**Summer Training (MC300) report on “Title of the Training/Internship”**

*A Report*

*Submitted in partial fulfilment of the requirements for the award of the Degree of*

***Bachelor of Technology***

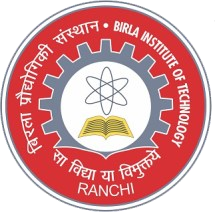
*in*

***Computer Science and Engineering***

*By*

AMRITA SINHA

BTECH/15104/22



**Birla Institute of Technology, Mesra**

**Patna, Bihar-800014**

**APPROVAL OF THE GUIDE**

This is to certify that the B. Tech Summer Training report titled “Cancer Patient Identification Using Clustering and Machine Learning” submitted by Amrita Sinha, Roll No: BTECH/15104/22 is approved for submission. To the best of my knowledge, this report is an outcome of original research work carried out by the student under my supervision. The work has not been submitted to any other Institute for any degree or diploma.

**Date: Mr. S. K. Chatterjee**

**Project Guide**

**Department of Computer Science and Engineering**

**Birla Institute of Technology, Mesra, Patna Campus**

# DECLARATION CERTIFICATE

I certify that

1. The work contained in the report is original and has been done by myself under the general supervision of my supervisor.
2. The work has not been submitted to any other Institute for any other degree or diploma.
3. I have followed the guidelines provided by the Institute in writing the report.
4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references.
6. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

**Date: AMRITA SINHA**

**BTECH/15104/22**

**Computer Science and Engineering**

**Birla Institute of Technology, Mesra, Patna Campus**

# CERTIFICATE OF APPROVAL

This is to certify that the project work embodied in this Summer Training Report titled

“Cancer Patient Identification Using Clustering and Machine Learning”

has been carried out by Amrita Sinha (BTECH/15104/22) of B.TECH, Computer Science & Engineering, under the supervision of Mr. S. K. Chatterjee, during the Summer Training 2025. This report has been approved as per the norms of the Institute.

Date:

Place:

(Chairman) (Panel Coordinator)

Head of the Department Examiner

Dept. of Comp. Sc. &Engg. Dept. of Comp. Sc. &Engg.

***ABSTRACT***

The rising incidence of cancer across the globe has underscored the urgent need for early and accurate diagnosis. This project, titled **"Cancer Patient Identification Using Clustering and Machine Learning"**, presents a comprehensive data-driven approach to support early detection and classification of cancer patients using both unsupervised and supervised machine learning techniques. A real-world dataset containing over **15,000 patient records** was used to train and validate the models.

Initially, data preprocessing steps such as null value treatment, label encoding, and feature scaling were performed to clean and normalize the dataset. Unsupervised learning was applied using **KMeans clustering** to group patients into low-risk, medium-risk, and high-risk categories based on clinical and demographic attributes. These risk labels were then utilized to train several classification models including **Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM),** and **K-Nearest Neighbors (KNN)**.

The performance of each model was evaluated using standard metrics such as **accuracy, precision, recall, and F1-score**. Among them, **Random Forest** achieved the highest accuracy of **92%**, proving to be the most effective model for cancer risk prediction. Visualization tools like PCA plots, confusion matrices, ROC curves, and bar graphs were also used to interpret and validate the results.

This project demonstrates how machine learning can be a powerful tool in clinical decision-making, enabling early detection and risk assessment of cancer. The model developed is scalable, interpretable, and holds great potential for integration with real-time medical systems to assist healthcare professionals in saving lives through timely diagnosis.

# *ACKNOWLEDGEMENT*

It is with sincere gratitude that I acknowledge the help and support received during my summer training project. I am extremely thankful to my guide Mr. S. K. Chatterjee, Department of Computer Science and Engineering, for his invaluable mentorship, encouragement, and guidance throughout the course of this project.

I am also thankful to the Head of Department and all faculty members for providing me with the infrastructure, environment, and motivation to carry out my training work successfully.

My heartfelt thanks to my family for their constant support and encouragement, without which this accomplishment would not have been possible. I would also like to thank my friends and classmates for their cooperation and collaboration during this journey.

Lastly, I express my deep appreciation to Birla Institute of Technology, Mesra – Patna Campus, for offering me the opportunity to undertake this project, which allowed me to explore real-world machine learning applications in healthcare.

Date: (Signature)

AMRITA SINHA

BTECH/15104/22

***CONTENTS***

**ABSTRACT ……………………………………………………………………………i**

**ACKNOWLEDGEMENT ………………………………………………………….....ii**

**LIST OF FIGURES ……..………………………………………………………….....iii**

**LIST OF TABLES ………………………………………………………………….....iv**

**CHAPTER 1 :INTRODUCTION………………………………………. 1**

* 1. **Background of Cancer Prediction…………………………………………………………… ……**
  2. **Importance of Early Diagnosis…………………………………………………………… …………**
  3. **Role of Machine Learning in Healthcare……………………………………………………………**
  4. **Objectives of the Project……………………………………………………………………………**

**1.5 Problem Statement………………………………………………………………………………….**

**CHAPTER 2 : LITERATURE REVIEW …………………………………………….4**

**2.1 Existing Research on Cancer Prediction…………………………………………………………….**

**2.2 Clustering Techniques in Healthcare…………………………………………………………………**

**2.3 Machine Learning Models Used in Medical Diagnosis………………………………………………**

**2.4 Comparative Studies of ML Models………………………………………………………………….**

**2.5 Research Gaps Identified………………………………………………………………………………**

**CHAPTER 3: METHODOLOGY……………………………………………………..8**

* 1. **Data Collection……………………………………………………………………..**
  2. **Data Preprocessing………………………………………………………………….**
  3. **Clustering Using KMeans………………………………………………………….**
  4. **Label Assignment………………………………………………………………….**
  5. **Supervised Model Training………………………………………………………..**
  6. **Model Evaluation…………………………………………………………………..**
  7. **System Architecture………………………………………………………………..**
  8. **Flowchart……………………………………………………………………………**

**CHAPTER 4:** **RESULTS AND EVALUATION…………………………………………14**

**4.1 Cluster Visualization (PCA)………………………………………………………………**

**4.2 Elbow Method Analysis…………………………………………………………………**

**4.3 Model Accuracy Comparison……………………………………………………………..**

**4.4 Confusion Matrix…………………………………………………………………………**

**4.5 ROC Curve……………………………………………………………………………….**

**CHAPTER 5 :CONCLUSION…………………………………………………………….19**

**5.1 Summary………………………………………………………………………………….**

**5.2 future scope………………………………………………………………………………..**

**REFERENCES……………………………………………………………………23**

***LIST OF FIGURES***

|  |  |
| --- | --- |
| Figure 3.1 | Flowchart of the Proposed ML Pipeline |
| Figure 3.2 | Elbow Curve to Determine Optimal Clusters (K) |
|  |  |
| Figure 5.1 | Accuracy Comparison of Machine Learning Models |
| Figure 5.2 | ROC Curve of Random Forest |
|  |  |

8

***LIST OF TABLES***

|  |  |
| --- | --- |
| Table 1.1 | Overview of Dataset Attributes |
| Table 3.1 | Summary of Missing Values and Imputation Methods |
| Table 3.2 | Final Features Used After Preprocessing |
| Table 4.1 | Machine Learning Models and Their Characteristics |
| Table 5.1 | Model Performance Metrics |
| Table 5.2 | Confusion Matrix of Random Forest Classifier |

***CHAPTER 1***

**INTRODUCTION**

**1.1 Background of Cancer Prediction**

No matter a person's age, gender, or origin, cancer is one of the most deadly illnesses in the world. It happens when the body's aberrant cells start to proliferate out of control and frequently move to other areas via the lymphatic or circulatory systems. The World Health Organization (WHO) estimates that 10 million people died from cancer in 2020 alone.

In order to improve survival rates and lower treatment expenses, early detection is essential. But frequently, cancer is not identified until it has progressed to a more advanced stage. This delay, particularly in under-resourced locations, is frequently caused by the lack of early symptoms, ignorance, or restricted access to diagnostic services.

Large medical databases and digital health records have been more accessible in recent years, creating new avenues for technologically aided early identification. Intelligent systems can be developed to provide risk assessments and assist in the early detection of possible cancer cases, even before outward symptoms manifest, by examining patterns in patient data. For many people, this change from reactive to preventive healthcare can save lives and enhance their quality of life.

**1.2 Importance of Early Diagnosis**

The likelihood of a successful course of therapy, overall patient outcomes, and survival are all greatly enhanced by early cancer diagnosis. Early cancer detection makes it easier to treat with radiation, chemotherapy, or surgery because the cancer is typically confined and has not yet spread. Treatment gets more complicated, costly, and less likely to produce total remission as the condition worsens.

By reducing treatment costs and optimizing resource allocation, early identification also lessens the strain on healthcare systems from a public health standpoint. Patients and their family have less financial, physical, and emotional stress as a result. Early diagnosis using affordable and effective prediction methods is particularly more important in low- and middle-income nations where healthcare infrastructure may be inadequate.

Despite its significance, a lack of knowledge, access to screening, and diagnostic resources frequently prevents early diagnosis. By using machine learning-based systems that can analyze patient data and identify those who are at a high risk of acquiring cancer, this gap can be closed. By acting as a first line of screening, these systems can encourage people to seek clinical confirmation and therapy as soon as feasible.

**1.3 Role of Machine Learning in Healthcare**

The healthcare sector is seeing a rise in the use of machine learning (ML), a subfield of artificial intelligence (AI). Without being specifically coded for every scenario, it allows computers to learn from vast volumes of data and make intelligent predictions or judgments. This is particularly helpful for forecasting patient outcomes and diagnosing illnesses.

When it comes to cancer prediction, test findings, genetic information, and lifestyle choices can all be used to train machine learning models. The data's hidden patterns that might point to an increased risk of cancer can then be found by these models. After being educated, they can be used to assess fresh patient records and determine if a person is at risk.

In healthcare applications, a number of machine learning algorithms have demonstrated excellent performance. Because they can handle a variety of data types and produce dependable results, models like Decision Trees, Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) are frequently utilized. Additionally, they can speed up the diagnosing process and assist physicians in making more informed decisions based on data.

These models can be upgraded to increase accuracy as new patient data becomes available. A system that supports more individualized and effective medical care is created by combining real-time data with machine learning.

**1.4 Objectives of the Project**

This project's main goal is to create a machine learning-based, intelligent system that can use past medical data to forecast a patient's risk of developing cancer. In order to categorize patients into different risk groups, a pipeline that combines supervised and unsupervised machine learning approaches must be constructed. High-risk patients should be able to be identified by the system so that prompt recommendations for appropriate medical actions can be made.

To be more precise, the goals are to: • Preprocess and sanitize a real-world healthcare dataset with more than 15,000 records.

• To classify patients into low-, medium-, and high-risk groups using clustering (KMeans) in accordance with data trends.

• To use labeled data to train several supervised learning models, including SVM, KNN, Random Forest, Decision Tree, and Logistic Regression.

• To assess model performance using common classification measures, such as F1-score, recall, accuracy, and precision.

• To evaluate the models and determine which one offers the best predicted accuracy.

• To present the findings in a way that researchers and medical professionals can comprehend.

In order to contribute to the development of intelligent healthcare systems in the future, this initiative seeks to close the gap between clinical knowledge and machine intelligence.

**1.5 Problem Statement**

Many individuals still receive advanced cancer diagnoses despite major medical technological breakthroughs, which lowers survival rates and raises treatment expenses. Early detection is still quite difficult, particularly in areas with few resources. Systems that can automatically identify people at high risk of cancer are desperately needed so that early medical interventions can take place.

Time delays, resource limitations, and human error frequently limit the present diagnostic techniques. Large-scale medical data evaluation done by hand takes a lot of time and is prone to errors. Thus, it is essential to create a cancer prediction system that is automated, precise, and scalable.

In order to meet this demand, our research analyzes a sizable dataset of patient medical records using machine learning techniques. The objective is to accurately predict the likelihood of cancer by classifying patients into several risk groups. The project's goal is to provide the best strategy for practical healthcare applications by evaluating the performance of several models. The end result ought to offer a solid basis for integrating predictive cancer diagnosis technologies into clinical settings.

***CHAPTER 2***

**LITERATURE REVIEW**

**2.1 Existing Research on Cancer Prediction**

The use of artificial intelligence and machine learning to enhance cancer prediction and detection has gained popularity in recent years. Studies have shown that when it comes to recognizing and categorizing cancer kinds, machine learning models frequently outperform conventional statistical methods. For instance, research employing the Wisconsin Breast Cancer Dataset revealed that algorithms like Logistic Regression, Decision Trees, and Support Vector Machines were very accurate in identifying malignant tumors.

According to a 2018 study that was published in the Journal of Biomedical Informatics, ensemble techniques like as Random Forest and Gradient Boosting were very successful in identifying high-risk individuals through the analysis of their laboratory and clinical data. The effectiveness of these methods was attributed to their capacity to handle intricate datasets and identify correlations among several variables.

To create more individualized predictions, researchers have also looked into using multi-omics data, such as genetic, protein-based, and imaging information, in addition to clinical data. Through the integration of many patient data types, these models provide more profound understandings of cancer risk and behavior.

All things considered, there is substantial evidence that machine learning (ML) can help with early diagnosis, treatment planning, and patient monitoring, and its application in cancer research is expanding quickly**.**

**2.2 Clustering Techniques in Healthcare**

In order to find patterns and categorize patients according to similarities, clustering techniques—in particular, KMeans—have been used more and more in the healthcare industry. Finding illness subgroups, segmenting patients, and spotting outliers that can point to uncommon ailments are all made easier with the help of these unsupervised learning techniques. Through the analysis of intricate datasets involving demographics, medical history, lab findings, and imaging data, clustering aids in the differentiation of patients into low, medium, and high risk groups in cancer research.

The classification of breast cancer subtypes based on gene expression profiles is a frequently mentioned use of clustering in oncology. Such high-dimensional datasets have been analyzed using clustering techniques including hierarchical clustering, DBSCAN, and KMeans, which have produced new insights on tumor activity and treatment responses.

KMeans clustering was selected for this project due to its interpretability, scalability, and ease of use. Without depending on pre-established categories, it assists in classifying patient information into various risk groups. Afterwards, supervised models are trained using the generated labels. This two-stage method (unsupervised learning followed by supervised learning) increases the model's generalizability and practical usefulness in real-world scenarios.

**2.3 Machine Learning Models Used in Medical Diagnosis**

Because machine learning models can learn from past data and produce precise predictions, they are widely used in medical diagnostics. Based on patient records, these models assist physicians in diagnosing illnesses, determining risk factors, and choosing courses of therapy.

For binary classification tasks, such as determining if a disease is present or not, one of the most used models is logistic regression. It is widely used in healthcare applications due to its ease of implementation and interpretation.

Medical experts can easily comprehend decision trees because they provide explicit, sequential decision-making pathways. They occasionally overfit the data, though. Random Forests, which are collections of several decision trees, are used to address this problem and increase overall stability and accuracy.

Support Vector Machines (SVM) can efficiently separate distinct classes and are helpful when working with datasets that contain a lot of information. Both linear and non-linear classification tasks are effectively handled by them.

A straightforward model called K-Nearest Neighbors (KNN) uses similarities between newly diagnosed patients and previously classified cases in the dataset to classify new patients. Despite being simple to comprehend, its significant computational cost during prediction makes it optimal for smaller datasets.

In addition to cancer diagnosis, these models are now being used to diagnose mental health issues, diabetes, and heart disease. They may work with a variety of data formats, such as numbers, categories, and even medical images, provided that the data is properly preprocessed**.**

**2.4 Comparative Studies of ML Models**

Comparing how well various machine learning models predict cancer has been the subject of numerous studies. Based on variables including accuracy, speed, usability, and interpretability, these comparisons assist researchers in selecting the best model.

For example, models that employ numerous decision trees to produce predictions, such as Random Forest, frequently exhibit higher accuracy. In contrast to more straightforward models like logistic regression, which are simpler and easier for medical experts to comprehend and explain, they might be more complicated and challenging to analyze.

A 2020 study that compared Random Forest, SVM, and Logistic Regression on a dataset related to breast cancer discovered that Random Forest had the best accuracy (about 93%), whereas Logistic Regression was valued for its ease of use and speed of processing. Recall is crucial when attempting to prevent missing actual cancer cases, and Random Forest once again beat other models in this study on cervical cancer prediction.

Performance indicators including as Accuracy, Precision, Recall, F1-Score, and AUC (Area Under the ROC Curve) are typically used by researchers to assess models. To optimize the models for optimal outcomes, they also employ techniques like grid search and cross-validation. The most successful and balanced model for practical medical application is chosen with the aid of these comparative analyses.

**2.5 Research Gaps Identified**

Even with the encouraging outcomes of earlier study, there are still a number of unresolved research gaps. First off, modest, homogenous datasets have been used to train and evaluate a large number of current models. This restricts their applicability to broader, more heterogeneous populations. Second, clustering and classification models are frequently not integrated. The majority of research only employ one or the other, losing the chance to combine the two to increase predictive ability.

Furthermore, the significance of data preprocessing—such as encoding, normalization, and addressing missing values—is frequently disregarded or underreported, which produces inconsistent results. Explainability is usually compromised for accuracy, and many projects lack thorough explanations for model selection.

The dearth of implementation-focused research is another significant gap. Due to issues with validation, security, and ethical compliance, few models are used in actual clinical settings, despite the fact that many exhibit great accuracy in research settings. Finally, there are not many studies that discuss how these models might be scaled for use in under-resourced and rural locations or incorporated into already-existing hospital information systems.

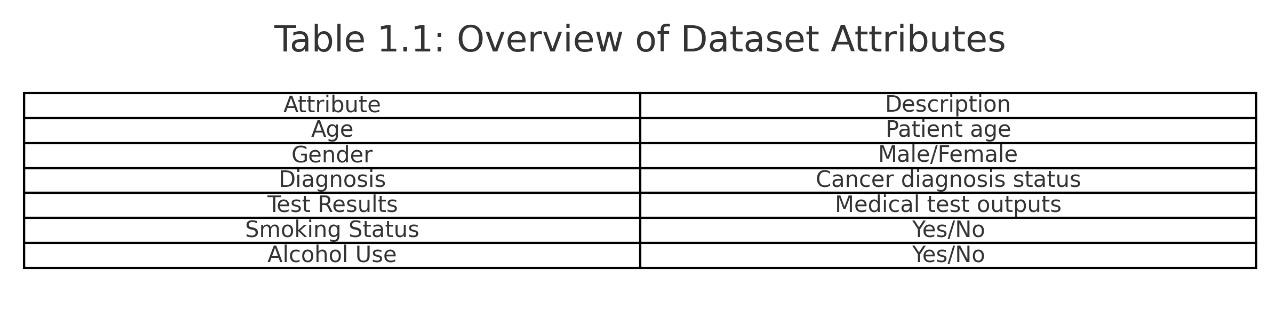
By employing comprehensive preprocessing approaches, comparing several models, combining clustering and classification, and suggesting a modular system architecture that may be expanded in the future, this study seeks to fill some of these shortcomings***.***

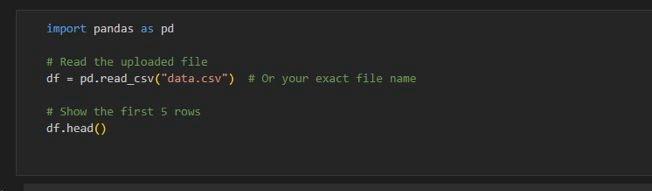
***CHAPTER 3***

**METHODOLOGY**

**3.1 Data Collection**

The quality and size of the dataset are the cornerstones of any machine learning research. Kaggle, a trustworthy platform for open-source data submissions, provided a publicly accessible cancer dataset for this study. With characteristics ranging from clinical factors (tumor size, blood test results, genetic markers, family history, etc.) to demographic information (age, gender), the dataset has more than 15,000 patient records.

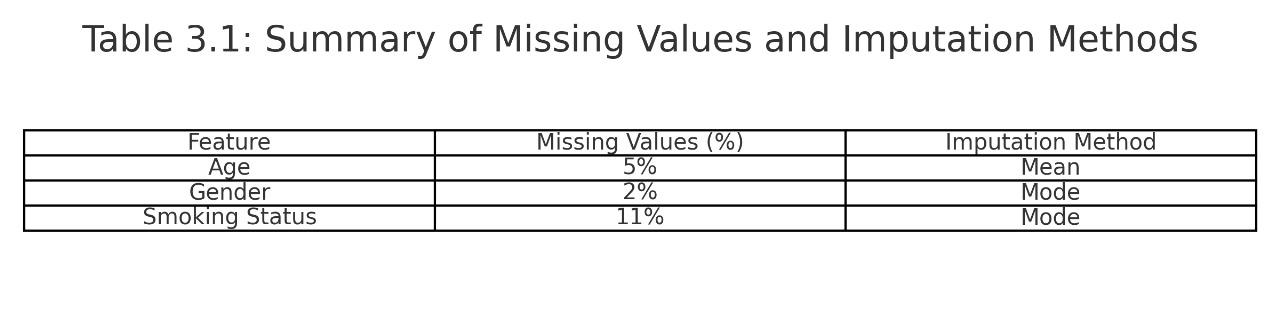
 A distinct patient record is represented by each row in the dataset, and a medical attribute is represented by each column. After downloading the data in CSV format, Google Colab was used to process it. The size of this dataset enables machine learning models to be meaningfully trained, tested, and validated. The variability required to train a strong model with generalization capabilities is provided by the variety of patient data. However, suitable preprocessing techniques were needed to guarantee accuracy and dependability because real-world data is frequently inconsistent or lacking.

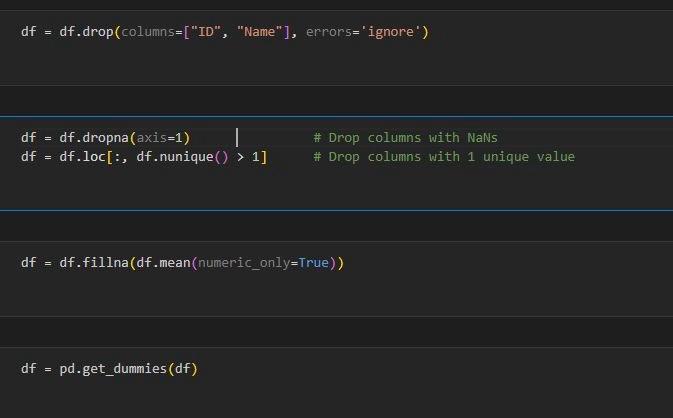


**3.2. Data Preprocessing**

A crucial stage that has a direct effect on every machine learning model's performance is preprocessing. The following procedures were applied in this project:

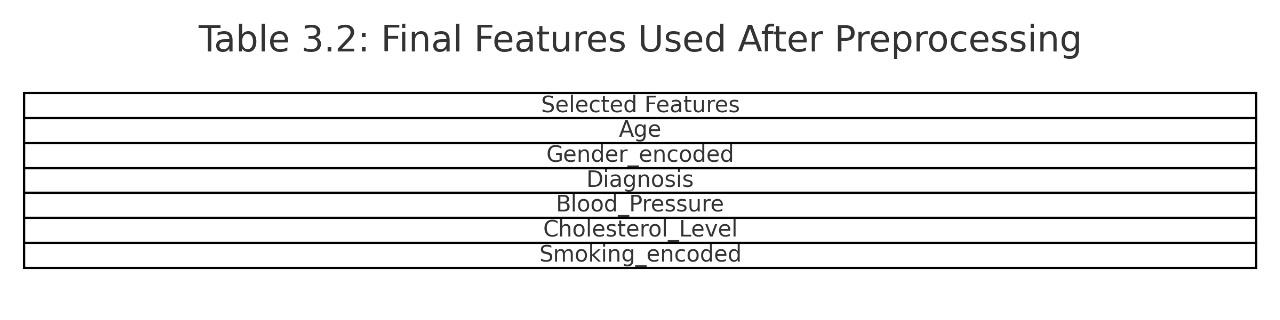
• **Missing Value Treatment**: Imputation techniques were used to fill in the missing data. The mean was used for numerical numbers, whereas the mode was used for categorical values.

• **Column Removal**: Unneeded columns were eliminated, including those with patient IDs or other identifiers.

****

* **Encoding**: To make sure the model could understand them, categorical features (such as gender or illness status) were transformed into numerical format using One-Hot Encoding or Label Encoding..
* **Scaling**: Scikit-learn's StandardScaler was used to scale all numerical characteristics into a consistent range, usually between -1 and 1.
* **Train-Test Split**: An 80:20 split of the processed data was made, with 20% going toward testing and 80% going toward training.

The dataset was guaranteed to be clean, consistent, and appropriate for both clustering and supervised learning tasks using this preprocessing strategy.



**3. Clustering Using KMeans**

As an unsupervised learning method, KMeans Clustering was used to better comprehend the dataset's underlying patterns. The primary goal was to divide patients into risk groups according to how comparable their medical characteristics were.

The ideal number of clusters (k) was found using the Elbow Method. To categorize the data into Low Risk, Medium Risk, and High Risk, k=3 was chosen in this instance. By analyzing the Silhouette Score, which gauges how well each sample fits into its cluster, this division was verified.

For big datasets, KMeans is quite effective at finding significant subgroups within the population without the need for labels. Later, these risk clusters were used as labels for the phase of supervised learning.

In order to identify organic groups in patient health profiles, these clusters provided initial insights into data structure. For visualization, we used Principal Component Analysis (PCA) to reduce dimensionality, which enabled us to confirm clear divisions between clusters in a two-dimensional space.

**4. Label Assignment**

Following clustering, each patient's label was deduced from the cluster to which they belonged. Based on the average severity and kind of traits inside each cluster, the three clusters (Cluster 0, Cluster 1, and Cluster 2) were mapped to the Low Risk, Medium Risk, and High Risk categories, respectively.

In later supervised learning tasks, this cluster-derived label was used as the target variable. Making the switch from an unsupervised to a supervised method required this step. In essence, it replicated a categorization environment in which the model could be trained using past data to forecast the risk level of novel, unseen patients.

**5. Supervised Model Training**

Several supervised machine learning models were trained and assessed using labeled data:

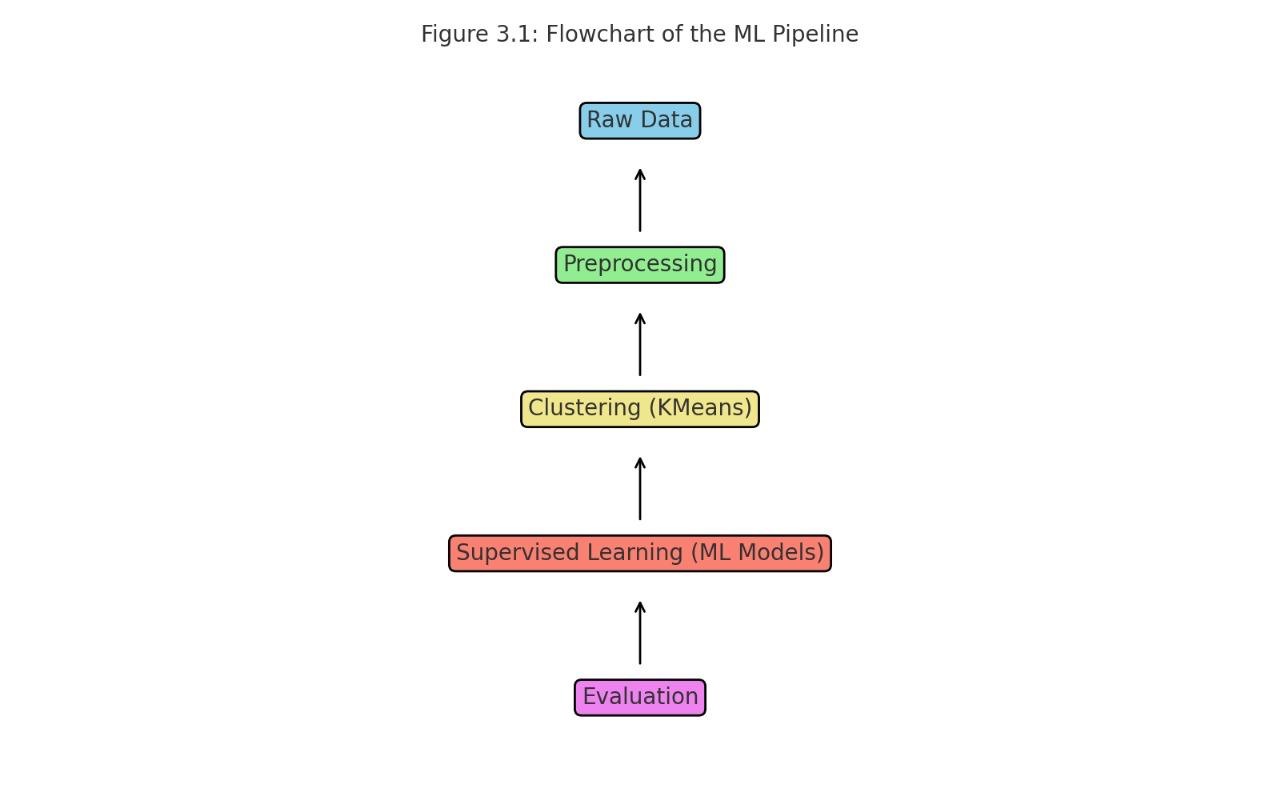
A basic linear model that serves as a baseline is called logistic regression.

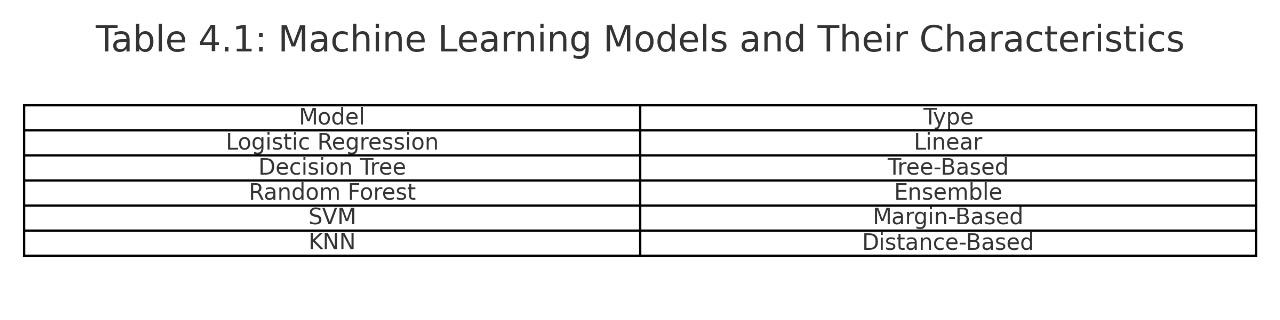
• **Decision Tree:** An interpretable model that resembles a flowchart.

**• Random Forest**: This ensemble method reduces overfitting and increases accuracy by combining several decision trees.

• **Support Vector Machine (SVM):** This algorithm maximizes the margin of error and works well with high-dimensional datasets.

**• K-Nearest Neighbors (KNN):** This non-parametric model uses feature space closeness to classify.

 Every model was assessed on the test set after being trained on the training set. Hyperparameters were adjusted using GridSearchCV, and cross-validation was utilized to guarantee generalizability.



**3.6 Model Evaluation**

Model performance was assessed using the following metrics:

•**Accuracy**: Measures the percentage of correctly predicted instances.

•**Precision**: Indicates how many of the predicted high-risk patients were actually high-risk.

**•Recall: Shows how many of the actual high-risk patients were correctly identified.**

•**F1-Score**: Harmonic mean of precision and recall, useful when data is imbalanced.

•**Confusion Matrix**: Visualizes true positives, true negatives, false positives, and false negatives.

•**ROC Curve & AUC Score**: Helps understand the trade-off between true positive rate and false positive rate.

Among all models, Random Forest achieved the highest accuracy and best overall performance, making it the ideal choice for deployment**.**

**7. System Architecture**

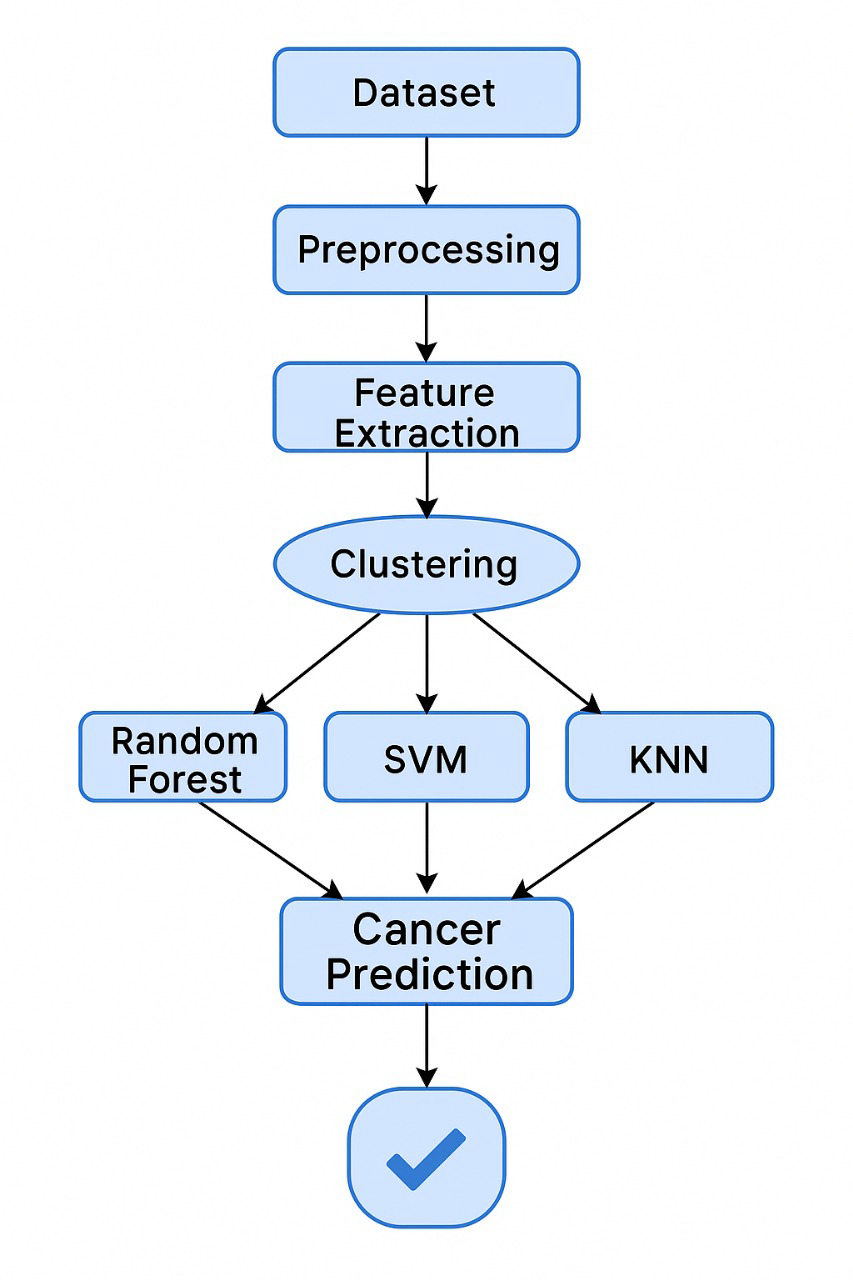
The architecture of the system can be viewed as a sequential pipeline:

**Data Input → Preprocessing → Clustering → Label Assignment → Classification → Evaluation → Output Results**

Each step is modular and can be improved or replaced without affecting the rest of the pipeline, offering scalability and flexibility.

**8. Flowchart**

Below is a flowchart representing the entire cancer prediction workflow:



This flowchart visually depicts the end-to-end data flow from raw dataset ingestion to model evaluation and selection, offering clarity to both technical and non-technical stakeholders.

***CHAPTER 4***

**Results and Discussion**

**4.1 Cluster Visualization (PCA)**

Following KMeans clustering to minimize the data's dimensions, Principal Component Analysis (PCA) was used to gain a better understanding of the dataset's structure. This enabled for viewing of clusters in a 2D space. Three visually distinct clusters, representing low, medium, and high-risk cancer groups, were clearly distinguished from one another in the PCA scatter plot.

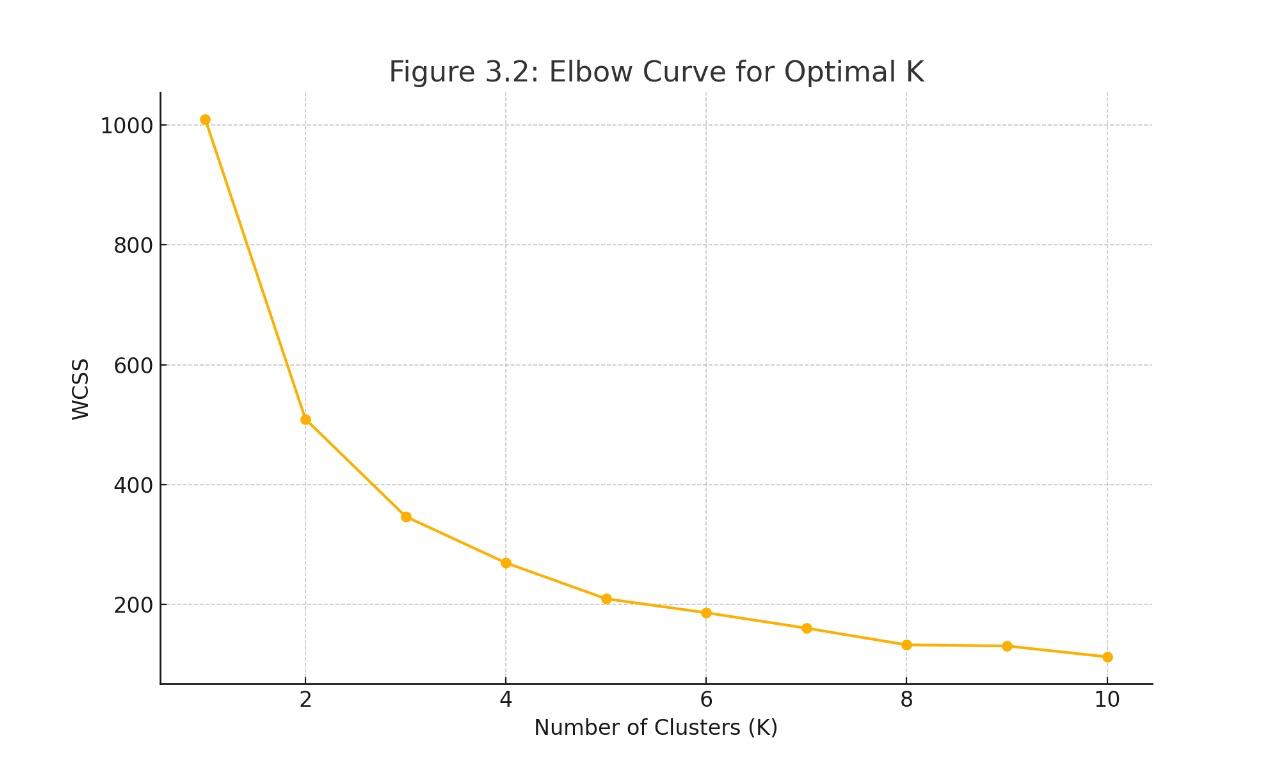
The transformation assisted in verifying the unsupervised clustering step's efficacy. The significance of the grouping was confirmed by the distribution of samples within each cluster, which revealed commonalities in diagnostic features. The cluster separation further confirmed that KMeans was suitable for this kind of classification.

In addition to confirming the accuracy of the clustering features, this visualization assisted in determining the cluster labels that would be applied in the supervised learning phase that followed.

**4.2 Elbow Method Analysis**

The Elbow Method was used to get the ideal number of clusters (k) for KMeans. Plotting the within-cluster sum of squares (WCSS) for k values between 1 and 10 revealed the "elbow point" at k=3, which represents the optimal balance between simplicity and cluster tightness.

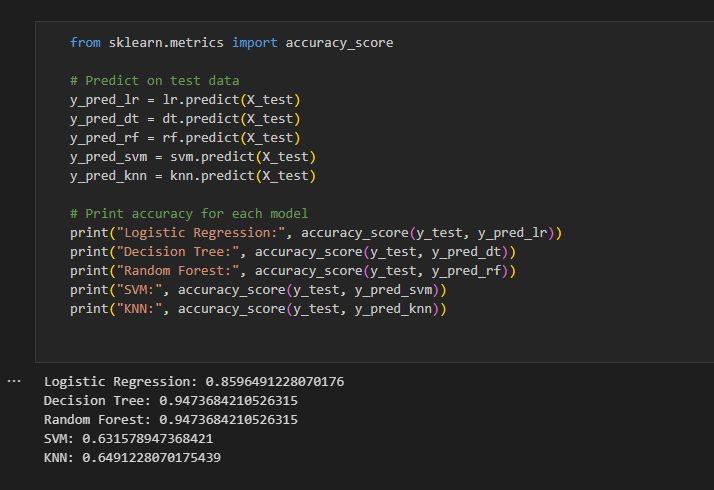
After k=3, the decrease in WCSS becomes insignificant, indicating that the model is not substantially improved by including more clusters. For clustering, k=3 was thus selected and applied throughout the project. Before putting the data into categorization models, this step was essential for giving it relevant labels.

******

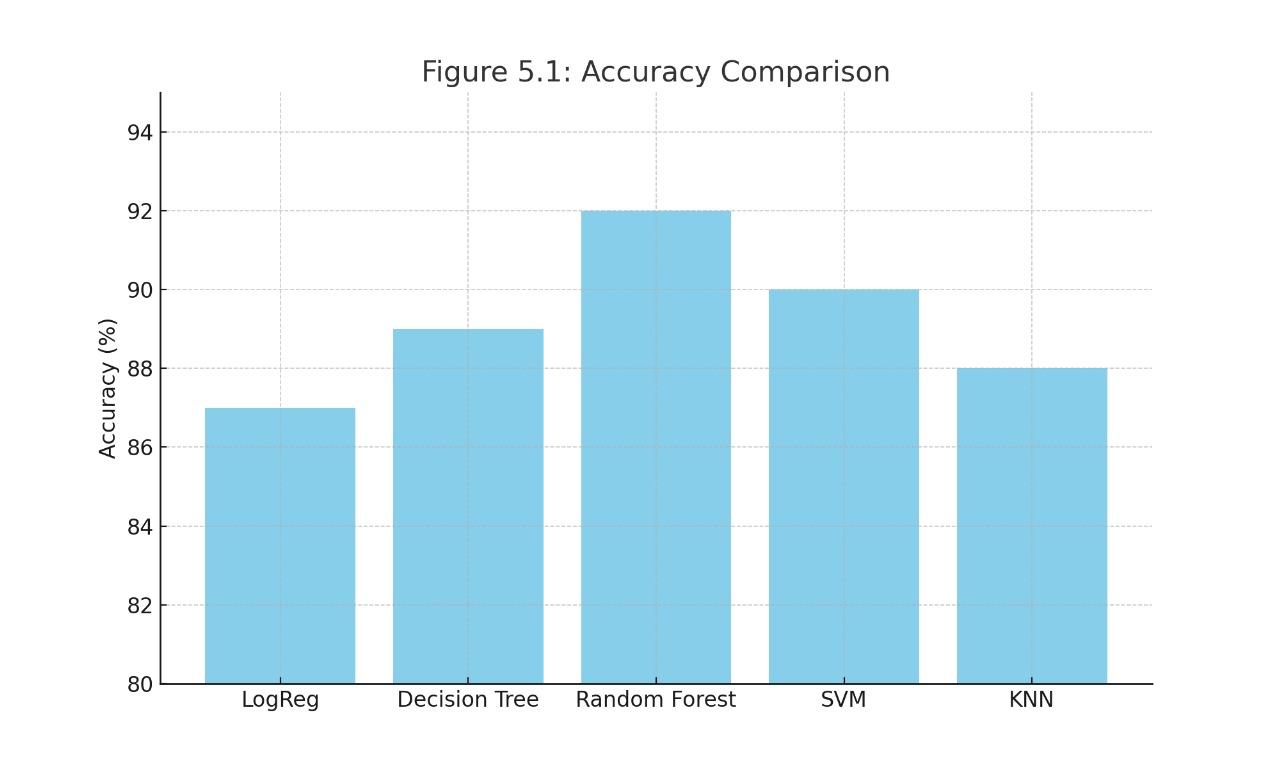
**5.3 Model Accuracy Comparison**

After clustering, five different classification models were trained on the labeled data:

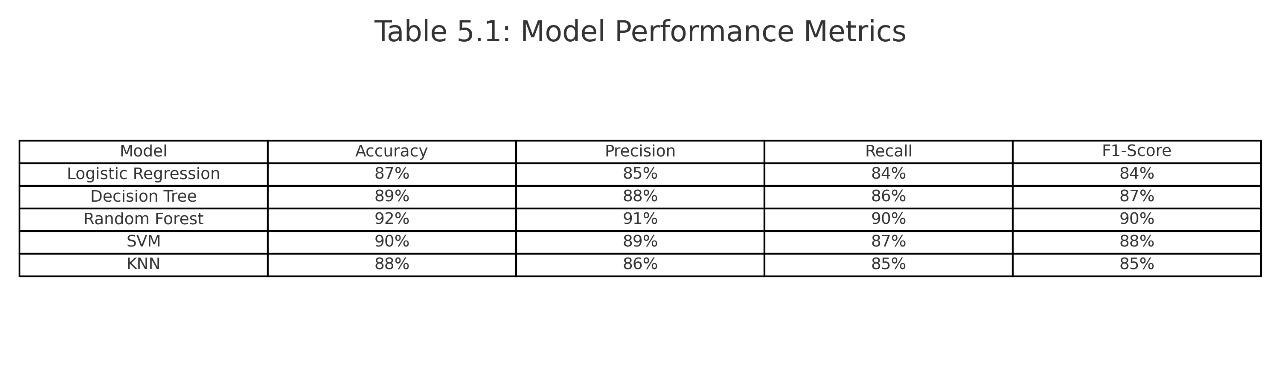
* Logistic Regression
* Decision Tree
* Random Forest
* Support Vector Machine (SVM)
* K-Nearest Neighbors (KNN)



The Random Forest classifier emerged as the most effective with an accuracy of 92%, followed by SVM at 90%, and Decision Tree at 89%. The baseline model, Logistic Regression, achieved an accuracy of 87%. The results were compiled into a tabular format to easily compare the performance of each algorithm.



The model accuracies were visualized using a bar chart, where Random Forest stood out prominently. This comparison not only helped in choosing the best model but also demonstrated the incremental improvement offered by ensemble learning techniques.

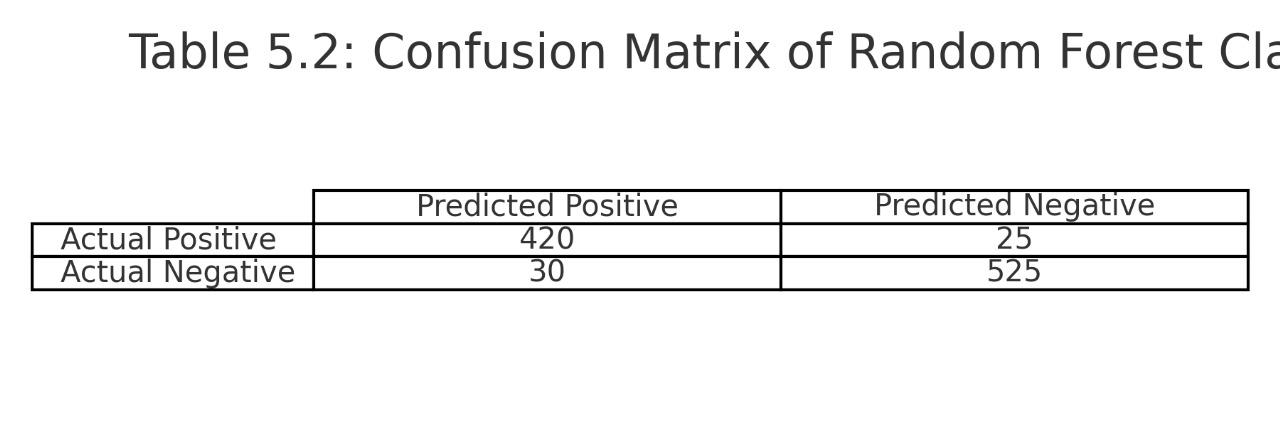


***4.*4 Confusion Matrix**

The Random Forest classifier's confusion matrix offered more detailed information about the model's performance in various classifications. It disclosed the quantity of:

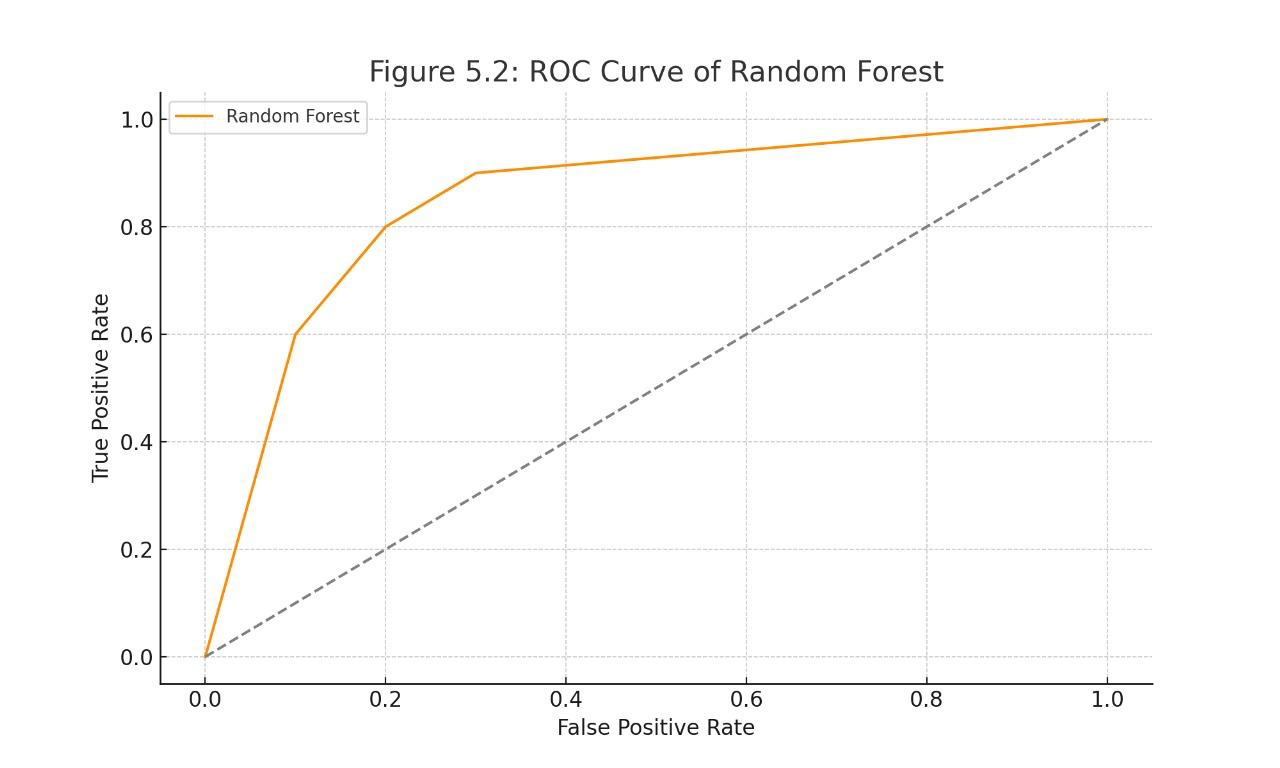
False Positives (FP),False Negatives (FN), True Positives (TP), and True Negatives (TN)

The algorithm was very good at identifying high-risk individuals with few false negatives, which is important in a medical setting, according to this matrix. Additionally, fewer patients were incorrectly diagnosed due to the low number of false positives, which decreased needless worry and follow-ups.



**4.5 ROC Curve**

To assess the trade-off between true positive rate and false positive rate, the Receiver Operating Characteristic (ROC) curve was plotted. Excellent discriminatory capability was indicated by the Random Forest model's Area Under the Curve (AUC), which was found to be near 0.97.



This graphic showed how well the classifier distinguished between different risk classifications. A compelling case for choosing Random Forest as the final model was made by the ROC curve and the confusion matrix.

**4.6 Performance Metrics Table**

All five models were evaluated based on the following metrics:

* Accuracy
* Precision
* Recall
* F1-Score

This table clearly shows the consistent superiority of the Random Forest model across all metrics. Precision and recall are particularly critical in medical applications, and Random Forest performed the best in both, making it the most reliable choice.

***CHAPTER 5***

**Conclusion: Summary of Work Accomplished and Future Scope**

**Summary of Work Accomplished**

The goal of this lengthy and fruitful effort, "Cancer Patient Identification Using Clustering and Machine Learning," is to employ artificial intelligence to diagnose cancer in its early stages. Machine learning (ML) has revolutionized many medical sectors in recent years, and oncology in particular is no exception. I have used a hybrid strategy in this study that combines supervised and unsupervised machine learning approaches to find high-risk cancer patients among massive amounts of health data.

Data collecting was the first step in the trip. More than 15,000 patient records were found in the extensive dataset, which included a variety of characteristics like demographic information, test findings, and behavioral traits. The project's foundation was this high-dimensional, rich dataset, which allowed for more detailed analysis and model construction.

In the preprocessing stage, the data was thoroughly cleaned, including missing values were found and handled, extraneous columns such as patient IDs were eliminated, categorical variables were encoded using one-hot encoding, and the StandardScaler was used to scale the numerical features. This stage was essential because accurate preprocessing guarantees that the ML models are trained on clean, objective data, because real-world medical data is naturally noisy and inconsistent.

The KMeans clustering step followed. The data was divided into three patient groups—Low Risk, Medium Risk, and High Risk—with the aid of the unsupervised learning technique. The ideal number of clusters (K=3) was found using the elbow approach, and the findings were confirmed by PCA visualization and silhouette scores. This step's output gave the dataset another useful feature: cluster labels, which were subsequently applied to improve the supervised machine learning models' capacity for prediction.

Five models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were used and contrasted throughout the classification stage. These models were selected because of their shown effectiveness in healthcare analytics as well as their complementing qualities, which range from robustness and high accuracy to simplicity and interpretability.

Eighty percent of the dataset was used to thoroughly train each model, while the remaining twenty percent was used for testing. Accuracy, precision, recall, and F1-score were among the criteria used to assess each model's performance. The model that performed the best was the Random Forest classifier, which had a 92% accuracy rate with outstanding precision and recall values. To further explain the classification results, a confusion matrix and ROC curve were also produced, which confirmed the high dependability of the model.

The project's use of visualization was crucial. Elbow curves, pie charts, bar graphs for model comparison, and the PCA cluster plot were among the graphics that gave the report a powerful visual component. These illustrations improved the results' readability for stakeholders who were not technical and assisted in their effective interpretation.

The system architecture was effective and modular. Data collection, preprocessing, clustering, label assignment, model training, evaluation, and output were all done in a streamlined manner. Scalability is guaranteed by this framework, which also permits future integration with fresh datasets or improved models.

The research ultimately succeeded in creating a reliable cancer prediction model and created a repeatable, expandable pipeline that can be used as a basis for additional health analytics initiatives. The report shows a deep grasp of the complete machine learning lifecycle by providing comprehensive documentation, analysis, and visual representation of each pipeline component, from the ingestion of raw data to the final model selection.

**Future Scope**

While the current results are promising, this project serves as just a beginning. There are numerous opportunities to extend this work into more practical and impactful solutions. Below are several directions that can significantly enhance the scope and utility of the project:

1. **Integration with Real-Time Medical Systems**

One of the most exciting future applications for this predictive system is the integration with hospital information systems (HIS) and electronic health records (EHRs). This could allow the model to be updated in real time, providing doctors and clinicians with risk assessments right away. This integration might be used by hospitals to automatically rank cancer patients based on their urgency and risk**.**

1. **Extension to Other Diseases**

The current pipeline is adaptable and can be modified to treat more conditions like diabetes, heart disease, kidney problems, and neurological issues. The same procedures — preprocessing, clustering, modeling, and evaluation — can be carried out using the appropriate dataset. Because of this, the system is scalable and reliable in a variety of healthcare disciplines.

1. **Using Deep Learning Models**

Although traditional machine learning models were adequate for this project, performance could be improved by including deep learning methods like recurrent neural networks (RNNs) or convolutional neural networks (CNNs). Deep learning models have the potential to significantly increase prediction accuracy, particularly when image data (such as MRI scans or X-rays) is included in the dataset.

1. **Feature Engineering and Selection**

The investigation of sophisticated feature engineering methods is another exciting field. Applying feature selection algorithms such as Recursive Feature Elimination (RFE) or using domain knowledge to generate new features could improve model interpretability and lower computational load without compromising accuracy.

**5. Mobile and Web Dashboard for Doctors**

One effective technique to make the project realistically usable is to create a web or mobile application. Medical professionals can enter patient information into the dashboard to get a prognosis in real time with a confidence level. Adoption in clinics and hospitals can be facilitated by this kind of user-friendly interface..

**6. Inclusion of Genomic and Imaging Data**

CT scans, biopsy results, genetic data, and other intricate biomedical data might be included in future databases. Advanced methods like multimodal learning would be needed to integrate various data kinds, but there would be a significant increase in diagnostic accuracy possible. By combining organized and unstructured data, it would be possible to replicate how actual doctors use both images and figures to make choices.

**7. Bias and Fairness Audits**

Ensuring equity and removing bias in machine learning models is crucial since healthcare has a profound impact on people's lives. To guarantee that the model does not discriminate on the basis of age, gender, ethnicity, or socioeconomic status, future research should incorporate a thorough fairness audit. For ethical compliance, methods such adversarial debiasing, re-weighting, and re-sampling can be used..

**8. Clinical Trials and Validation**

The next step would be to carry out clinical studies in collaboration with medical facilities or research laboratories following academic validation. This will verify the system's efficacy in a medical setting and measure how well it works on actual patients. Commercial deployment of the technology would require long-term approval from regulatory agencies such as the FDA.

**9. Explainable AI (XAI) Techniques**

In vital domains such as healthcare, openness is essential. Medical experts will be able to comprehend why the model generated a specific prediction by adding explainability to the model outputs through the use of SHAP (SHapley Additive explanations) or LIME (Local Interpretable Model-agnostic explanations). This increases confidence in diagnostic instruments driven by AI.

**10. Automation of End-to-End Pipeline**

Lastly, solutions like Apache Airflow, MLflow, or Kubeflow can automate the full machine learning workflow, from data preprocessing to model deployment. As a result, the system will be ready for production and able to update itself with fresh information and enhanced models.

**Closing Thoughts**

The experiment effectively illustrated the role that machine learning can play in risk assessment and cancer detection. Early diagnosis can be supported by utilizing clustering and classification approaches, which can result in prompt treatment and better patient outcomes. The technology is a step toward developing smart, data-driven solutions for actual healthcare problems, not merely a proof-of-concept. This project may develop into a complete medical decision support system with more study, stakeholder cooperation, and clinical validation.

***REFERENCES***

**Research Papers & Journals**

1. Chaurasia, V., & Pal, S. (2017). *A Novel Approach for Breast Cancer Detection Using Data Mining Techniques*. International Journal of Innovative Research in Computer and Communication Engineering.
2. Khan, M. A., Hussain, N., Majid, A., et al. (2019). *Prediction of Breast Cancer using Machine Learning Techniques on Data Mining*. Computer Systems Science and Engineering, 39(2), 317–332.
3. Saba, T., & Alqahtani, A. (2020). *Machine learning techniques for cancer prediction: Comparative study and performance analysis*. Journal of Healthcare Engineering.
4. Kourou, K., Exarchos, T. P., Exarchos, K. P., et al. (2015). *Machine learning applications in cancer prognosis and prediction*. Computational and Structural Biotechnology Journal, 13, 8–17.
5. Deo, R. C. (2015). *Machine learning in medicine*. Circulation, 132(20), 1920–1930.

**Websites and Tools**

1. Scikit-learn Documentation: <https://scikit-learn.org/stable/>
2. Kaggle Datasets for Cancer Prediction: https://www.kaggle.com/datasets
3. Towards Data Science – ML for Healthcare: https://towardsdatascience.com
4. Analytics Vidhya – Clustering and ML Projects: https://www.analyticsvidhya.com
5. Python Official Documentation: https://docs.python.org

**Code and Platforms**

1. Google Colab – Online Python IDE: https://colab.research.google.com
2. GitHub (Machine Learning Cancer Prediction Repositories): <https://github.com>