Machine Learning Algorithms and Applications

Assessment 1 | Part D

Nathan Collins

12062131

Assessment 1:	Regression Models
Type:	Individual Assessment
Length:	Jupyter Notebook (x3) and Report 1642 words
Weight:	30%
Due:	Friday, 31 March, 23:59

Assessment Criteria:

- Quality of data exploration (visual + summary stats)
- Strength of justification for features selected and model used
- Quality of code and accuracy of results
- Appropriateness of the CRISP-DM framework usage
- Depth of discussion of ethics/privacy issues, value, benefits and recommendation for the business

Section 1: Business Understanding

What does the business need?

[1.1] Objective

The business objective was to **accurately** predict cancer mortality rates in different United States (US) counties through a 33-feature dataset sourced from these counties. By accurately predicting cancer mortality rates, counties at risk of cancer could be identified and appropriate measures were taken to reduce the incidence of cancer in those counties.

[1.2] Situation

Cancer is a major public health concern in the US, and identifying counties at high risk of cancer can help prioritise intervention efforts. The American Cancer Society describes cancer as the second leading cause of death in the US, accounting for nearly 1 in 4 deaths. In 2021, it is estimated that there will be over 1.9 million new cancer cases and over 600,000 cancer deaths in the US. Cancer mortality rates vary significantly depending on demographic and geographic factors. The National Cancer Institute emphasises increased mortality in rural locations than in urban areas, with certain ethnic backgrounds experiencing elevated rates of cancer mortality than others. By developing an understanding of such trends, the identification of areas requiring more interventional measures can reduce the incidence of cancer, preserve lives, and be applied internationally.

[1.3] Datamining Goals

The data mining goals for this project were as follows:

1. Identify important predictors of cancer mortality rates and assess their impact on the target variable.
2. Develop a regression model that can accurately predict cancer mortality rates based on various features related to US counties. This was modelled with features at first and progressively expand as more information is revealed.
3. Perform feature engineering to derive new insights.
4. Evaluate the performance of various regression models and identify the best-performing model.
5. Predict cancer mortality rates with the best-performing model for new data.

By achieving these goals, stakeholders are provided with the capacity to create informed decisions and act to reduce cancer mortality rates.

[1.4] Project Plan

1. Data Collection: Collect publicly available and contextually similar data to mitigate privacy issues. While sources such as “US Census Bureau” or the “National Cancer Institute”, a data set was already sourced.

2. Data Cleaning and Pre-processing: Data was cleaned for suitability for analysis. This step included handling missing values, encoding a categorical variable, handling NaN values, and dropping columns.

3. Feature Selection: Identify the most relevant features that can help in predicting cancer mortality rates. This step involves analysing the correlation between each feature and the target variable.

4. Model Selection and Instantiation: Appropriate models were constructed, starting with Univariate linear regression, then to Multivariate linear regression, Lasso, Ridge, KNN, and Random Forest.

5. Model Evaluation: Performance evaluated using metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE).

[1.5] Ethical Considerations

As with any data analysis project, it is crucial to consider the ethical and privacy implications of the data used. In this project, we used a public dataset that did not contain any personal information about individuals. However, it is still essential to be mindful of how the results of our analysis might be used and interpreted.

- An ethical issue that could arise from this project is the potential for the models to perpetuate existing biases and discrimination in the housing market. For example, if the dataset is biased towards certain neighbourhoods or demographics, the models could learn to discriminate against certain groups. It is essential to evaluate the models for any such biases and address them accordingly.

- Another privacy concern is the potential for the models to reveal sensitive information about individuals or groups. For example, if the models were used to predict the price of a specific property, it could reveal personal information about the property owner's financial status. It is essential to be transparent about the models' use and limitations and take steps to protect individuals' privacy.

Section 2: Data Understanding

Exploratory Data Analysis

[2.1] Collection of Initial Data

While already sourced, additional publicly-available data was sourced and examined, originating from “US Census Bureau” or the “National Cancer Institute”.

[2.2] Description of Data

The original dataset contained information on 4876 individuals, deriving from a range of counties across the US. The features included demographic and socioeconomic variables. There are 34 predictor variables and 1 target variable (TARGET_deathRate, the cancer mortality rate per 100,000 people). The target variable ranges from 293 to 66, with a mean of 178 and a standard deviation of 27.5.

There was no supplementary data provided on human behaviour or health, such as physical activity, air pollution and access to health services. Two variables in the original data frame were objects, four were integers, and the remainder were floats. Three columns contained missing values, one significant and the other two under a quarter of the total values.
Each feature consisted of a range of outliers, some realistic, while others had errors.

[2.3] Exploration

A variety of takes were implemented to inspect and analyse the dataset better. Examining the distributions of the predictor variables and target identified mostly normal distributions. Relationships between the target variable and the predictor features were also correlated, identifying several features to be moderately correlated with the target variable, such as” PctPublicCoverageAlone” and “povertyPercent.” These were visualised through scatter plots and a heatmap. The use of box plots were also deployed to discern outliers distribution of the features. The data exploration phase facilitated a stronger understanding of the dataset, a necessity prior to modelling.

[2.4] Verifying Quality

Table 1: Methods of verifying data quality.

Steps of verifying data quality	Method
Check for missing values	Imputation, or deletion
Check for duplicate entries	Removal, to avoid bias
Check for outliers	Removal, to avoid skewing
Check for consistency	Manually fixing entry errors, with caution.
Check for normalisation	Visualise.

Section 3: Data Preparation

<80% of the Project>

[3.1] Select data

Dual .csv files containing training and test datasets were concatenated prior to data cleaning, so the process was uniform and accounted for all variables. All columns were included except the object columns, "binnedInc" and "Geography" were not included in Parts A and B, however, Geography was integrated into Part C, feature engineering. “ID” was also dropped.

[3.2] Clean data

The two variables provided as objects, “binnedInc”, (Median Income Per capita), and “Geography”, (county location), needed to be converted to numerical values for the data to be considered clean. As “binnedInc” is a reduced dataset of an already existing feature, it was dropped, while geography was reduced to its corresponding State and processed with One-Hot-Encoding (qualitative nominal data), for Part C.

The NaN values were visualised with missingno, where the mostly empty column “PctSomeCol18_24” was dropped, while “PctEmployed16_Over” and “PctPrivateCoverageAlone” were imputed with the average. While unconventional and warped each distribution, their missing values were a fraction of their total, and the data was considered too valuable to discount.

While outliers were apparent in all columns, the “medianAge” outliers were deemed the only error cases, as datasets can’t have values >300. There were no data duplicates to remove.

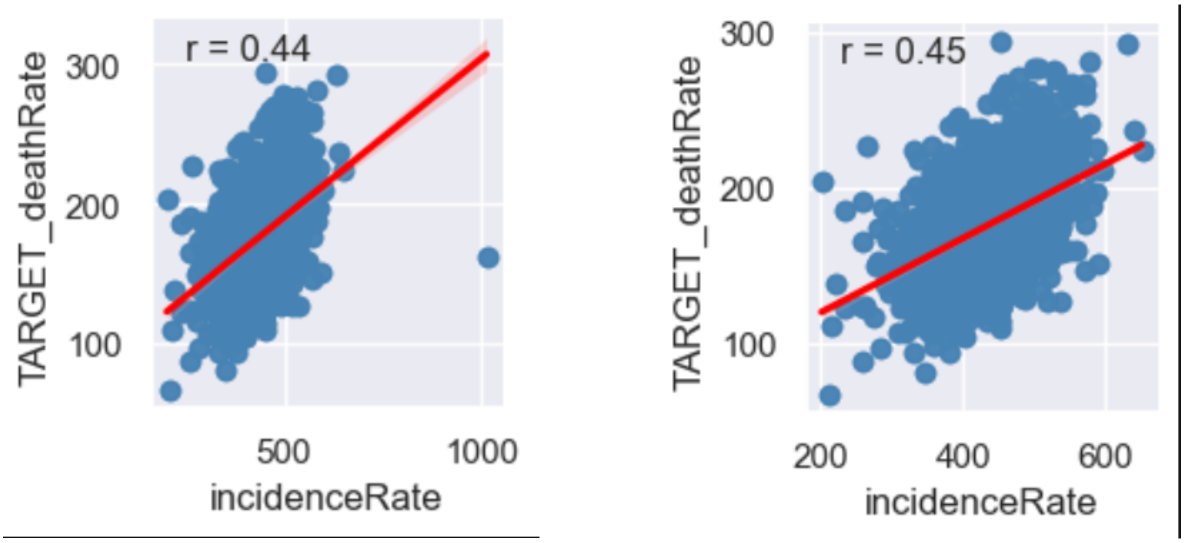


Figure 1: The impact on correlation with removing one outlier.

[3.3] **Construct** and **Format data**

Additional features were constructed as a component of Part C, feature engineering.

These included: “

- Geography”, reduced to state and converted to numerical values through One-Hot-Encoder.
- “Higher Tertiary Education”, a concatenated column of "PctBachDeg18_24" and "PctBachDeg25_Over".
- “Non-Caucasian Descent”, derived through the literature review stating certain ethnic backgrounds have increased cancer mortality rates. This was created by concatenating "PctBlack", "PctAsian" and "PctOtherRace".
- “Male-Female Median Age Gap”, subtracting the mean of the "MedianAgeMale" column from the mean "MedianAgeFemale" column.
- “Percentage Uninsured” which concatenated the "PctPrivateCoverageAlone" and "PctPublicCoverageAlone" columns, subtracted from 100.

[3.4] **Integrate data**

While data from the “US Census Bureau” or the “National Cancer Institute” was scraped and explored, it was removed from the final project on the grounds of misaligning with existing features. This data included CO2 emissions, diet and dates.

Prior to integration into models, the data was split into features and the target variable. This meant extracting the target variable from the data frame through (“.pop”).

Section 4: Modelling

<80% of the Project>

[4.1] Selecting techniques

Univariate Linear Regression

Two models were constructed: one with incidence rate as the independent variable and death rate as the dependent variable, and the other with poverty percent as the independent variable and death rate as the dependent variable.

1. Incidence Rate vs Death Rate
2. Poverty Percent vs Death Rate

Multivariate Linear Regression

One model was constructed, including all numerical features of the data frame, against the dependent variable, death rate.

Multivariate Linear Regression with feature engineering

Lasso Regression

Lasso regression was selected, as it is a type of linear regression that uses L1 regularization to perform feature selection by shrinking the coefficients of less important features to zero, resulting in a simpler and more interpretable model.

Ridge Regression

Ridge regression was selected as it is a regularization technique used in linear regression to prevent overfitting. It adds a penalty term to the cost function, which restricts the magnitude of the coefficients, resulting in a simpler and more stable model.

KNN Regression

(K-Nearest Neighbours) was selected, as it is a non-parametric machine learning algorithm that can be used for classification or regression tasks. It works by finding the k closest training data points to a given test data point and using their labels or values to make a prediction.

Random Forest Regression

As Random Forest is an ensemble learning method for classification, regression, and other tasks that constructs a multitude of decision trees and combines their predictions to improve accuracy and prevent overfitting, it was selected as a model to include. It randomly selects a subset of features for each tree and aggregates their results to obtain the final prediction.

[4.2] Generate test design

As the data was concatenated for cleaning, the data was split once more into a conventional 80% training set and 20% testing set through the Scikit-learn library. Testing data was screened as a representation of the population to avoid overfitting or underfitting, as well as cross-validated to evaluate the performance on the split data. The metrics used to evaluate the model were Mean Squared Error and Mean Absolute Error.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

[4.3] Build model

Scikit-learn library tools were utilised to construct each model, defining the X, y, X_train and y_train variables. The process of building a model involves creating an instance of the chosen model, fitting it to the training data, and then using it to make predictions on new data. The fit() method is used to train the model on the training data, while the predict() method is used to generate predictions on new data.

[4.4] Assess model

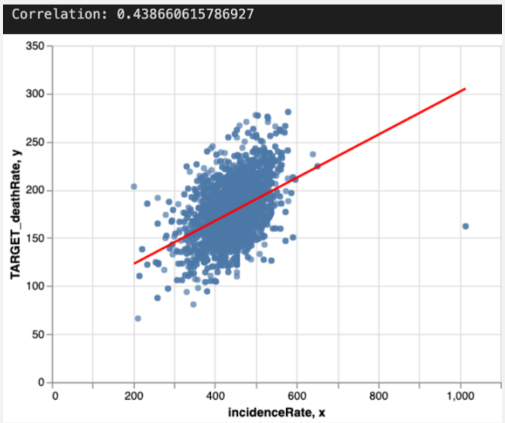
In regression tasks, common evaluation metrics include mean absolute error (MAE), mean squared error (MSE), and R-squared (R2). MAE and MSE measure the difference between the predicted and actual values, while R2 measures the proportion of variance in the dependent variable that is explained by the independent variables. For this reason, MSE and MAE were the evaluation metrics.

Section 5: Evaluation

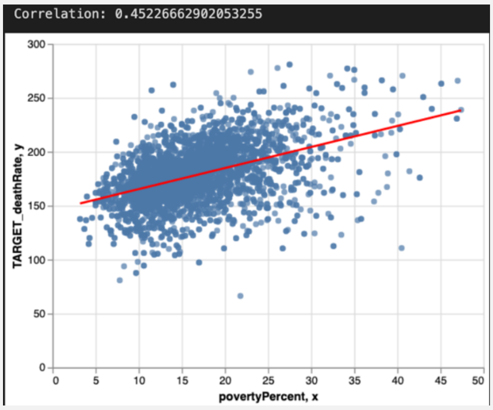
<80% of the Project>

[5.1] **Results**

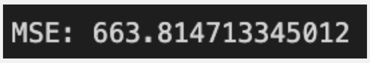
Part A: Univariate Linear Regression:



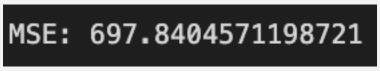
```
Correlation: 0.438660615786927
```



```
Correlation: 0.45226662902053255
```

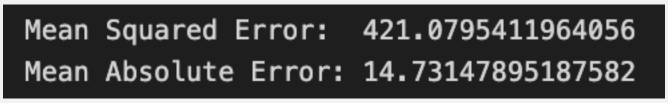


```
MSE: 663.814713345012
```



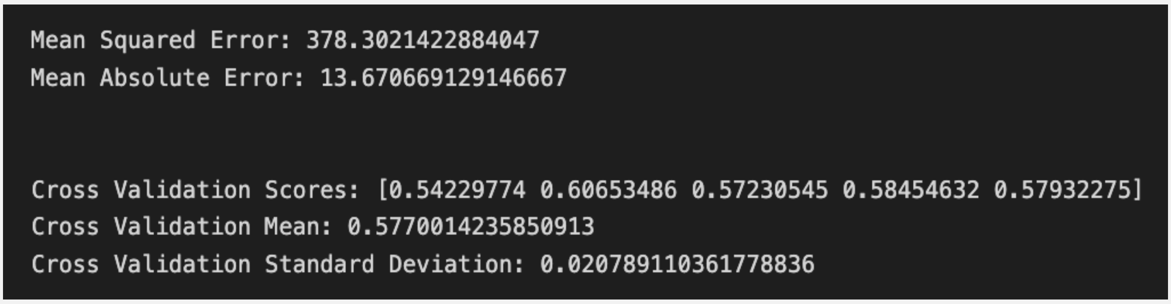
```
MSE: 697.8404571198721
```

Part B: Multivariate Linear Regression:



```
Mean Squared Error:  421.0795411964056
Mean Absolute Error: 14.73147895187582
```

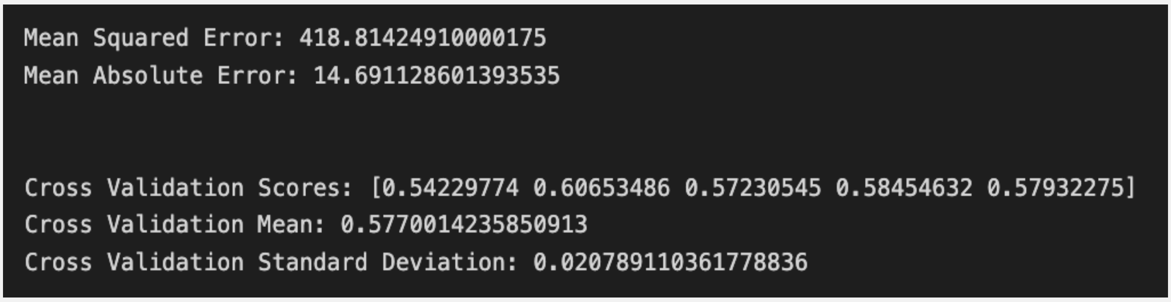
Part C: Multivariate Linear Regression with feature Engineering:



```
Mean Squared Error: 378.3021422884047
Mean Absolute Error: 13.670669129146667


Cross Validation Scores: [0.54229774 0.60653486 0.57230545 0.58454632 0.57932275]
Cross Validation Mean: 0.5770014235850913
Cross Validation Standard Deviation: 0.020789110361778836
```

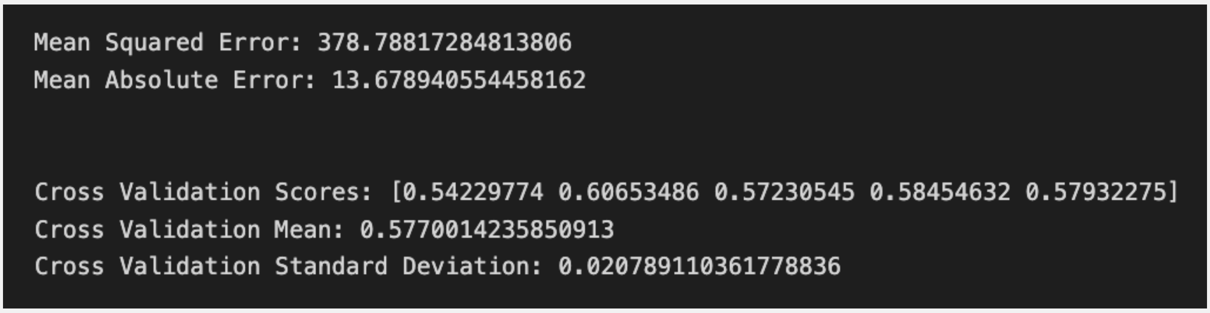
Lasso Regression:



```
Mean Squared Error: 418.81424910000175
Mean Absolute Error: 14.691128601393535


Cross Validation Scores: [0.54229774 0.60653486 0.57230545 0.58454632 0.57932275]
Cross Validation Mean: 0.5770014235850913
Cross Validation Standard Deviation: 0.020789110361778836
```

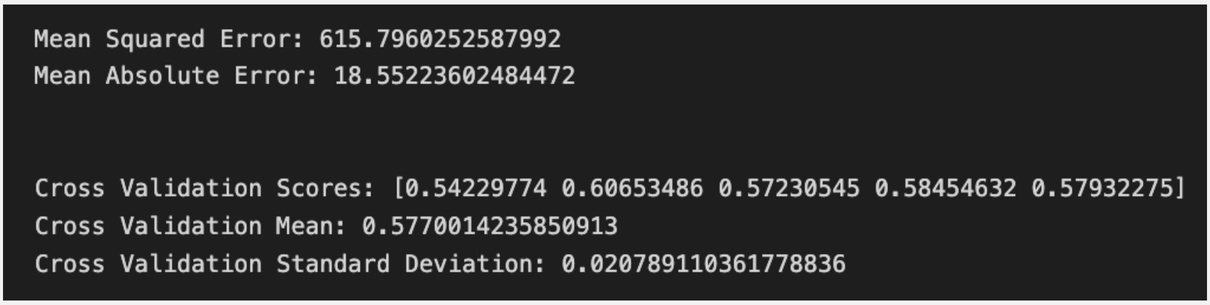
Ridge Regression



```
Mean Squared Error: 378.78817284813806
Mean Absolute Error: 13.678940554458162


Cross Validation Scores: [0.54229774 0.60653486 0.57230545 0.58454632 0.57932275]
Cross Validation Mean: 0.5770014235850913
Cross Validation Standard Deviation: 0.020789110361778836
```

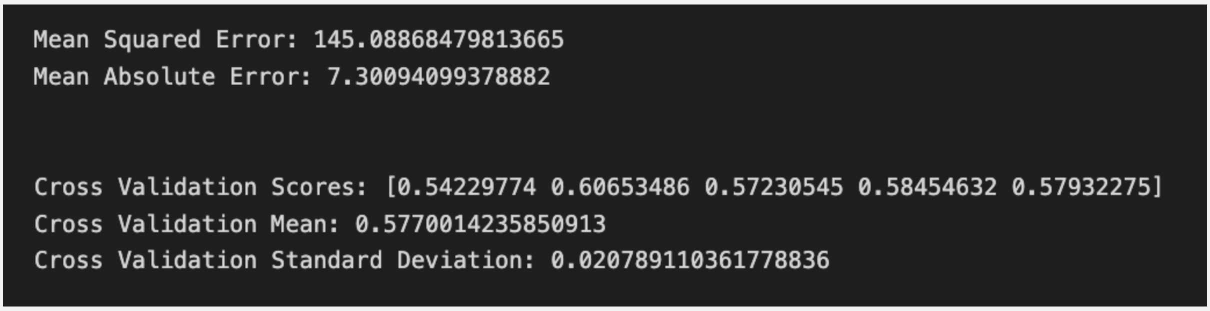
KNN Regression:



```
Mean Squared Error: 615.7960252587992
Mean Absolute Error: 18.55223602484472


Cross Validation Scores: [0.54229774 0.60653486 0.57230545 0.58454632 0.57932275]
Cross Validation Mean: 0.5770014235850913
Cross Validation Standard Deviation: 0.020789110361778836
```

Random Forest Regression:



```
Mean Squared Error: 145.08868479813665
Mean Absolute Error: 7.30094099378882


Cross Validation Scores: [0.54229774 0.60653486 0.57230545 0.58454632 0.57932275]
Cross Validation Mean: 0.5770014235850913
Cross Validation Standard Deviation: 0.020789110361778836
```

[5.2] Reflecting on the Process

Post-evaluation, it was determined that each model was incapable of accurately predicting cancer mortality. Univariate linear regression provided an insignificant foundation, with an MSE of 663 and 697 for Incidence Rate, against Death Rate and Poverty Percent against Death Rate, respectively.

Multivariate linear regression was able to improve the model by a third, to an MSE of 421, where the added feature engineering progressively enhanced and detracted from the model performance throughout testing.

Experimentation with Ridge and Lasso regression did not lead to further clear improvements, while KNN regression proved to be the poorest performing of these, with an MSE of 615. As KNN attempts to find the closest data point and does not consider underlying relationships, it was not a suitable choice. Random forest regression performed the best among all models, with an MSE of 145, representing a significant improvement over the other models.

To improve model performance, further feature engineering could be pursued, and different machine learning algorithms, such as gradient boosting, could be explored. Collecting more data and refining the existing data could also be beneficial for improving the model's performance.

[5.3] Future Iterations

While the random forest regression model showed the best performance in terms of MSE, it still fell short of meeting the business goal. The data used in the model is also not a complete representation of the population, which could lead to biased results.

Given these limitations, it would not be prudent to deploy the model without further refinement and validation. Instead, the model can be used as a tool for generating insights and guiding decision-making, but caution should be exercised when interpreting the results. Further research and data collection may be necessary to build a more robust and accurate model that can meet the business goal and be deployed in the real world.

References

1. American Cancer Society. (2022). Cancer Facts and Figures 2022. Atlanta: American Cancer Society, Inc.
2. National Cancer Institute. (2021). Cancer Statistics. Retrieved March 31, 2023, from https://www.cancer.gov/about-cancer/understanding/statistics
3. Miller, K. D., Nogueira, L., Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Alfano, C. M., ... & Jemal, A. (2019). Cancer treatment and survivorship statistics, 2019. CA: a cancer journal for clinicians, 69(5), 363-385.
4. Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. CA: a cancer journal for clinicians, 70(1), 7-30.
5. Howlader, N., Noone, A. M., Krapcho, M., Miller, D., Brest, A., Yu, M., ... & Cronin, K. A. (2021). SEER Cancer Statistics Review, 1975-2018. National Cancer Institute.
6. Henley, S. J., Ward, E. M., Scott, S., Ma, J., Anderson, R. N., Firth, A. U., ... & Jemal, A. (2020). Annual report to the nation on the status of cancer, part I: National cancer statistics. Cancer, 126(10), 2225-2249.
7. Kohler, B. A., Ward, E., McCarthy, B. J., Schymura, M. J., Ries, L. A., Eheman, C., ... & Edwards, B. K. (2015). Annual report to the nation on the status of cancer, 1975-2011, featuring incidence of breast cancer subtypes by race/ethnicity, poverty, and state. Journal of the National Cancer Institute, 107(6), djv048.
8. Institute of Medicine (US) Committee on Cancer Control in Low- and Middle-Income Countries. (2007). Cancer control opportunities in low-and middle-income countries. National Academies Press (US).
9. DeSantis, C. E., Siegel, R. L., Sauer, A. G., Miller, K. D., Fedewa, S. A., Alcaraz, K. I., ... & Jemal, A. (2016). Cancer statistics for African Americans, 2016: progress and opportunities in reducing racial disparities. CA: a cancer journal for clinicians, 66(4), 290-308.
10. Edwards, B. K., Ward, E., Kohler, B. A., Eheman, C., Zauber, A. G., Anderson, R. N., ... & Fritz, A. (2010). Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer, 116(3), 544-573.