KLE Technological University

Huballi

A Course Project Report on

"Equity in Healthcare: Tackling the differences between socio-economic aspects and health equity"

A Course Project Report Submitted in Partial Fulfillment of the Requirement for the Course of

Exploratory Data Analysis

in

4th Semester of Computer Science and Engineering

by

Soham Mali	02FE22BCI025
Manish Sabnis	02FE22BCI026
Snehal Gujjar	02FE22BCI046
Shreya Rokade	02FE22BCI042

Under the guidance of

Dr. Prema T. Akkasaligar

Professor,

Department of Computer Science and Engineering,

KLE Technological University's Dr. MSSCET, Belagavi.

KLE Technological University's

Dr. M. S. Sheshgiri College of Engineering and Technology,

Belagavi – 590 008.

June 2024

DECLARATION

We hereby declare that the matter embodied in this report entitled "**Equity in Healthcare: Tackling the differences between socio-economic aspects and health equity** " submitted to KLE Technological University for the course completion of Exploratory Data Analysis (22ECAC210) in the 4th Semester of Computer Science and Engineering (Artificial Intelligence) is the result of the work done by us in the Department of Computer Science and Engineering, KLE Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Dr. Prema T. Akkasaligar, Professor, Department of Computer Science and Engineering. We further declare that to the best of our knowledge and belief, the work reported here doesn't form part of any other project based on which a course or award was conferred on an earlier occasion on this by any other student(s). Also, the results of the work are not submitted for the award of any course, degree, or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590 008
Date : 19th June 2024

Soham Mali
(02FE22BCI025)

Manish Sabnis
(02FE22BCI026)

Snehal Gujjar
(02FE22BCI046)

Shreya Rokade
(02FE22BCI042)

CERTIFICATE

This is to certify that the project entitled "Equity in Healthcare: Tackling the differences between socio-economic aspects and health equity" submitted to KLE Technological University's Dr. MSSCET, Belagavi for the partial fulfillment of the requirement for the course - Exploratory Data Analysis (22ECAC210) by Soham Mali, Manish Sabnis, Snehal Gujjar, and Shreya Rokade, students in the Department of Computer Science and Engineering, KLE Technological University's Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590 008
Date : 19th June 2024

Dr. Prema .T Akkasaligar
(Course Teacher)

Prof. Priyanka Gavade
(Course Coordinator)

Dr. Rajashri Khanai
(Head of the Department)

Abstract

Problem Statement: Healthcare inequity is a global challenge. Addressing this challenge has an extensive impact on any patient's life. This project aims to tackle the inequity in healthcare and find the disparity in treatments and the drivers of those biases, such as demographic or societal factors.

Triple Negative Breast Cancer (TNBC) is a very aggressive form of breast cancer that accounts for nearly 10-15% of the reported breast cancers in the world. The metastasis of this TNBC is considered the most aggressive and requires urgent care and treatment. Delaying the diagnosis and hence the treatment can have devastating effects on women.

Given the socio-economical and demographic details of a patient, find out the relationship the diagnosis period has with these aspects. Primarily, build a model that can detect the relationship between the demographics of the patient and the amount of time it took them to receive a diagnosis. Secondarily, find out if the diagnosis period and treatment was impacted by climate patterns recorded from the year 2013-2018 over the patient's residential region.

Solution :

1. Data Cleaning and Preprocessing: The dataset given is vast and has a lot of attributes. Identifying the appropriate attributes that can effectively predict the target variable is necessary. To address this challenge, data cleaning and preprocessing are required which includes methods like handling missing values, normalizing variables, removing outliers, resolving inconsistencies, and ensuring that our data is in a suitable format for analysis

2. Exploratory Data Analysis (EDA): The analysis of the dataset, distribution of the attributes, identifying any underlying relationships or patterns, and finding the correlation of the variables and trends in the data will be completed.

3. Data Visualization: Employ data visualization tools to gain insights into the data. Use charts, maps, and plots to determine the relationships between variables in the dataset by visual analysis.

4. Feature Engineering: Identify important and irrelevant features that add no power to the prediction of the target variable. Find out the dependencies of the attributes on the target variable that will boost the performance of the models intended to be applied.

5. Model Selection: Test several models to find the best model to fit the processed and cleaned data. Use cross-validation methods to evaluate several models on their Root Mean Squared Error (RMSE) and hence employ the best model.

Results: The final objective is to predict the time (in days) that the patient had to wait to get a diagnosis of metastatic TNBC. The predictions will be based on several aspects that were deemed to be impacting the cancer diagnosis period.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Triple-negative Negative Breast Cancer is mainly caused by the absence of 3 common receptors estrogen receptors, Progesterone receptors, and human epidermal growth factor receptor 2 (HER2).

TNBCis also known as Basal breast Cancer.

- **Receptors**: TNBC lacks the Estrogen Receptors(ER),Progesterone Receptors(PR) and HER2 [2]
- **Aggressiveness**: It tends to grow and spread more quickly compared to other breast cancer types.

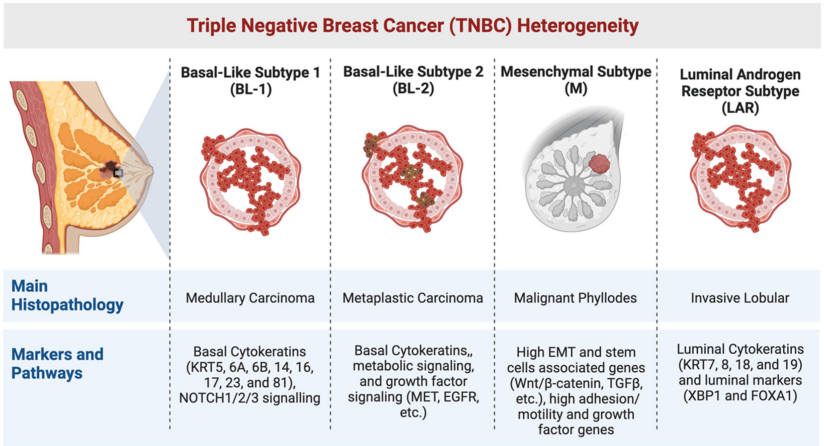

FIGURE 1.1: Triple-Negative Breast cancer

- **Prevalence**: It is observed more common in younger women,African Descent and accounts for 10–20 percent of all breast cancers.

- **Prognosis**:Here, the patient's is chance of recovery is generally less due to its aggressive nature and limited treatment opinions.

1.2 Problem Statement

Given a patient's socioeconomic and demographic details, build a predictive model that can accurately predict the time it takes for the patient to get a diagnosis of the metastasis of the TNBC. Find the relationship of the diagnosis period with the zip code-level climate patterns. Find if any biases played a role in getting the patient her quick diagnosis and timely treatment.

1.2.1 Objectives

1. Predicting metastatic Diagnosis Period
3. Derive relationship between diagnosis period and demographics of the patient
3. Model Development and Evaluation.
4. Calculating metrics - evaluate using Root Mean Square Error(RMSE)
5. Submission Compliance

Chapter 2

Knowing the Dataset

2.1 Dataset

The dataset is a part of WiDS Datathon 2 2024 [1]. Our Dataset is split into the Train Dataset and the Test Dataset.

1. Train Dataset
    - The Train dataset consists of 13173 entries with 152 attributes.
    - The data types are float64 (137 columns), int64 (4 columns), and object (11 columns).
    - The memory size is 15.3+ MB
2. Test Dataset
    - The Test dataset consists of 5646 entries with 151 attributes.
    - The data types are float64(137 columns), int64 (3 columns), and object(11 columns).
    - The memory size is 6.5+ MB

2.2 Features of the Dataset

TABLE 2.1: Features of the Dataset

Feature Name	**Data Type**	**Distinct Values**	**Missing Values**
patient_id	Numerical	13173	0
patient_race	Categorical	5	6657
payer_type	Categorical	3	1765
patient_state	Categorical	44	0
patient_zip3	Numerical	751	0
Region	Categorical	4	0
Division	Categorical	8	0
patient_age	Numerical	67	0
patient_gender	Categorical	1	0
bmi	Numerical	1304	9071
breast_cancer_diagnosis_code	Categorical	47	0
breast_cancer_diagnosis_desc	Categorical	47	0
metastatic_cancer_diagnosis_code	Categorical	43	0
metastatic_first_novel_treatment	Categorical	2	13162
metastatic_first_novel_treatment_type	Categorical	1	13162
population	Numerical	751	0
density	Numerical	747	0
age_median	Numerical	594	0
age_under_10	Numerical	450	0
age_10_to_19	Numerical	472	0

age_20s	Numerical	519	0
age_30s	Numerical	457	0
age_40s	Numerical	405	0
age_50s	Numerical	460	0
age_60s	Numerical	524	0
age_70s	Numerical	493	0
age_over_80	Numerical	368	0
male	Numerical	440	0
female	Numerical	441	0
married	Numerical	657	0
divorced	Numerical	509	0
never_married	Numerical	632	0
widowed	Numerical	439	0
family_size	Numerical	109	5
family_dual_income	Numerical	666	5
income_household_median	Numerical	750	5
income_household_under_5	Numerical	381	5
income_household_5_to_10	Numerical	377	5
income_household_10_to_15	Numerical	451	5
income_household_15_to_20	Numerical	430	5
income_household_20_to_25	Numerical	418	5
income_household_25_to_35	Numerical	501	5

income_household_35_to_50	Numerical	503	5
income_household_50_to_75	Numerical	510	5
income_household75_to_100	Numerical	487	5
income_household_100_to_150	Numerical	566	5
income_household_150_over	Numerical	624	5
income_household_six_figure	Numerical	670	5
income_individual_median	Numerical	751	0
home_ownership	Numerical	683	5
housing_units	Numerical	750	0
home_value	Numerical	750	5
rent_median	Numerical	748	5
rent_burden	Numerical	596	5
education_less_highschool	Numerical	591	0
education_highschool	Numerical	671	0
education_some_college	Numerical	617	0
education_bachelors	Numerical	630	0
education_graduate	Numerical	586	0
education_college_or_above	Numerical	670	0
education_stem_degree	Numerical	610	0
labor_force_participation	Numerical	656	0
unemployment_rate	Numerical	469	0
self_employed	Numerical	561	5

farmer	Numerical	451	5
race_white	Numerical	666	0
race_black	Numerical	565	0
race_asian	Numerical	380	0
race_native	Numerical	217	0
race_pacific	Numerical	72	0
race_other	Numerical	430	0
race_multiple	Numerical	511	0
hispanic	Numerical	621	0
disabled	Numerical	608	0
poverty	Numerical	615	5
limited_english	Numerical	384	5
commute_time	Numerical	587	0
health_uninsured	Numerical	586	0
veteran	Numerical	515	0
Average of Jan_13	Numerical	672	3
Average of Feb_13	Numerical	680	3
Average of Mar_13	Numerical	661	0
Average of Apr_13	Numerical	649	0
Average of May_13	Numerical	630	3
Average of Jun_13	Numerical	624	20
Average of Jul_13	Numerical	589	0

Average of Aug_13	Numerical	624	17
Average of Sep_13	Numerical	632	27
Average of Oct_13	Numerical	645	59
Average of Nov_13	Numerical	646	3
Average of Dec_13	Numerical	682	3
Average of Jan_14	Numerical	686	4
Average of Feb_14	Numerical	677	9
Average of Mar_14	Numerical	682	29
Average of Apr_14	Numerical	650	180
Average of May_14	Numerical	614	0
Average of Jun_14	Numerical	632	152
Average of Jul_14	Numerical	603	0
Average of Aug_14	Numerical	616	0
Average of Sep_14	Numerical	621	0
Average of Oct_14	Numerical	642	0
Average of Nov_14	Numerical	672	24
Average of Dec_14	Numerical	673	0
Average of Jan_15	Numerical	672	6
Average of Feb_15	Numerical	682	12
Average of Mar_15	Numerical	680	12
Average of Apr_15	Numerical	656	28
Average of May_15	Numerical	615	0

Average of Jun_15	Numerical	628	0
Average of Jul_15	Numerical	617	0
Average of Aug_15	Numerical	597	22
Average of Sep_15	Numerical	606	0
Average of Oct_15	Numerical	638	16
Average of Nov_15	Numerical	667	16
Average of Dec_15	Numerical	661	18
Average of Jan_16	Numerical	677	16
Average of Feb_16	Numerical	679	16
Average of Mar_16	Numerical	665	0
Average of Apr_16	Numerical	660	0
Average of May_16	Numerical	629	19
Average of Jun_16	Numerical	627	0
Average of Jul_16	Numerical	616	16
Average of Aug_16	Numerical	612	0
Average of Sep_16	Numerical	640	0
Average of Oct_16	Numerical	647	0
Average of Nov_16	Numerical	643	3
Average of Dec_16	Numerical	671	13
Average of Jan_17	Numerical	679	9
Average of Feb_17	Numerical	668	0
Average of Mar_17	Numerical	671	0

Average of Apr_17	Numerical	646	0
Average of May_17	Numerical	642	0
Average of Jun_17	Numerical	619	1
Average of Jul_17	Numerical	631	31
Average of Aug_17	Numerical	611	0
Average of Sep_17	Numerical	620	10
Average of Oct_17	Numerical	628	21
Average of Nov_17	Numerical	673	5
Average of Dec_17	Numerical	662	0
Average of Jan_18	Numerical	673	0
Average of Feb_18	Numerical	688	5
Average of Mar_18	Numerical	677	6
Average of Apr_18	Numerical	654	0
Average of May_18	Numerical	634	0
Average of Jun_18	Numerical	643	9
Average of Jul_18	Numerical	603	46
Average of Aug_18	Numerical	613	16
Average of Sep_18	Numerical	639	7
Average of Oct_18	Numerical	664	7
Average of Nov_18	Numerical	661	12
Average of Dec_18	Numerical	663	33
metastatic_diagnosis_period	Numerical	366	0

2.3 Observations

- How are the features? All categorical? Mix? The dataset has 140 numerical columns and 11 categorical columns. Hence we can conclude that the dataset is a mixed dataset
- Are there any missing values? If yes, are they large or small? Yes, there are columns with heavy missing values and there are columns with less missing values (section 2.4)
- Are there any outliers? The given dataset has outliers
- Are any of the features skewed? All the variables are either positively skewed or negatively skewed
- Does any of the features require normalization, scaling? No. None of the features require scaling or normalization.
- Overall what are the characteristics of your dataset? The dataset is diverse and has a lot of inconsistencies. Data preprocessing and cleaning are required for a data equivalent for any analysis.

2.4 Statistical Data Analysis

TABLE 2.2: Statistical Data Analysis Table

Feature Name	**Mean**	**Median**	**Standard Deviation**
patient_id	555441.784939	555769.00	259476.503094

patient_zip3	568.530859	557.00	275.758485
patient_age	59.271313	59.00	13.218883
bmi	29.168808	28.58	5.752820
population	20651.373928	18952.78	13840.379638
density	1776.872376	700.34	3876.061897
age_median	40.542676	40.64	4.031027
age_under_10	11.104988	11.00	1.511534
age_10_to_19	12.857587	12.90	1.952248
age_20s	13.297375	12.53	3.390472
age_30s	12.893557	12.40	2.410845
age_40s	12.063957	12.12	1.248652
age_50s	13.458827	13.57	1.671822
age_60s	12.631076	12.52	2.571409
age_70s	7.670396	7.33	2.138788
age_over_80	4.021445	3.82	1.257552
male	50.094310	49.98	1.694808
female	49.905878	50.02	1.694915
married	47.645569	49.43	7.528704
divorced	12.688142	12.72	2.055966
never_married	33.819417	32.01	8.126657
widowed	5.846155	5.55	1.556496
family_size	3.196401	3.16	0.222907

family_dual_income	51.800184	52.59	6.696196
income_household_median	74149.173616	69729.95	20425.924830
income_household_under_5	3.286397	2.88	1.430226
income_household_5_to_10	2.535715	2.20	1.328466
income_household_10_to_15	4.159681	3.79	1.751091
income_household_15_to_20	3.943212	3.79	1.402426
income_household_20_to_25	4.081337	4.04	1.327044
income_household_25_to_35	8.428095	8.43	2.212647
income_household_35_to_50	11.586349	11.83	2.576229
income_household_50_to_75	16.906539	17.08	2.726104
income_household_75_to_100	12.671315	12.68	1.849686
income_household_100_to_150	15.806551	15.94	3.139121
income_household_150_over	16.595873	14.65	8.893866
income_household_six_figure	32.402666	30.52	10.980561
income_individual_median	36606.228237	35211.02	8513.235050
home_ownership	65.895238	69.91	14.491534
housing_units	7589.209077	6994.41	4967.148033
home_value	337379.068122	241157.10	253061.144879
rent_median	1230.443146	1155.43	428.828282
rent_burden	31.233619	30.83	4.772054
education_less_highschool	11.933390	10.75	5.036011
education_highschool	27.687036	27.48	8.058160

education_some_college	28.874659	29.29	5.072259
education_bachelors	19.263585	18.87	6.255266
education_graduate	12.243182	10.78	5.984538
education_college_or_above	31.506428	29.79	11.801702
education_stem_degree	43.302040	42.99	4.567257
labor_force_participation	61.633658	62.78	5.977344
unemployment_rate	5.943257	5.49	1.920640
self_employed	13.186546	12.73	3.388123
farmer	1.931520	0.45	3.194201
race_white	69.975747	70.90	17.941299
race_black	11.440878	6.41	12.403575
race_asian	5.367018	2.82	6.635362
race_native	0.880295	0.43	2.318380
race_pacific	0.139133	0.05	0.529791
race_other	5.628344	3.52	6.194343
race_multiple	6.569015	5.65	3.539595
hispanic	18.144554	11.98	16.829748
disabled	13.418155	12.96	3.698844
poverty	13.417748	12.21	5.105035
limited_english	4.401387	2.75	4.782378
commute_time	27.975049	27.79	5.089195
health_uninsured	8.511348	7.36	4.169557

veteran	7.097288	6.99	3.101826
Average of Jan_13	38.959688	35.41	13.339969
Average of Feb_13	39.218355	36.71	13.701362
Average of Mar_13	44.483714	40.59	12.311727
Average of Apr_13	54.841429	53.65	9.525471
Average of May_13	64.484898	63.89	6.098297
Average of Jun_13	72.307275	71.18	6.307366
Average of Jul_13	75.560733	74.46	5.293818
Average of Aug_13	74.047466	72.51	5.951861
Average of Sep_13	69.473466	68.27	6.894506
Average of Oct_13	58.808620	57.17	8.166959
Average of Nov_13	47.098565	43.37	11.252402
Average of Dec_13	38.919349	36.49	14.383550
Average of Jan_14	34.256369	31.10	17.630565
Average of Feb_14	36.708865	34.69	17.617165
Average of Mar_14	43.766486	41.96	14.365601
Average of Apr_14	55.890057	55.35	9.154345
Average of May_14	64.994554	64.03	6.697788
Average of Jun_14	72.397563	71.41	6.068137
Average of Jul_14	74.343185	73.96	5.978399
Average of Aug_14	74.317533	73.23	5.859982
Average of Sep_14	69.491319	67.59	7.315835

Average of Oct_14	60.632386	58.05	9.001259
Average of Nov_14	45.586518	41.86	12.318052
Average of Dec_14	42.954875	39.63	11.516327
Average of Jan_15	36.974767	34.30	15.403094
Average of Feb_15	34.759469	33.39	19.205324
Average of Mar_15	47.521080	45.21	14.035280
Average of Apr_15	57.004955	55.41	9.204489
Average of May_15	65.244507	64.96	6.306477
Average of Jun_15	72.620689	71.14	6.313503
Average of Jul_15	75.560679	74.72	5.747061
Average of Aug_15	75.125510	74.45	6.168244
Average of Sep_15	72.271894	71.18	6.390686
Average of Oct_15	61.211773	57.61	9.372343
Average of Nov_15	51.595938	48.96	9.506538
Average of Dec_15	47.010274	46.32	10.443744
Average of Jan_16	37.472639	33.12	13.126782
Average of Feb_16	43.529698	39.46	13.532730
Average of Mar_16	51.631609	50.11	9.779115
Average of Apr_16	56.419806	55.78	9.225037
Average of May_16	63.375224	61.86	6.424771
Average of Jun_16	73.644230	72.58	6.170639
Average of Jul_16	77.270195	76.48	5.909783

Average of Aug_16	76.562339	76.37	5.246549
Average of Sep_16	71.436420	70.89	6.236389
Average of Oct_16	61.991704	60.21	8.112799
Average of Nov_16	52.612660	49.15	9.176596
Average of Dec_16	40.281865	36.82	13.667811
Average of Jan_17	40.719157	37.94	12.611809
Average of Feb_17	46.279066	44.27	11.402153
Average of Mar_17	48.901873	47.79	12.622846
Average of Apr_17	59.008388	57.60	8.165056
Average of May_17	63.444372	62.72	7.218949
Average of Jun_17	72.434105	71.21	5.969036
Average of Jul_17	76.757869	75.78	6.064217
Average of Aug_17	73.938746	72.31	6.152335
Average of Sep_17	70.144358	69.37	5.909376
Average of Oct_17	61.822501	60.65	8.173483
Average of Nov_17	49.864691	46.50	12.007462
Average of Dec_17	40.152299	35.90	14.337058
Average of Jan_18	36.800997	33.93	14.943633
Average of Feb_18	43.061795	42.02	14.423773
Average of Mar_18	46.074308	43.24	11.974793
Average of Apr_18	52.411982	50.29	11.069900
Average of May_18	67.932236	66.12	6.301740

Average of Jun_18	73.074635	71.64	6.581630
Average of Jul_18	77.120929	76.65	5.335091
Average of Aug_18	76.281156	76.08	5.282156
Average of Sep_18	71.417027	70.88	6.502875
Average of Oct_18	59.323286	57.45	9.261934
Average of Nov_18	46.100150	42.43	12.245594
Average of Dec_18	42.096213	38.50	11.209011
metastatic_diagnosis_period	96.515221	44.00	108.969873

Chapter 3

Implement Framework

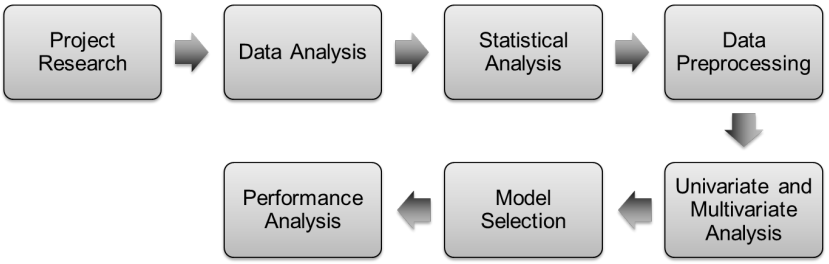

FIGURE 3.1: Implementation Framework Flow

1. Project Research
    - Gather Domain Knowledge and Relevant Information.
    - Review Literature and Existing Solutions.
    - Define the Project's Deliverables.
2. Data Analysis
    - Understand the Dataset to gain Insights.
    - Perform initial Exploratory Data Analysis(EDA)
    - Summarize Data with Descriptive Statistics.
3. Statistical Analysis
    - Apply Statistical Methods to Understand the data patterns and relationships.

- Calculation of key statistical measures(Mean,Median,Standard Deviation,Correlation)
- Conduct Hypothesis test to validate Assumptions.

4. Data Preprocessing

   - Preparation of data for Analysis and Modelling.
   - Cleaning the data by Handling Missing values,Outliers,Inconsistent Data.
   - Encoding Categorical Variables.

5. Univariate and Multivariate Analysis

   - Examining the data in depth and understanding the relationships between variables.
   - Conduct Univariate Analysis to understand individual variable distributions.
   - Conduct Multivariate analysis to explore Interactions and Correlations between multiple variables.

6. Model Selection

   - Define the criteria for model selection (i.e,Accuracy,Interpratability, interpratability, computational efficiency)
   - Experimenting different Algorithms (i.e,Regression, Decision Trees, SVM)
   - Selecting the best performing model on Evaluation Metrics.

7. Performance Analysis

   - Validate the models on Test Dataset.
   - Calculate Performance metrics.
   - Analyze residuals or errors to identify any patterns or biases.

Chapter 4

Data Pre-processing

Data Preprocessing is a crucial step that needs to be performed in Data Analysis. It involves Cleaning, Transforming, and Organizing Raw data into a suitable format.

1. Data Cleaning
    - NLP-based Replacement: Here we have used Natural Language Processing (NLP) Techniques to replace any incorrect descriptions and their equivalent cancer codes with female cancer descriptions and respective cancer codes from the ICD10 website.
2. Data Transformation
    - Imputation with Mean: Here we imputed Missing Values with Mean.
    - By imputing with mean values we notice that the BMI column exploded.
    - So, Now we try imputing values with KNN imputation and apply the same for categorical columns.
3. Feature Engineering
    - LOFO(Leave One Feature Out) calculates the importance of features and guides us in deciding which column to drop.
    - We used LOFO importance to calculate important columns and dropped the columns that had a negative importance Mean score

4. Capping Outliers

    - We capped the outliers with extreme values in the dataset that can skew the model performance.
    - By capping the outliers the extreme values will fall within a suitable range which in turn will improve the Reliability and Accuracy of our Model.

5. Skewness Comparison

    - We Compared the skewness of the columns before capping the outliers and after capping the outliers.
    - There was no significant change in the values which determines that the outliers are not heavily influencing the overall distribution of the Dataset.

6. Encoding Categorical Columns

    - There are various Encoding methods such as One-hot Encoding, Binary Encoding, Frequency Encoding, Integer Encoding, etc.
    - We opted for One-hot Encoding method for the Categorical Columns, making them suitable for Machine Learning Algorithms.

The data preprocessing pipeline includes steps for data cleaning, imputation, feature engineering, data transformation, and handling categorical data, all integral to the process.

Chapter 5

Exploratory Data Analysis

5.1 Hypothesis on the Problem Statement

1. Given that patients are mostly ordinary people, the fact that the commercial payer type is dominant in the dataset indicates that most patients prefer private insurance, indicating that most patients are predominantly ordinary people who do not enjoy any government benefits.

2. **Why so? How can you prove it?** LOFO importance model was plotted, which indicates that occupations like farmer or veteran had a negative importance mean on the target variable, which brings us to the theory that the diagnosis period did not depend on the occupation that dominated the state or region the patient is from, which explains why the commercial payer type is dominant because most farmers or veterans use Medicaid or the Medicare-advantage payer type.

3. **How can you prove that insurance played a role in determining the diagnosis period?** The LOFO model shows a positive importance mean of the 'uninsured health' attribute of the patient's region, with the target variable meaning if a patient lived in a region with more uninsured health clients, it probably affected the diagnosis period of the patient.

4. **So who are these patients?** The LOFO models show higher importance mean for the education_bachelors attribute, which means that the diagnosis period of the patient was dependent on how well the people in the patient's

region are educated. Since most Americans have a bachelor's degree, it could simply mean that areas It could be that these people must have advised the patient to opt for certain tests that may have prolonged the diagnosis period, or maybe otherwise.

5. **Does that mean that the patients were uneducated or perhaps not from the USA?** Patient education was not specified in the dataset, but it could be that the patients were not originally from the USA. They could be migrants, as the LOFO importance model shows that the race of the patient influences the target variable as well.

6. **What else indicates that the patients were not from USA?** The dependence of the diagnosis period on the attribute 'limited English', as shown by the LOFO model, means that the patients may not have been from the USA.

7. **Could there be a possibility that the dataset is biased?** There seems to be no possibility that the dataset is biased. The OLS Model shows that races like Black or Hispanic occurred frequently in the dataset, meaning that patients identified as Black or Hispanic, which could follow our Hypothesis Number 4

8. **How was the dependence of the diagnosis period on the patients' health?** As shown by both, the LOFO model, the scatter plot, and the OLS model, the patient's BMI did not play any role in the diagnosis period of the patient which could be true since BMI could be important in determining if a patient has cancer or not, it may not be important in predicting how long it took for the patient to get diagnosed. It is also scientifically proven that BMI is not a good measure of health as it does not assess body fat, muscle fat and bone density.

9. **So in terms of patient's health, what contributed to the diagnosis period?**

Patient's age showed a significant dominance with the target variable. This could be due to the tests that may have been excluded if not for the patient's age.

10. **Is the BMI dependent on the patient's age?**

    No. The scatter plot shows no relationship between age of the patient and her BMI.

11. **And what about the marital status?**

    Patients marital status was not specified in the dataset, but the marital statuses of the people in the patient's region played an important role in determining the diangosis period. Statuses 'married', 'never married', 'widowed' or 'divorced' playing an important role in determining how quickly the patient was diagnosed. This may be because the patient is involved in some community activities that could postpone the tests that the patient was scheduled for.

12. **If marital status played a role in the diagnosis period of the patient, does that mean the family size did as well?**

    No. The average family size of the patient's region did not determine the diagnosis period of the patient.

13. **So how did the people in the region of the patient affect the diagnosis period?**

    The average household income of the patient's region influences the diagnosis period. Like what hypothesis 1 says, most patients are middle-class, earning between $35k - $155k (per SmartAsset.com). So the regions with family incomes that lie between that range played a role in the diagnosis period of the patient as shown by the OLS Model.

14. **In terms of the cancers, who was responsible in the late or quick diagnosis period of the patients?**

Both the cancer codes and diagnosis descriptions of the patients were responsible for the diagnosis periods of the patients, as shown by the OLS, and LOFO models. This could be because of the mutations in each cancer that could have required more or fewer tests to get a diagnosis.

15. **So their cures played a role too?** No. Both, the OLS and LOFO model showed that the treatments did not play a role in the patient's diagnosis period. The treatment and the treatment types had heavy missing values and did not affect the target variable

16. **What else played a role in diagnosing a patient?** The average temperatures of the regions played a role in diagnosing the patient. As shown by the LOFO model and the objective of the dataset, average temperatures did play a role in diagnosing a patient. Maybe due to extreme heat or cold, patients postponed their tests.

5.2 Analysis

Since the dataset has too many attributes, an interpretable heatmap could not be plotted to understand the correlation of attributes. However, the following are a few graphs plotted concerning the Hypothesis mentioned above

1. Scatter Plot of Age Vs BMI

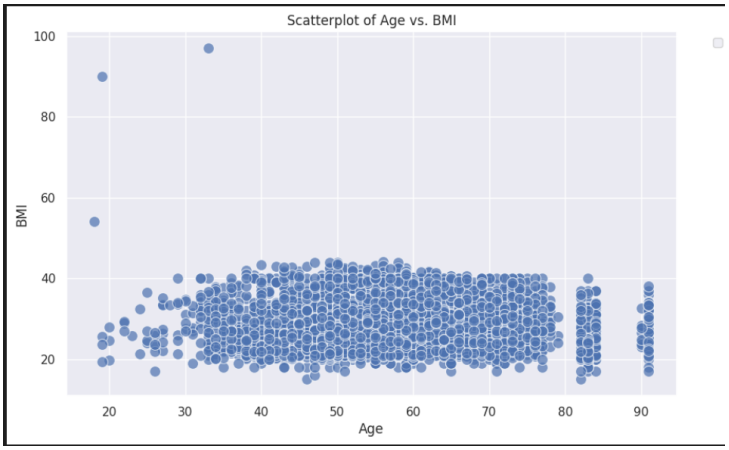

FIGURE 5.1: Age Vs BMI

2. Scatter Plot of Average Male Vs Female Numbers by state

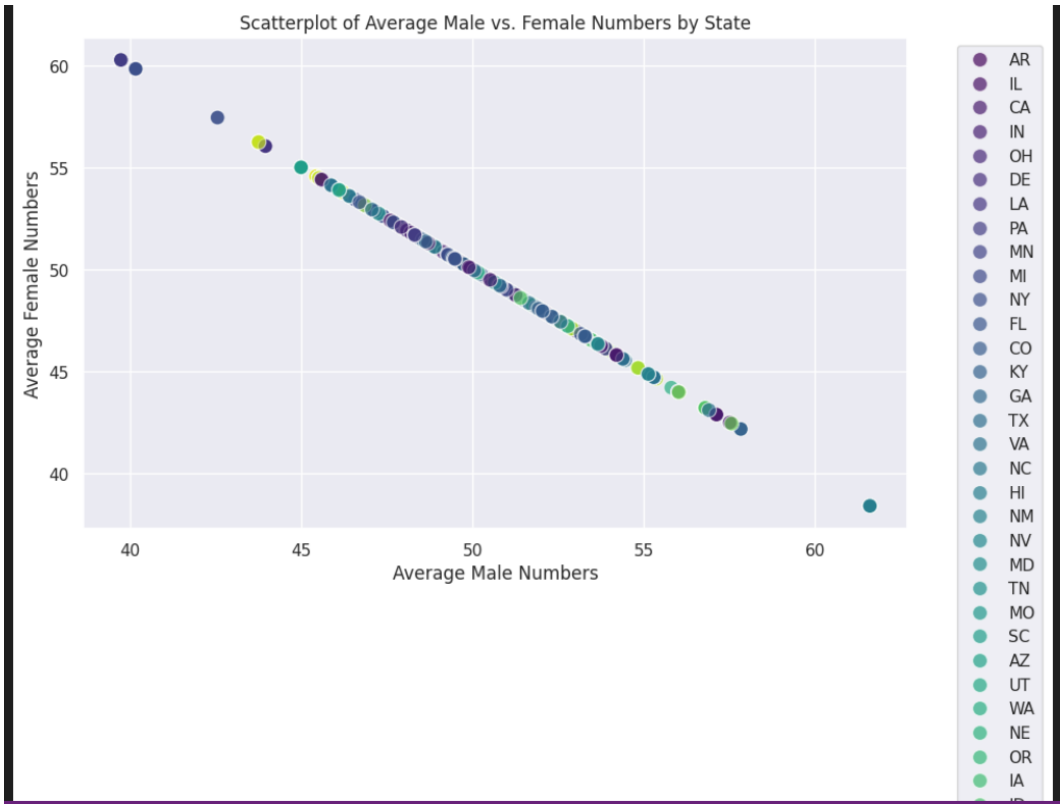

FIGURE 5.2: Average Male Vs Female Numbers by state

3. Scatter Plot between BMI Vs Target Attribute

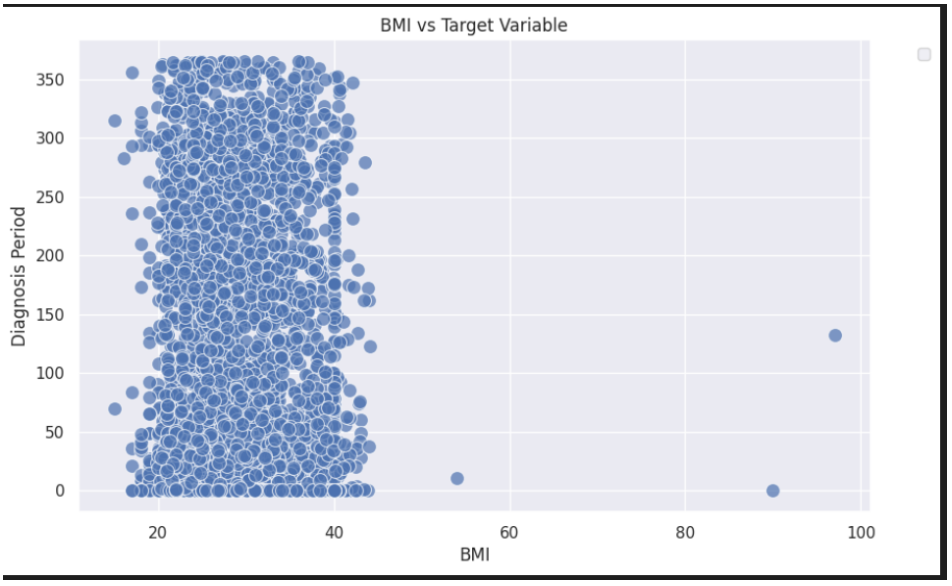

FIGURE 5.3: BMI Vs Target Attribute

Chapter 6

Results and Outcomes

Based on the provided dataset, the outcomes that can be derived from the analysis are as follows:

1. The patient's diagnosis period depended on each of the following features:
2. A relationship between the patient's socio-economic aspects and the patient's diagnosis period was found.
3. Considering that most patients are regular people.Since commercial payers make up the majority of the dataset, most patients are likely regular people without access to government benefits, which suggests that most patients prefer private insurance. .
4. A farmer or a veteran, for example, had a negative impact on the diagnosis period, according to the Leave-One-Feature-Out (LOFO) importance model, which suggests that occupation did not significantly influence the diagnosis period. Given that these groups frequently utilize government benefits, this lends credence to the observation that commercial payer types predominate.
5. The 'uninsured health' attribute of the patient's region has a positive importance mean according to the LOFO model, suggesting that areas with higher rates of uninsured people may also have longer patient diagnosis periods, thereby connecting insurance status and diagnosis period.
6. The LOFO model indicates that regions with higher levels of education have an impact on the diagnosis period, as indicated by a higher importance

mean for the 'education_bachelors' attribute. This could be the result of knowledgeable people telling patients to get certain tests done or not, which affects how quickly a diagnosis is made.

7. Although patient education levels are not specified in the dataset, it is possible that some patients are not native to the United States given the impact of the 'race' attribute on the diagnosis period. This is corroborated by the 'limited English' attribute, which suggests that linguistic obstacles may cause delays in diagnosis.

8. Patient age shows a significant impact on the diagnosis period, potentially due to age-related test exclusions or inclusions.

9. The diagnosis period was influenced by the distribution of marital status in the patient's region, even though the patient's individual marital status was not specified. This could be because the scheduling of medical tests is impacted by community involvement.

10. The average family size of the patient's region did not affect the diagnosis period, indicating that family size is not a significant factor.

11. The duration of the diagnosis was influenced by the average household income in the patient's area. Influential regions with middle-class incomes (between $35k and 155k) lend credence to the theory that the majority of patients come from middle-class families.

12. The diagnosis period was unaffected by the types or extent of treatments received. The large percentage of missing values and the absence of influence on the target variable were clear indicators of this.

13. The diagnosis timeline was influenced by the average temperature of the various regions; high temperatures may have caused patients to postpone their tests.

Conclusions

The primary aim of this experiment was to derive a relationship between the socio-economic factors of a patient and the time it took for the patient to receive a metastatic diagnosis of the Triple Negative Breast Cancer that she is suffering from. Additionally, the secondary objective of this experiment was to find if the region's climate patterns influenced this diagnosis period.

Importance of solving this problem can result in a clearer picture of what can affect the diagnosis of this aggressive form of cancer which can lead to possible changes that need to be brought in the patients' lifestyle or surroundings that can lead the patient to get diagnosed quickly so that effective action would be taken against this metastatic TNBC.

To solve this concern, data analysis was done which included various steps like data collection; data pre-processing, with cleaning; correlation analysis; data visualization alongside feature engineering; and ultimately, model selection. By following these steps, a comprehensive understanding of the relationship between the target variable i.e., the diagnosis period, and the other aspects of the patient was deduced using which, we can perhaps improve the lifestyle condition of the patient.

Conclusively, the analysis of this dataset of breast cancer patients has revealed relationships that influenced the diagnosis period of the metastasis of this cancer. Upon improving these relationships, we can see that there may be considerable improvements in the time it takes to get a diagnosis of this metastatic cancer which may save the life of a woman suffering from one of the most common yet aggressive and deadly cancers that exist in this world. Through data analysis, we saw that

a patient must know what she is suffering from. Her social conditions and health conditions influenced the amount of time it took for her to get diagnosed with metastatic TNBC. Her education, her social conditions like her marital status or race, her regional conditions like the most common occupations of the residents in her area, her financial conditions like the insurance, and her health conditions like age, all affect the diagnosis period. If these conditions are not improved, it may be far too late to get an appropriate treatment and hence save the life of the patient. With the help of government agencies and social communities, awareness of this relationship is imperative.

Bibliography

[1] WiDS Datathon 2024. "wids datathon 2024 challenge #2".

[2] American Cancer Society. "triple-negative breast cancer".