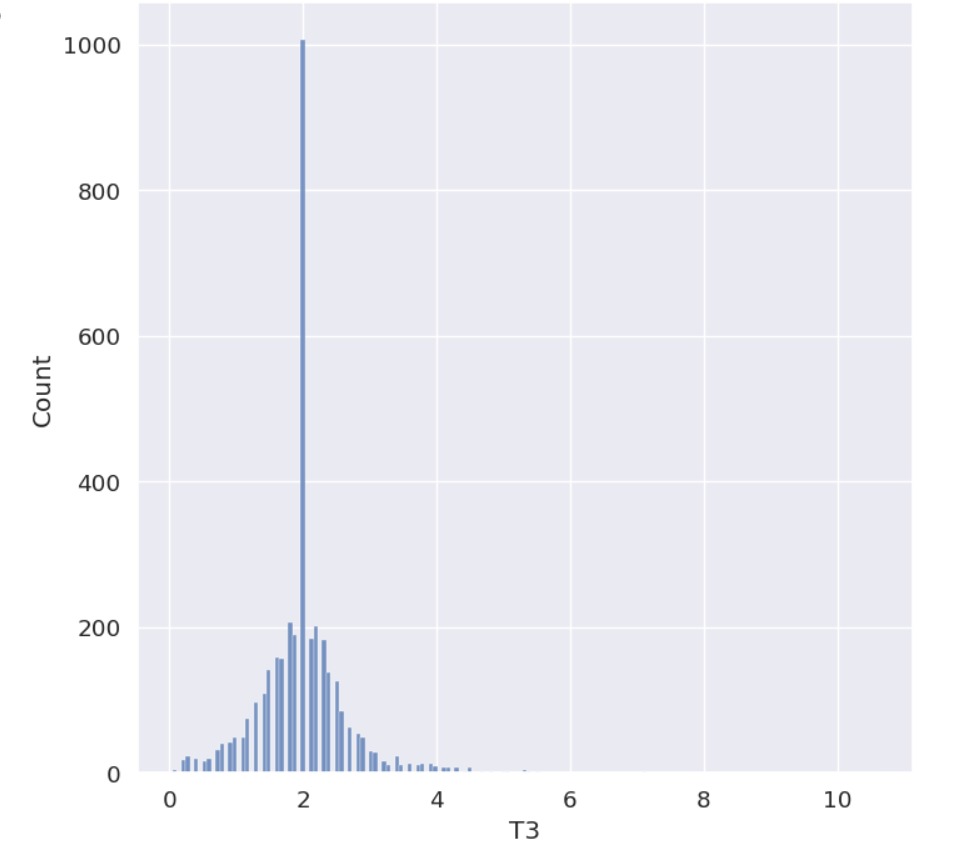


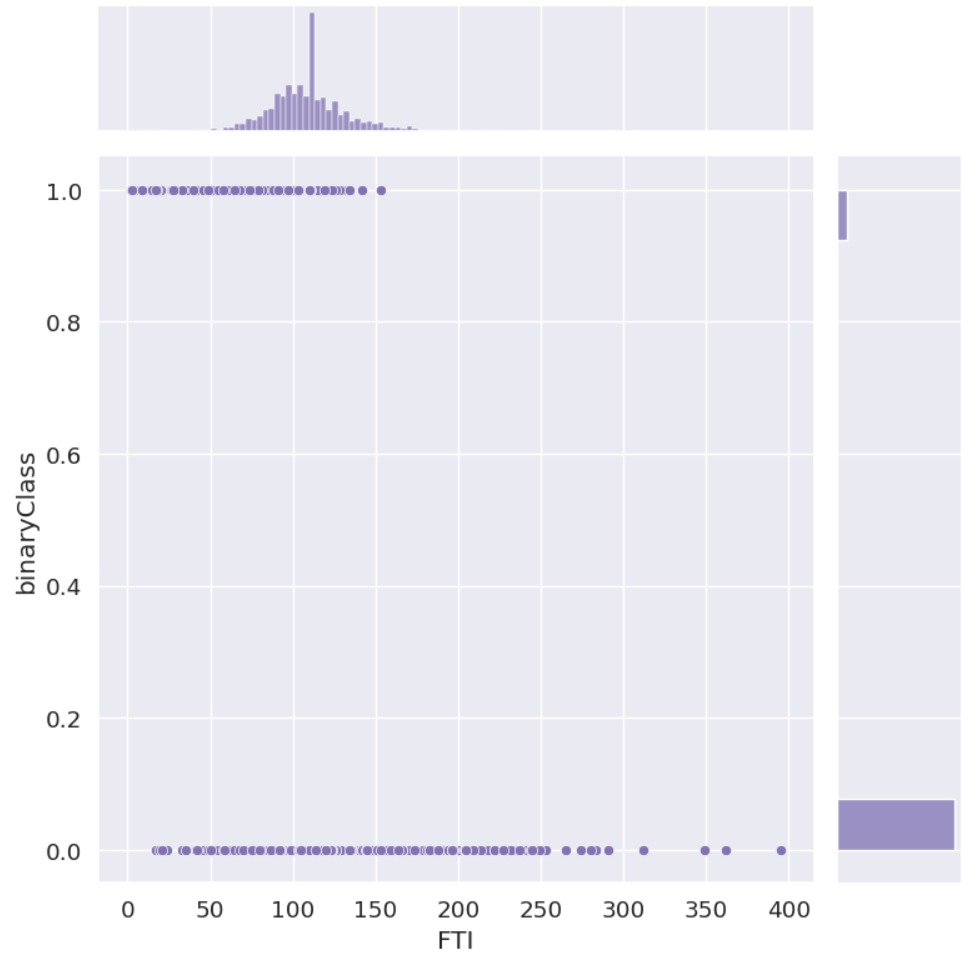


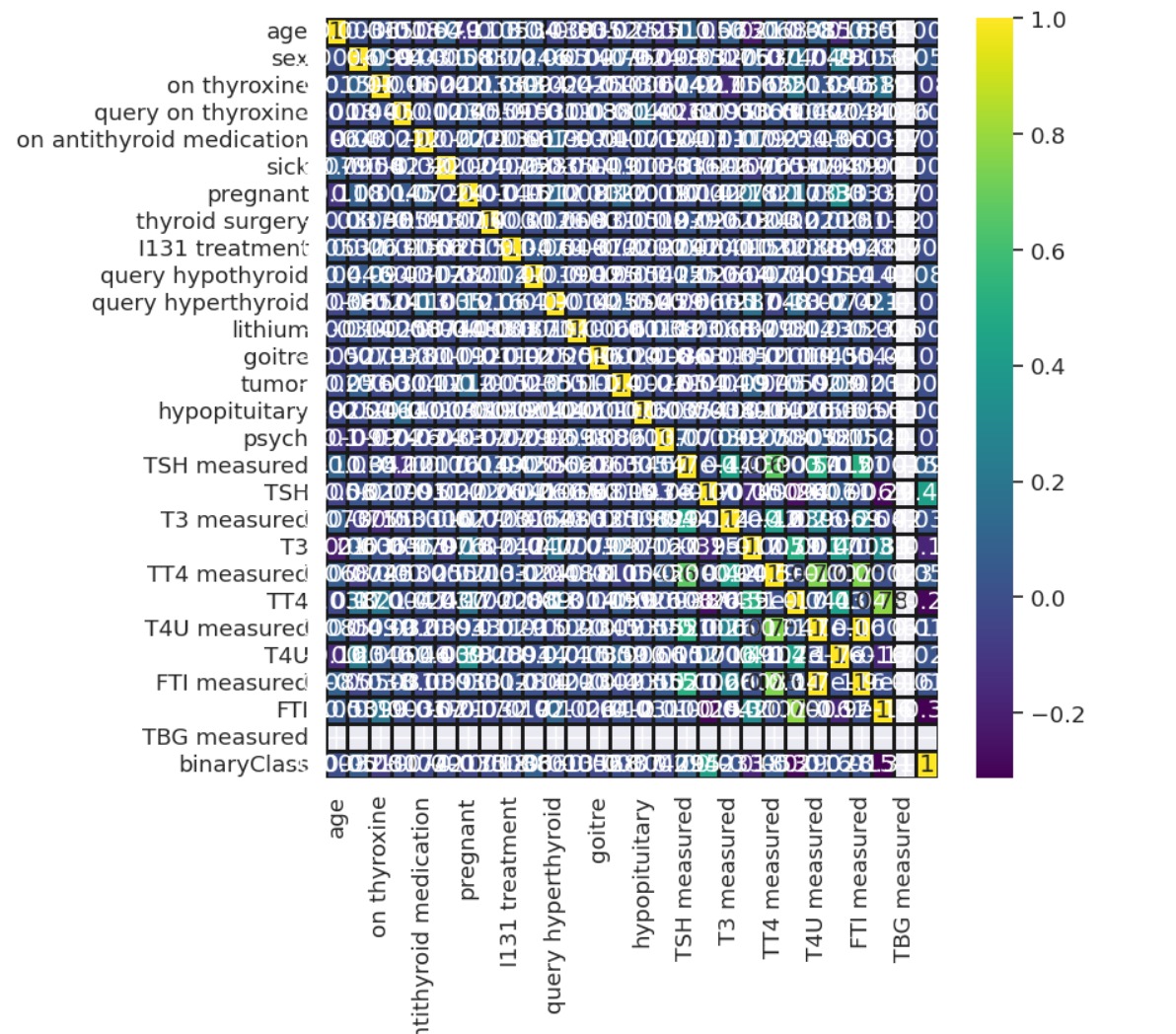


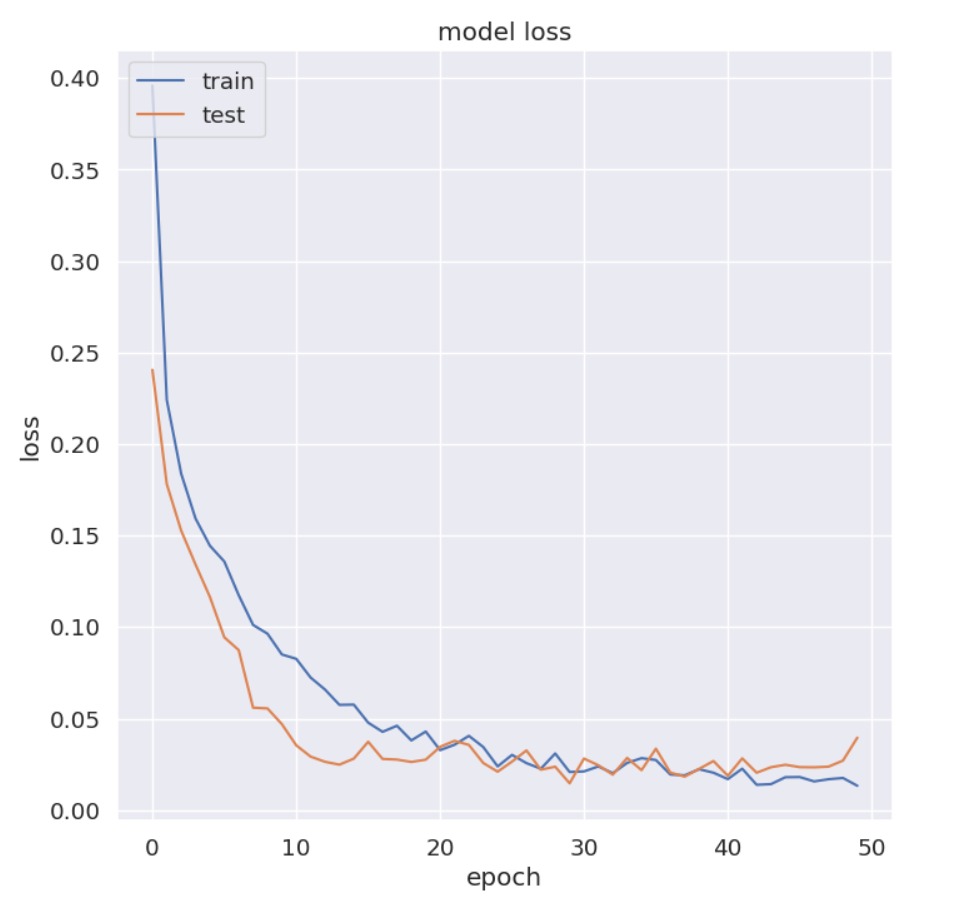


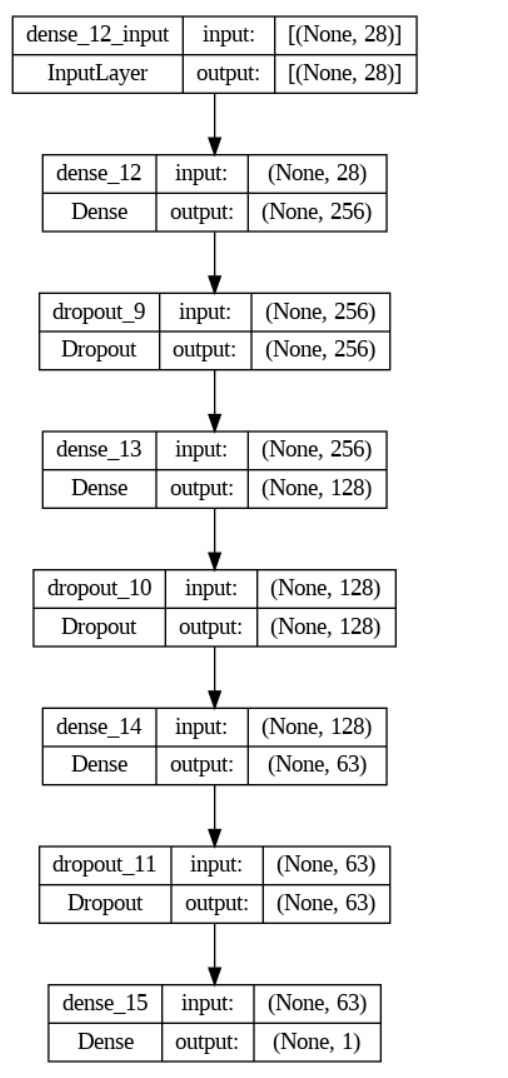


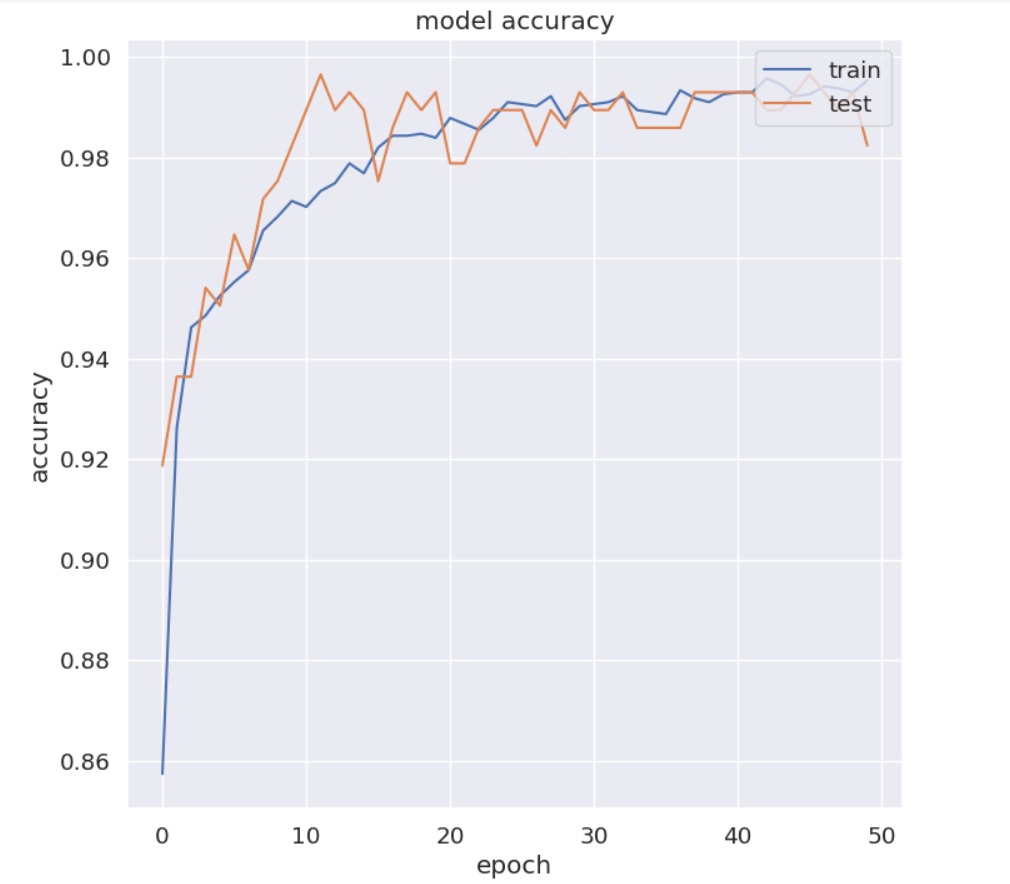












**4.METHODOLOGY**

Enough methods are performed on the data to evaluate the data set and gather knowledge about the data. Let's perform some Machine Learning model and Experimentation to create amodel that helps us to achieve our goal we state in the problem definition. In this we talks about the various machine learning algorithms used for the project. They are Logistic Regression, KNN, SVM, Naïve Bayes.

### 4.1 LOGISTICREGRESSION

Logistic regression is a popular method in machine learning for binary classification tasks, where the goal is to predict one of two possible outcomes. In the case of throat cancer detection, the two outcomes could be "cancerous" or "non-cancerous".

Here are the steps for using logistic regression methodology for throat cancer detection in an AIML project:

1. Data collection: Collect a dataset that includes information about patients who have undergone tests for throat cancer, including their symptoms, test results, and whether they were diagnosed with throat cancer or not.

2. Data preparation: Prepare the data by cleaning and preprocessing it, which may involve removing missing values, normalizing or standardizing the data, and encoding categorical variables.

3. Splitting the data: Split the data into training and testing sets. The training set will be used to train the logistic regression model, while the testing set will be used to evaluate the model's performance.

4. Model training: Train the logistic regression model on the training set. The logistic regression model uses a sigmoid function to output a probability that a given input belongs to one of the two classes.

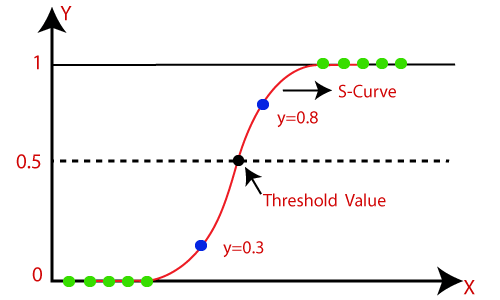
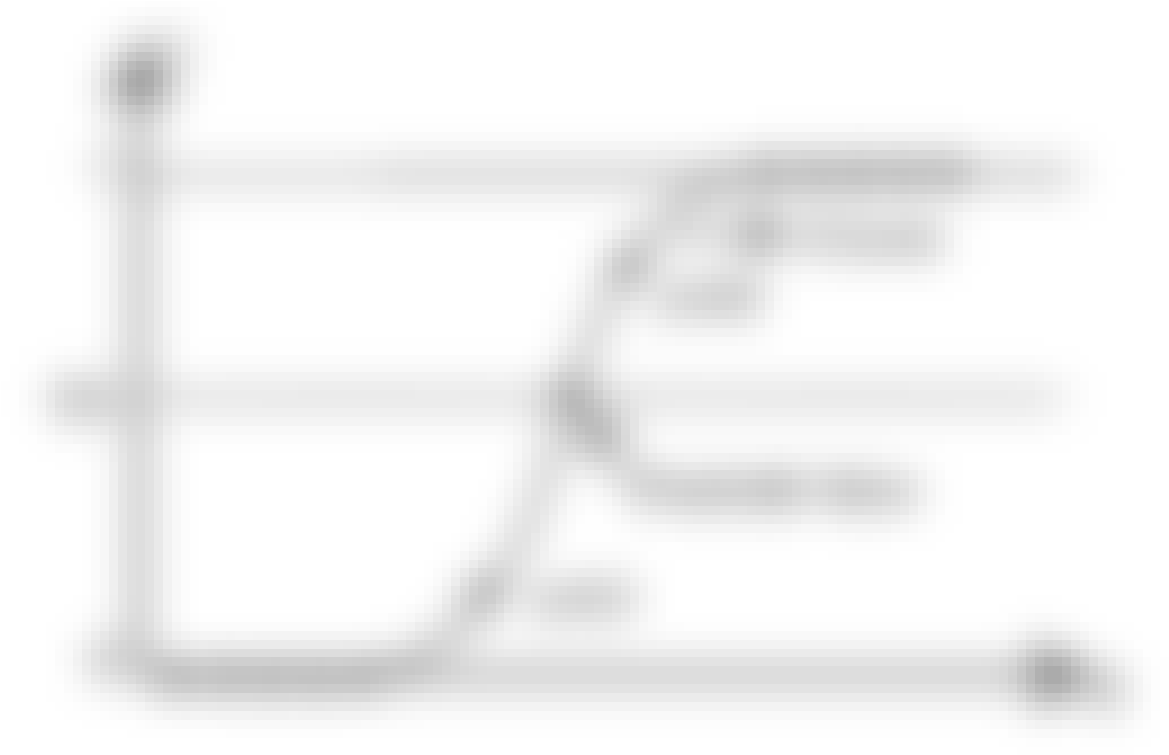
5. Model evaluation: Evaluate the performance of the model on the testing set using metrics such as accuracy, precision, recall, and F1-score. These metrics will help you understand how well the model is performing and whether it is generalizing well to new data.

6. Model optimization: Depending on the performance of the model, you may want to optimize it by adjusting hyperparameters such as the regularization strength or the learning rate.

7. Deployment: Once you have a well-performing logistic regression model for throat cancer detection, you can deploy it in your application or integrate it into a larger machine learning system.

Overall, logistic regression is a simple yet effective method for binary classification tasks, such as throat cancer detection. By following these steps, you can build a robust and accurate model for detecting throat cancer in patients.

### Fig4.1:Logistic Regression



**RESULT:** 25%

### 4.2 K-Nearest Neighbors (KNN)

KNN (k-nearest neighbors) is a popular methodology in machine learning for classification tasks. For throat cancer detection, KNN can be used as follows:

1. Collect and preprocess data: Gather data on patients with and without throat cancer. Preprocess the data by cleaning, normalizing, and transforming it as necessary.

2. Split the data: Divide the data into training and testing sets. The training set will be used to train the KNN algorithm, while the testing set will be used to evaluate its performance.

3. Determine the value of k: Choose a value for k, which determines the number of nearest neighbors to consider when classifying a new instance. This value can be determined using cross-validation or other techniques.

4. Train the KNN model: Train the KNN model on the training set by storing the feature vectors of each instance along with its class label.

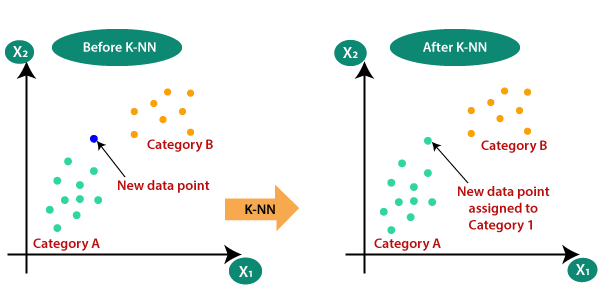
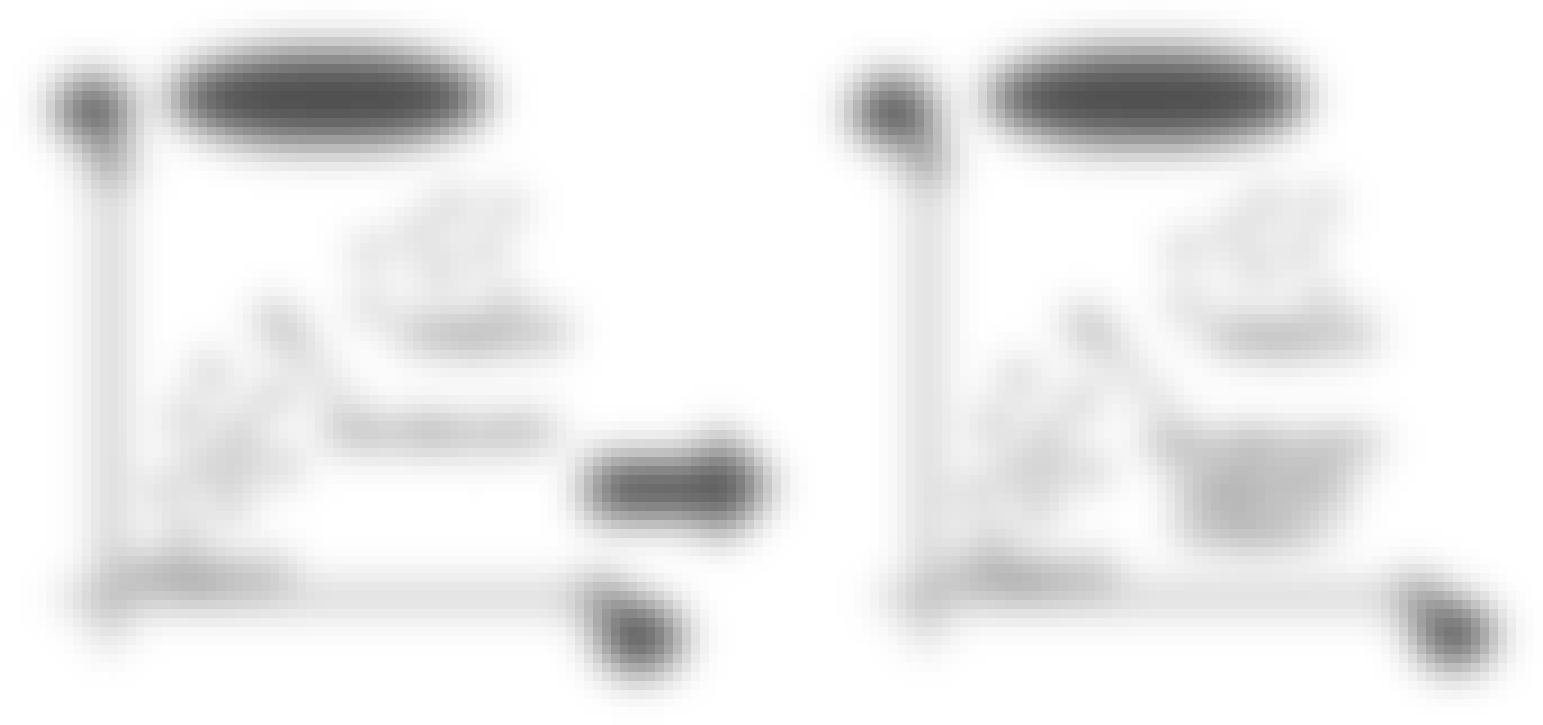
5. Classify new instances: When a new instance is presented to the KNN model, it identifies the k nearest neighbors to that instance in the training set and assigns the new instance to the class that is most common among those neighbors.

6. Evaluate the model: Evaluate the performance of the KNN model on the testing set by calculating metrics such as accuracy, precision, recall, and F1 score.

7. Optimize the model: Experiment with different values of k and other parameters to optimize the performance of the KNN model.

Keep in mind that KNN is just one of many machine learning methodologies that can be used for throat cancer detection. Other approaches, such as decision trees, random forests, or support vector machines,

may also be effective depending on the specifics of the problem and the available data



### Fig4.2: KNN

**RESULT :** 96%

### 4.3 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machines (SVM) is a machine learning algorithm that can be used for binary classification, which can be applied to the task of throat cancer detection. Here is a general methodology that can be used for an AIML project:

Data collection: Collect a dataset of throat images, either from a medical database or by collaborating with a medical institution. The dataset should include images of patients with throat cancer and images of patients without throat cancer.

Data pre-processing: Pre-process the images to remove any noise, resize the images to a standard size, and normalize the pixel values.

Feature extraction: Extract features from the pre-processed images, such as texture, color, and shape. One approach is to use a pre-trained convolutional neural network (CNN) to extract features automatically.

Data splitting: Split the dataset into training, validation, and test sets. The training set will be used to train the SVM model, the validation set will be used to tune the hyperparameters, and the test set will be used to evaluate the performance of the final model.

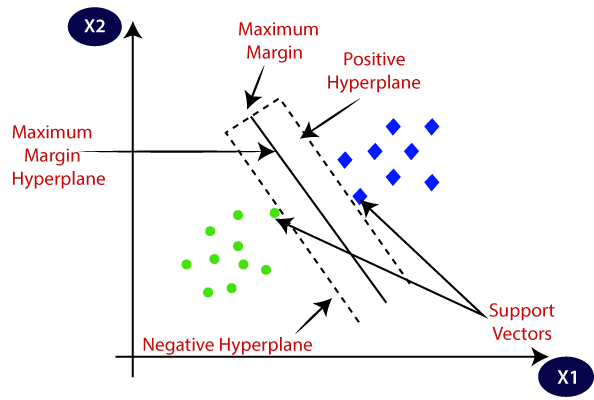
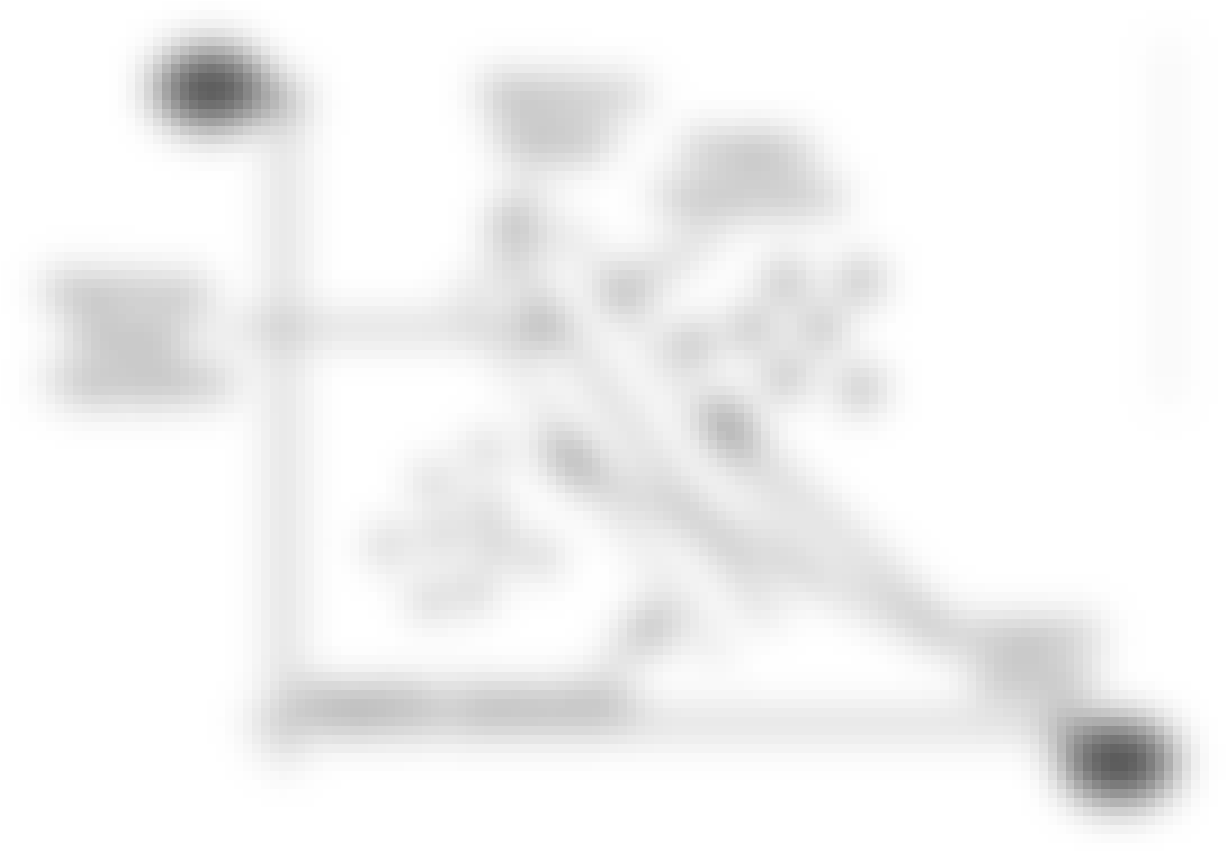
Model training: Train an SVM model on the training set using the extracted features. One approach is to use the radial basis function (RBF) kernel, which is suitable for non-linear classification tasks.

Hyperparameter tuning: Tune the hyperparameters of the SVM model using the validation set. The two main hyperparameters are the regularization parameter (C) and the kernel parameter (gamma).

Model evaluation: Evaluate the performance of the final model on the test set using metrics such as accuracy, precision, recall, and F1-score. The confusion matrix can also be used to visualize the performance of the model.

Model deployment: Deploy the final model in a web application or a mobile application to allow medical professionals to use it for throat cancer detection.

It is important to note that this methodology is just a general guideline and may need to be adapted



**Fig4.3:SVM**

## RESULT : **96%**

### 4.4 Naïve Bayes

Naive Bayes is a commonly used classification algorithm in machine learning that can be used for medical diagnosis, including throat cancer detection. Here is a general outline of how to approach using Naive Bayes for throat cancer detection:

1. Collect and prepare data: Collect a dataset of patients with throat cancer and without throat cancer. The dataset should include relevant features such as age, gender, smoking status, alcohol consumption, family history of cancer, symptoms, and diagnostic test results. Preprocess the data to ensure it is clean and consistent.

2. Split the dataset: Split the dataset into training and testing sets. The training set will be used to train the Naive Bayes model, and the testing set will be used to evaluate the model's performance.

3. Train the Naive Bayes model: Use the training set to train the Naive Bayes model. Naive Bayes assumes that all features are independent, which is often not true in medical diagnosis. However, it can still perform well in practice due to its simplicity and efficiency. There are different types of Naive Bayes models, such as Gaussian, Multinomial, and Bernoulli Naive Bayes. Choose the appropriate model based on the type of data.

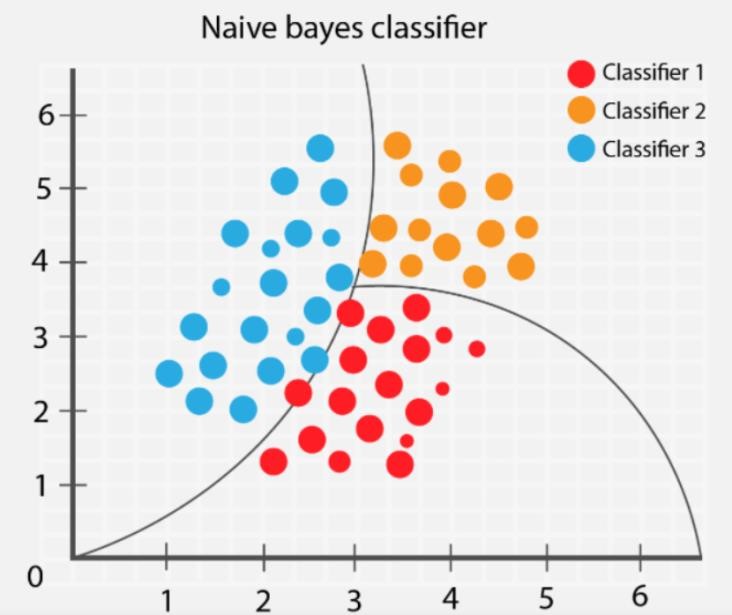
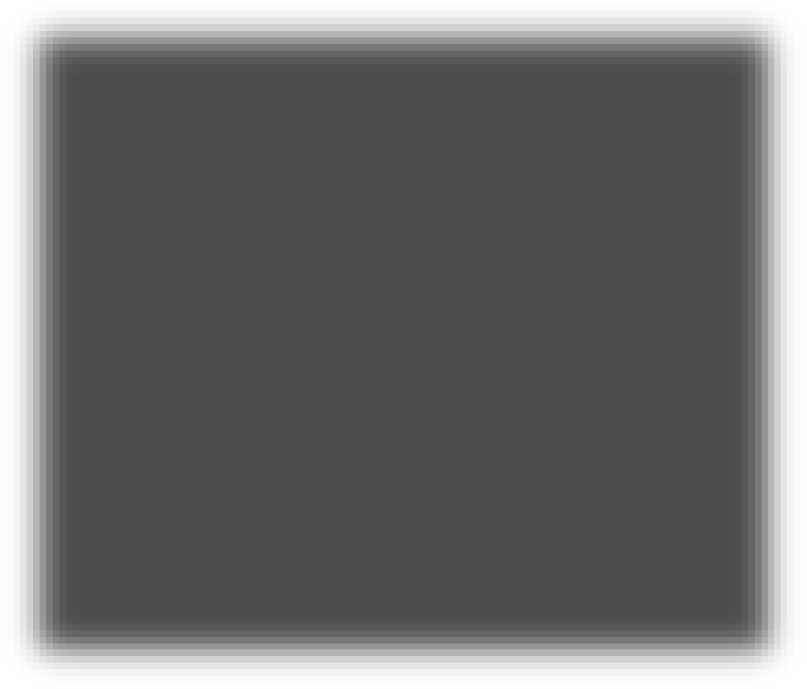
4. Test the Naive Bayes model: Use the testing set to evaluate the performance of the Naive Bayes model. Calculate metrics such as accuracy, precision, recall, and F1-score to measure the model's performance.

5. Improve the Naive Bayes model: If the Naive Bayes model's performance is not satisfactory, try improving it by tuning the model's hyperparameters or by using more advanced machine learning algorithms.

6. Deploy the Naive Bayes model: Once the Naive Bayes model is trained and tested, it can be deployed for throat cancer detection. Users can input patient data into the model, and the model will output a prediction of whether the patient has throat cancer or not.

7. Monitor and update the Naive Bayes model: It is important to monitor the performance of the Naive Bayes model over time and update it as necessary to ensure it remains accurate and relevant.

Note that developing a machine learning model for medical diagnosis is a complex task that requires careful consideration of ethical and legal issues, such as data privacy and patient safety. It is important to consult with healthcare professionals and follow established ethical guidelines and regulations when working on such projects.

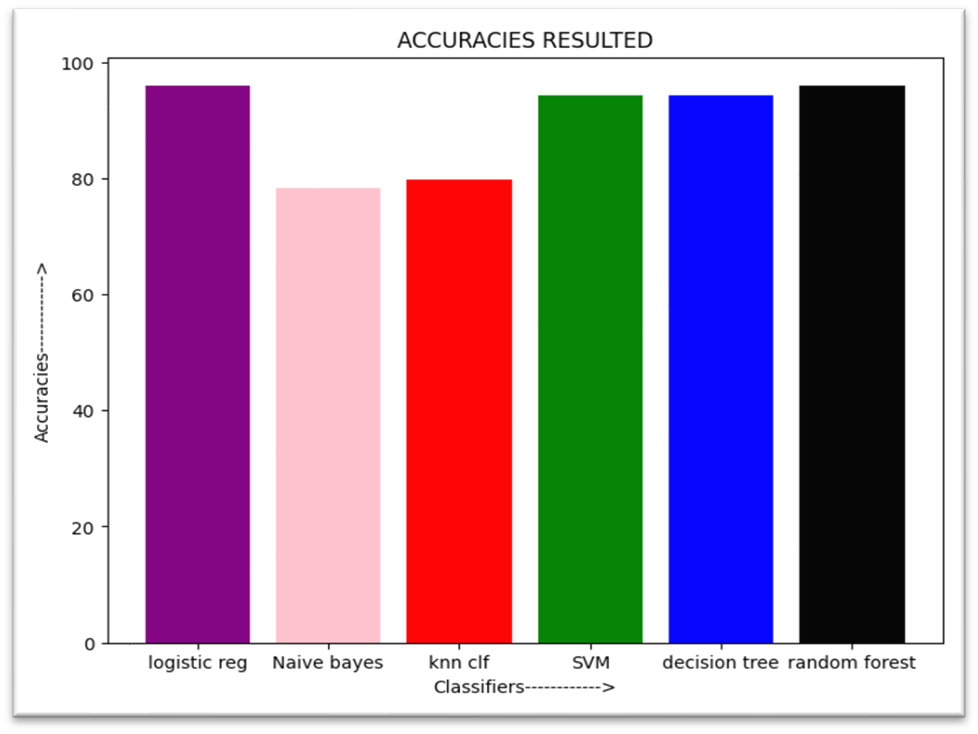


### Fig4.4:Naïve Bayes

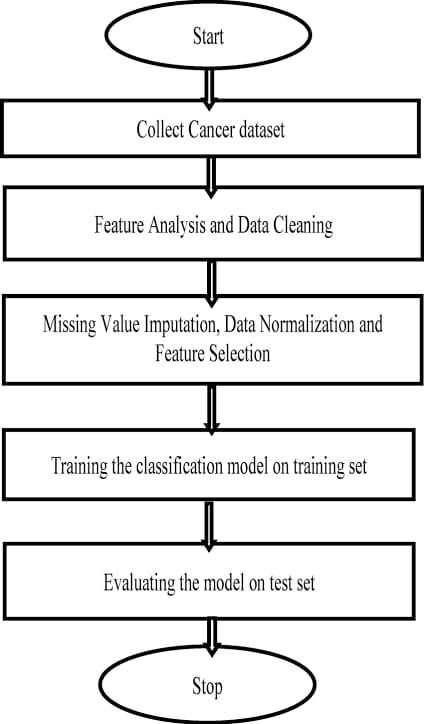
**RESULT :**96%

##### CHAPTER 6 RESULTS

|  |  |
| --- | --- |
| **Machine Learning Model** | **Accuaracy** |
| Logistic Regression | 96% |
| KNN | 95% |
| Naïve Bayes | 75% |
| Support Vector System | 96% |



**4.2 : MODEL ARCHITECTURE:**



**4.4:SOFTWARE DESCRIPTON:**

By analyzing large Throat cancer detection software is a computer program designed to assist doctors in identifying potential cases of throat cancer in patients. The software uses artificial intelligence and machine learning algorithms to analyze patient data and detect early signs of cancer.

The software is typically fed with data from a patient's medical history, such as their age, gender, smoking history, and family history of cancer. The software also analyzes patient symptoms, such as difficulty swallowing, persistent sore throat, or changes in voice quality.

The software uses this information to create a risk profile for each patient, which indicates the likelihood of developing throat cancer. This information can help doctors to prioritize patients for further testing and potentially catch cases of throat cancer earlier, when treatment is more effective.

One of the main advantages of this software is that it can analyze large amounts of data quickly and accurately, which can be challenging for human doctors to do. It can also learn from new data over time, improving its accuracy and ability to identify potential cases of throat cancer.

Overall, the throat cancer detection software is an innovative and valuable tool for doctors to identify potential cases of throat cancer early and improve patient outcomes. By leveraging the power of artificial intelligence and machine learning, this software can help doctors make more informed decisions and provide better care to their patients.

**CHAPTER 6 CONCLUSION AND FUTURESCOPE:**

**Conclusion:**

In conclusion, the use of artificial intelligence and machine learning techniques for the detection of throat cancer shows great promise. By analyzing large amounts of data, such as medical images and patient histories, these tools can help identify potential cases of throat cancer at an earlier stage, when treatment is more likely to be effective.

One of the main advantages of using AIML for throat cancer detection is that it can help reduce the reliance on subjective human interpretation. With the help of sophisticated algorithms, patterns and anomalies in the data can be identified with much greater accuracy and speed, leading to more accurate diagnoses and better patient outcomes.

However, it is important to note that while AIML can help support clinical decision-making, it is not a substitute for human expertise. Medical professionals must still be involved in the process of diagnosis and treatment, and should use AIML tools as a supplement to their own knowledge and experience.

Overall, the use of AIML in throat cancer detection represents an exciting development in the field of healthcare, with the potential to improve patient outcomes and save lives. As research in this area continues, we can expect to see further advancements in the use of AI and machine learning for the early detection and treatment of various forms of cancer.

**FUTURE SCOPE:**

Throat cancer is a serious and life-threatening condition that affects millions of people worldwide. Early detection of throat cancer is essential to improving patient outcomes and survival rates. One promising approach to improving throat cancer detection is through the use of artificial intelligence and machine learning (AIML) techniques.

AIML algorithms can be trained on large datasets of medical images and patient data to recognize patterns and identify early indicators of throat cancer. These algorithms can then be used to analyze new patient data and identify individuals who may be at high risk for developing throat cancer.

Some potential future directions for AIML-based throat cancer detection include:

1. Developing more accurate and efficient algorithms: As the field of AIML continues to evolve, researchers will likely develop new and more effective algorithms for detecting throat cancer. These algorithms may be able to analyze more complex and varied types of medical data, such as MRI and CT scans, to provide more accurate and detailed diagnoses.

2. Implementing more accessible and affordable diagnostic tools: AIML-based throat cancer detection could potentially be implemented in a variety of settings, including low-resource and rural areas where access to medical technology and expertise is limited. This could help to improve early detection rates and ultimately save lives.

3. Incorporating AI into treatment planning: In addition to improving diagnostic accuracy, AIML could also be used to help guide treatment decisions for throat cancer patients. By analyzing patient data and treatment outcomes, algorithms could help doctors identify the most effective treatment plans for individual patients.

Overall, the future of AIML-based throat cancer detection is promising, with the potential to improve early detection rates and ultimately save lives. As researchers continue to develop and refine these techniques, we may see significant improvements in the way we diagnose and treat this deadly disease.

##### **CHAPTER 7 REFERENCES**

1. https://www.kaggle.com/datasets/kumar012/hypothyroid