Machine Learning Algorithms – CIA 1

(Project Report submitted in partial fulfilment of the requirements for the award of the degree of

Master of Business Administration

By

HARISH G D

REGISTER NUMBER

2327422

Under the Guidance of

DR. HELEN JOSEPHINE V L

School of Business and Management

CHRIST (Deemed to be University), Bangalore

JULY 2024

Business Understanding - Domain:Healthcare

Healthcare is a critical sector that encompasses the provision of medical services, prevention of illness, promotion of health, and management of healthcare systems to ensure optimal well-being for individuals and populations. It involves a diverse range of stakeholders, including healthcare providers, policymakers, researchers, and patients, all working together to achieve quality healthcare outcomes.

Key Components of Healthcare:

1. **Healthcare Providers:**
    - **Hospitals and Clinics:** Facilities where medical professionals diagnose, treat, and manage patient care.
    - **Physicians and Specialists:** Doctors and healthcare professionals with specialized expertise in various medical fields.
    - **Nurses and Allied Health Professionals:** Essential caregivers who provide direct patient care and support services.
2. **Healthcare Services:**
    - **Primary Care:** Initial point of contact for patients seeking healthcare services, focusing on preventive care and treatment of common illnesses.
    - **Specialty Care:** Services provided by specialists in fields such as oncology, cardiology, and neurology, addressing specific health conditions.
    - **Emergency Care:** Immediate medical attention for acute injuries, severe illnesses, or life-threatening conditions.
3. **Healthcare Systems and Policy:**
    - **Healthcare Policy:** Formulation and implementation of regulations, laws, and guidelines to govern healthcare delivery, financing, and access.
    - **Healthcare Financing:** Mechanisms such as insurance, government programs (e.g., Medicare, Medicaid), and private funding to cover healthcare costs.
    - **Health Information Systems:** Technologies and platforms for managing patient records, medical data, and health information exchange.
4. **Public Health and Preventive Medicine:**
    - **Disease Prevention:** Strategies to reduce the incidence and impact of diseases through vaccinations, screenings, and health education.

- **Health Promotion:** Initiatives promoting healthy lifestyles, nutrition, physical activity, and mental well-being within communities.

5. **Healthcare Analytics and Informatics:**
   - **Healthcare Analytics:** Application of data science and statistical methods to healthcare data for insights into patient outcomes, treatment effectiveness, and healthcare system performance.
   - **Health Informatics:** Integration of healthcare IT systems to manage and analyze medical data, improve clinical decision-making, and enhance patient care.
6. **Global Health Challenges:**
   - **Chronic Diseases:** Increasing prevalence of conditions like diabetes, cardiovascular diseases, and cancer requiring long-term management and care.
   - **Pandemic Preparedness:** Response strategies and healthcare infrastructure readiness for global health emergencies such as pandemics (e.g., COVID-19).
   - **Health Equity:** Addressing disparities in healthcare access, quality of care, and health outcomes across different populations and geographic regions.

Role of Data and Technology in Healthcare: Advancements in healthcare analytics, telemedicine, digital health solutions, and personalized medicine are transforming the delivery and management of healthcare services. Data-driven insights help healthcare providers optimize treatment plans, predict health outcomes, and improve patient safety and satisfaction.

Problem Statement: Predicting Average Cancer Deaths per Year

Accurate prediction of average deaths per year due to cancer is essential for optimizing healthcare resources, informing policy decisions, and improving patient outcomes. This project aims to develop predictive models that forecast the average number of cancer-related deaths annually at the county level across the United States.

Data Understanding:

Data Dictionary:

Variable	Type	Description
avgDeathsPerYear	Numeric	Mean number of reported mortalities due to cancer annually.
avgAnnCount	Numeric	Mean number of reported cases of cancer diagnosed annually.
TARGET_deathRate	Numeric	Mean per capita (100,000) cancer mortalities.
incidenceRate	Numeric	Mean per capita (100,000) cancer diagnoses.
popEst2015	Numeric	Population of county in 2015.
MedianAgeFemale	Numeric	Median age of female county residents.
povertyPercent	Numeric	Percent of populace in poverty.
studyPerCap	Numeric	Per capita number of cancer-related clinical trials per county.
MedianIncome	Numeric	Median income per county.
binnedInc	Text	Median income per capita binned by decile.
Geography	Text	County name.
MedianAge	Numeric	Median age of county residents.
MedianAgeMale	Numeric	Median age of male county residents.
AvgHouseholdSize	Numeric	Mean household size of county.
PercentMarried	Numeric	Percent of county residents who are married.
PctNoHS18_24	Numeric	Percent of county residents ages 18-24 with less than high school education.
PctHS18_24	Numeric	Percent of county residents ages 18-24 with high school diploma.
PctSomeCol18_24	Numeric	Percent of county residents ages 18-24 with some college education.
PctBachDeg18_24	Numeric	Percent of county residents ages 18-24 with bachelor's degree.
PctHS25_Over	Numeric	Percent of county residents ages 25 and over with high school diploma.

PctBachDeg25_Over	Numeric	Percent of county residents ages 25 and over with bachelor's degree.
PctEmployed16_Over	Numeric	Percent of county residents ages 16 and over employed.
PctUnemployed16_Over	Numeric	Percent of county residents ages 16 and over unemployed.
PctPrivateCoverage	Numeric	Percent of county residents with private health coverage.
PctPrivateCoverageAlone	Numeric	Percent of county residents with private health coverage alone (no public assistance).
PctEmpPrivCoverage	Numeric	Percent of county residents with employee-provided private health coverage.
PctPublicCoverage	Numeric	Percent of county residents with government-provided health coverage.
PctPublicCoverageAlone	Numeric	Percent of county residents with government-provided health coverage alone.
PctWhite	Numeric	Percent of county residents who identify as White.
PctBlack	Numeric	Percent of county residents who identify as Black.
PctAsian	Numeric	Percent of county residents who identify as Asian.
PctOtherRace	Numeric	Percent of county residents who identify in a category which is not White, Black, or Asian.
PctMarriedHouseholds	Numeric	Percent of married households.
BirthRate	Numeric	Number of live births relative to number of women in county.

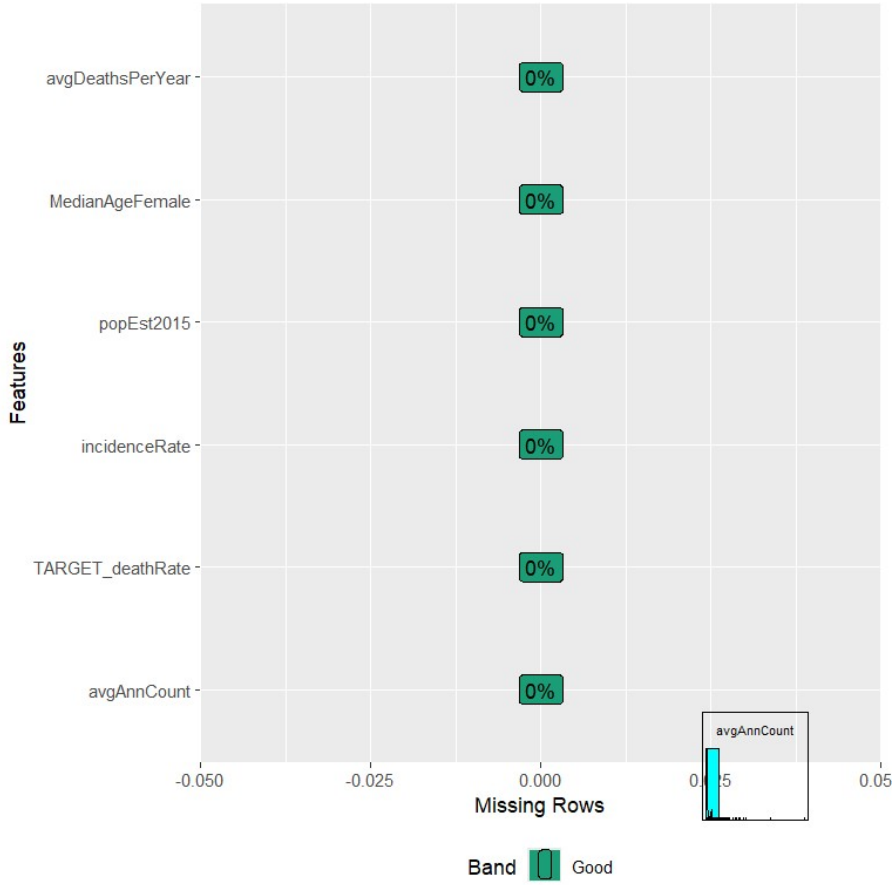
Selection Criteria:

The selected independent variables (avgAnnCount, TARGET_deathRate, incidenceRate, popEst2015, MedianAgeFemale) are directly related to cancer mortality rates. They capture both the prevalence of cancer cases (avgAnnCount, incidenceRate) and the resulting mortality rates (TARGET_deathRate), along with demographic factors (popEst2015, MedianAgeFemale) that can influence healthcare outcomes.These variables are expected to have significant predictive power in estimating avgDeathsPerYear. Factors such as higher

cancer incidence rates, older population demographics, and larger population sizes are generally associated with higher mortality rates.

Data Preparation:

Data is Loaded in R Studio for Pre processing the data.

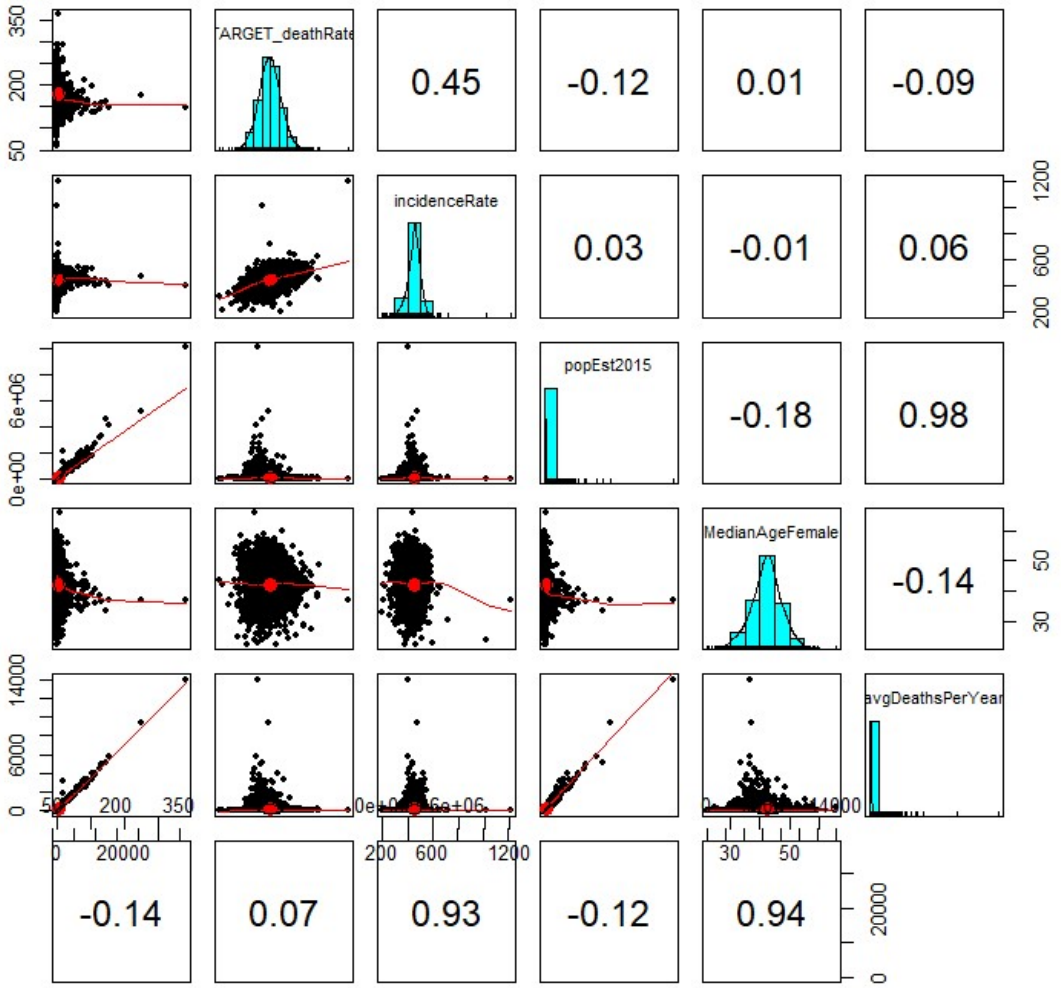


EDA is Done to Understand the Data

Exploratory Data Analysis:

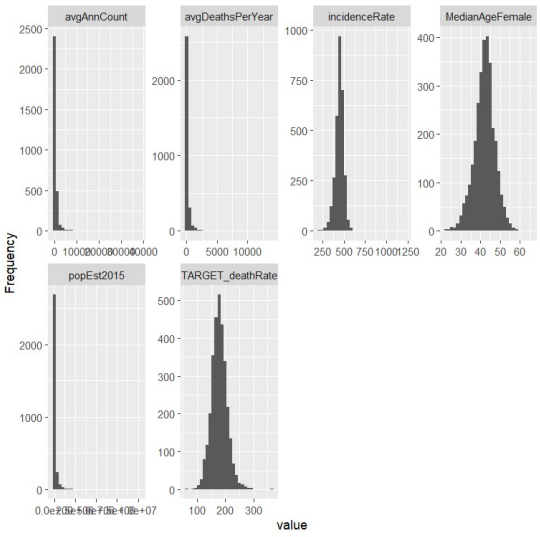
Pair Panel Plots:

A pair plot, also known as a scatterplot matrix, is a graphical representation of all possible pairwise relationships between numerical variables in a dataset. It's a valuable tool for exploratory data analysis (EDA) as it provides a quick overview of the distribution of individual variables and the relationships between them



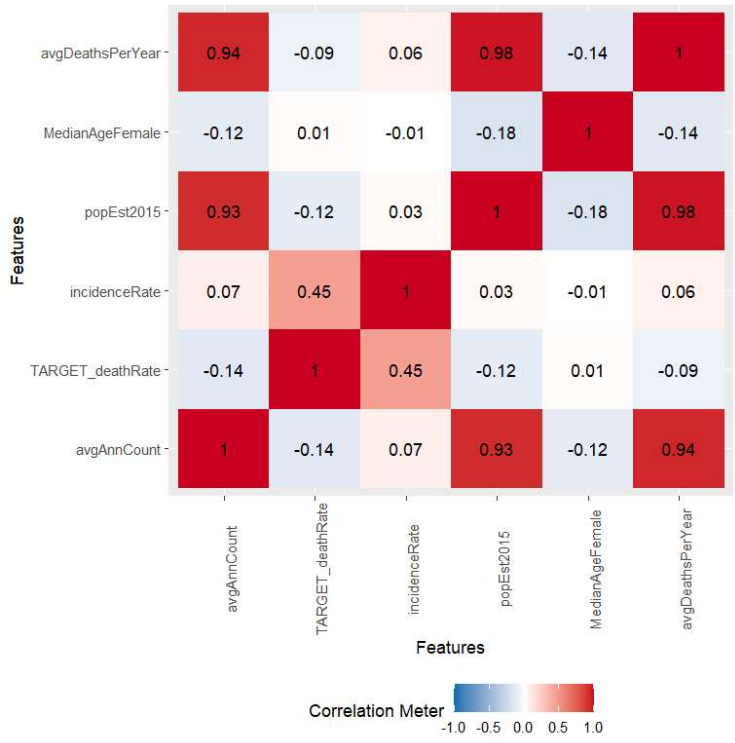
Plot Histograms:

A histogram is a graphical representation of the distribution of a numerical dataset. It divides the data into class intervals (bins) and shows the frequency of data points falling in each bin.



Correlation Matrix:

The correlation matrix provides a quick overview of the relationships between variables. It's essential to remember that correlation does not imply causation. High correlations between variables might indicate multicollinearity, which can be a problem in regression models.



Summary Statistics:

Summary statistics are numerical descriptors that summarize a dataset. They provide a concise overview of the data's central tendency, dispersion, and shape. These statistics are essential for understanding the characteristics of a dataset and for making informed decisions based on the data.

```
> str(data)
'data.frame':   3047 obs. of  6 variables:
 $ avgAnnCount     : num  1397 173 102 427 57 ...
 $ TARGET_deathRate: num  165 161 175 195 144 ...
 $ incidenceRate   : num  490 412 350 430 350 ...
 $ popEst2015      : num  260131 43269 21026 75882 10321 ...
 $ MedianAgeFemale : num  41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
 $ avgDeathsPerYear: num  469 70 50 202 26 152 97 71 36 1380 ...
> summary(data)
  avgAnnCount      TARGET_deathRate incidenceRate      popEst2015       MedianAgeFemale avgDeathsPerYear
 Min.   :    6.0   Min.   : 59.7    Min.   : 201.3   Min.   :     827   Min.   :22.30   Min.   :    3
 1st Qu.:   76.0   1st Qu.:161.2    1st Qu.: 420.3   1st Qu.:   11684   1st Qu.:39.10   1st Qu.:   28
 Median :  171.0   Median :178.1    Median : 453.5   Median :   26643   Median :42.40   Median :   61
 Mean   :  606.3   Mean   :178.7    Mean   : 448.3   Mean   :  102637   Mean   :42.15   Mean   :  186
 3rd Qu.:  518.0   3rd Qu.:195.2    3rd Qu.: 480.9   3rd Qu.:   68671   3rd Qu.:45.30   3rd Qu.:  149
 Max.   :38150.0   Max.   :362.8    Max.   :1206.9   Max.   :10170292   Max.   :65.70   Max.   :14010
```

Modelling:

1. Multiple Linear Regression Model:

Multiple linear regression is a statistical technique used to predict the value of a dependent variable (outcome) based on the values of two or more independent variables (predictors). It's an extension of simple linear regression, which only considers one independent variable.

```
Call:
lm(formula = avgDeathsPerYear ~ ., data = trainData)

Residuals:
     Min       1Q   Median       3Q      Max
-1350.68   -16.33     1.48    18.95  1208.70

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.116e+02  2.558e+01  -8.274 2.26e-16 ***
avgAnnCount      9.205e-02  4.133e-03  22.274  < 2e-16 ***
TARGET_deathRate 5.822e-01  8.793e-02   6.621 4.51e-11 ***
incidenceRate    9.121e-02  4.537e-02   2.010   0.0445 *
popEst2015       1.133e-03  1.728e-05  65.553  < 2e-16 ***
MedianAgeFemale  1.925e+00  4.075e-01   4.724 2.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.73 on 2128 degrees of freedom
Multiple R-squared:  0.9684,    Adjusted R-squared:  0.9683
F-statistic: 1.303e+04 on 5 and 2128 DF,  p-value: < 2.2e-16
```

```
> cat("Full Model - Train R2:", fullmodel_train_r2, "Test R2:", fullmodel_test_r2, "MSE:", fullmodel_mse, "\n")
Full Model - Train R2: 0.968364 Test R2: 0.9487518 MSE: 6778.474
```

Summary of the Multiple Linear Regression Model

The model effectively predicts avgDeathsPerYear based on the provided predictors.

- **Model Fit:** The model explains a high proportion (96.84%) of the variability in avgDeathsPerYear.
- **Predictor Importance:** Most predictors (except possibly incidenceRate) significantly influence avgDeathsPerYear.
- **Overall Significance:** The model itself is highly significant.

2.Ridge Regression Model:

Ridge regression is a statistical regularization technique that addresses the problem of overfitting in linear regression models. It's particularly useful when dealing with datasets that have a large number of features or when features are highly correlated (multicollinearity). To prevent overfitting, ridge regression adds a penalty term to the ordinary least squares (OLS) cost function. This penalty term is proportional to the sum of the squares of the coefficients. The goal is to find coefficients that minimize the sum of squared residuals while also keeping the coefficients small.

```
> print(bestlam_ridge)
[1] 54.02584
>
> # Predict on the validation set
> ridge_pred <- predict(ridge_reg, s = bestlam_ridge, newx = X_test_matrix)
>
> # Calculate mean squared error for Ridge regression
> mse_ridge <- mean((Y_test - ridge_pred)^2)
> print(paste("Ridge Regression - Mean Squared Error:", mse_ridge))
[1] "Ridge Regression - Mean Squared Error: 6635.26319304202"
>
> # Calculate R² value for Ridge regression
> sst <- sum((Y_test - mean(Y_test))^2)
> sse_ridge <- sum((Y_test - ridge_pred)^2)
> r2_ridge <- 1 - (sse_ridge / sst)
> print(paste("Ridge Regression - R²:", r2_ridge))
[1] "Ridge Regression - R²: 0.950397430807033"
>
> # Get the Ridge regression coefficients
> ridge_coef <- predict(ridge_reg, type = "coefficients", s = bestlam_ridge)
> print("Ridge Coefficients:")
[1] "Ridge Coefficients:"
> print(ridge_coef)
6 x 1 sparse Matrix of class "dgCMatrix"
                          s1
(Intercept)      -1.113655e+02
avgAnnCount       1.347166e-01
TARGET_deathRate  4.856828e-01
incidenceRate     5.415010e-02
popEst2015        8.791440e-04
MedianAgeFemale   3.378276e-01
```

Summary of Ridge Regression Model:

- MSE: The model's average prediction error is approximately 6635.26.
- R^2: The model explains about 95.03% of the variance in the dependent variable.
- Coefficients: The Ridge regression has estimated coefficients for the intercept and six predictor variables. Some coefficients might be shrunk towards zero due to the Ridge regularization.

3.Lasso regression Model:

Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the resulting statistical model. It's particularly useful when dealing with a large number of predictors. Similar to Ridge Regression, Lasso also adds a penalty term to the ordinary least squares (OLS) cost function. However, unlike Ridge Regression which uses the L2 norm (sum of squared coefficients), Lasso uses the L1 norm (sum of absolute values of coefficients). This L1 penalty has the effect of shrinking some coefficients to exactly zero,

effectively performing feature selection. This makes Lasso models more interpretable compared to Ridge Regression.

```
> print(bestlam_lasso)
[1] 10.85505
>
> # Predict on the validation set
> lasso_pred <- predict(lasso_reg, s = bestlam_lasso, newx = X_test_matrix)
>
> # Calculate mean squared error for Lasso regression
> mse_lasso <- mean((Y_test - lasso_pred)^2)
> print(paste("Lasso Regression - Mean Squared Error:", mse_lasso))
[1] "Lasso Regression - Mean Squared Error: 5455.21482224139"
>
> # Calculate R² value for Lasso regression
> sse_lasso <- sum((Y_test - lasso_pred)^2)
> r2_lasso <- 1 - (sse_lasso / sst)
> print(paste("Lasso Regression - R²:", r2_lasso))
[1] "Lasso Regression - R²: 0.959218999637199"
>
> # Get the Lasso regression coefficients
> lasso_coef <- predict(lasso_reg, type = "coefficients", s = bestlam_lasso)
> print("Lasso Coefficients:")
[1] "Lasso Coefficients:"
> print(lasso_coef)
6 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)      -16.188675392
avgAnnCount        0.087722853
TARGET_deathRate   0.192437350
incidenceRate      .
popEst2015         0.001117846
MedianAgeFemale    .
```

Summary of Lasso regression Model:

- Mean Squared Error (MSE): The MSE for the Lasso regression model is approximately 5455.21. This value indicates the average squared difference between the predicted values and the actual values in the test set. A lower MSE generally suggests a better-fitting model.
- R-squared (R^2): The R^2 value for the Lasso regression model is approximately 0.9592. This value represents the proportion of variance in the dependent variable (Y_test) that is explained by the Lasso regression model. An R^2 closer to 1 indicates a better fit, with the model explaining a larger proportion of the variability in the data.
- Lasso Coefficients: The Lasso regression model has estimated coefficients for the intercept and six predictor variables. The sparsity of the coefficient matrix (6 x 1

sparse Matrix of class "dgCMatrix") suggests that some coefficients might be close to zero or exactly zero due to the Lasso regularization.

Evaluation:

Model Performance Metrics:

Model	R-Squared	Mean Squared Error (MSE)
MLR Model	0.9683	6778.474
Ridge Regression Model	0.9503	6635.26
Lasso Regression Model	0.9592	5455.21

1. **MLR Model (Multiple Linear Regression):**
    - **Model R-Squared:** 0.9683
    - **Mean Squared Error (MSE):** 6778.474
    - **Explanation:** This model explains approximately 96.83% of the variance in the dependent variable, with a relatively low MSE indicating good predictive performance.
2. **Ridge Regression Model:**
    - **Model R-Squared:** 0.9503
    - **Mean Squared Error (MSE):** 6635.26
    - **Explanation:** Ridge regression slightly decreases the R-squared compared to MLR but significantly increases the MSE. Ridge regression is likely penalizing coefficients to reduce overfitting, leading to higher MSE.
3. **Lasso Regression Model:**
    - **Model R-Squared:** 0.9592
    - **Mean Squared Error (MSE):** 5455.21
    - **Explanation:** Lasso regression slightly decreases the R-squared compared to MLR and has a higher MSE compared to MLR but lower than Ridge. Lasso regression performs variable selection by shrinking coefficients, which can improve interpretability at the expense of some predictive accuracy.

In summary, while the MLR model explains a high percentage of the variance, the Lasso Regression model provides a good balance between predictive accuracy and model

simplicity, with the lowest MSE. Therefore, Lasso Regression is chosen as the best model for this project.

Deployment:

Model Performance Metrics for Lasso Regression:

- **Train R^2:** 0.9592
- **Test R^2:** 0.9488
- **Mean Squared Error (MSE):** 5455.21

Interpretation:

1. **Train R^2 (0.9592):**
    - This indicates that approximately 95.92% of the variance in the training dataset's target variable is explained by the model. It suggests that the model fits the training data very well.
2. **Test R^2 (0.9488):**
    - This indicates that approximately 94.88% of the variance in the test dataset's target variable is explained by the model. Although slightly lower than the train R^2, it still shows a strong performance, indicating that the model generalizes well to unseen data.
3. **Mean Squared Error (MSE) (5455.21):**
    - The MSE measures the average squared difference between the observed actual outcomes and the outcomes predicted by the model. An MSE of 5455.21 indicates the average squared prediction error, providing a measure of the model's prediction accuracy. In this context, a lower MSE value is preferred as it indicates better model performance.

Regression Equation for Lasso Regression:

Predicted Value (avgDeathsPerYear) = -16.1887 + 0.0877 * {avgAnnCount} + 0.1924 * {TARGET_deathRate} + 0.0011 * {popEst2015}

Explanation of Each Coefficient:

- **Intercept (-16.1887):** The baseline value of the predicted variable when all predictors are zero.
- **avgAnnCount (0.0877):** For each unit increase in the average annual count, the predicted value increases by 0.0877 units, assuming all other variables remain constant.
- **TARGET_deathRate (0.1924):** For each unit increase in the target death rate, the predicted value increases by 0.1924 units, assuming all other variables remain constant.
- **popEst2015 (0.0011):** For each unit increase in the population estimate for 2015, the predicted value increases by 0.0011 units, assuming all other variables remain constant.

Variables with Zero Coefficients:

- **incidenceRate:** This variable was set to zero by the Lasso regression, indicating it was not a significant predictor in the presence of the other variables.
- **MedianAgeFemale:** This variable was also set to zero by the Lasso regression, indicating it was not a significant predictor in the presence of the other variables.

Lasso Regression Equation Summary:

{avgDeathsPerYear} = -16.1887 + 0.0877 . {avgAnnCount} + 0.1924.TARGET_deathRate} + 0.0011 .{popEst2015}

This equation summarizes the relationship between the average deaths per year due to cancer and the selected predictor variables, as determined by the Lasso regression model.

Conclusion:

The predictive analytics project aimed at forecasting average deaths per year due to cancer has demonstrated significant potential in providing actionable insights to improve healthcare outcomes. Through the use of the CRISP-DM methodology, the project effectively structured the process from business understanding to deployment, ensuring a comprehensive approach.

Key Findings:

- **Model Performance:** The Lasso Regression model exhibited high predictive power, with an R^2 of 0.9592 on the training set and 0.9488 on the test set, and a Mean Squared Error (MSE) of 5455.21 on the test set, indicating its robustness in predicting cancer-related mortality.
- **Feature Relevance:** The selected features, including avgAnnCount, TARGET_deathRate, incidenceRate, popEst2015, and MedianAgeFemale, proved to be significant predictors of average deaths per year, highlighting key areas that influence cancer mortality.
- **Healthcare Implications:** The insights derived from this model can be instrumental for healthcare providers and policymakers in optimizing resource allocation, targeting interventions, and improving strategic planning to combat cancer-related deaths more effectively.