**36106 Machine Learning Algorithms and Applications-**

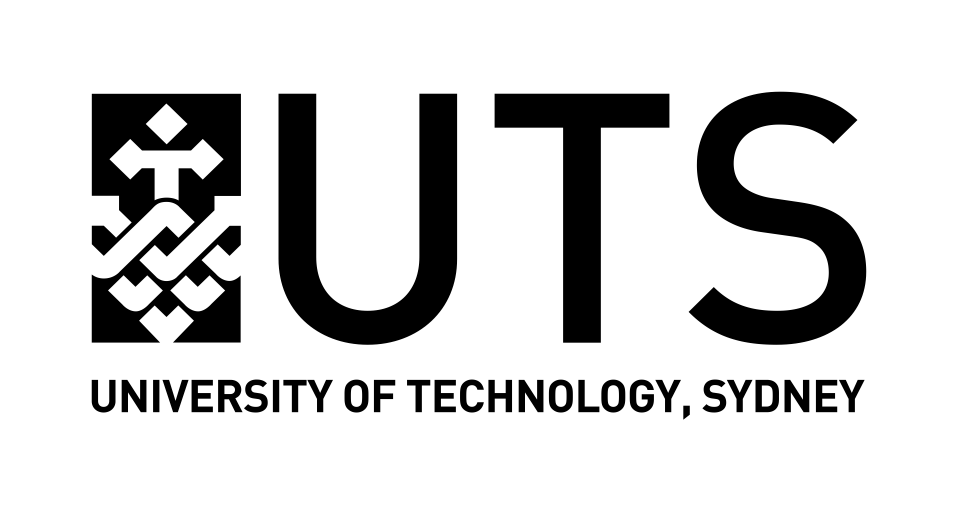
**Autumn 2023**

**Assessment Task 1**

Regression Models

**Student Name:** Shalimar Chalhoub

**Student ID:** 14071892



Contents

[Introduction 2](#_Toc131192442)

[Business Understanding 2](#_Toc131192443)

[Data Understanding 3](#_Toc131192444)

[Data Preparation 5](#_Toc131192445)

[Modeling 6](#_Toc131192446)

[Evaluation 6](#_Toc131192447)

[Deployment 9](#_Toc131192448)

[References 10](#_Toc131192449)

[Appendix 10](#_Toc131192450)

# Table of Figures

[Figure 1- Heat Map of Correlation Between the Data 4](file:///C:\Users\lenovo\Desktop\MLAA\AT1\MLAA%20Assignment%201%20Report.docx#_Toc131192451)

[Figure 2- Scatter plot of The Data 5](#_Toc131192452)

[Figure 3- Model Performance of Part A1 7](#_Toc131192453)

[Figure 4- Model Performance of Part A2 8](#_Toc131192454)

[Figure 5- Model Performance of Part B 8](#_Toc131192455)

[Figure 6- - Model Performance of Part C 9](#_Toc131192456)

Final Report for Cancer Death Rate Prediction among US Counties

# Introduction

The aim of this project is to build a regression model that can predict the cancer death rate among different US counties using several features present in the dataset provided by the company. The assignment consisted of 3 parts, the first, testing univariate linear regression, the second. testing the multivariate linear regression and the third had no constraint and allowed all feature engineering as well as the choice of any model without restraints. All of the above-mentioned experiments were done in Colab, using Python for data cleaning and exploration, visualizations, data analysis and building the models.

This report presents the results of the project according to the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.

# Business Understanding

Cancer is of the leading causes of death worldwide, and especially in the US. That being said, it is very important for the business to try and predict the cancer death rates in different US counties in order to see how resources should be distributed among said counties in order to try and reduce those rates.

If there are US counties with high cancer death rates, the organization could allocate funds towards cancer research facilities to help decrease these rates. Alternatively, the organization could invest in improving healthcare plans for individuals with cancer, as well as providing better access to medicine and comfortable accommodations. Whilst investing these resources nation wide can be very costly, using a regression model that predicts the death rate can help the business better distribute these funds and lower their cost margins.

A successful model would give an error rate of less than 10% as we are looking at means rather than actual values and a 10% error rate in this case would be acceptable.

By using python to preprocess our data and create our Linear Regression models, we are able to engineer new features, manipulate the data and create a model that can accurately predict cancer death rates in US counties which will assist the business develop strategies to reduce the burden of cancer in the US.

# Data Understanding

The data used in this project was provided in the initial phase of this project and it consist of 33 different features including information about the state’s demography, healthcare plans percentages, household incomes as well as medical information such as mean number of reported cancer cases and mean number of death due to cancer among others. The full Data Dictionary can be inspected in Appendix A. The target variable to predict is label ‘TARGET\_deathRate’ and is the Mean per capita (100,000) cancer mortalities in a county. All of the demographical data in the dataset is from the 2013 Census Estimates, whilst all of the medical information are from the years 2010-2016.

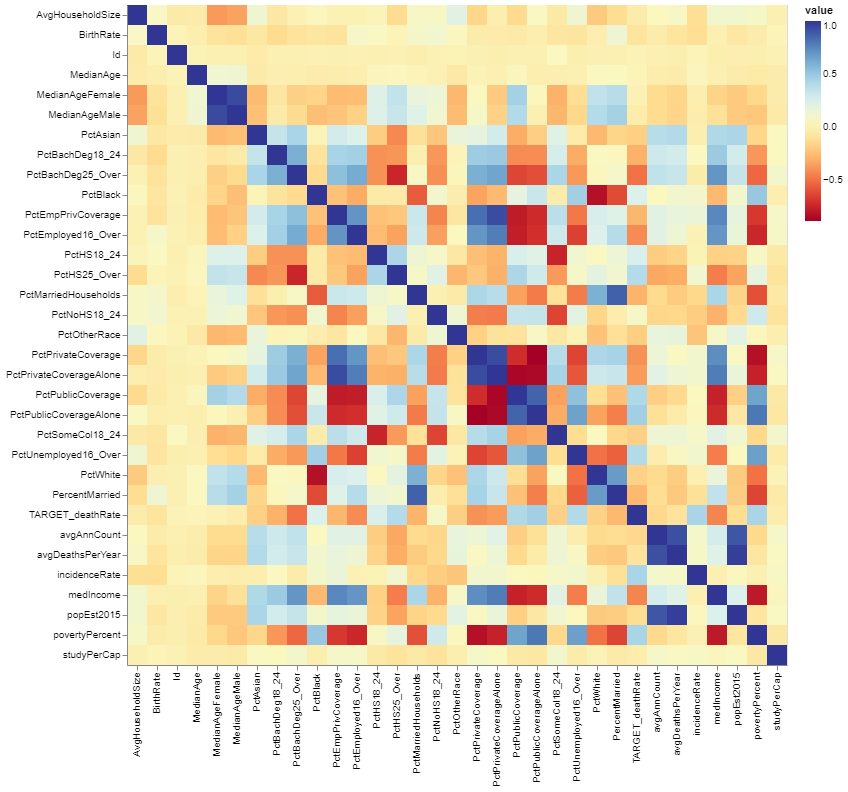
The heatmap below in figure 1 shows the correlation between all of my data in the dataset.

Figure 1- Heat Map of Correlation Between the Data

As can be seen from the heatmap in figure 1, the data is not very correlated with each other, with the warm yellowish colors and the light blue colors representing a low relationship between the data hence, my data is not too linear.

Figure 2 below shows the scatter plots of all of the medical data as well as income, population and poverty rate against each other

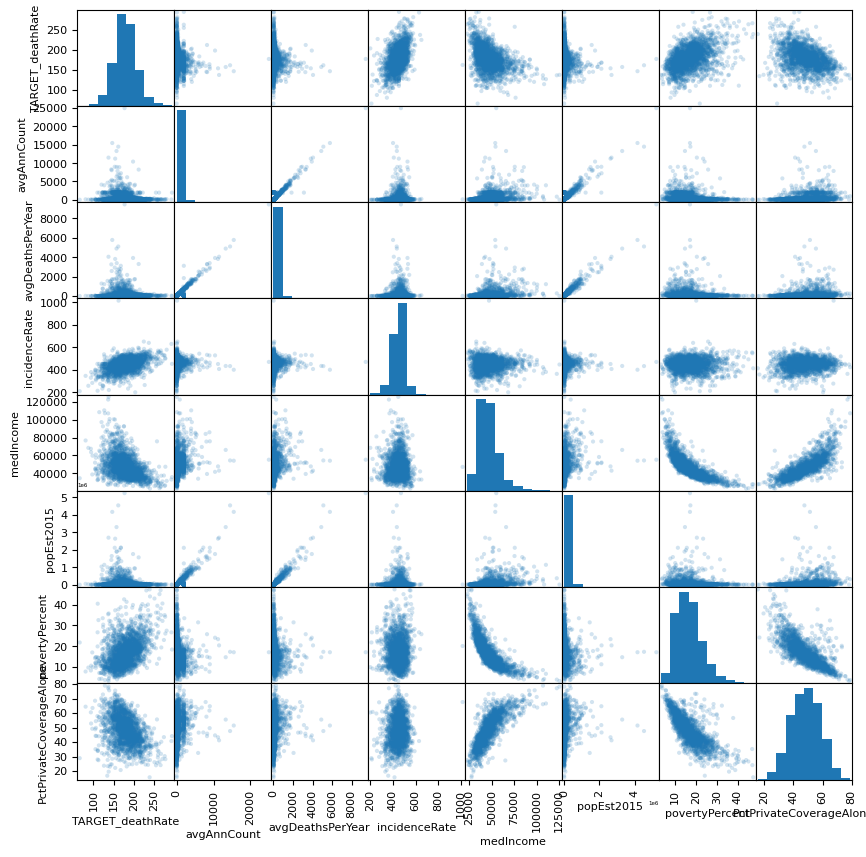


Figure 2- Scatter plot of The Data

# Data Preparation

Not all of the experiments had the same data preparation, as the three were different.

The data preparation for the Univariate Linear regression was very minimal consisting only of setting boundaries for my data as well as splitting the train dataset into train and dev.

For the multivariate linear regression, some more data preparation was done which consisted of the following:

1. Setting bounds for the deathrate between 130 and 240
2. Dropping all rows with an avgAnnCount of 1962.6676684 because, by looking at the dataset, it looked like an outlier
3. Removing nonnumeric columns
4. Removing PctSomeCol18\_24 from the dataset because it had more than three fourth of its data missing
5. Removing columns with low correlation such as PctBlack, PctAsian and PctWhite as well as Id column
6. Replacing PctEmployed16\_Over missing values with the mean
7. Splitting the train set to dev and train (20-80)

As for the lasso regression, the following preprocessing was done:

1. Removing nonnumeric columns
2. Dropping PctSomeCol18\_24 for having too many missing values and the column Id because it is irrelevant
3. Creating a new feature and adding it to the model which is the average death per year per county because of cancer multiplied by 100k and divided by the population. This was done in hopes of getting a close replica of the target value.
4. Splitting the train set to dev and train (20-80)

# Modeling

For Part A, univariate Linear Regressions, I did a simple linear regression model which I fit with one chosen feature from my dataset which was incident rate for Part A1 and PctPublicCoverageAlone for Part A2.

Fort Part B, multivariate Linear Regression, the same linear regression model was used but this time the model was fitted with all the cleaned data variables.

For Part C I used the Lasso regression which I fit with all my preprocessed training data in order to predict the target variable.

# Evaluation

Below are the results for each model

* Part A – 1 MSE: 616.2699362503008
* Part A -2 MSE: 507.2823423425601
* Part B MSE: 306.90807156480156
* Part C MSE: 166.0555473398766

The graphs below show the visual performance of each of these above-mentioned models

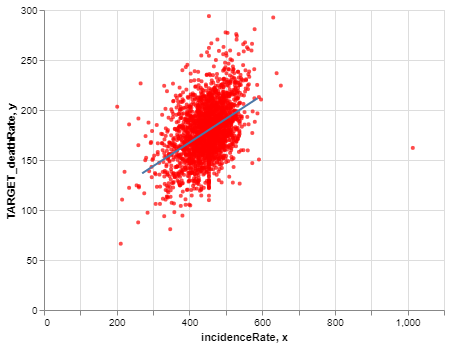


Figure 3- Model Performance of Part A1

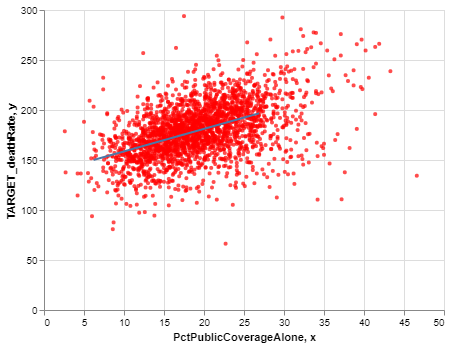


Figure 4- Model Performance of Part A2

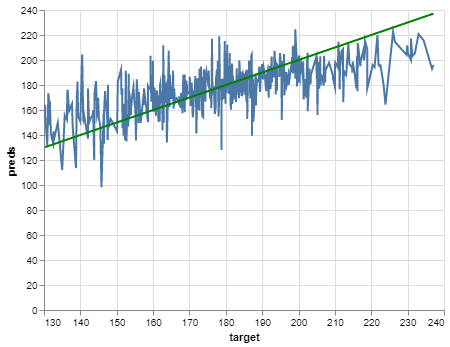


Figure 5- Model Performance of Part B

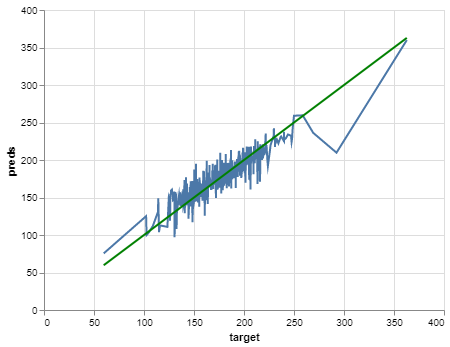


Figure 6- - Model Performance of Part C

As can be seen above, part C presented the best results with the difference between predictions and actual values around 12.5 data points which indicates very good results as it goes with the business’s goal of having less than 10% error.

That being said, there might be a way to make this data prediction better by maybe setting bounds for the data as the model does not perform well for values under 100 and over 250, however, I tested setting bounds for my data and the model did not perform much better.

By this data being this accurate, it can correctly predict the cancer death rates and thus be able to control the cancer epidemic better.

# Deployment

This particular model is designed to be implemented as a Python code, which can serve as a powerful tool for various stakeholders in the healthcare industry, such as cancer research companies, policymakers, and healthcare providers. By leveraging the model's predictive capabilities, these entities can estimate the potential cancer death rates for different counties across the United States.

This information can be incredibly valuable as it can help to identify areas that may require additional resources, such as medical facilities, staffing, and funding. By accurately predicting death rates, stakeholders can prioritize the allocation of resources in areas that need it the most, ultimately improving the quality of healthcare for patients and potentially reducing mortality rates.

# References

For all of the models I have used the templates provided in the MLAA sessions

# Appendix

Appendix A – Data Dictionary

**TARGET\_deathRate:** Dependent variable. Mean *per capita* (100,000) cancer mortalities(*a*)

**avgAnnCount:** Mean number of reported cases of cancer diagnosed annually(*a*)

**avgDeathsPerYear:** Mean number of reported mortalities due to cancer(*a*)

**incidenceRate:** Mean *per capita* (100,000) cancer diagoses(*a*)

**medianIncome:** Median income per county (*b*)

**popEst2015:** Population of county (*b*)

**povertyPercent:** Percent of populace in poverty (*b*)

**studyPerCap:** *Per capita* number of cancer-related clinical trials per county (*a*)

**binnedInc:** Median income per capita binned by decile (*b*)

**MedianAge:** Median age of county residents (*b*)

**MedianAgeMale:** Median age of male county residents (*b*)

**MedianAgeFemale:** Median age of female county residents (*b*)

**Geography:** County name (*b*)

**AvgHouseholdSize:** Mean household size of county (*b*)

**PercentMarried:** Percent of county residents who are married (*b*)

**PctNoHS18\_24:** Percent of county residents ages 18-24 highest education attained: less than high school (*b*)

**PctHS18\_24:** Percent of county residents ages 18-24 highest education attained: high school diploma (*b*)

**PctSomeCol18\_24:** Percent of county residents ages 18-24 highest education attained: some college (*b*)

**PctBachDeg18\_24:** Percent of county residents ages 18-24 highest education attained: bachelor's degree (*b*)

**PctHS25\_Over:** Percent of county residents ages 25 and over highest education attained: high school diploma (*b*)

**PctBachDeg25\_Over:** Percent of county residents ages 25 and over highest education attained: bachelor's degree (*b*)

**PctEmployed16\_Over:** Percent of county residents ages 16 and over employed (*b*)

**PctUnemployed16\_Over:** Percent of county residents ages 16 and over unemployed (*b*)

**PctPrivateCoverage:** Percent of county residents with private health coverage (*b*)

**PctPrivateCoverageAlone:** Percent of county residents with private health coverage alone (no public assistance) (*b*)

**PctEmpPrivCoverage:** Percent of county residents with employee-provided private health coverage (*b*)

**PctPublicCoverage:** Percent of county residents with government-provided health coverage (*b*)

**PctPubliceCoverageAlone:** Percent of county residents with government-provided health coverage alone (*b*)

**PctWhite:** Percent of county residents who identify as White (*b*)

**PctBlack:** Percent of county residents who identify as Black (*b*)

**PctAsian:** Percent of county residents who identify as Asian (*b*)

**PctOtherRace:** Percent of county residents who identify in a category which is not White, Black, or Asian (*b*)

**PctMarriedHouseholds:** Percent of married households (*b*)

**BirthRate:** Number of live births relative to number of women in county (*b*)

(*a*): years 2010-2016

(*b*): 2013 Census Estimates