DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

FINAL YEAR PROJECT 2022-2023

“Normalization and Classification of Histopathology Electronic Health Record on Breast Cancer using NLP and ML Approaches.”

Under the guidance of

Dr. R J Prathibha

Associate Professor

Team members:

Ananya Tomar	01JST19IS006
Anup Shandilya	01JST19IS008
Chandana M	01JST19IS013
Pruthvi Bhat	01JST19IS039

Contents:

- Introduction
- Literature Survey
- Limitations
- Problem Statement
- Relevance of the domain
- Aim and objective
- Requirement Specification
- System Architecture
- Methodology
- Implementation
- Results and Discussions
- Conclusion
- Future scope
- References

Introduction:

- Histopathology is the diagnosis and study of disorders of the tissues.
- The necessary characteristics that aid in determining whether cells are malignant or not can be found in histopathological data. EHR is made up of text-based, unstructured data.
- NLP approaches to create a structured system for cancer staging.
- The processed structured data is classified using various ML algorithms to predict the stage of cancer.

Literature Survey:

Author	Year	Title	Proposed work	Limitations
David S Carrell, Scott Halgrim, Diam Thi Tran	2013	Using NLP to improve the efficiency of manual chart abstraction in research : The case of Breast Cancer Recurrence	Developed an NLP based system using open source software to process EHR from 1995 to 2012 for women with early stage breast cancer to identify whether and when recurrences were diagnosed.	The study was limited to Stage one or two cancer and machine learning techniques were not incorporated which could have enhanced the accuracy of status annotations.
Wang Liwei, Sunyang Fu, Andrew Wen	2020	Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing.	A comprehensive review of cancer NLP for research and patient care using EHRs data. Issues and barriers for wide adoption of cancer NLP were identified and discussed.	Cancer research and patient care require some data elements beyond mCODE as expected. Transparency and reproductivity are not sufficient in NLP methods, and inconsistency in NLP evaluation exists.

Alexander W. Forsyth MEng, Regina Barzilay, Kevin S. Hughes	2018	Machine Learning methods to extract documentation of breast cancer symptoms from electronic health records.	Manually annotated sentences and trained a conditional random field model to predict words indicating symptoms. Sentences labeled by human coders were divided into raining, validation, and test data sets. Final model performance was determined on test data unused in model development or tuning.	Ambiguous descriptions in the free text which is in narrative manner. Accuracy could be increased if more stringent labeling was used. Used manual labeling which is time consuming. The small and restricted size
Matthias Becker, Stefan Kasper, Britta lockmann	2019	Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation	NLP pipeline developed to identify specific guideline-based patient information and to annotate it with Unified Medical Language System concepts for manual evaluation by physicians.	NLP techniques alone cannot aid in identifying the groups. A combination of risk factors and laboratory parameters were not considered. Falsely detected negation among the data increased false negative rate. Incomplete data in German language available.

Imon Banerjee, Selen Bozkurt, Jennifer Lee Caswell-Jin	2019	Natural Language Processing approaches to detect the timeline of metastatic recurrence of breast cancer.	Curated vocabulary by processing radiology and pathology reports. Developed and evaluated 2 NLP approaches to analyse free-text notes. Trained the NLP models and extracted results for future data.	Limited documentation of metastatic recurrence in clinical notes. Relatively short median follow-up time of 5 years. Understanding the basis for determinations made by neural networks is obscure.
Yoojoong Kim,Jeong Hyeon Lee, Sunho Choi	2020	Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records	Proposed a keyword extraction method for pathology reports based on the deep learning models for NLP, BERT, to extract pathological keywords, namely specimen, procedure, and pathology, from pathology reports.	The algorithm was developed using the pathology reports of a single institution, which might limit the generalizability of its application to other institutions.

Limitations:

- Large portion of this data is unstructured and stuck in free-text documents.
- Identification of medical terms related to cancer and extracting features for the structured data.
- To make classification easier, these retrieved features must be annotated.
- The order of occurrence of feature values in the text data should be managed to provide a universal approach for further generalization.

Problem Statement:

“Normalization and Classification of Histopathology Electronic Health Record on Breast Cancer using NLP and ML Approaches.”

Relevance of the domain:

- Most of the pertinent data needed to make reliable forecasts and suggestions, when working in healthcare can be found in free-text clinical notes. A large portion of this data is unstructured and stuck in free-text documents.

- NLP involves feeding an algorithm large amounts of EHR notes from which it “learns” a set of rules to identify what is meaningful.

- Machine learning can make patterns evident but only if the data used is clean, normalized and complete.

Aim and Objectives:

Aim:

To determine whether a new sample is malignant or not, as well as to assist the hospital in storing fresh samples in a tabular format.

Objectives:

- Collection of raw data.
- Preprocessing and Normalization of unstructured text data into structured data.
- Classification of new samples into stages of breast cancer using machine learning techniques.

Requirement Specification:

Software Requirements

IDE: Jupyter Notebook or Google Colab (efficient tool used for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.)

Programming Language: Python (Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation via the off-side rule. Python is dynamically typed and garbage-collected.

Operating Systems : The software will run on all kinds of OS. There is no limitation on OS, but it must have an updated version of web browsers.

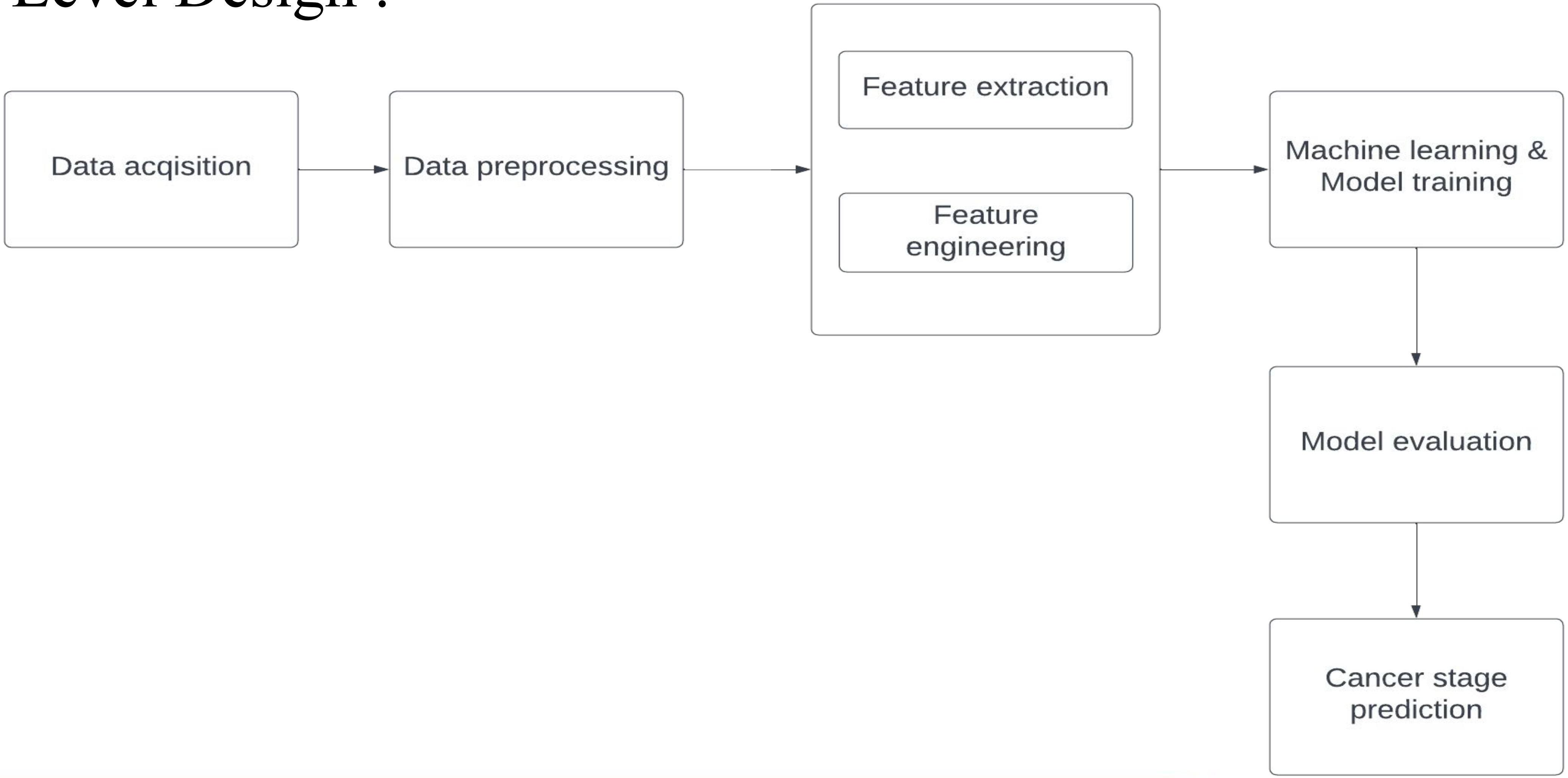
Hardware Requirements

Since the ML algorithms are running over the internet(Colab), all the connections to the internet will be hardware interface for the system

e.g., Modem, WAN - LAN, Ethernet Cross-Cable.

Design:

High Level Design :



Data acqisition

Data preprocessing

Feature extraction

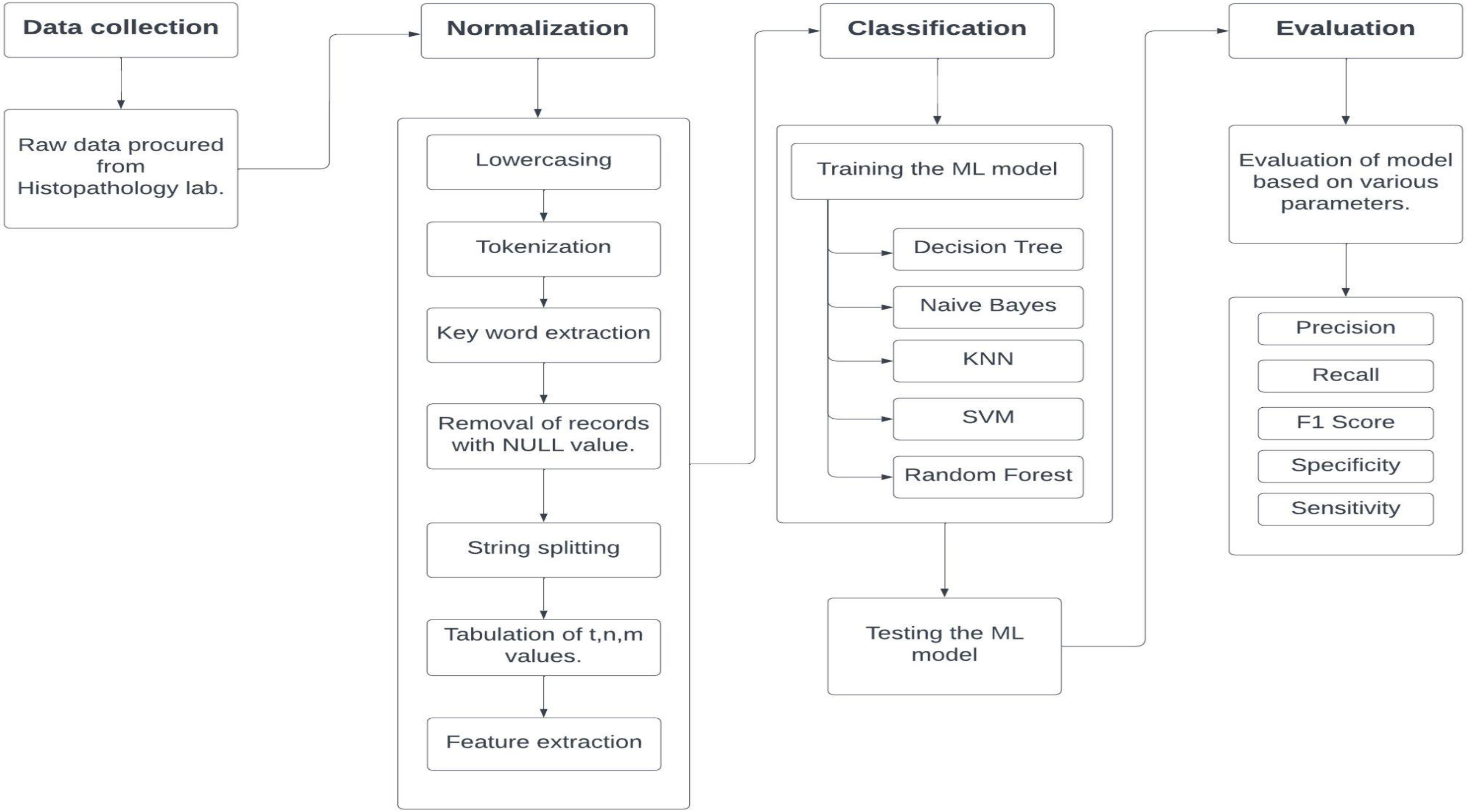
Feature engineering

Machine learning & Model training

Model evaluation

Cancer stage prediction

Low Level Design:



Data collection

Raw data procured from Histopathology lab.

Normalization

Lowercasing

Tokenization

Key word extraction

Removal of records with NULL value.

String splitting

Tabulation of t,n,m values.

Feature extraction

Classification

Training the ML model

Decision Tree

Naive Bayes

KNN

SVM

Random Forest

Testing the ML model

Evaluation

Evaluation of model based on various parameters.

Precision

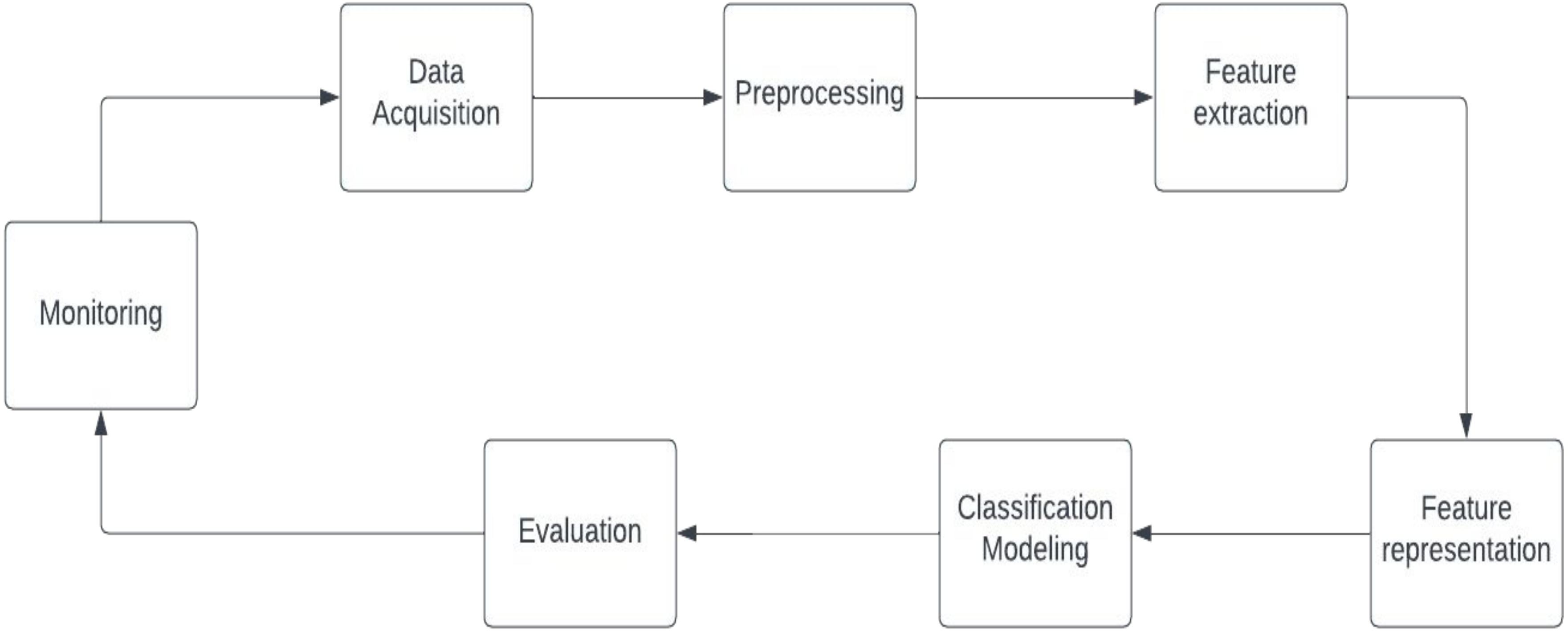
Recall

F1 Score

Specificity

Sensitivity

Methodology:



Data Acquisition

Preprocessing

Feature extraction

Feature representation

Classification Modeling

Evaluation

Monitoring

1. **Collection of raw data** : In this analytical study, EHRs are collected from the hospital. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users.

2. **Pre-processing of the data**: Later, NLP techniques are used to extract the useful information from the EHR. NLP is a form of machine learning which enables the processing and analysis of free text. When used with medical notes, it can aid in the prediction of patient outcomes and generate diagnostic models that detect early-stage chronic disease.

3. **Feature extraction**: Identifying the important keywords that helps determine whether the sample is malignant or not.

ex: number size and color of the cells.

4. **Feature representation**: included tabulation of the t, n, and m values.

5. **Classification modeling**: Various ML algorithms are used to classify the new sample with it's stage of cancer.

6. **Evaluation**: The most accurate algorithm is chosen after a comparison of the findings' accuracy from the different methods is done. The data is validated by comparing the obtained result with the ground truth value.

7. **Monitoring**: To achieve the best results, fresh samples are fed to the trained model.

Implementation:

The input raw text: (one record)

Left breast lump- carcinoma.Received two containers:1] Container 1: Labelled as left modified radical mastectomy.- Received a specimen of left modified radical mastectomy, measuring 15.5 X 12.5 X 3 cm, with overlying skin flap measuring 13 X 6.8 cm. - The overlying skin, nipple and areola are unremarkable.- On serial sectioning of the breast a unifocal, well circumscribed grey white tumour is identified, measuring 4.2 X 3.6 X 3 cm, located in the inner upper quadrant. - The tumour is located 0.6 cm from the skin and 0.1 cm from the deep resection margin.- The remaining breast tissue shows grey white rubbery areas.2] Container 2: Labelled as Axillary lymph node: Specimen consists of multiple grey yellow fibrofatty tissue fragments measuring 9.5 x 3 cm.Isolated 32 lymphnodes, largest measuring 1.8 x 0.8 x 0.1 cm and smallest 0.3 x 0.3 cm. Cut section- Grey white.[Bits from:Container 1:Tumor: A-F; Deep resected margin: G-H; Adjacent breast: I-K; Nipple areola: L; Overlying skin: M Container 2:Lymphnodes: N-V]. MICRO:1] Sections from tumor in the left breast show cells arranged in sheets, clusters, lobules, cords and trabeculae. Individual cells have moderate amount of cytoplasm, vesicular nuclei with some showing prominent nucleoli with 25-30 mitoses per ten HPF. Foci of necrosis and areas of desmoplasia seen.- Lymphovascular invasion noted.- Perineural invasion not identified.- Adjacent breast shows fibrocystic change with foci of usual ductal hyperplasia.- Over lying skin is free of tumor.- Deep resected margin is free of tumor (Closest margin- 1 mm).- 3/32 lymph nodes show metastasis.Features are those of Invasive Ductal Carcinoma with lobular features. Grade 3 (Bloom Richardson score 3+2+3= 8), pT2N1Mx.

Preprocessing steps:

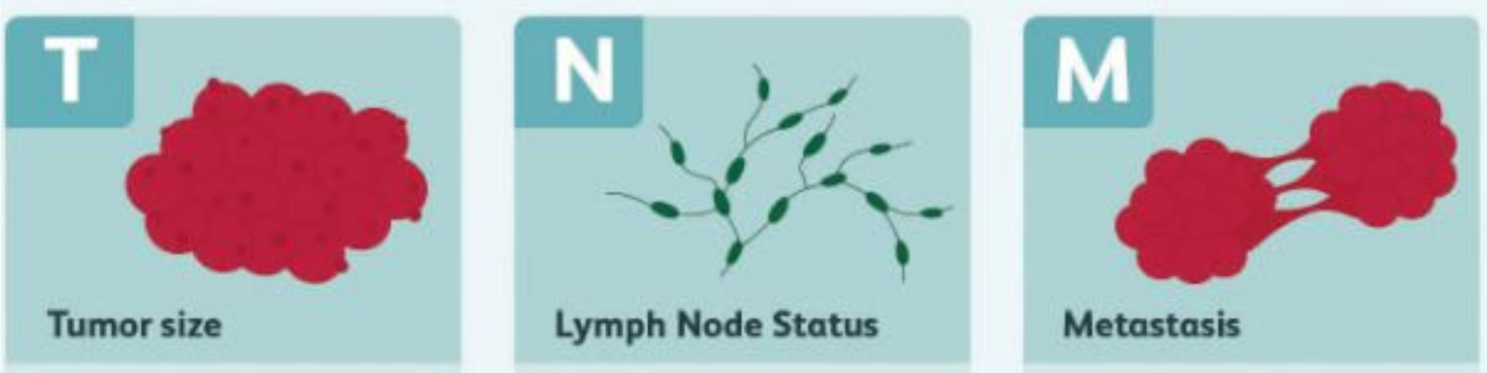
- Conversion of raw text to lowercase.
- Extracting keywords with p, t, n, m character in it.
- Removing the NULL records.
- Splitting the keyword into sub-strings as to extract the separate feature values.
- Tabulating the features against their values.
- Conversion of data frame into a csv file for further processing.

The labels in our dataset are:

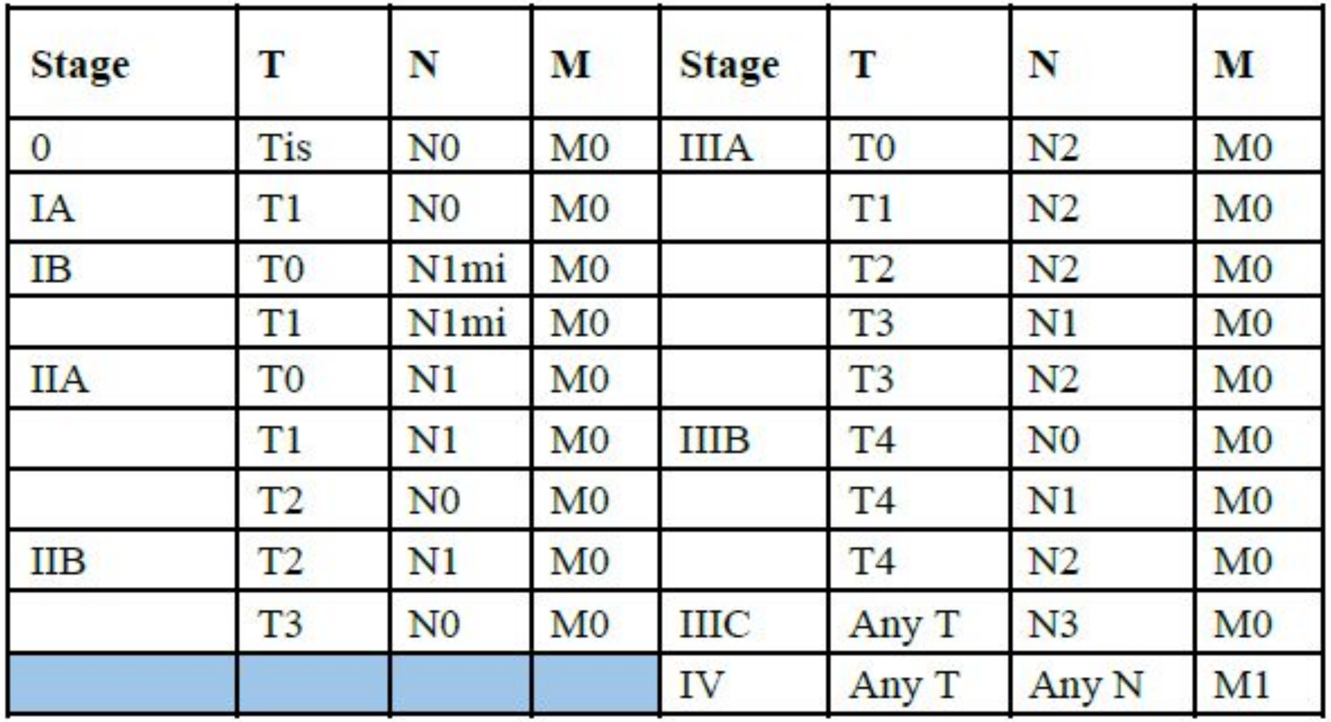
t (tumor size) → 0, 1a, 1b, 1c, 2, 3, 4a, 4b, 4c

n (lymph node) → 0, 1a, 1b, 1c, 2a, 2b, 3a, 3b, 3c

m (metastasis) → x, 1



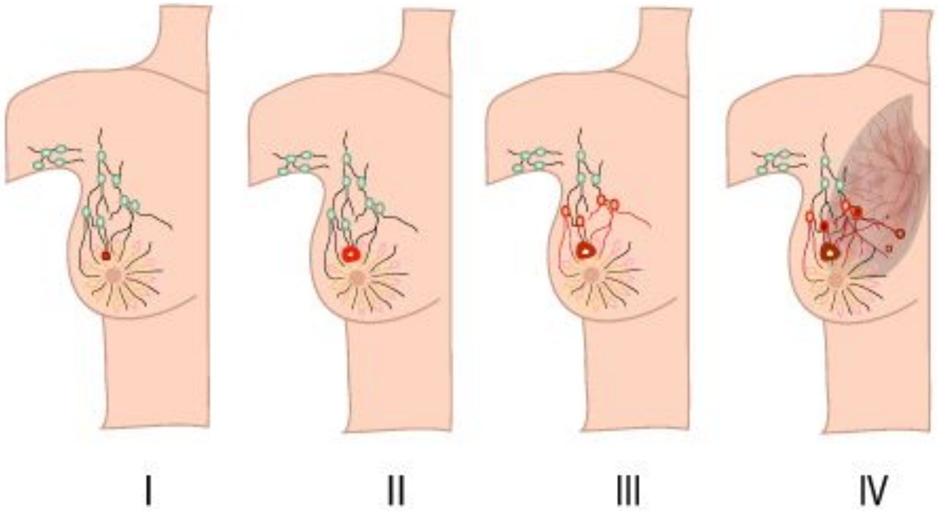
The reference table for labelling the tnm dataset by American Joint Committee on Cancer (AJCC).



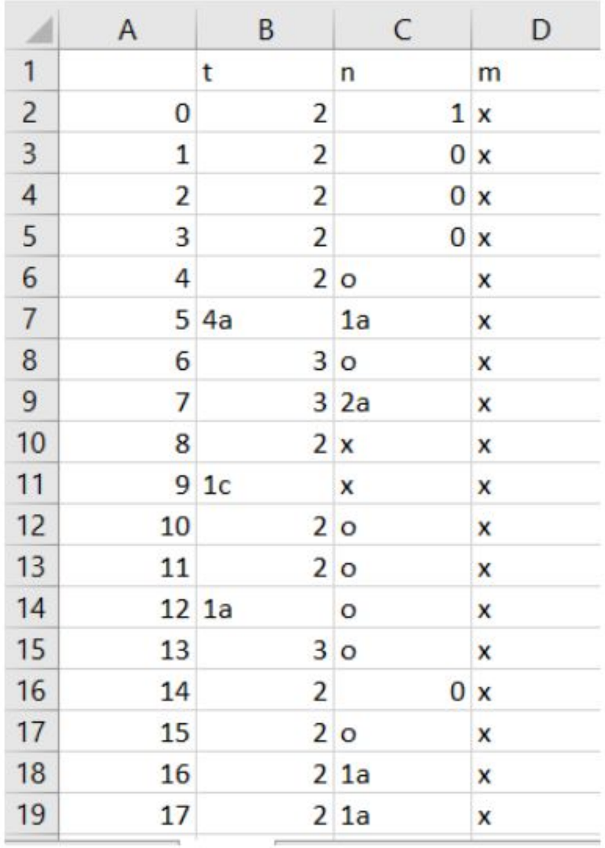
Stage	T	N	M	Stage	T	N	M
0	Tis	N0	M0	IIIA	T0	N2	M0
IA	T1	N0	M0		T1	N2	M0
IB	T0	N1mi	M0		T2	N2	M0
	T1	N1mi	M0		T3	N1	M0
IIA	T0	N1	M0		T3	N2	M0
	T1	N1	M0	IIIB	T4	N0	M0
	T2	N0	M0		T4	N1	M0
IIB	T2	N1	M0		T4	N2	M0
	T3	N0	M0	IIIC	Any T	N3	M0
				IV	Any T	Any N	M1

Classes for labelling:

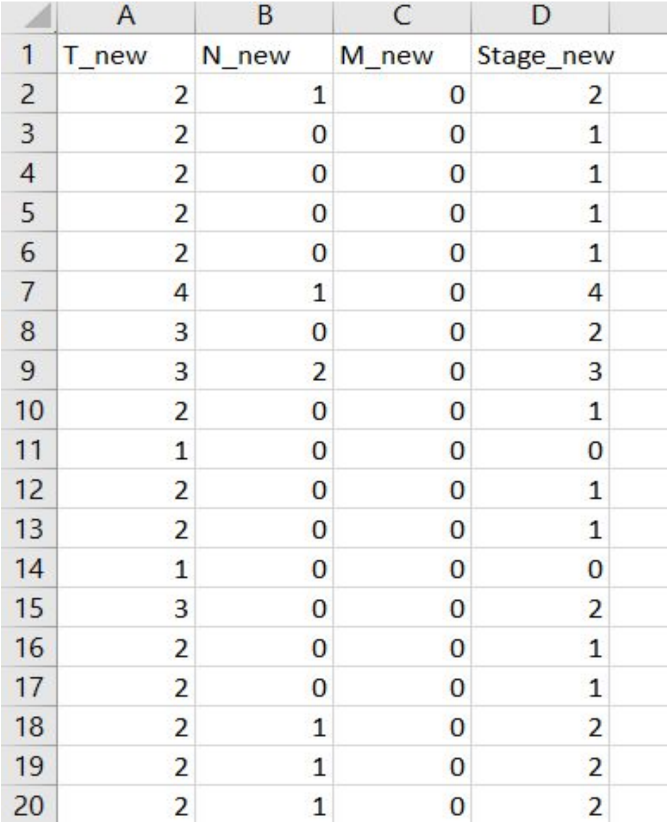
STAGE	LABEL
IA	1
IIA	2
IIB	3
IIIA	4
IIIB	5
IIIC	6
IV	7



Converting the t, n, m values from the raw text as per the AJCC standard and labelling the dataset.



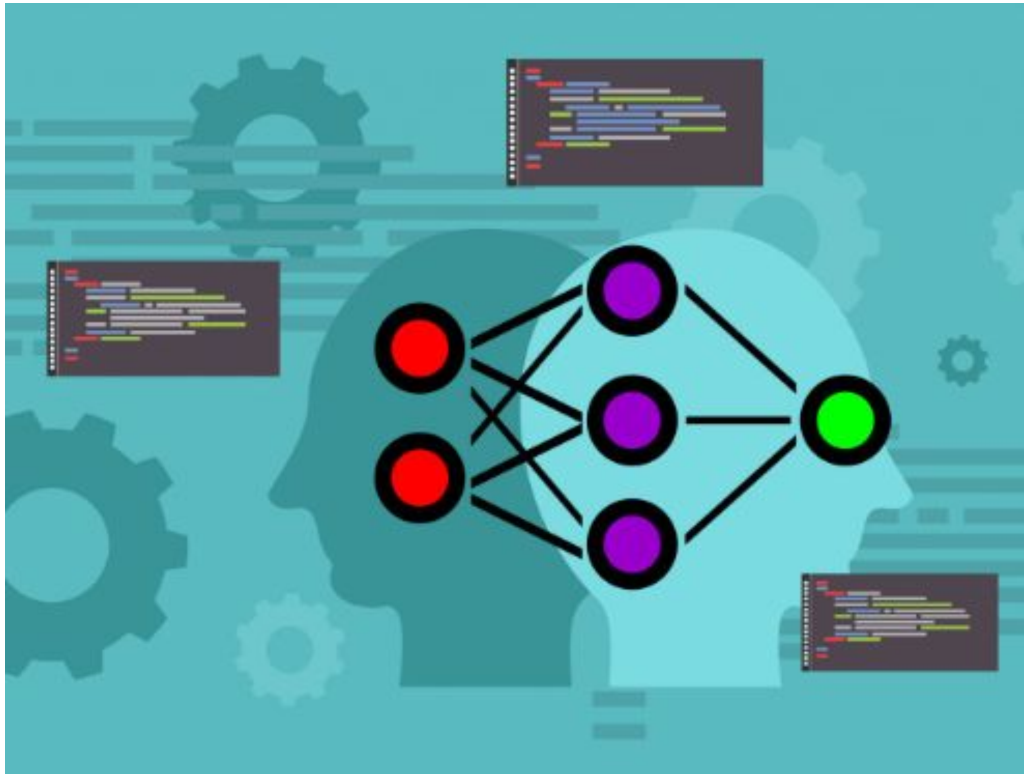
	A	B	C	D
1		t	n	m
2	0	2	1	x
3	1	2	0	x
4	2	2	0	x
5	3	2	0	x
6	4	2	o	x
7	5	4a	1a	x
8	6	3	o	x
9	7	3	2a	x
10	8	2	x	x
11	9	1c	x	x
12	10	2	o	x
13	11	2	o	x
14	12	1a	o	x
15	13	3	o	x
16	14	2	0	x
17	15	2	o	x
18	16	2	1a	x
19	17	2	1a	x



	A	B	C	D
1	T_new	N_new	M_new	Stage_new
2	2	1	0	2
3	2	0	0	1
4	2	0	0	1
5	2	0	0	1
6	2	0	0	1
7	4	1	0	4
8	3	0	0	2
9	3	2	0	3
10	2	0	0	1
11	1	0	0	0
12	2	0	0	1
13	2	0	0	1
14	1	0	0	0
15	3	0	0	2
16	2	0	0	1
17	2	0	0	1
18	2	1	0	2
19	2	1	0	2
20	2	1	0	2

Classification algorithms used:

- Decision Tree
- Gaussian Naive Bayes
- K-Nearest Neighbors (KNN)
- Random Forest
- Support vector Machine (SVM)
- Gradient Boosting



Evaluation metrics:

- **Accuracy:** A measure of how well a classification model predicts labels. It's the ratio of correctly predicted instances to total instances.
- **Recall:** Measures correctly identified positive instances. Shows how well the model finds all positives.
- **Precision:** Measures true positives among predicted positives. Reflects accuracy of positive predictions.
- **F1 Score:** Combines precision and recall into one metric. It’s the harmonic mean of Precision and Recall.
- **Sensitivity:** Indicates the model's ability to correctly identify positive instances.
- **Specificity:** Measures the model's ability to identify negative instances. High specificity means low false positives.

Tabulated result for all algorithms:

Training: Testing 90:10	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
Decision tree	72.189	0.684	0.722	0.698	0.452	0.945
Gaussian naive bayes	57.396	0.419	0.574	0.469	0.486	0.892
KNN	94.67	0.938	0.947	0.941	0.772	0.986
Random forest	**96.44**	**0.961**	**0.959**	**0.959**	**0.949**	**0.989**
SVM	85.57	0.938	0.947	0.941	0.695	0.977
Gradient Boosting	95.23	0.967	0.959	0.959	0.949	0.989

For the 90:10 split of the dataset into training and testing, **Random forest** approach gives the best accuracy of **96.44%** .

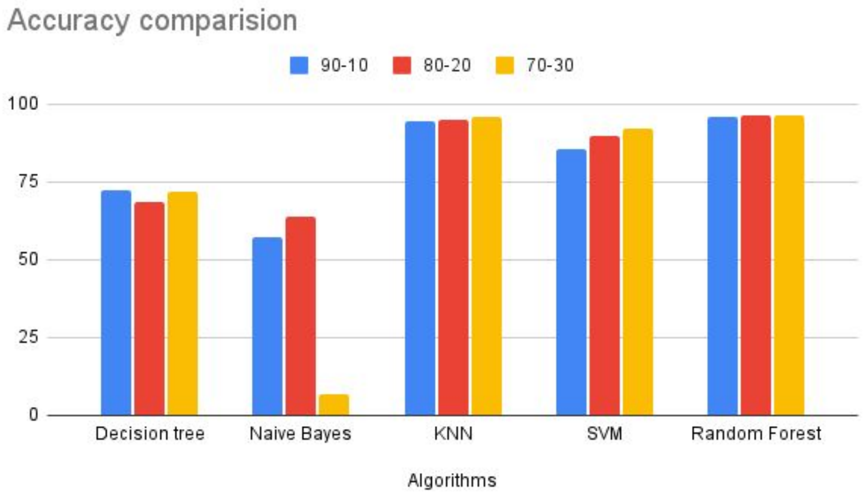
Training: Testing 80:20	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
Decision tree	68.639	0.609	0.686	0.641	0.487	0.957
Gaussian naive bayes	63.905	0.595	0.639	0.564	0.536	0.914
KNN	94.97	0.95	0.95	0.949	0.792	0.990
Random forest	**96.153**	**0.965**	**0.962**	**0.962**	**0.941**	**0.991**
SVM	89.94	0.95	0.95	0.949	0.698	0.982
Gradient Boosting	96.015	0.962	0.961	0.961	0.917	0.991

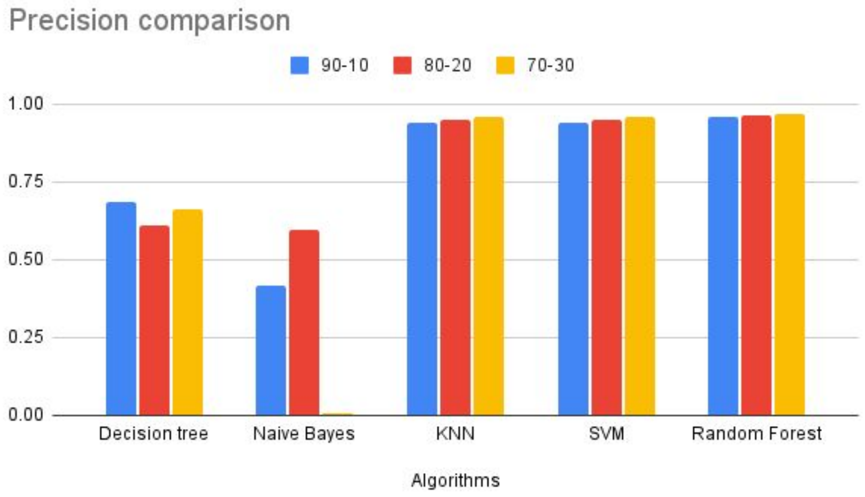
For the 80:20 split of the dataset into training and testing, **Random forest** approach gives the best accuracy of **96.153%**.

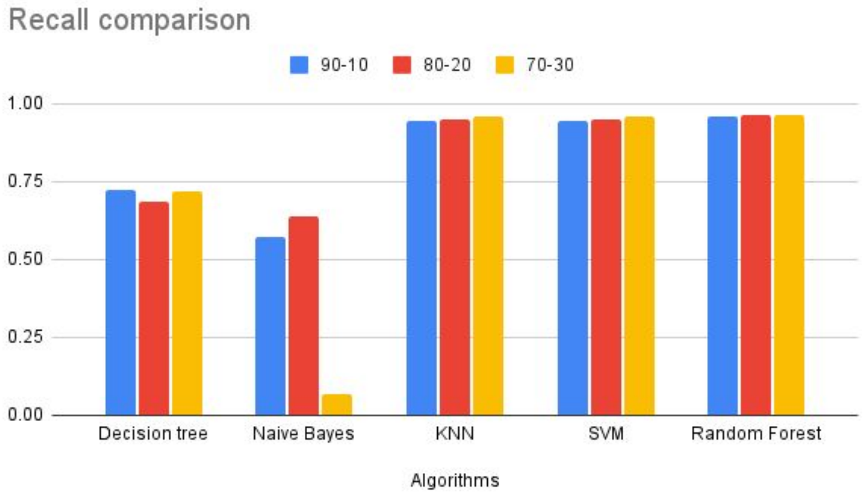
Training: Testing 70:30	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
Decision tree	71.992	0.660	0.720	0.686	0.451	0.943
Gaussian naive bayes	6.706	0.008	0.067	0.012	0.281	0.857
KNN	95.660	0.959	0.957	0.957	0.881	0.991
Random forest	**96.449**	**0.967**	**0.964**	**0.965**	**0.954**	**0.991**
SVM	92.110	0.959	0.957	0.957	0.821	0.987
Gradient Boosting	96.05	0.964	0.961	0.961	0.97	0.992

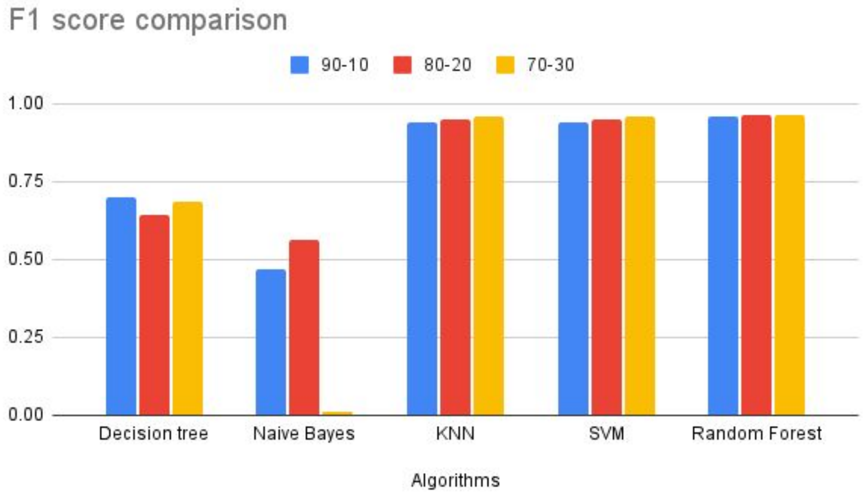
For the 70:30 split of the dataset into training and testing, **Random forest** approach gives the best accuracy of **96.449%**.

Comparison graphs on results obtained using different ML algorithms:

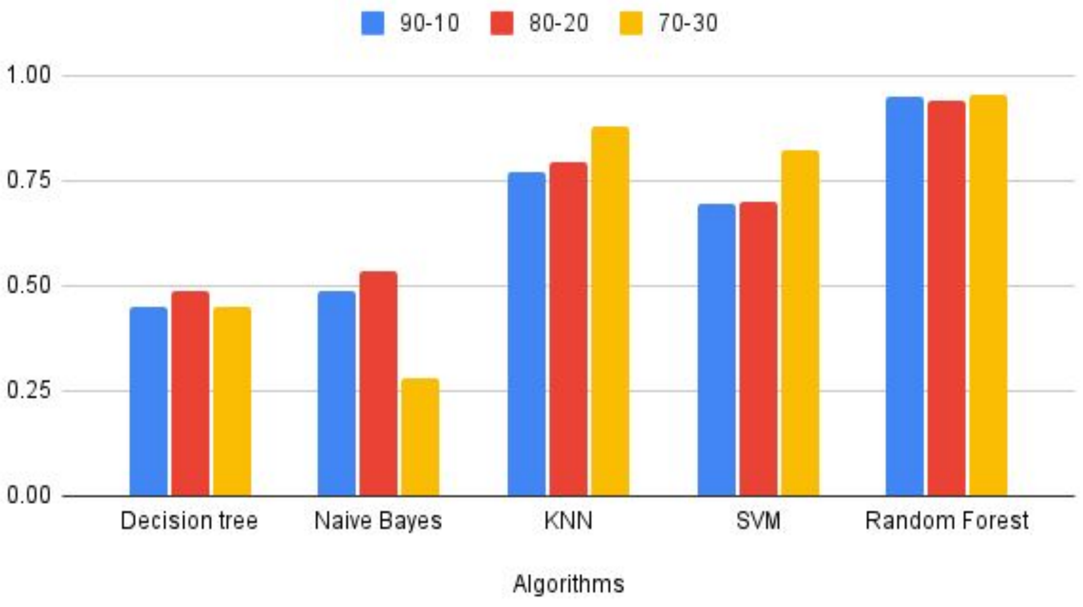




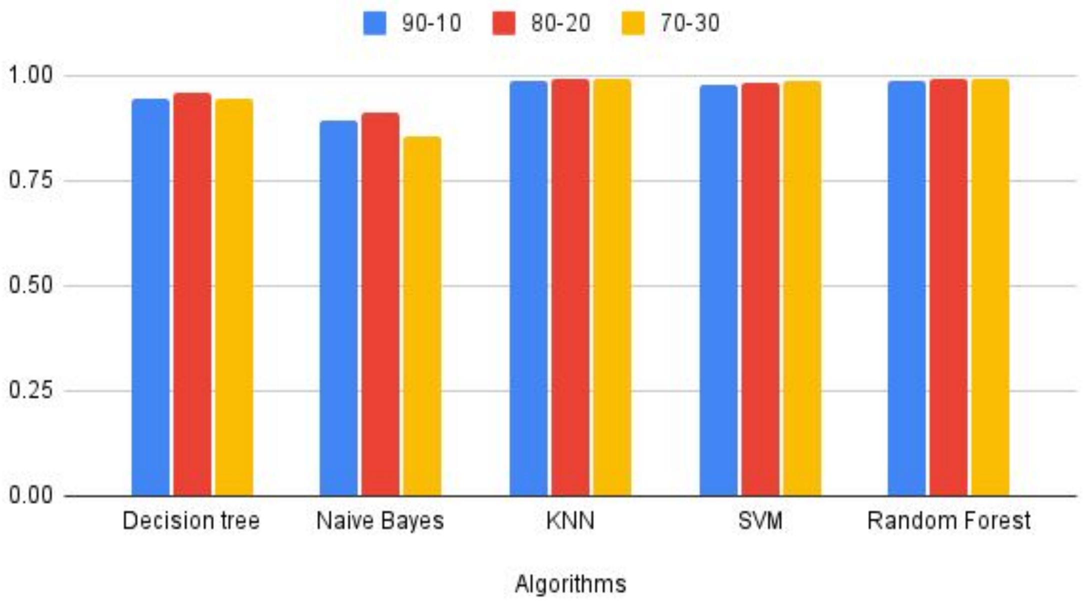




Sensitivity comparison



Specificity comparison



Conclusion:

- **Utilization of digital EHRs:** The project acquired digital EHRs from the hospital and extracted relevant attributes for cancer determination.

- **Importance of data normalization and feature extraction:** Standardizing and structuring EHRs enabled efficient analysis and decision-making in breast cancer-related data.

- **NLP-based feature extraction:** NLP techniques extracted cancer-related features from unstructured textual data, aiding in diagnosis and staging.

- **Enhanced cancer stage classification:** Integration of ML techniques improved predictive power and efficiency, providing valuable tools for patient care.

- Several preprocessing techniques were employed, including tokenization, case conversion, and string splitting, to clean and standardize the EHR data. Additionally, specific information such as the t, n, and m values, which are important for cancer staging, were extracted and tabulated.

- A reference table from the American Joint Committee on Cancer (AJCC) was obtained to assign appropriate cancer staging labels to the data records. This step involved matching the extracted values with the predefined staging criteria, thereby labeling each record accordingly.

- Next, the labeled dataset was split into training and testing sets using different ratios, specifically 70:30, 80:20, and 90:10, to evaluate the performance of various machine learning models.

- After evaluating the results, the Random Forest model performed the best when using the 70:30 split. Random Forest is an ensemble learning method that combines multiple decision trees. It excels in handling complex datasets, reduces overfitting, and provides robust predictions.

Future Scope:

- **Deep Learning Techniques:** Explore CNNs or RNNs to extract intricate patterns from EHRs, enhancing the model's predictive capabilities.

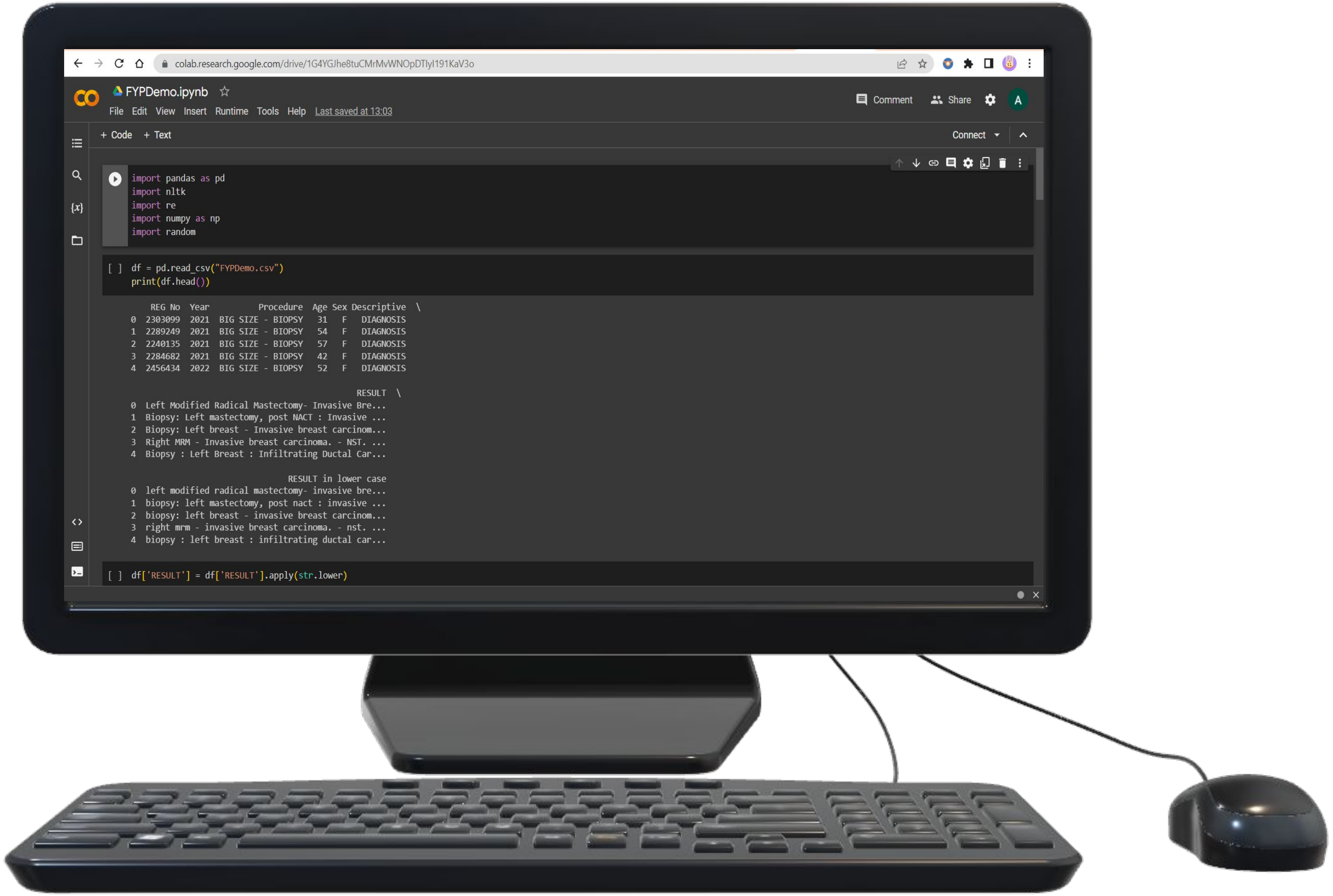
- **Expansion to Other Cancers:** Adapt the methodology to encompass different cancer types for a comprehensive diagnostic and staging system.

- **Real-Time Monitoring and Decision Support:** This systems can aid healthcare professionals in treatment planning and disease progression monitoring.

- **Collaborative Research and Data Sharing:** Collaborate with research institutions and healthcare organizations to share anonymized EHR data for analysis and model validation.

- **Integration with Clinical Workflow Systems:** Integrate the project's outcomes into existing clinical workflow systems for seamless access to predictive models and decision support tools.

Project Demo:


colab.research.google.com/drive/1G4YGJhe8tuCMrMvWNOpDTlyI191KaV3o
FYPDemo.ipynb
File Edit View Insert Runtime Tools Help Last saved at 13:03
Comment Share
A
+ Code + Text
Connect

```
import pandas as pd
import nltk
import re
import numpy as np
import random
```

```
[ ] df = pd.read_csv("FYPDemo.csv")
    print(df.head())
```

```
     REG No  Year          Procedure  Age Sex Descriptive  \
0  2303099  2021  BIG SIZE - BIOPSY   31   F   DIAGNOSIS
1  2289249  2021  BIG SIZE - BIOPSY   54   F   DIAGNOSIS
2  2240135  2021  BIG SIZE - BIOPSY   57   F   DIAGNOSIS
3  2284682  2021  BIG SIZE - BIOPSY   42   F   DIAGNOSIS
4  2456434  2022  BIG SIZE - BIOPSY   52   F   DIAGNOSIS

                                              RESULT  \
0  Left Modified Radical Mastectomy- Invasive Bre...
1  Biopsy: Left mastectomy, post NACT : Invasive ...
2  Biopsy: Left breast - Invasive breast carcinom...
3  Right MRM - Invasive breast carcinoma. - NST. ...
4  Biopsy : Left Breast : Infiltrating Ductal Car...

                                 RESULT in lower case
0  left modified radical mastectomy- invasive bre...
1  biopsy: left mastectomy, post nact : invasive ...
2  biopsy: left breast - invasive breast carcinom...
3  right mrm - invasive breast carcinoma. - nst. ...
4  biopsy : left breast : infiltrating ductal car...
```

```
[ ] df['RESULT'] = df['RESULT'].apply(str.lower)
```

Thank you