Devising a solution to the problems of Cancer awareness in Telangana

Priyanka Avhad
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
pravhad_b19@ce.vjti.ac.in

Vedanti Kshirsagar
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
vakshirsagar_b19@ce.vjti.ac.in

Mahek Nakhua
Department of Computer Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
msnakhua_b19@ce.vjti.ac.in

Urvi Ranjan
Department of Electronics Engg.
Veermata Jijabai Technological Institute
Mumbai, India 400019
urranjan_b19@el.vjti.ac.in

***Abstract*—According to the data[1], the percent of women who underwent screening for cervical cancer, breast and oral cancer in Telangana in the year 2019-2020 was 3.3 per cent, 0.3 per cent and 2.3 per cent respectively. Although early detection is the only way to reduce morbidity and mortality, people have very low awareness about cervical and breast cancer signs and symptoms and screening practices. We developed an ML classification model to predict if a person is susceptible to breast or cervical cancer based on demographic factors. We devised a system to provide suggestions for the nearest hospital or Cancer treatment centres based on the user's location or address. In addition to this, we can integrate the health card to maintain medical records of all individuals and conduct awareness drives and campaigns. For ML classification models, we used decision tree classification and support vector classification algorithms for cervical cancer susceptibility and breast cancer susceptibility respectively. Thus, by devising this solution we come one step closer to our goal which is spreading cancer awareness, thereby, decreasing the cancer mortality and increasing cancer literacy among the people of Telangana.**

I. Introduction

Cancer Literacy is essential to reduce global cancer mortality and crucial for the detection of cancer at an early stage. Breast Cancer is the most common type of cancer women are diagnosed with worldwide, including in India where advanced stages of diagnosis and rising incidence and mortality rates make it imperative to increase cancer literacy in women.

The NFHS dataset[1] was examined for the indicators, Screening for Cancer among Women (age 30-49 years) in case of cervical cancer, breast cancer, oral cancer. This was classified into two categories - Residence in rural and urban areas, District-wise.

The percentages of screening for Cancer among Women (age 30-49 years) in rural and urban areas are as shown in Fig.1

To make a comparison based on districts in Telangana, a new combined dataset was designed with all latitudes and longitudes. Data visualization of 3 indicators (Ever undergone a screening test for cervical cancer (%), Ever undergone a breast examination for breast cancer (%), Ever undergone an oral cavity examination for oral cancer (%)) was performed. Fig. 2, Fig. 3, and Fig.4 shows results mapped onto the district along with a tabular representation of all the districts in decreasing order of percentages.

NFHS		Urban	Rural
Women of Telangana in the age group of 30-49	Who have ever undergone a screening test for cervical cancer	2.3 %	3.9 %
	Who have ever undergone a breast examination for breast cancer	0.3 %	0.4 %
	Who have ever undergone an oral cavity examination for oral cancer	3.2 %	2.1 %

Fig. 1. Summary of NFHS dataset highlighting the small percentage of women undergoing these tests

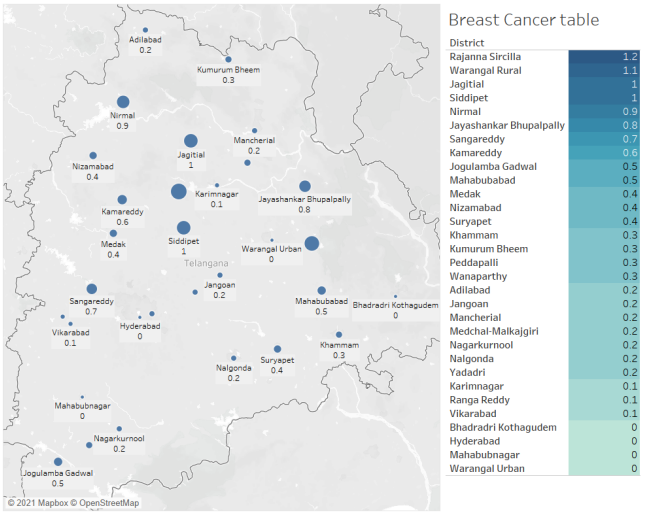

Fig. 2. Percentage distribution of Breast Cancer examination tests taken by Women in Telangana (District wise)

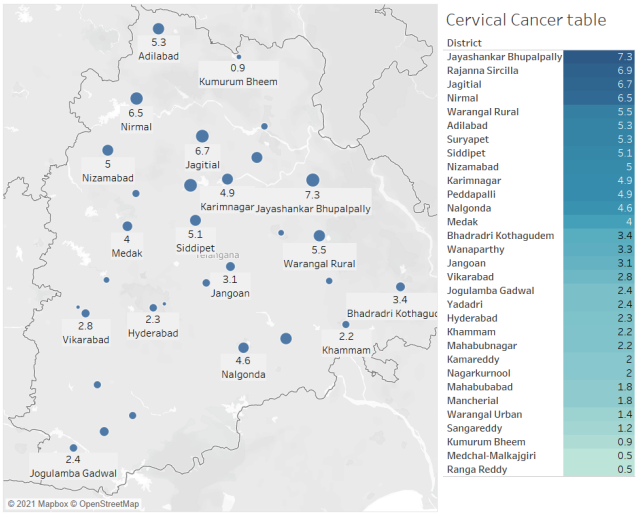

Fig. 3. Percentage distribution of screening tests for Cervical Cancer taken by Women in Telangana (District wise)

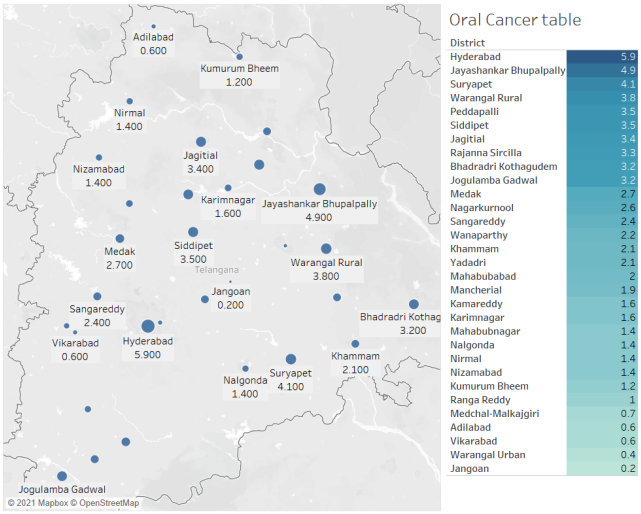

Fig. 4. Percentage distribution of oral cavity examination for Oral Cancer taken by Women in Telangana (District wise)

According to the report "Profile of Cancer and Related Factors - Telangana, 2021"[2], the five leading sites of cancers in females are Breast (35.5%), Cervix Uteri (8.7%), Ovary (6.9%), Corpus Uteri (5.5%), Lung (4.1%). In the age group of 0 to 74 years, the cumulative risk of developing cancer in females is 1 in every 7. (Cumulative risk is the probability that an individual will be diagnosed with cancer in the absence of any competing cause of death and assuming that the current trends prevail over time).

On further studying the clinical extent of disease for cancer of selected anatomical sites in females, the following proportion shown in Fig.5 was found considering all the cases.

A limitation of the Clinical Extent of Disease at presentation (%) for cancers of selected anatomical sites is that it has been calculated from the HBCRs in the state which may not represent the entire state.

	Localized Only (%)	Loco Regional (%)	Distant Metastasis (%)	Unknown Extent (%)
Breast	32	60	7	1
Cervix Uteri	71	29	-	-
Lung	30	47	21	2
Stomach	30	50	19	1
Head & Neck	44	51	4	1

Fig. 5. Summary of cases with anatomical sites in females in Telangana

Projected incidence of cancer cases for the state for the years 2020 and 2025 was calculated according to gender using incidence data from the composite period of 2012-2016 was used as a reference. In the case of females, it was found to be 25434 cases in 2020 and 28708 cases in 2025.

From both of these studies/reports, we can conclude that a minuscule percentage of women undergo screening tests despite the high-risk factor. The probable reasons for this could be unawareness of the severity of existing risk, lack of accessibility to medical facilities especially in rural areas, and financial constraints. There is an immediate need for state-level and PAN India awareness programs, involving multiple stakeholders from society and the health system to improve cancer literacy in India. Thousands of lives could be saved each year if people were more aware of the signs and symptoms of cancer and people looked for help as soon as possible as treatment is usually more effective in the early stages of cancer.

Some awareness projects that already took place in Telangana are as follows:

A free cancer screening and training camp was organized at the Appolo Hospital as part of the international cancer conference going on at the hospital. As part of the camp over 60 DWCRA women were trained in procedures like self-breast examination for the early detection of cancer. According to Dr. Nalini, an oncologist, who was one of the trainers, the program aims at spreading the message across the state that cancer is a curable disease if detected early.

Dr. Charanjith Reddy Veeramalla, Managing Director of Omega Bannu Hospitals, said breast and cervical cancer were most common among Indian women and the number is increasing in recent times. However, they can be preventable if detected at an early stage through the screening tests and that there is a need to conduct more screening tests for the women for the prevention of breast and cervical cancer. The District Legal Services Authority (DLSA) of Warangal, in association with Omega Bannu Hospitals, had organized a free cancer screening test camp at the DLS building in Aug 2021 for the benefit of female judicial officers, staff members, and advocates. A total of 80 women underwent tests at the camp. Tests like Mammography for Breast Cancer(for females age 40 years above), Pap smear for Cervical Cancer(for females age 18 years above), Anaemia Screening by Blood Test and ECG, were conducted at the camp.

II. Model Function

The main goal of our model is the classification of patients as per their susceptibility to breast and cervical cancer. Cancer, to a certain extent, depends on environmental and demographic factors as well as genetic predisposition. Hence early screening of individuals susceptible to cancer can greatly reduce treatment needed, recovery time and provide a better chance at survival. The model developed in this project learns the relationship between these factors and the risk of developing cancer. It classifies individuals into two categories, susceptible to breast/cervical cancer and not susceptible to it. Susceptible individuals can then go through clinical breast exams, mammographies and ultrasounds to detect cancer early.

III. Data

A. Cervical Cancer

The data set was obtained from Kaggle[3], which is gratefully acknowledged. The data was gathered at the 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset includes 858 patients' demographics, habits, and medical records from the past. Because of privacy concerns, several patients chose not to answer certain questions (missing values). This data set describes the risk factors for cervical cancer that lead to a biopsy. This dataset is made up of 36 columns and 858 rows. It is a multivariate dataset that contains both real and integer values.

B. Breast Cancer

The data set was obtained from BCSC Research[4] and is kindly appreciated. Originally collected by the Breast Cancer Surveillance Consortium in 1996-2002, this dataset specifies risk factors for susceptibility to breast cancer. This data record consists of 16 columns and 4,62,563 rows. It is a multivariate data set with real and integer values.

IV. Preprocessing and features

A. Cervical Cancer

All numerical data provided in the dataset was processed before use. There were a lot of missing values and 22 duplicate values in the dataset which were completely removed to avoid ambiguity. Certain columns with exceeding null values and no correlation were completely excluded from the dataset. All the numerical values were cast to float type using astype() in python. After cleaning the dataset, patients with no null values and duplicate values were retained in the dataset. After this processing, a dataset of 688 patients was obtained which was subsequently used for model training. The data set is split for training and validation in a ratio of 3:1 (75% and 25% respectively). The age, number of sexual partners, first sexual encounter, number of pregnancies, smokes(years), Hormonal Contraceptives (years), IUD (years), Dx: HPV, etc are the primary features used for modelling. The independent parameters consisted of all the columns except 'Dx: Cancer' from the processed dataset. The dependent parameter was the column 'Dx: Cancer'. The independent features were fitted to avoid variance.

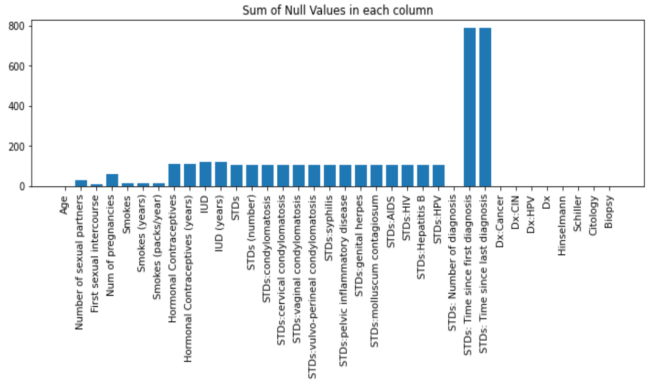

Fig. 6. Number of null values in every feature of the dataset

```
Index(['Age', 'Number of sexual partners', 'First sexual intercourse',
       'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)',
       'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD',
       'IUD (years)', 'STDs', 'STDs (number)', 'STDs:condylomatosis',
       'STDs:vaginal condylomatosis', 'STDs:vulvo-perineal condylomatosis',
       'STDs:syphilis', 'STDs:pelvic inflammatory disease',
       'STDs:genital herpes', 'STDs:molluscum contagiosum', 'STDs:HIV',
       'STDs:Hepatitis B', 'STDs:HPV', 'STDs: Number of diagnosis',
       'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller',
       'Citology', 'Biopsy'],
      dtype='object')
```

Fig. 7. Features used for training the ML classification model for cervical cancer

B. Breast Cancer

The data collected as mentioned above was cleaned and processed before it was used for model fitting. The data set contained a 'training' column that divided the data into validation and training categories, which was removed. In addition, missing values in the data set were identified and rows with missing values were removed. Duplicate samples (14655) have also been removed.

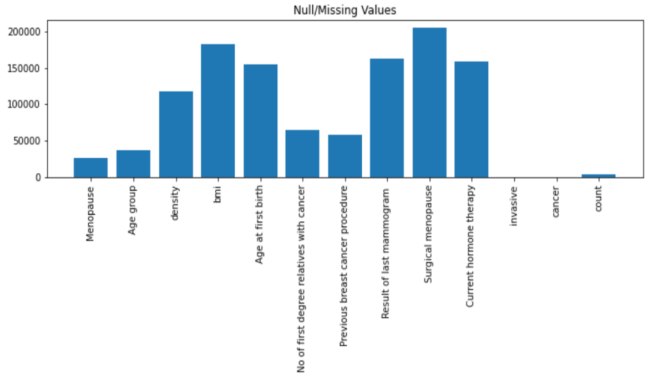

Fig. 8. Number of missing values in every feature of the dataset

After cleaning, 15203 rows were left. The data was also normalised by scaling to unit variance using StandardScaler() before training. The data set was split for training and testing in a ratio of 1:1 (50% and 50% respectively).

The primary features used for modelling are the age group, menopause, invasive tumour, basal metabolic index, age at first birth, number of relatives with first-degree cancer, current

hormone therapy, BI-RADS density, previous breast cancer procedure, the result of the mammogram, surgical menopause and current hormone therapy. The dependent feature to be predicted is cancer.

```
Features:
Menopause
Age group
density
bmi
Age at first birth
No of first degree relatives with cancer
Previous breast cancer procedure
Result of last mammogram
Surgical menopause
Current hormone therapy
invasive
count
```

Fig. 9. Features used for training the machine learning algorithm

V. Exploratory Data Analysis

A. Cervical Cancer

After loading the dataset we first look at the dimensions which are (650, 32). Looking at the information of the dataset to get insights to the data like its features, data types of the feature, etc. Thereafter, we preprocess and clean the data. Statistical summary of the features can be useful in inspecting the feature distribution and anomalies, if any.

The data given in Fig. 11, Fig. 12, and Fig. 13 gives us a lot of valuable information about the data.

- The maximum value in the 'Age' column is 84, but the maximum value in the other columns is much lower, which could lead to poor model performance because the 'Age' column has more influence than the other columns. To avoid differences in influence when training the model, standardise the values of all columns.The maximum value in the 'Age' column is 84, but the maximum value in the other columns is much lower, which could lead to poor model performance because the 'Age' column has more influence than the other columns. To avoid differences in influence when training the model, standardise the values of all columns.
- The maximum value in the 'Num of pregnancies' column is 11, which is a very high number of pregnancies, and it is possible that this is an outlier that affects all of the other values in this column. The solution could be to delete these rows, but we will only do so if our performance is poor.
- The columns 'STDs:cervical condylomatosis' and 'STDs:AIDS' contain only zeros and are thus useless. Getting rid of them is the solution.
- Since the mean is 0.0255, the 'Dx:Cancer' column (which will be our dependent variable or variable to predict) is very unbalanced. The mean would be close to 0.5 if the class was well balanced. For a better understanding, we will depict this with a plot. Solving this problem is extremely difficult; the best solution would be to obtain more positive data to train our model, but this is not possible in our case; another solution could be to remove some negative cases to balance them with the positive cases, but this would result in a significant loss of information.

Before we can standardise our data, we need to know if we have columns that provide the same (or very similar) information, which could cause our model to perform poorly. This information can be obtained by creating a correlation matrix. We then plot a few columns against the Dx:Cancer column individually to check dependence of Cancer on each of the features.

```
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           832 non-null    object
 2   First sexual intercourse            851 non-null    object
 3   Num of pregnancies                  802 non-null    object
 4   Smokes                              845 non-null    object
 5   Smokes (years)                      845 non-null    object
 6   Smokes (packs/year)                 845 non-null    object
 7   Hormonal Contraceptives             750 non-null    object
 8   Hormonal Contraceptives (years)     750 non-null    object
 9   IUD                                 741 non-null    object
 10  IUD (years)                         741 non-null    object
 11  STDs                                753 non-null    object
 12  STDs (number)                       753 non-null    object
 13  STDs:condylomatosis                 753 non-null    object
 14  STDs:cervical condylomatosis        753 non-null    object
 15  STDs:vaginal condylomatosis         753 non-null    object
 16  STDs:vulvo-perineal condylomatosis  753 non-null    object
 17  STDs:syphilis                       753 non-null    object
 18  STDs:pelvic inflammatory disease    753 non-null    object
 19  STDs:genital herpes                 753 non-null    object
 20  STDs:molluscum contagiosum          753 non-null    object
 21  STDs:AIDS                           753 non-null    object
 22  STDs:HIV                            753 non-null    object
 23  STDs:Hepatitis B                    753 non-null    object
 24  STDs:HPV                            753 non-null    object
 25  STDs: Number of diagnosis           858 non-null    int64
 26  STDs: Time since first diagnosis    71 non-null     object
 27  STDs: Time since last diagnosis     71 non-null     object
 28  Dx:Cancer                           858 non-null    int64
 29  Dx:CIN                              858 non-null    int64
 30  Dx:HPV                              858 non-null    int64
 31  Dx                                  858 non-null    int64
 32  Hinselmann                          858 non-null    int64
 33  Schiller                            858 non-null    int64
 34  Citology                            858 non-null    int64
 35  Biopsy                              858 non-null    int64
dtypes: int64(10), object(26)
memory usage: 241.4+ KB
```

Fig. 10. Information of the dataset

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs	STDs (number)
count	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000
mean	27.264970	2.523952	17.142216	2.323353	0.143713	1.235524	0.458953	0.643713	2.290037	0.112275	0.530030	0.097305	0.166168
std	8.727432	1.640299	2.852046	1.465319	0.351061	4.193611	2.336308	0.479260	3.724400	0.315942	2.001308	0.296595	0.551073
min	13.000000	1.000000	10.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	21.000000	2.000000	15.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	26.000000	2.000000	17.000000	2.000000	0.000000	0.000000	0.000000	1.000000	0.500000	0.000000	0.000000	0.000000	0.000000
75%	33.000000	3.000000	18.000000	3.000000	0.000000	0.000000	0.000000	1.000000	3.000000	0.000000	0.000000	0.000000	0.000000
max	84.000000	28.000000	32.000000	11.000000	1.000000	37.000000	37.000000	1.000000	22.000000	1.000000	19.000000	1.000000	4.000000

Fig. 11. Statistical Summary of the dataset

STDs:condylomatosis	STDs:cervical condylomatosis	STDs:vaginal condylomatosis	STDs:vulvo-perineal condylomatosis	STDs:syphilis	STDs:pelvic inflammatory disease	STDs:genital herpes	STDs:molluscum contagiosum	STDs:AIDS	STDs:HIV	STDs:Hepatitis B	STDs:HPV
668.000000	668.0	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.0	668.000000	668.000000	668.000000
0.055389	0.0	0.005988	0.053892	0.022455	0.001497	0.001497	0.001497	0.0	0.019461	0.001497	0.002994
0.228910	0.0	0.077208	0.225974	0.148269	0.038691	0.038691	0.038691	0.0	0.138242	0.038691	0.054677
0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
1.000000	0.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.0	1.000000	1.000000	1.000000

Fig. 12. Statistical summary of the dataset

STDs: Number of diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000	668.000000
0.092814	0.025449	0.004491	0.023952	0.023952	0.044910	0.094311	0.058383	0.067365
0.310355	0.157603	0.066915	0.153015	0.153015	0.207262	0.292480	0.234642	0.250841
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Fig. 13. Statistical summary of the dataset

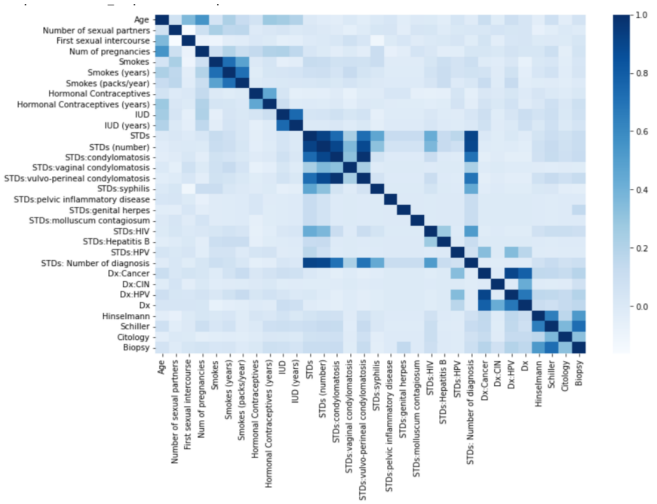

Fig. 14. Heatmap depicting correlation between all the dataset features

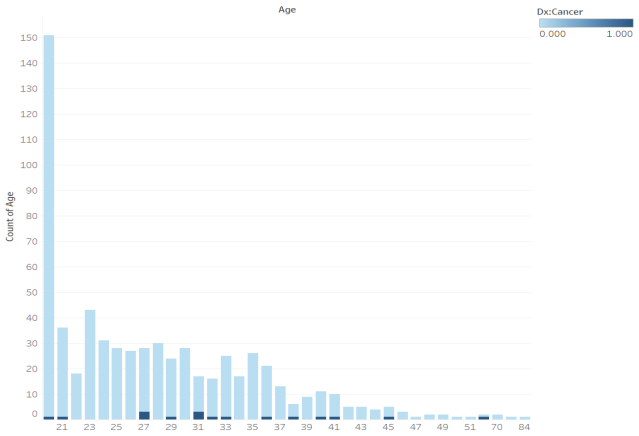

Fig. 15. Bar plot depicting the ratio cancerous and non-cancerous samples against age

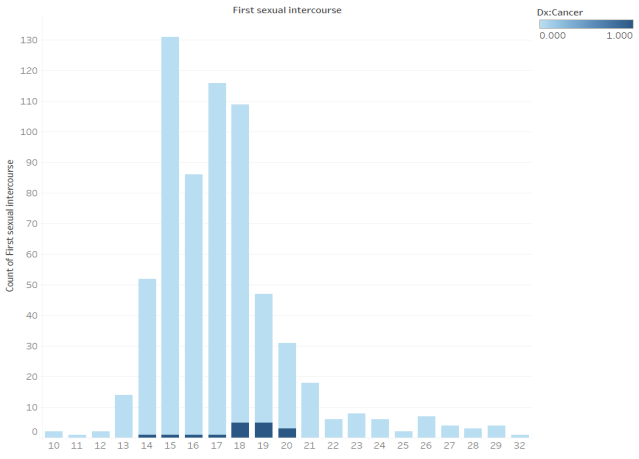

Fig. 16. Bar plot depicting the ratio cancerous and non-cancerous samples against first sexual intercourse

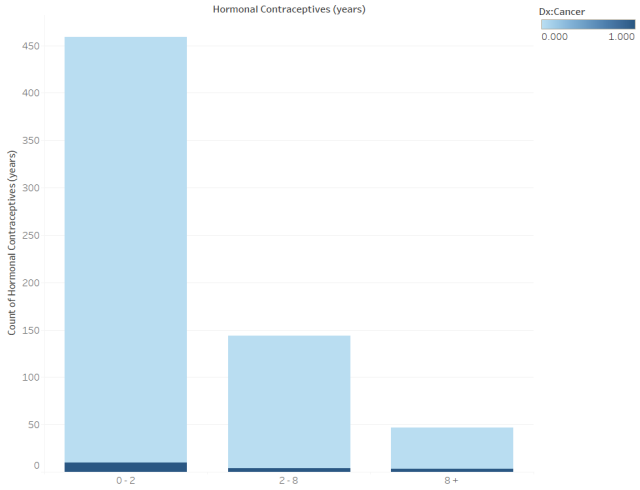

Fig. 17. Bar plot depicting the ratio cancerous and non-cancerous samples against Hormonal Contraceptives(years)

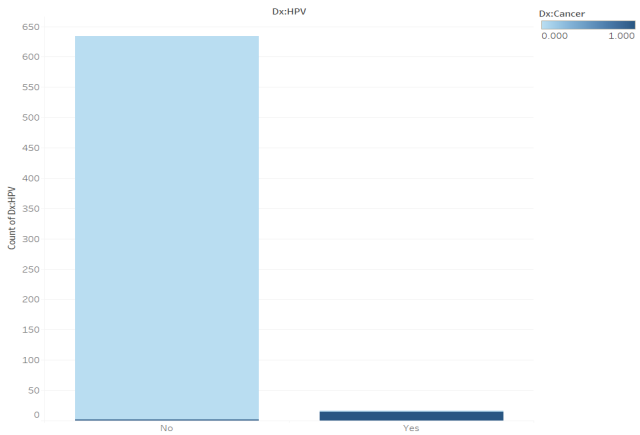

Fig. 18. Bar plot depicting the ratio cancerous and non-cancerous samples against Dx:HPV

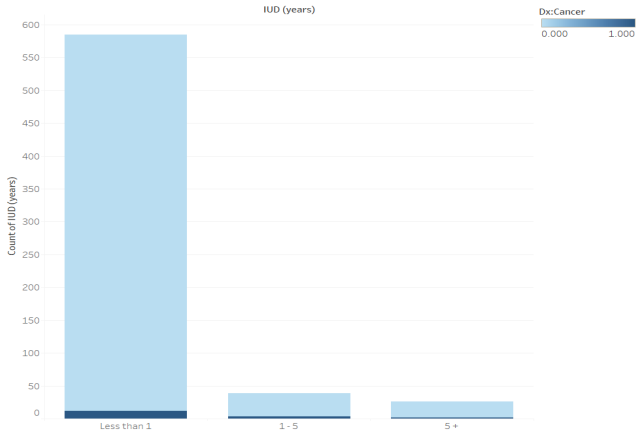

Fig. 19. Bar plot depicting the ratio cancerous and non-cancerous samples against IUD (years)

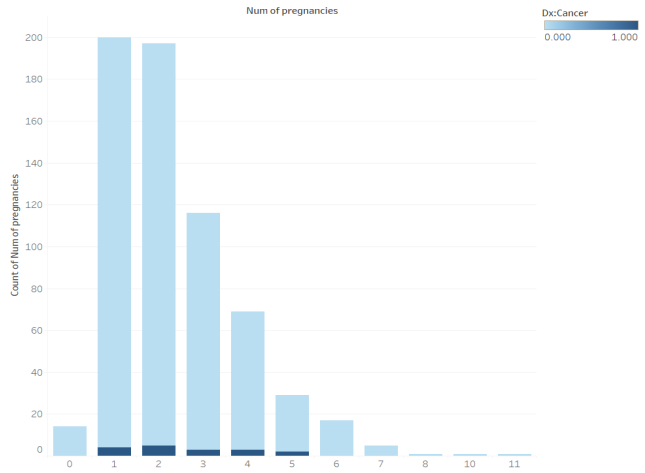

Fig. 20. Bar plot depicting the ratio cancerous and non-cancerous samples against Number of pregnancies

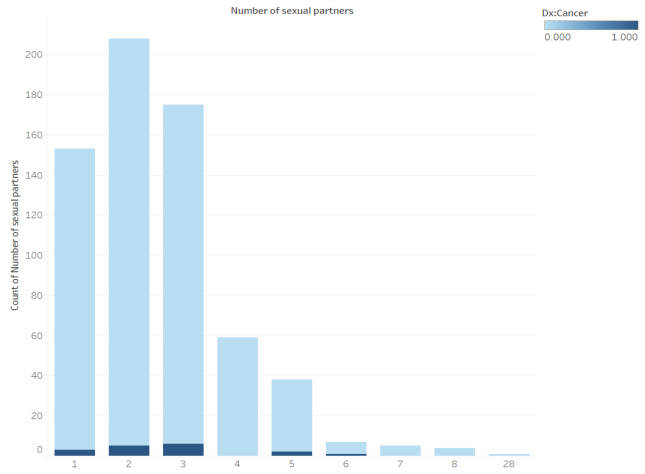

Fig. 21. Bar plot depicting the ratio cancerous and non-cancerous samples against Number of Sexual partners

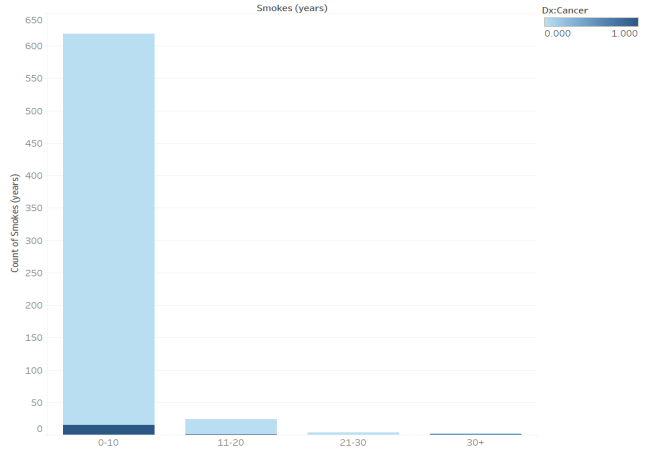

Fig. 22. Bar plot depicting the ratio cancerous and non-cancerous samples against Smokes (years)

B. Breast Cancer

After loading the dataset, we look at the dimensions. Looking at the information of the dataset to get insights into the data like its features and data types.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15203 entries, 29 to 181193
Data columns (total 13 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Menopause                             15203 non-null  float64
 1   Age group                             15203 non-null  float64
 2   density                               15203 non-null  float64
 3   bmi                                   15203 non-null  float64
 4   Age at first birth                    15203 non-null  float64
 5   No of first degree relatives with cancer 15203 non-null  float64
 6   Previous breast cancer procedure      15203 non-null  float64
 7   Result of last mammogram              15203 non-null  float64
 8   Surgical menopause                    15203 non-null  float64
 9   Current hormone therapy               15203 non-null  float64
 10  invasive                              15203 non-null  int64
 11  cancer                                15203 non-null  int64
 12  count                                 15203 non-null  float64
dtypes: float64(11), int64(2)
memory usage: 1.6 MB
```

Fig. 23. Basic information about the features used for training

Thereafter, we preprocess and clean the data. We drop columns that are not useful and rows with missing values. Statistical summary of the features can be useful in inspecting the feature distribution and anomalies if any.

The summary given below gives us a lot of valuable information about the data.

- The maximum value in the 'count' column is 1128, significantly higher than the maximum of the other columns, which could lead to poor model performance because this column has more influence than the others. To avoid differences in influence when training the model, we standardise the values of all columns.
- Since the mean is 0.043, the 'cancer' column (which will be our dependent variable) is very unbalanced. If the class was balanced, the mean would have been 0.5. For a better understanding, we will depict this with a plot. Solving this problem is extremely difficult - the best solution would be to obtain more positive data to train our model, but this is not possible in our case. Another solution could be to remove some negative cases to balance them with the positive cases, but this would result in a significant loss of information.

Before we can standardise our data, we need to know if we have columns that provide the same (or very similar) information, which could cause our model to perform poorly. This information can be obtained by creating a correlation matrix. We then plot a few columns against the cancer column individually to check the dependence of cancer on each of the features.

	Menopause	Age group	density	bmi	Age at first birth	No of first degree relatives with cancer	Previous breast cancer procedure	Result of last mammogram	Surgical menopause	Current hormone therapy	invasive	cancer	count
count	15203.0	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000	15203.000000
mean	1.0	5.753272	2.423732	2.239295	0.643557	0.421693	0.378017	0.077287	0.393738	0.472209	0.033217	0.043544	11.453726
std	0.0	1.967195	0.883353	1.264345	0.829635	0.615414	0.484908	0.267056	0.488594	0.499244	0.179209	0.204085	38.214238
min	1.0	3.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	1.0	4.000000	2.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
50%	1.0	6.000000	2.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000
75%	1.0	7.000000	3.000000	3.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	7.000000
max	1.0	10.000000	4.000000	4.000000	2.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1128.000000

Fig. 24. Statistical summary of the dataset features

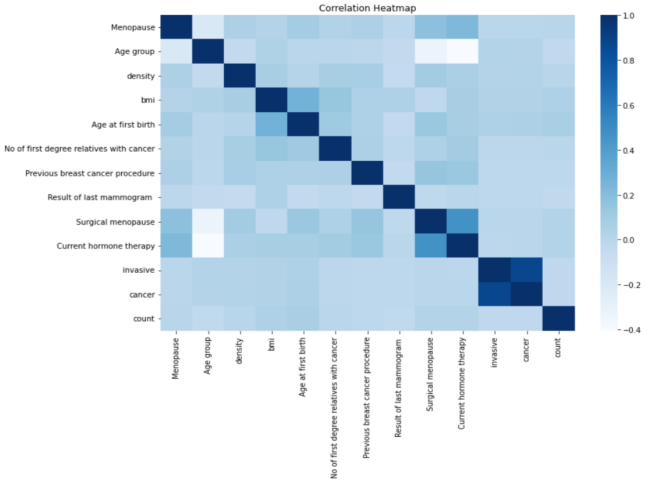

Fig. 25. Heatmap depicting the correlation between all the dataset features

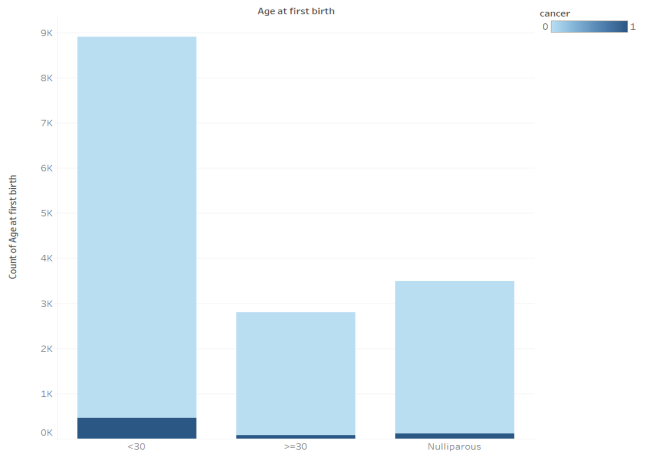

Fig. 26. Bar chart depicting the ratio cancerous and non-cancerous samples against of first birth

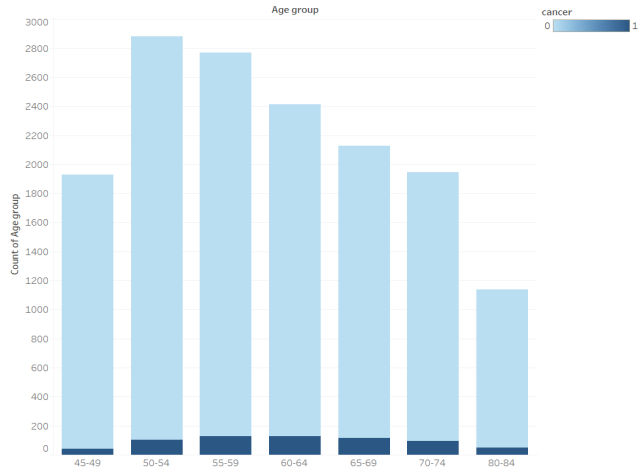

Fig. 27. Bar chart depicting the ratio cancerous and non-cancerous samples against age group

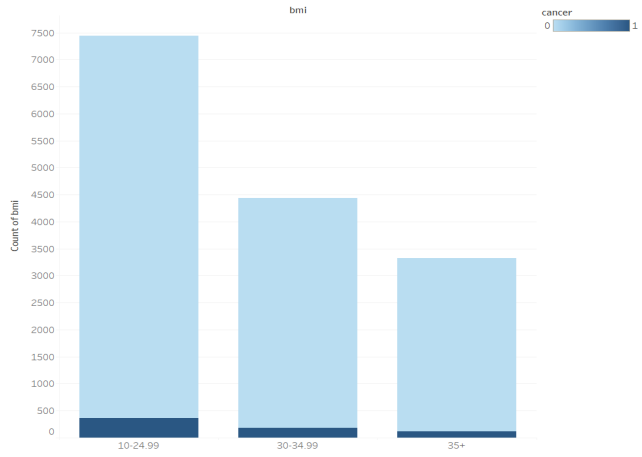

Fig. 28. Bar chart depicting the ratio cancerous and non-cancerous samples against basal metabolic index

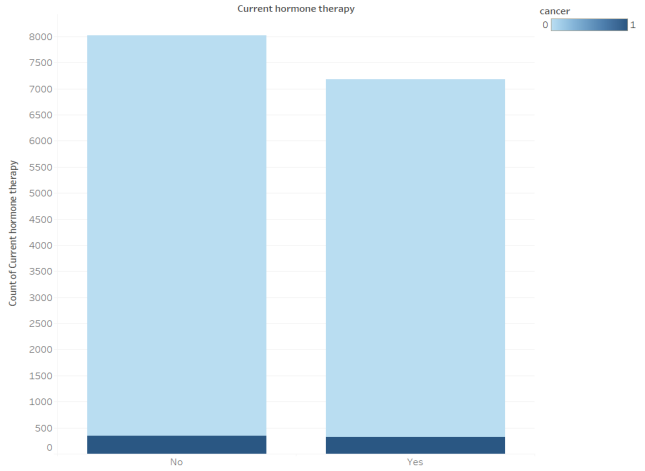

Fig. 29. Bar chart depicting the ratio cancerous and non-cancerous samples against current hormone therapy

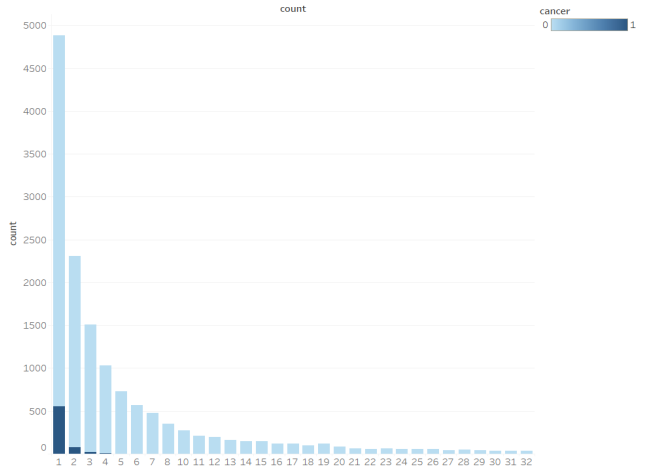

Fig. 30. Bar chart depicting the ratio cancerous and non-cancerous samples against covariate frequency count

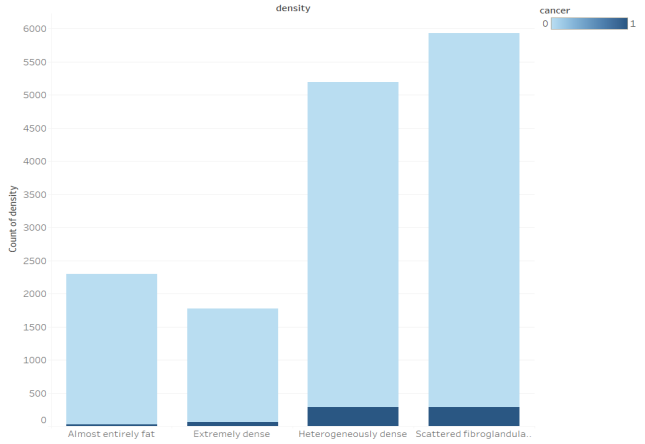

Fig. 31. Bar chart depicting the ratio cancerous and non-cancerous samples against BI-RADS breast density

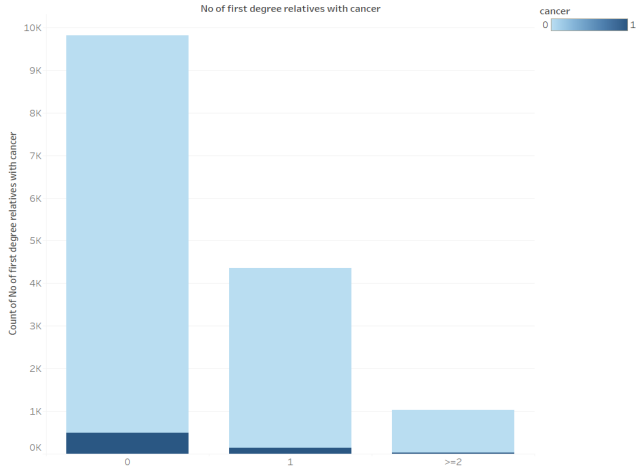

Fig. 32. Bar chart depicting the ratio cancerous and non-cancerous samples against first degree relatives with breast cancer

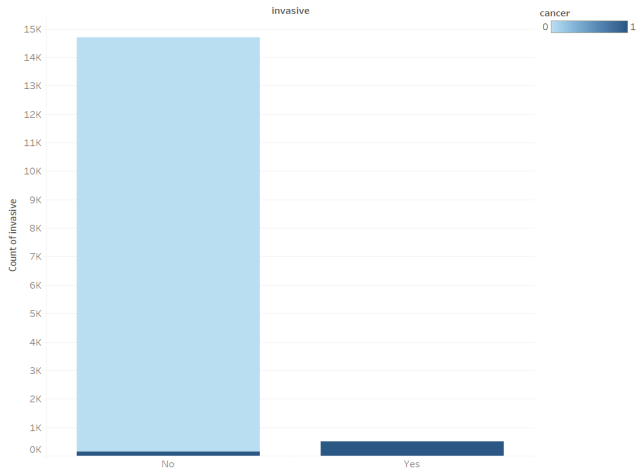

Fig. 33. Bar chart depicting the ratio cancerous and non-cancerous samples against diagnosis of invasive cancer within a year of screening

VI. Classification Model

A. Support Vector Classification

A support vector machine (SVM) is a supervised machine learning model that solves two-group classification problems using classification algorithms. They have two major advantages over newer algorithms, such as neural networks, in terms of speed and performance with a limited number of samples (in the thousands). SVMs are particularly suitable for this task since they perform well in a high-dimensional space. SVMs are versatile in terms of the types of classification issues they are suited for since they allow the user to specify the kernel function that will act as the decision function in the model. SVMs become less successful as the number of features exceeds the number of samples, notwithstanding their effectiveness and versatility. Because this was not the case for both of our classification issues (36 labels, 688 training samples in the cervical cancer data set and 13 labels, 15203

training samples in the breast cancer data set), the group began by categorising using Support Vector Classification.

B. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a simple yet effective method for fitting linear classifiers and regressors to convex loss functions such as (linear) Support Vector Machines and Logistic Regression. SGD has been successfully applied to large-scale and sparse machine learning problems. Because the data is sparse, the classifiers in this module easily scale to problems with more than 105 training examples and more than 105 features. SGD, strictly speaking, is an optimization technique that does not correspond to a specific family of machine learning models. It is simply a method for training a model. SGDs are useful for this problem because they are very efficient and easy to implement.

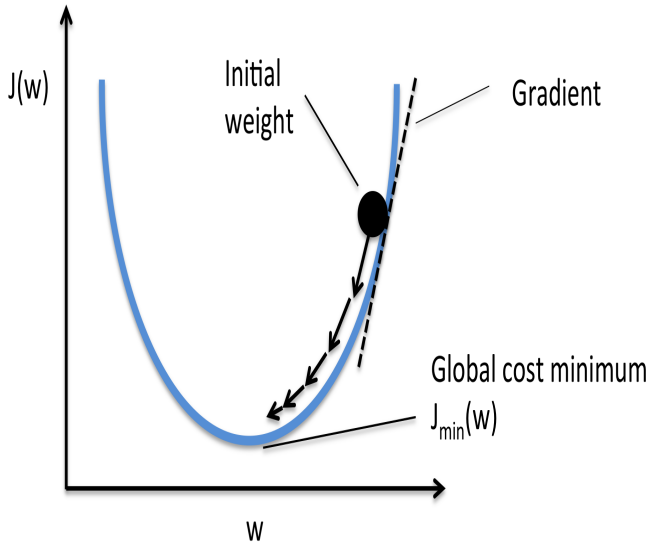

Fig. 34. Example of gradient descent along cost function

SGDClassifier facilitates multi-class classification by combining multiple binary classifiers in an OVA ("one in all") scheme. A binary classifier that distinguishes between each of the k classes is learned for each of the k classes. We compute the confidence score (i.e. the signed distances to the hyperplane) for each classifier at the testing time and select the class with the highest confidence.

C. Decision Tree Classifier

The classification technique is a systematic approach to building classification models from a set of input data. Decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayesian classifiers, for example, are different techniques for solving a classification problem. Each technique adopts a learning algorithm to identify the model that best fits the relationship between the attribute set and the class label of the input data. Therefore, the main goal of the learning algorithm is to create a predictive model that accurately predicts the class names of previously unknown records.

Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.

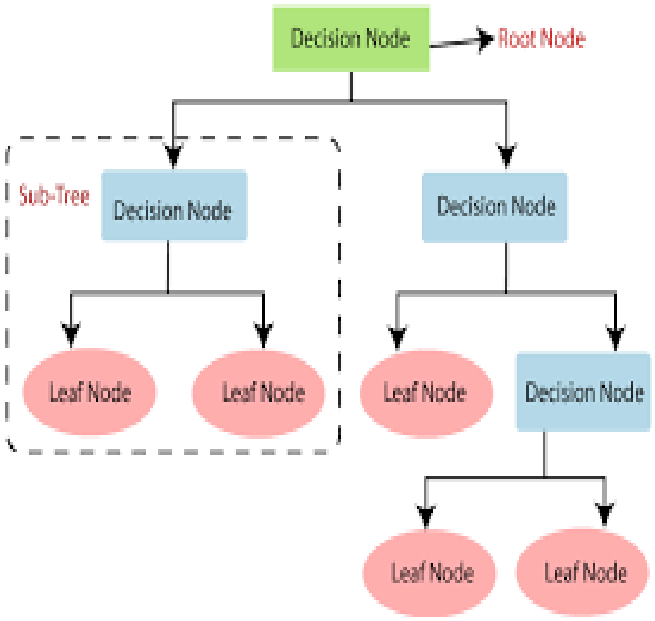

Fig. 35. Example of decision tree

D. Random Forest Classifier

Random Forest is a supervised learning algorithm. The forest built consists of multiple decision trees that are generally trained using the 'bagging' method. The general idea of the bagging method is that a combination of learning models increases the overall result. Each tree in the random forest generates a class prediction and the class with the most votes becomes the model prediction.

Random Forest adds extra randomness to the model as the trees grow. Instead of looking for the most important feature when splitting a node, the best feature is sought out of a random subset of features. This leads to a great variety, which generally leads to a better model.

VII. RESULT

We are very pleased with the final results of our classification models. All the four classification models mentioned produced the greatest train and test accuracies. The models used to train the cervical cancer dataset were stochastic gradient descent, support vector classification and decision tree classification. In the same way, the models used to train breast cancer data set were support vector classification, decision tree classification and random forest classification. In the end, we

chose the best classifier for each of the datasets, the details of which are explained further.

A. Cervical Cancer

Out of the three classification models mentioned above, the decision tree classification algorithm produced the best results. The test data (25%) accuracy for the same is 99.39 percent while the train data (75%) accuracy is 100 percent. The confusion matrix and the classification report of the results produced by the decision tree algorithm is as shown in Fig. 36.

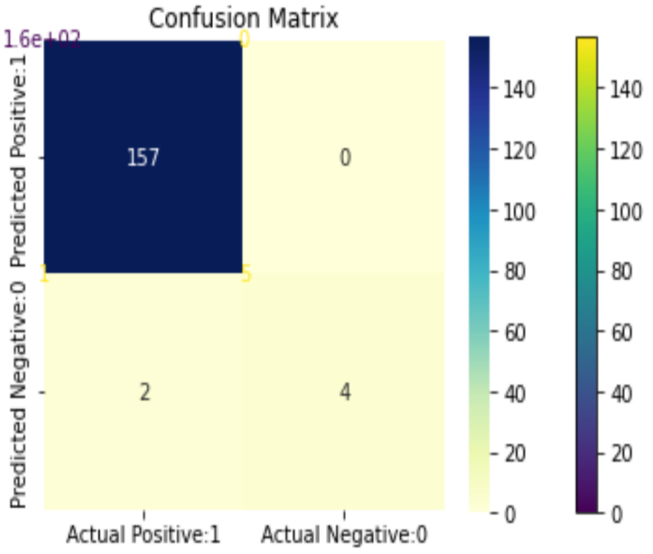

Fig. 36. Confusion matrix obtained after training the model using Decision Tree Classification Algorithm

	precision	recall	f1-score	support
0.0	0.99	1.00	1.00	157
1.0	1.00	0.83	0.91	6
accuracy			0.99	163
macro avg	1.00	0.92	0.95	163
weighted avg	0.99	0.99	0.99	163

Fig. 37. Classification Report obtained after training the model using Decision Tree Classification Algorithm

B. Breast Cancer

Out of the three classification models, the Support Vector Classification produced the best results. The test data (50%) accuracy for the same is 98.89 per cent while the train data (50%) accuracy is 99.04 per cent. The confusion matrix, classification report and feature importances obtained as shown in Fig. 38.

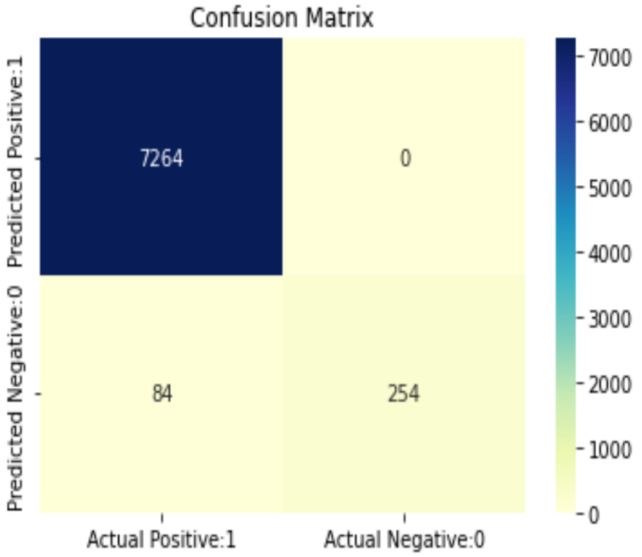

Fig. 38. Confusion matrix obtained after training the model

	precision	recall	f1-score	support
0	0.99	1.00	0.99	7264
1	1.00	0.75	0.86	338
accuracy			0.99	7602
macro avg	0.99	0.88	0.93	7602
weighted avg	0.99	0.99	0.99	7602

Fig. 39. Classification report obtained after training the model

VIII. Continued Work

The work presented in this paper devises a classification model which can be used to check one's susceptibility for Breast and Cervical cancer. The model can be an effective tool as an open-source application available to everyone. Additionally, the application would consist of other approaches supporting our attempt towards spreading awareness and highlighting its importance.

A. Nearest Hospital/Centre suggestions

The application will be accompanied by our system which provides suggestions for the nearest hospital or Cancer treatment centres based on the user's location or address. The system incorporates 2 APIs which help find the best suggestions. The first API from Position Stack (https://positionstack.com/) helps find the latitude and longitude of the user and the second API from MapMyIndia (https://www.mapmyindia.com/api/) helps find the nearest hospital or Cancer treatment centres. These suggestions will ensure that everyone also knows the correct source to contact in case of concern.

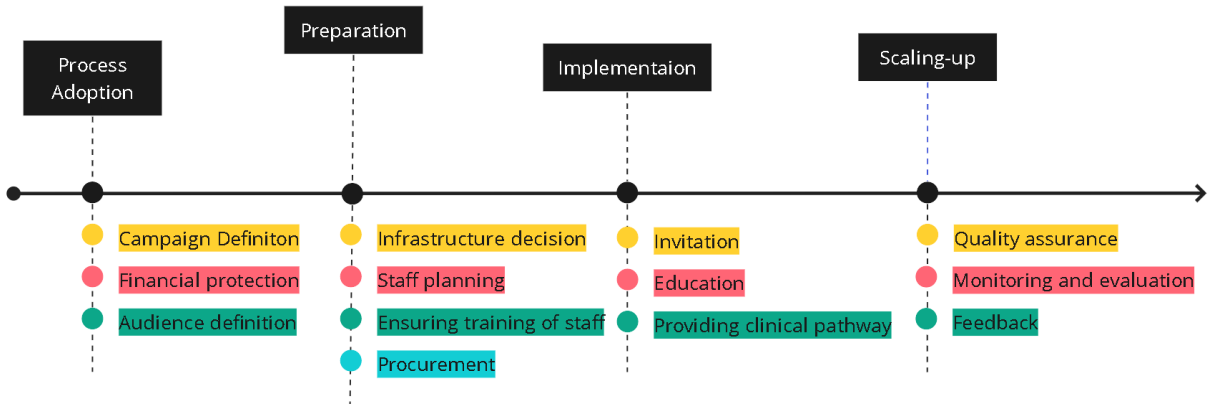

Fig. 40. Cancer awareness campaign planning

B. Awareness drives and Campaigns

Cancer awareness campaigns are crucial in cancer prevention programs. The aim of these campaigns is to create cancer awareness amongst the population of Telangana. It is important to dispel the myths that people wrongly believe, inform them about the signs and symptoms, and the importance of screening for early detection. Moreover, knowledge of cancer risk factors is a determinant element in this process. It can be implemented using the process timeline shown in Fig. 40.

C. Integration with Health Card

On September 27, 2021 Prime Minister Narendra Modi introduced the digital health id card which will be provided to all people. It will create a seamless online platform that will make all the health-related information portable and easily accessible to doctors. This can be used to integrate with our system for easy access. Everyone's health records will be maintained and used to identify the need for campaigns and drives based on locations and demographics using a k-means clustering algorithm as shown in Fig. 41.

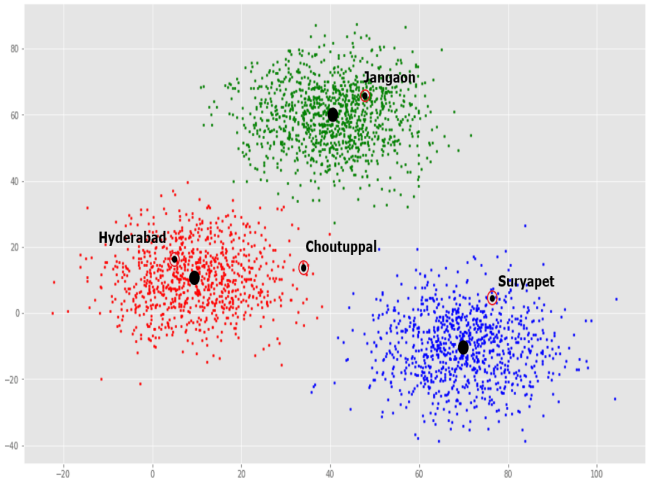

Fig. 41. Example of k-means clustering algorithm to determine the locations of Campaigns

IX. CONCLUSION

After analyzing the current statistics of cancer screening it was found that there is a critical need for Cancer Literacy in females. Awareness drives and training programs need to be frequently held for this purpose. We have devised an ML classification model to predict if a person is susceptible to breast or cervical cancer based on demographic factors. A suggestion system is then triggered which directs the user to their nearest hospital. Since this system which is integrated with susceptibility calculation and hospital suggestion is open source, people will be easily able to check their risk and get a cancer screening test if required. To take this a step further, we can integrate it with the Health Card which would make patient data more accessible (based on the privacy settings of the patients) which can help in making new policies considering real-time statistics and thereby increasing cancer literacy amongst women. Further, looking at the district-wise cancer cases and mortality, new localized schemes, awareness drives, training camps, and financial support policies can be organized which can be tailored to the level of cancer literacy and degree of urbanization in that district in Telangana.

REFERENCES

[1] NHFS dataset
[2] https://ncdirindia.org/All_Reports/State_Factsheet_21/Factsheet/FS_Telangana.pdf
[3] https://www.kaggle.com/franxi/predicting-cervical-cancer/data
[4] https://www.bcsc-research.org/data/rfdataset/dataset