Devising a solution to the problems of Cancer awareness in Telangana

This report is inspired by Sustainable Development Goal 3, regarding "Good Health and Well-being". Specifically, we focus on the lack of Cancer Literacy in women, suggest ways for its betterment, and perform data analysis and visualization.

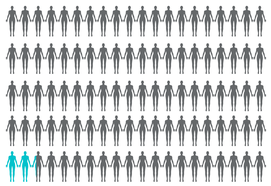
INTRODUCTION

According to a report "Profile of Cancer and Related Factors - Telangana, 2021", 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.The projected incidence of cancer cases in Telangana for the years 2020 and 2025 was calculated and found to be 25434 cases and 28708 cases for females respectively.

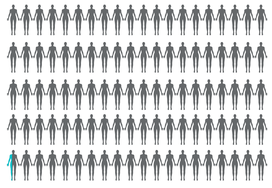
The NFHS dataset was examined for the indicators: Screening for Cancer among Women (30-49 years) for cervical cancer, breast cancer, oral cancer. This was then classified into two categories - Residence in rural and urban areas, District-wise.

Rural and Urban Areas

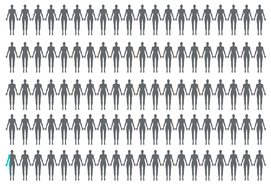
3.9% Percentage of women from Rural Telangana who have ever undergone a screening test for cervical cancer



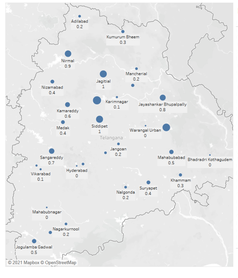
2.3% Percentage of women from Urban Telangana who have ever undergone a screening test for cervical cancer



0.4% Percentage of women from Rural Telangana who have ever undergone a breast examination for breast cancer



0.3% Percentage of women from Urban Telangana who have ever undergone a breast examination for breast cancer



District-wise in Telangana

To make a comparison based on districts in Telangana, a new combined dataset was designed appending all the latitudes and longitudes of the respective districts.

Data visualization on the indicators (ever undergone a screening test for cervical cancer (%), ever undergone a breast examination for breast cancer (%)) was done such that the results were mapped onto the district along with a tabular representation of the percentages in all the districts in decreasing order.

Thus after analyzing the data we can conclude that a minuscule percentage of women undergo screening tests despite the high-risk factor. There is an immediate need for state-level and PAN India awareness programs, involving multiple stakeholders and the health system to improve cancer literacy.

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

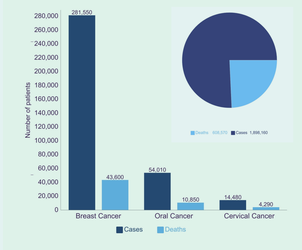
A free cancer screening and training camp was organized at the Appolo Hospital as part of the international cancer conference going on at the hospital. As part of the camp over 60 DWCRA women were trained in procedures like self-breast examination for the early detection of cancer. According to **Dr. 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 screening tests**. The District Legal Services Authority (DLSA) of Warangal, in association with Omega Bannu Hospitals, organized a free cancer screening test camp in Aug 2021. 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 for a total of 80 women.

Importance of Cancer Awareness

A dataset containing the estimated number of new cancer cases and deaths in 2021 by ASC was considered and the following patterns were observed. The dataset consisted of multiple types of cancers and after data cleaning, the necessary fields were examined.

After multiple comparisons with the data from previous years it is observed that the incidence of cancer is increasing rapidly; therefore it is important to step up cancer literacy and knowledge amongst the population.



METHODOLOGY

We developed an ML classification model to predict if a person is susceptible to breast or cervical cancer based on demographic factors. We then devised a system to provide suggestions for the nearest hospital or cancer treatment centers 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.

Model Function

The model developed in this project learns the relationship between demographic factors as well as genetic predisposition 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.

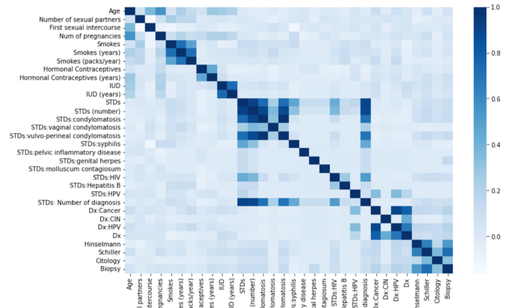
Data Preprocessing & Features

Cervical Cancer:
The data set was obtained from Kaggle, gathered at the 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset includes 858 patients' demographics, habits, and medical records from the past. This dataset is made up of 36 columns and 858 rows. This data set describes the risk factors for cervical cancer that lead to a biopsy. Missing and duplicate values, columns with exceeding null values, and no correlation were excluded from the dataset. Subsequently, a dataset of 688 patients was obtained and used for model training. The feature to be predicted is 'Dx: Cancer'.
Breast Cancer:
The data set was obtained from BCSC Research. 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.
Rows with missing values and duplicate samples in the data set were removed. After cleaning, 15203 rows were left. The data was also normalized by scaling to unit variance using StandardScaler() before training. The dependent feature to be predicted is cancer.

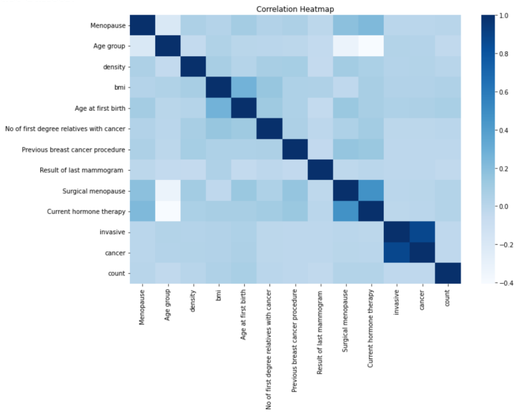
Exploratory Data Analysis

After loading the dataset we first look at the dimensions. Looking at the information of the dataset to get insights into 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. And before we can standardize 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 as shown in the 2 heatmaps below

Cervical Cancer:



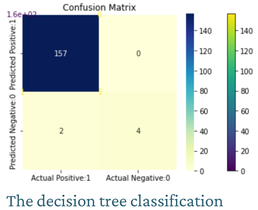
Breast Cancer:



Classification Model Results

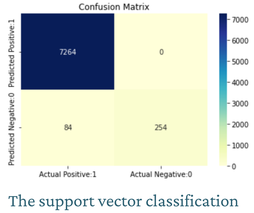
We are very pleased with the final results of our classification models. The models used to train the cervical cancer dataset were stochastic gradient descent, support vector classification, and decision tree classification. In the same way, the models used to train breast cancer data set were support vector classification, decision tree classification, and random forest classification. In the end, we chose the best classifier for each of the datasets.

Cervical Cancer



The decision tree classification algorithm produced the best results. For test data (25%) **accuracy = 99.39 %** while the train data (75%) **accuracy = 100 %**

Breast Cancer



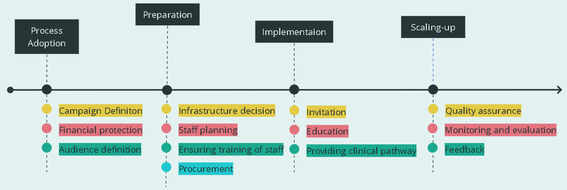
The support vector classification algorithm produced the best results. For test data (50%) **accuracy = 99.89 %** while the train data (50%) **accuracy = 99.04%**

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

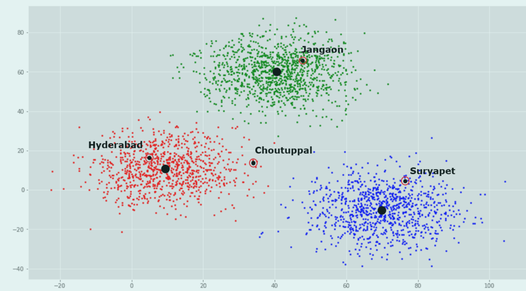
The work presented in this paper devises a classification model which can be used to check one's susceptibility for Breast and Cervical cancer. The **model can be an effective tool as an open-source application** available to everyone. Additionally, the application would consist of other approaches supporting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

1. Nearest Hospital/Centres suggestions
The application will be accompanied by our system which provides suggestions for the nearest hospital or Cancer treatment centers 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 helps find the latitude and longitude of the user and the second API from MapMyIndia helps find the nearest hospital or Cancer treatment centers. These suggestions will ensure that everyone also knows the correct source to contact in case of concern.

2. 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 below.



3. 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 figure.



CONCLUSION

Considering the district-wise cancer cases and mortality rates in Telangana with an amalgamation of the abovementioned suggestions; it could aid in creation 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. This will take us a step closer to our goal of increasing cancer literacy among women.

~ Team MetAFour