**Technical Design Document (TDD)**

**Title: RELAPSE CASE PREDICTION AND SURVIVAL ANALYSIS IN TNBC CASES**

|  |  |
| --- | --- |
| **Mentors*:*** | **Authors:** |
| *…..*  **Mentor (Client):**  *Sharat Chandra*  *Sandeep* | DEVANAND R |

**Table of Contents**

***[1.](#_heading=h.gjdgxs)******Introduction 4***

**[1.1 Purpose 4](#_heading=h.30j0zll)**

**[1.2](#_heading=h.1fob9te)****Scope 4**

**[1.3 Document Organization 4](#_heading=h.3znysh7)**

**[1.4](#_heading=h.2et92p0)****Audience 4**

***[2.](#_heading=h.tyjcwt)******Data Understanding 5***

**[2.1 Context 5](#_heading=h.3dy6vkm)**

**[2.2 Terminologies/Jargons used 5](#_heading=h.1t3h5sf)**

***[3.](#_heading=h.2s8eyo1)******Exploratory Data Analysis* 6**

**[3.1 Tools](#_heading=h.17dp8vu) 6**

**[3.2 Data Preprocessing](#_heading=h.3rdcrjn) 7**

**[3.2 Feature Selection 7](#_heading=h.26in1rg)**

***[4.](#_heading=h.lnxbz9)******Model Building 8***

**[4.1 Classification Model 8](#_heading=h.35nkun2)**

**[4.2 Models 8](#_heading=h.1ksv4uv)**

***[5.](#_heading=h.2jxsxqh)******Model Evaluation and Deployment 9***

**[5.1 Performance parameter 9](#_heading=h.z337ya)**

**[5.2 Deployment 9](#_heading=h.3j2qqm3)**

***[6.](#_heading=h.4i7ojhp)******Conclusion 10***

**Revision History**

|  |  |  |
| --- | --- | --- |
| Version | Date | Description |
| 1.0 | 21/11/2021 | Initial Draft |
| *2.0* | 07/12/2020 | Final Draft |

# Introduction

## 1.1 Purpose

The purpose of this document is to outline the high-level design of the ***Project Title*** and provide an overview for the tool implementation.

Its main purpose is to –

* Recommend relevant subscribed course module for student query

## Scope

***Project Title*** is Triple Negative breast cancer relapse prediction model based on some given data

.1.3 Document Organization

This document is organized into the following section:

|  |  |
| --- | --- |
| Introduction | Provides information related to the document |
| System Overview | Describes the approach, architectural goals and constraints, Guiding principles |
| Application Architecture | Describe the application architecture in terms of different layers of application. Description of the presentation layer, business layer, data access layer and resource layer and their relationship to each other. |
| Database Architecture | Describes the overall Data model and entity relationship diagram |
| Assumptions and Constraints | Details various assumptions made during design and development of the Online Screening tool |

## Audience

The intended audiences for this document are: -

* Innodatatics Inc.
* The project development team
* Mentors
* Health Professionals

# Data Understanding

## 2.1 Context

Triple-negative breast cancer (TNBC) is considered an aggressive cancer because it grows quickly, is more likely to have spread at the time it’s found and is more likely to come back after treatment than other types of breast cancer. So in this project we focus on what are the key features that would help to predict the relapse of tnbc using various statistical tool and ML algorithms.

## 2.2 Terminologies/ Medical Jargons Used

The various medical jargons that are involved in the data are given below:

**Menaupausalstatus** – Based on the menstrual cycle of the women, they are mainly classified into 3 types - Premanopausal state, perimenopausal state and post-menopausal state.

postmenopausal if they have not had a menstrual period for >12 months due to natural causes

Perimenopause means "around menopause" and refers to the time during which your body makes the natural transition to menopause, marking the end of the reproductive years.

Premenopause is when you have no symptoms of perimenopause or menopause.

**Histology** - is the study of tissues Invasive ductal carcinoma (IDC), also known as infiltrating ductal carcinoma, is cancer that began growing in a milk duct and has invaded the fibrous or fatty tissue of the breast outside of the duct.

**Stages -**

|  |
| --- |
| stage 0 – indicates that the cancer is where it started (in situ) and hasn't spread stage I – the cancer is small and hasn't spread anywhere else stage II – the cancer has grown, but hasn't spread stage III – the cancer is larger and may have spread to the surrounding tissues and/or the lymph nodes (part of the lymphatic system)  stage IV – the cancer has spread from where it started to at least one other body organ; also known as "secondary" or "metastatic" cance |

**Grade** -

|  |
| --- |
| grade I – cancer cells that resemble normal cells and aren't growing rapidly grade II – cancer cells that don't look like normal cells and are growing faster than normal cells grade III – cancer cells that look abnormal and may grow or spread more aggressively |

**Nodal Status** - This is referred to as lymph node status. If the breast cancer has not spread to nearby lymph nodes, the status is referred to as node-negative.

**LVI -** Lymphovascular invasion (LVI) has been a predictor of worse survival outcomes in breast cancer.

Lymphovascular invasion (LVI) is defined as the presence of tumor cells within a definite endothelial-lined space (lymphatics or blood vessels) in the breast surrounding invasive carcinoma. Lymphovascular invasion (LVI) has been a predictor of worse survival outcomes in breast cancer. .4

url:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6078671/

**Margins -** The edge or border of the tissue removed in cancer surgery. The margin is described as negative or clean when the pathologist finds no cancer cells at the edge of the tissue, suggesting that all of the cancer has been removed.

**AR –** The androgen receptor (AR) is discussed as a prognostic and/or predictive marker in breast cancer patients.

**RT -**Radiation therapy may be used to treat breast cancer at almost every stage. Radiation therapy is an effective way to reduce your risk of breast cancer recurring after surgery. In addition, it is commonly used to ease the symptoms caused by cancer that has spread to other parts of the body (metastatic breast cancer)

# Exploratory Data Analysis

## 3.1 Tools

Numerous plots are generated using various libraries for EDA

Categorical distribution of each variable were identified using various categorical plots and histograms.

Some of the major insights extracted from the EDA are given in the PPT

**Survival Analytics**

**The plots are included in the PPT**

In order to perform the survival analytics, Kaplan Meier plot was made and was found that the probability of survival for TNBC cancer patients is less than 0.6 from the data given.

## 3.2 Data Preprocessing

Data cleaning was necessary for various features in the datasets.

In Tumor size column some of the cells were having unwanted characters, spaces and letters which were removed and was converted to int datatype.Then area is converted to length by finding the diagonal length.This was done after doing some research in the web.

NA values in the tumor size is imputed with 0.

NA in outcome time is imputed with median imputation

Nil value in chemo given column were converted to no chemo given and was treated as another category.

The categorical variables are encoded by one hot encoding

The stage of cancer feature is label encoded

## 3.3 Feature Selection

From the dataset given, not all the features can be used for making the predictive model as some variable like relapse time highly influence the output variable. So, by performing various data pre-processing the following features were selected for training the predictive model:

* Age
* HPE
* Stage
* Surgery
* Tumor\_Size
* Surgerylevel
* Chemo given initially
* Relapse

The dataset was split into two for training and testing in the ratio 7:3.

.The data was very unbalanced.So SMOT technique is used to resample the data.

# Model Building

## 4.1 Classification Model

Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under. Here in this project our objective is to classify whether the patient would relapse or not.

## Models

Various Classification models used in this projects are:

**Naïve Bayes Model –**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes algorithm based on Bayes’ theorem with the assumption of independence between every pair of features. Naive Bayes classifiers are extremely fast compared to more sophisticated methods. In this study Multinomial bayes model was selected for the model building.

**KNN -**

K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint**.**Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. The hyperparameter for this ML algorithm is taken by plotting elbow plot.

**Xtreme gradient Boosting -**

Extreme Gradient Boosting is a tree-based algorithm, which sits under the supervised branch of Machine Learning. While it can be used for both classification and regression problems, all of the formulas and examples in this story refer to the algorithm’s use for classification**.**Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. The hyper parameters was found by using grid search CV

**Logistic regression -**

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. ... The hypothesis of logistic regression tends it to limit the cost function between 0 and 1 **.**Very effective model for classification problem.The accuracy was not upto the mark while using regression model for this problem

**SVM -**

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.Support vector machines are also used for classification.SVM with kernel linear is used for model building

# Model Evaluation and Deployment

## 4.1 Performance parameters

The performance of various model used for classification are compared with respect to various performance parameters. These performance parameters help choose the best model based on our objective for the model prediction. These performance parameters of each model can be printed by the function classification\_report from sklearn library. Various performance parameters include:

**Accuracy** - It may be defined as the number of correct predictions made as a ratio of all predictions made.

**Precision** - It may be defined as the number of correct documents returned by our ML model. A better precision would also mean that the false positive cases in the model are minimized.

**Recall** - Recall may be defined as the number of positives returned by our ML model.

**F1 score** - This score will give us the harmonic mean of precision and recall. Mathematically, F1 score is the weighted average of the precision and recall.

In this case, the precision is preferred to have more score than accuracy or other parameters because the cost of a False positive prediction will be very large as compared with the other business problem.

The performance evaluation of the models are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **Logistic Regression** | **XG Boost**  **(Best Model)** | **KNN** | **SVM** | **Naive bayes** |
| **Accuracy** | 64 | 74 | 72 | 72 | 65 |

XGB and KNN were giving good results.XGB is selected for deployment

## 4.2 Deployment

In this file, we will use the flask web framework to deploy the app

* The final model, ie. xgbt model was saved and dumped as .pkg file format using pickle library. This model can be later loaded using the load function in pickle.
* A function was then defined input all the variable and predict the final output.
* Once the request is run, the local server starts to function. The url of the deployment model is <http://localhost:80/tool> .This can run on any web browser in the system

# Conclusion

The dataset after performing the EDA and preprocessing several insights were drawn. These insights can be used by the customer of this model to implement necessary action plan to reduce the relapse cases.

The model was not giving satisfactory result but with the limited data access , the model is performing well