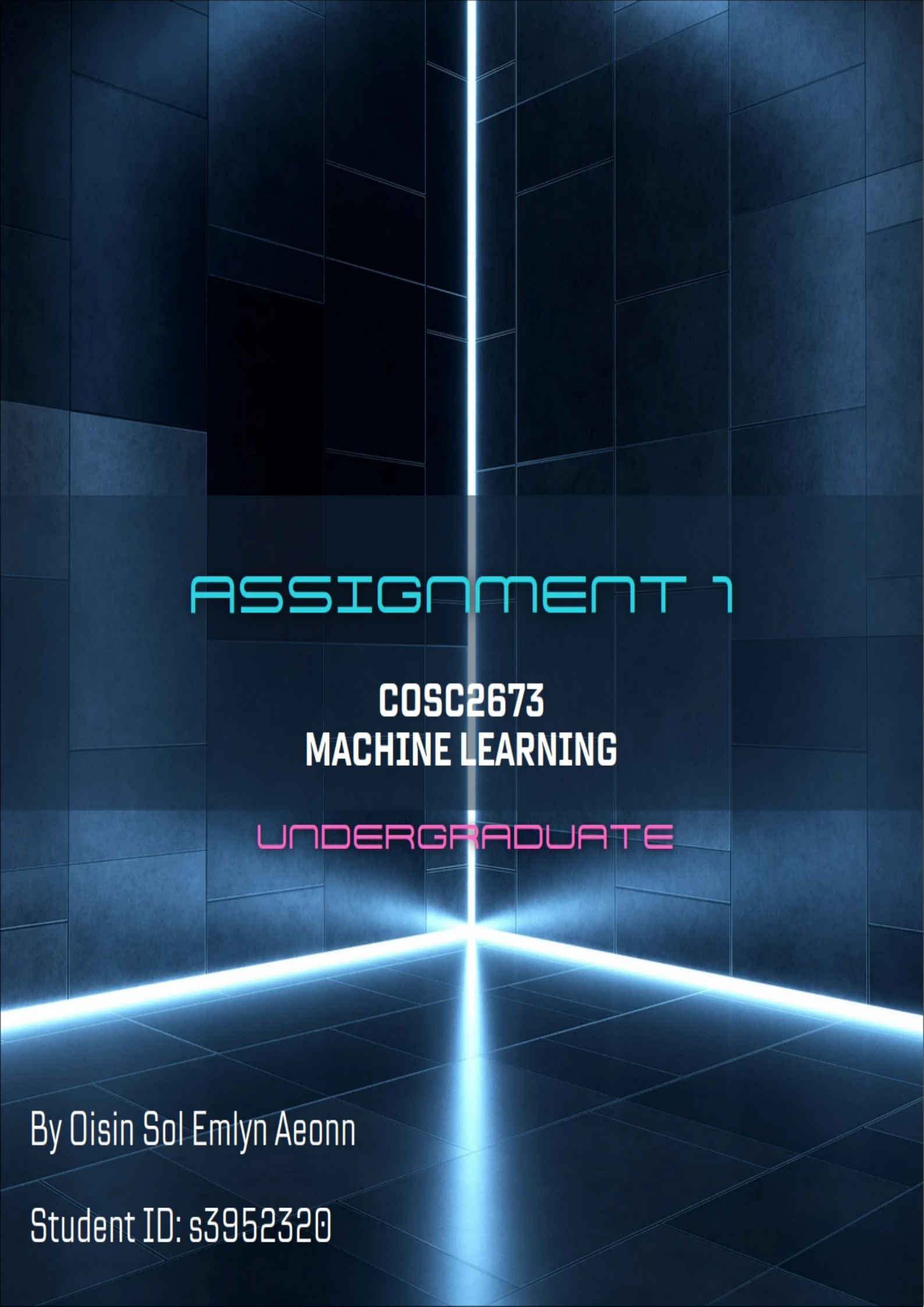


ASSIGNMENT 1

COSC2673
MACHINE LEARNING

UNDERGRADUATE

By Oisin Sol Emlyn Aeonn

Student ID: s3952320

This Page was intentionally left blank.

Assignment 1:
Introduction to Machine Learning

Published COSC2673 Machine Learning (Undergraduate) CANVAS

April 8, 2024 RMIT University

Oisin Sol Emlyn Aeonn

Student ID: 3952320

X_____

For Dr. Azadeh Alavi, Dr. Pubudu Sanjeevani, and Ms. Rumin Chu

Table of Contents

Abstract

This study aims to develop a regression model to predict life expectancy. The Dataset we are using is a modified World Health Organisation dataset, which covers many features of nations. Employing the machine learning techniques learned in weeks 1-4 of RMIT's **COSC2673 Machine Learning** Course, the research involves a comprehensive exploratory data analysis (EDA), data pre-processing, iterative model refinement starting from the initial baseline, attempting to enhance the predictive accuracy using validation to ensure the model's generalizability on an unseen dataset.

Research Methodology

This paper will utilise Python libraries such as pandas for data manipulation, seaborn, Plot Express, & Matplotlib for exploratory data analysis (EDA), and sklearn for data transformations, as well as model development and evaluation. Doing so will uncover the difficult to see relationships between various features and their impact on TARGET_LifeExpectancy. By employing sklearn's pre-processing tools for data cleaning and feature scaling, with the robust suite of algorithms for fine-tuning predictive models, we aim to accurately predict life expectancy with a high R^2, as well as low RMSE, and MAE scores. The effectiveness of these models will be evaluated using these metrics to ensure high predictive accuracy and reliability in estimating TARGET_LifeExpectancy.

Significance Statement

By identifying and understanding the key factors influencing life expectancy, we will learn how to: apply, and evaluate regression techniques in a real-world setting. This project also contributes 30% of our grade in **COSC2673 Machine Learning**.

Introduction

Problem Statement:

- The goal of this project is to develop a regression model to predict Life Expectancy using a modified WHO dataset, focusing on understanding key factors that influence it.
- Starting with a baseline model incorporating all features, subsequent optimisation will be pursued through advanced machine learning techniques such as Regularisation, Feature Scaling / Selection, and Hyperparameter Tuning.

Dataset:

- The analysis is based on a modified version of the WHO Life Expectancy dataset from Kaggle, which has numerous features that could impact life expectancy.

Objectives:

- Master the core principles of Machine Learning covered in weeks 1-4.
- Perform a comprehensive Exploratory Data Analysis (EDA) to uncover insights within the dataset, and identify required changes / transformation.
- Implement data preprocessing to ensure the dataset's integrity for modelling by identifying outliers, skews, normal-distributions, different data scales, etc.
- Iteratively refine the regression model to enhance its accuracy in predicting Life Expectancy by training, and validating your methodology.
- Apply and interpret evaluation metrics to systematically improve the model.

Scope:

- Focus on predicting Life Expectancy leveraging methods and insights from weeks 1-4 of the **COSC2673 Machine Learning** Course.
- Adopt strategies such as regularisation and normalisation to refine the model, maintaining the integrity of the feature set.
- Strive for the optimal model performance within the established constraints.
- Undertake predictions on a separate, unseen dataset (predictions.csv) to validate the model's generalisability.

Exploratory Data Analysis (EDA)

- First we must ingest the data into a Pandas DataFrame. Let's call it: 'lifeExpectancyFrame'. Pandas DataFrames allow us to efficiently explore and manipulate data. Now let's examine the dataset!
- Using the lifeExpectancyFrame.info() we can see some key insights into the data like the datatype, as well as seeing if there are any missing values. However, as we will find out the absence of null values isn't indicative that there aren't any incorrect values as many of the data points originally null have been 'cleaned' to 0, or averaged. We can view if our data contains any duplicate rows by using the has_duplicates = lifeExpectancyFrame.duplicated().any(), fortunately again we are returned Duplicate Data: False meaning our dataset is free from any missing values, and doesn't contain any duplicate data. We can also use lifeExpectancyFrame.describe() which will return some very important IQR values. This information will be critical for identifying, removing outliers and verifying data transformations.
- **Identified Outliers:** Outliers were detected in the majority of the dataset columns by comparing the IQR 25% and 75% to the minimum, and maximum values in the following: AdultMortality, AdultMortality-Female, AdultMortality-Male, SLS, Alcohol, PercentageExpenditure, Measles (very large outliers), BMI, Under5LS, TotalExpenditure, HIV-AIDS (very large outliers), GDP (large outliers), Population (large outliers), Thinness1-19years, & Thinness5-9years.

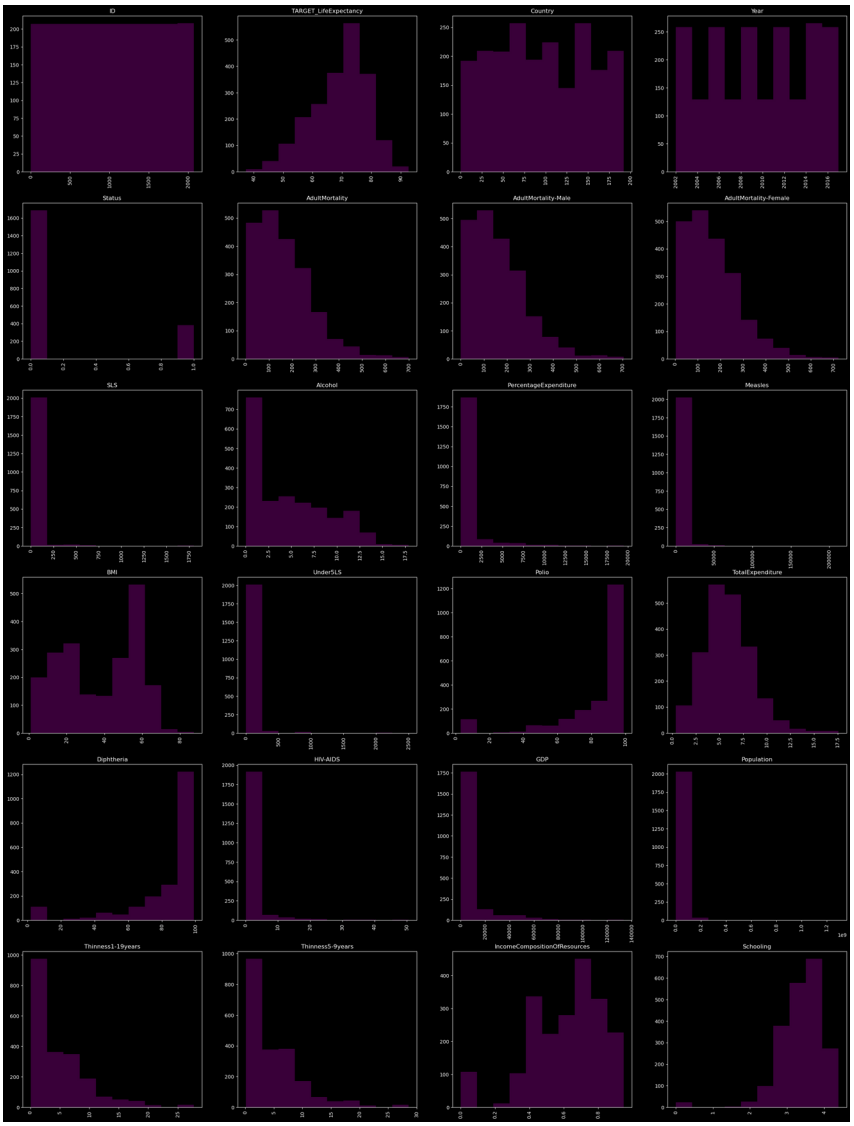


Figure 1: Histogram of Distributions for all Features

- It is clear in Figure 1: which shows histograms of the distributions for all features that there is a high amount of variance showing which are skewed, Gaussian, multi-modal, contain outliers, etc. They also enable a closer comparison of the value representations in the data, such as how large or small the values are. Given the considerable variation in our data, I hypothesise that transformations like power transforming, min max scaling as well as removal of outliers will definitely lead to considerable improvements in our prediction metrics.

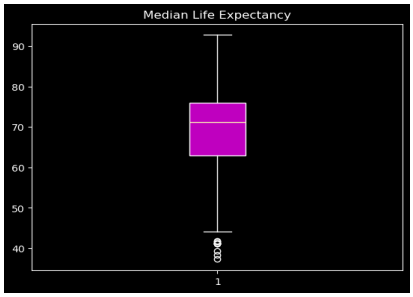


Figure 2: IQR Box Plot of TARGET_LifeExpectancy

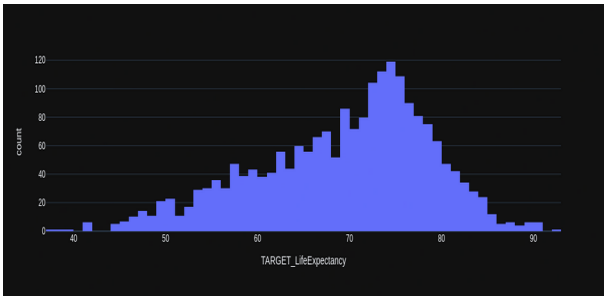


Figure 3: Histogram of TARGET_LifeExpectancy.

Taking a closer look at our target variable 'TARGET_LifeExpectancy' we can observe that there are some outliers which are especially visible in Figure 2: Boxplot. However, the distribution of Life Expectancy is near normal (Gaussian) see Figure 3.

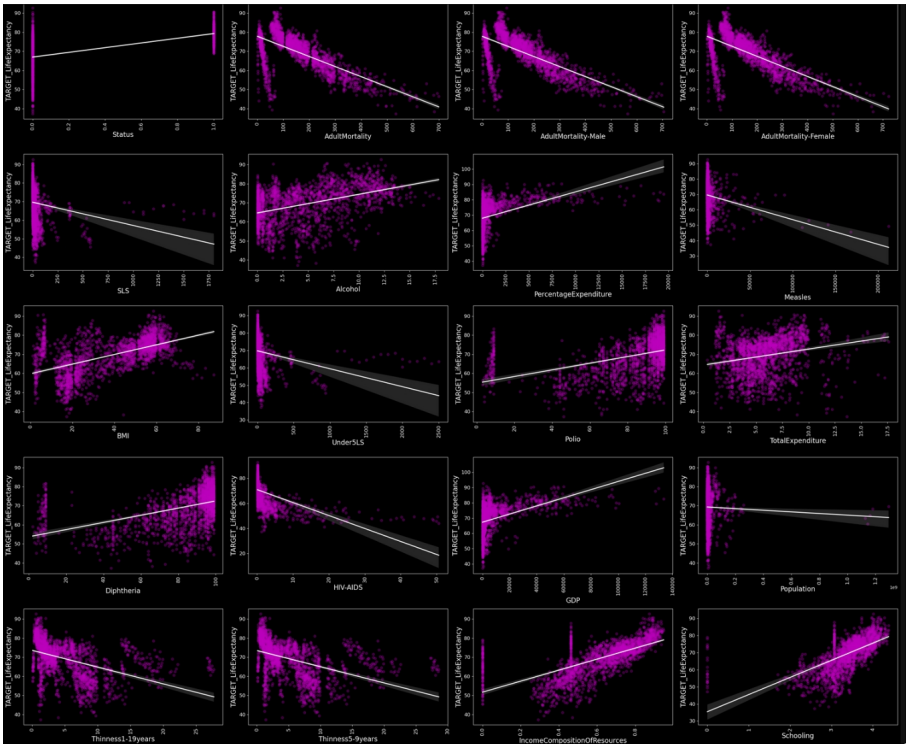


Figure 4: Reg plot of trends from our target: TARGET_LifeExpectancy to all features.

Regression plots like seen in Figure 4 combine trend analysis with a scatter plot to graph trends between two variables. In our case, I've used this to plot trends between LifeExpectancy, and all other features. As observed, there are quite a few trends, which is promising suggesting we should be able to achieve high evaluation scores. Here are some of my other key observations I made:

- There is a slight upward trend observed over the years, indicating an improvement in Life Expectancy over time.
- Status Indicator: The distinction between 'developed' and 'developing' status is a significant indicator of Life Expectancy, potentially useful for logistic regression using a sigmoid function. However, because the data is heavily skewed as there are more examples of Developing Countries this makes it less usable.
- Mortality and Infant Deaths: Adult mortality rates (for both males and females) and infant deaths are obviously linked to Life Expectancy.
- Alcohol Consumption: Surprisingly, there seems to be a positive correlation between alcohol consumption and Life Expectancy. This observation might be counterintuitive and warrants further investigation.
- Outliers in Expenditure: PercentageExpenditure has notable outliers that might not be appropriate for the dataset. These will be explored in more detail later.
- Disease-Related Trends: Clear trends are observed in data related to diseases such as Diphtheria, HIV-AIDS, and Polio, which are expected to impact Life Expectancy.
- BMI and Thinness: Both BMI and Thinness exhibit trends, suggesting their influence on Life Expectancy.
- Economic Factors: IncomeCompositionOfResources and GDP show trends, indicating their correlation with Life Expectancy.
- Schooling: Schooling also demonstrates a trend, further supporting its role in influencing Life Expectancy.
- Overall Trends: Most of the data exhibits a trend (either a negative or positive correlation) with the target Life Expectancy, except for categorical or nominal variables which are not good predictors (i.e., not correlated to the target), such as ID and Country.

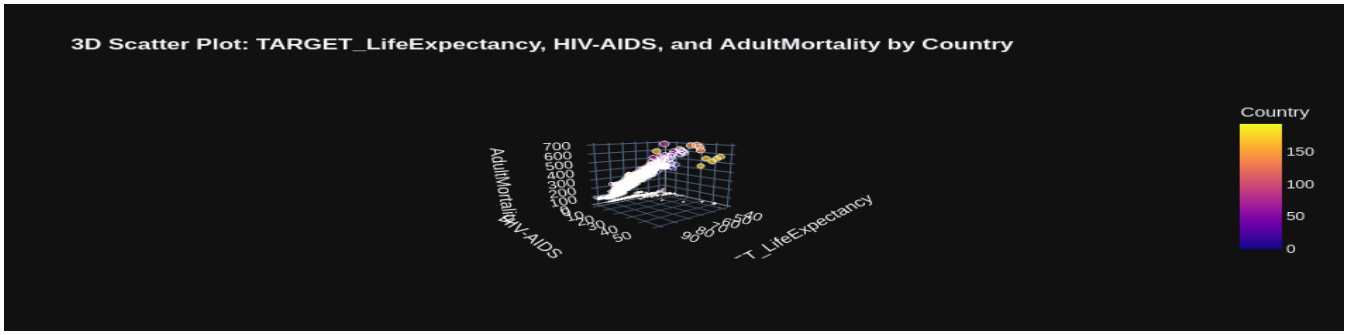


Figure 5: 3D Scatter Plots on HIV-AIDS, AdultMortality, and LifeExpectancy.

I even extended upon EDA techniques creating 3D visualisations of data. See video or Notebook for Figure 6!

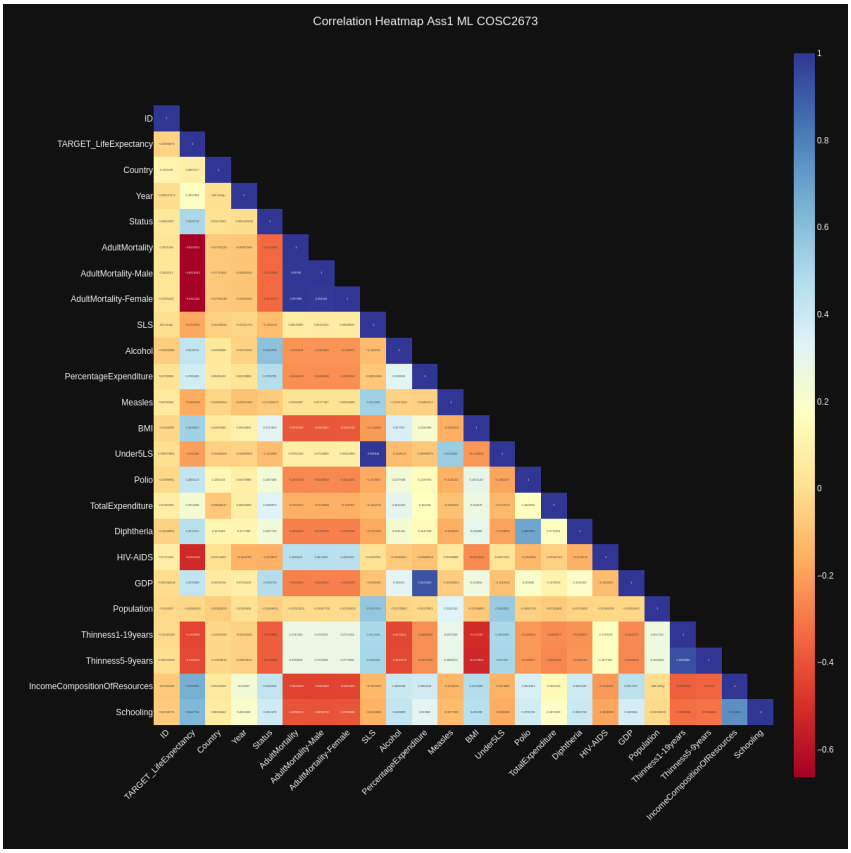


Figure 6: Heatmap of all feature combination coefficients

Following the analysis of Figure 6 (heatmap), we gain comprehensive insights that reinforce our initial observations and unveil additional relationships across the dataset: The heatmap effectively highlights a near 1:1 relationship among various mortality indicators, as well as between infant mortality and short lifespan indicators, underscoring their significant impact on life expectancy.

A surprising near 1:1 correlation is observed between PercentageExpenditure and GDP, which seems counterintuitive given it is only supposed to be a percentage. This warrants a closer examination to understand the underlying factors and potential data anomalies.

The strength of a heatmap lies in its ability to visualise all feature relationships simultaneously, offering a holistic view beyond the pairwise correlations previously examined with the target variable. This comprehensive perspective is invaluable for identifying both expected and unexpected correlations within the dataset. To enhance readability and focus on unique information, I designed the heatmap to display only half of the data matrix, omitting the symmetrically duplicated half. This decision streamlines the visualisation without sacrificing analytical value. The inclusion of exact value labels within the heatmap further refines our analysis, enabling precise identification of correlation strengths and facilitating a more nuanced interpretation of the data relationships. This heatmap analysis not only corroborates our earlier findings but also broadens our understanding of the intricate interplay among various features, setting the stage for deeper investigation into specific areas of interest. I also used plotly express so it’s interactive.

Data Inconsistencies Identified

As previously identified and noted, certain aspects of our data appear to be questionable. In this section, we will delve deeper into a few of these observations, critically assess potential biases, and devise a strategic plan for mitigating these issues in our dataset moving forward.

Percentage Expenditure Column

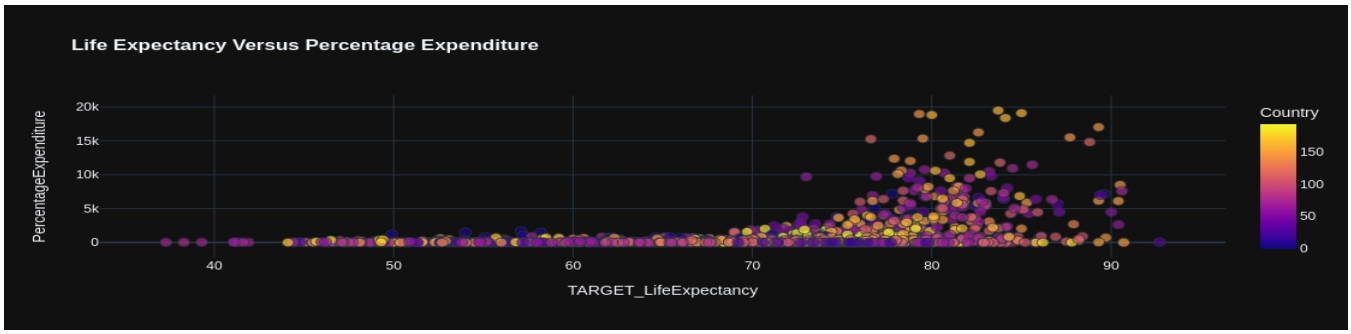
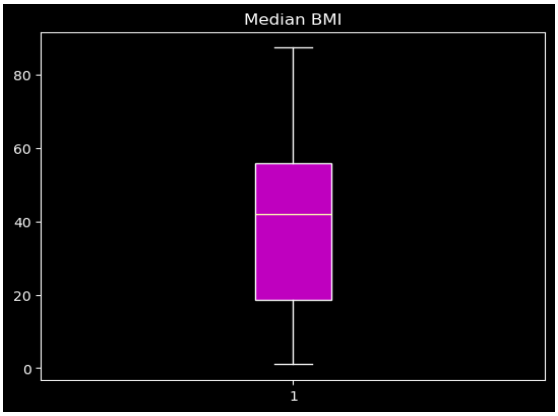


Figure 7: Outliers / wrong data in PercentageExpenditure column

As shown in Figure 7 the values for PercentageExpenditure, significantly exceed 100%, which clearly does not align with the feature's metadata description. The metadata tells us this metric is a percentage of GDP allocated to health — a figure that historically has never approached levels as high as 19000%. My initial instincts lean towards capping the data, removing outliers, or applying a log transformation, however the exceptionally high correlation with GDP (0.93) signals a more fundamental issue. Further examination of the original WHO Dataset, which shares many of these values, revealed that the data in question actually represents the raw monetary amount spent on health per capita within each country. To rectify this, we will adjust the health expenditure figures by dividing them by the GDP per capita and subsequently multiplying by 100 to convert them into the correct percentage format. There are two equally plausible solutions which are correct. Firstly we could alter the metadata label so when we discuss this feature we know what the values actually are, or to as above apply the transformation to make our features independent of one another. Verification against authentic WHO data through a side-by-side comparison confirmed the accuracy of these adjusted PercentageExpenditure values, suggesting an oversight or mislabelling by the dataset's creator. Warning, multicollinearity, such as the observed correlation between PercentageExpenditure and GDP, can compromise the integrity of our model by causing inflated variance in coefficient estimates and obscuring the distinct influence of correlated predictors on the target variable. However it is consistent with the Mortality features as they are essentially duplicated.

BMI Column



- Figure 8: BMI Boxplot

According to WHO 2015 records, the lowest BMI recorded is 20.5 (Eritrea) and the highest is 32.5 (Nauru). In contrast, our dataset (see Figure 8) shows an unrealistic mean BMI of ~47, with extreme min and max values ranging from nearly 0 to 90. Despite this, the distribution correlates with our target variable, Life Expectancy, indicating the data distribution is relatively coherent. To address the inflated mean BMI, I plan to use scaling techniques. This will adjust the mean to a more realistic range while keeping the distribution's shape intact. I could then apply some outlier capping, to further remedy this issue. Visual analyses have also uncovered biologically improbable fluctuations in BMI, such as dramatic year-over-year changes within some countries which if I was to do the project again I would fix. These findings highlight the importance of thorough data validation and correction, ensuring the dependability of our analyses and predictive models.

Population Column

Population is a noteworthy example within our dataset where the data is represented by significantly large numbers. To prevent the Population feature from unduly influencing our model due to its large scale I plan to apply min max scaling techniques during the Data Preprocessing phase to preserve the distribution, but make it only from 0 to 1. It is also a good approach to perhaps normalise Population values, as it is heavily skewed.

We will address these concerns, and many other identified outliers we identified earlier in our EDA during the Data Preprocessing phase.

Data Splitting

In this phase, we're partitioning the `training.csv` file into distinct training and validation subsets, a critical step that lays the groundwork for model training and preliminary evaluation. Adopting the strategy recommended by Azadeh and favouring an 80/20 split caters to the dataset's scale and aligns with best practices for achieving a balance between learning complexity and validation accuracy. This structured separation is pivotal, ensuring our model is rigorously trained and evaluated in a controlled environment before facing the unpredictability of the 'blind' test data, which mimics the real-world environment, and tests our model’s generalisability.

A novel insight, inspired by a peer's methodology, involves the potential reintegration of the validation set for training purposes post-validation. This technique, particularly advantageous for smaller datasets, could further enhance the model's performance by maximising the learning data volume.

It is imperative to maintain the independence and identical distribution of the test data. This safeguards against data leakage and guarantees the model's evaluation remains impartial. The careful division of `training.csv` is not just a procedural necessity; it's instrumental in refining our models, ensuring they're not just tailored to the training data but are truly robust and capable of generalising well.

Utilising `train_test_split` from `sklearn.model_selection`, with a randomised shuffling mechanism, facilitates this process efficiently. The original dataset, comprising 2071 instances, is divided into 1656 training and 415 validation instances. This allocation preserves the dataset's integrity, ensuring the combined total of training and validation instances mirrors the original count, thereby maintaining data consistency and reliability throughout the model development lifecycle. However, we still need to conduct a bit more EDA to ensure our Data is being represented by both sets.

Checking for Data Leaks

Random splitting is essential but can lead to leakage if two splits are not truly independent.

- We will utilise insights from Exploratory Data Analysis (EDA) to detect any hidden sources of leakage in the dataset.
- We will examine histograms for each attribute in the training and validation sets, using different colours, to ensure the splits are identically distributed.
- The distributions of attributes seen in Figure 9 show that the training, and validation datasets align closely. This indicates that the random splitting process has effectively preserved the dataset's overall statistical properties, meaning we can move on to applying some pre-processing.

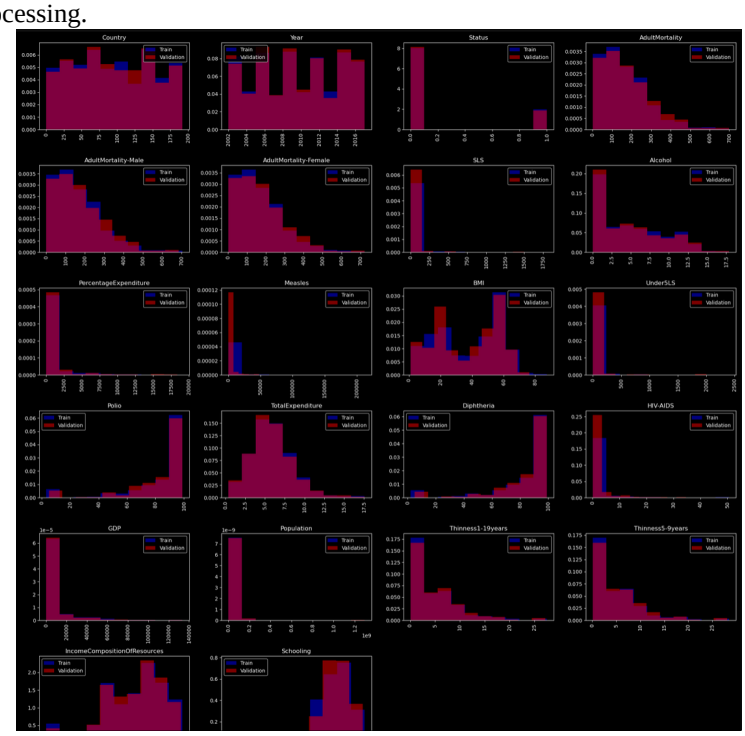


Figure 9: Checking for Data Leaks

Data Preprocessing

Uniform data preprocessing across training, validation, and blind test sets is crucial for predictive consistency. Min Max scaling normalises values to a 0-1 range, while preserving the distribution. Power scaling, such as Yeo-Johnson, corrects data to approximate normal distribution, suitable for both positive and negative skewed distributions. Log scaling reduces skewness by applying logarithmic transformations, primarily to positive data. These techniques, along with outlier management via the IQR method, enhance model reliability by standardising feature scales and distributions.

Min Max Scaling: Applied universally to ensure all features are normalised to a range between 0 and 1, facilitating equal consideration during model training.

I performed Power Scaling (Yeo-Johnson) on the following attributes: SLS, HIV-AIDS, Adult Mortality, Adult Mortality-Male, Schooling, Thinness 5-9 years, Polio, Diphtheria, Alcohol, Under5LS, Income Composition of Resources. These features were selected for normalisation to manage skewness and improve model interpretability.

The PercentageExpenditure adjustment produced negligible R^2 score reduction, this transformation aimed to remove the large outliers present in the dataset, while reducing the dimensions of the scale.

Similarly, but reflected the BMI adjustments which planned to shift values downwards for a more realistic distribution gave slight positive R^2 improvements.

Baseline Multivariate Linear Regression Model

The methodology I adopted for the Baseline approach to this project involved starting with a multivariate regression model that included all features. I sought to see the effect of my transformations on the dataset so included the pre-processed data verses the original un-altered dataset to see the potential imrovements. The Baseline served to gauge feature correlations before proceeding with feature transformation, regularisation, feature selection, hyperparameter tuning, and overall model Optimisation. This approach aligns with the assignment's explicit requirement aimed at enhancing fundamental ML/Statistical skills, emphasising a comprehensive understanding before delving into more advanced topics. Incorporating methodologies similar to those encountered in lab exercises, While various metrics were considered to evaluate model performance, particular emphasis was placed on RMSE and R^2. These metrics are favoured due to me being more comfortable with them, having a solid grasp of their implications, what constitutes a good score, and their operational principles. Metrics Overview: RMSE (Root Mean Square Error) measures the model's prediction error magnitude, while R^2 (Coefficient of Determination) assesses the proportion of the variance in the dependent variable that is predictable from the independent variables. I also included Mean Absolute Error (MAE).

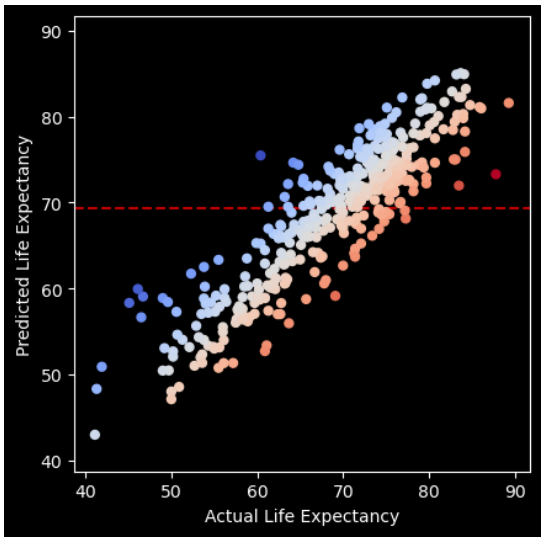


Figure 10: Predicted verses Actual Life Expectancy Baseline

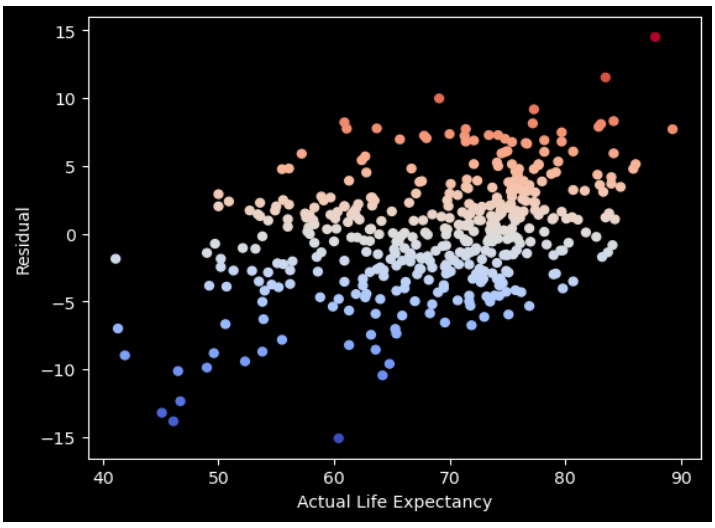


Figure 11: Residual Error for actual vs predicted Life Expectancy Baseline

The process of developing the baseline model was both challenging and enlightening, involving extensive data preprocessing and addressing various issues to obtain the desired output in a .csv format from the blind test data. The unscaled data still presents a reasonable predictability, with an R^2 score of ~0.75, indicating that a significant proportion of the variance in life expectancy is predictable from the features. For RMSE, and MAE we achieved ~4.5, and ~3.5 respectively. See Figure 10, and 11 for more info.
Performance Improvement Through Preprocessing: Scaling, outlier removal, and normalisation contributed to performance improvements, as evidenced by an increased R^2 score to ~0.8 and reduced RMSE to ~4 and MAE to ~3 in the scaled model.
Further Exploration: The Baseline doesn't end our exploration we will continue with hyperparameter tuning and regularisation techniques aimed at reducing overfitting and enhancing the model's generalizability to unseen data as well as Polynomial Regression.

Prediction Output

I created a python function to perform the same transformations we made on our data, and then created a new data frame containing only the ID, and TARGET_LifeExpectancy as instructed by the sample prediction output. This prediction can be made at many points in my model development.

Advanced Model with Feature Selection Model

Advanced Model Development: This phase involves refining our initial model by integrating more sophisticated techniques and methodologies.

Feature Selection Based on EDA: Decisions on feature inclusion or exclusion will be guided by insights gained from Exploratory Data Analysis (EDA) and comprehensive data analysis.

Rationale for Feature Removal: Each feature considered for removal will be accompanied by a clear justification, focusing on its relevance and impact on the model's performance.

Performance Comparison: We will systematically evaluate the model's performance before and after each change to assess the impact of these changes.

Exploring Advanced Techniques: The model enhancement process will include the exploration of regularisation techniques to prevent overfitting, the application of polynomial features to capture non-linear relationships, and other advanced machine learning methods.

Feature Scaling: As previously explored about Feature Scaling, and the removal of outliers, Min Maxing our data, and normalising it we will be applying similar features to all our models to see how well different models cope with different types of data too.

Regularisation

Regularisation is a technique used in machine learning to prevent overfitting. It does so by adding a penalty on the larger coefficients in the model such as the ones we just observed from the heatmap in the previous section.

Purpose of Using Regularisation: We employ regularisation to enhance the generalizability of our model, ensuring it performs well not just on the training data but also on unseen real-world data. Regularisation keeps the model simpler and more robust, thereby reducing the risk of overfitting and improving model performance on new data.

I explored Ridge Regression (L2 Regularisation), Lasso Regression (L1 Regularisation), and Elastic Net (Combined Ridge, and Lasso). I found that Regularisation techniques, including Lasso and Ridge, yielded slightly lower performance metrics compared to the baseline scaled model against my validation set. However, note that the primary objective of regularisation is not to enhance performance metrics but to increase the model's generalizability by preventing under or overfitting. Elastic Net, which combines the strengths of both Lasso and Ridge regularisation, was employed to leverage a more balanced regularisation approach, aiming for improved model stability and generalizability while not necessarily getting the best performance metrics.

However, I did not focus on Regularisation as a technique as it was not the primary focus on this assignment.

Hyperparameter Tuning & Polynomial Regression

Both Hyperparameter-Tuned Models utilised a 0.1 learning rate, and a polynomial degree of 2, trying to reach a R^2 score of 0.86 as also found during our GridSearch. For the unscaled version we achieved an R^2 score of ~0.8012, meaning that 80.12% of the variance in the life expectancy variable can be explained by the model. Other metrics were: Root Mean Squared Error (RMSE) of 4.3093 and a Mean Absolute Error (MAE) of 3.1022, which reflect the model's prediction errors. The scaled dataset achieved a higher R^2 score of ~0.8632, marking an improvement over the unscaled model. This score aligns with that which was found in our GridSearch. Other metrics also scored lower with an RMSE of 3.5742 and an MAE of 2.6137. We can see an example of the models performance in Figure 12 showing 50 real verses predicted samples.

Summary: The hyperparameter-tuned model, which used scaled data and underwent optimisation, outperformed the unscaled tuned model in terms of both the R^2 score and error metrics (RMSE and MAE). The tuning process led to the selection of a polynomial degree of 2 and a learning rate of 0.1 as optimal parameters through a GridSearch, with a notable improvement in predictive accuracy and precision. This underscores the importance of hyperparameter tuning and feature scaling in enhancing the performance of machine learning models.

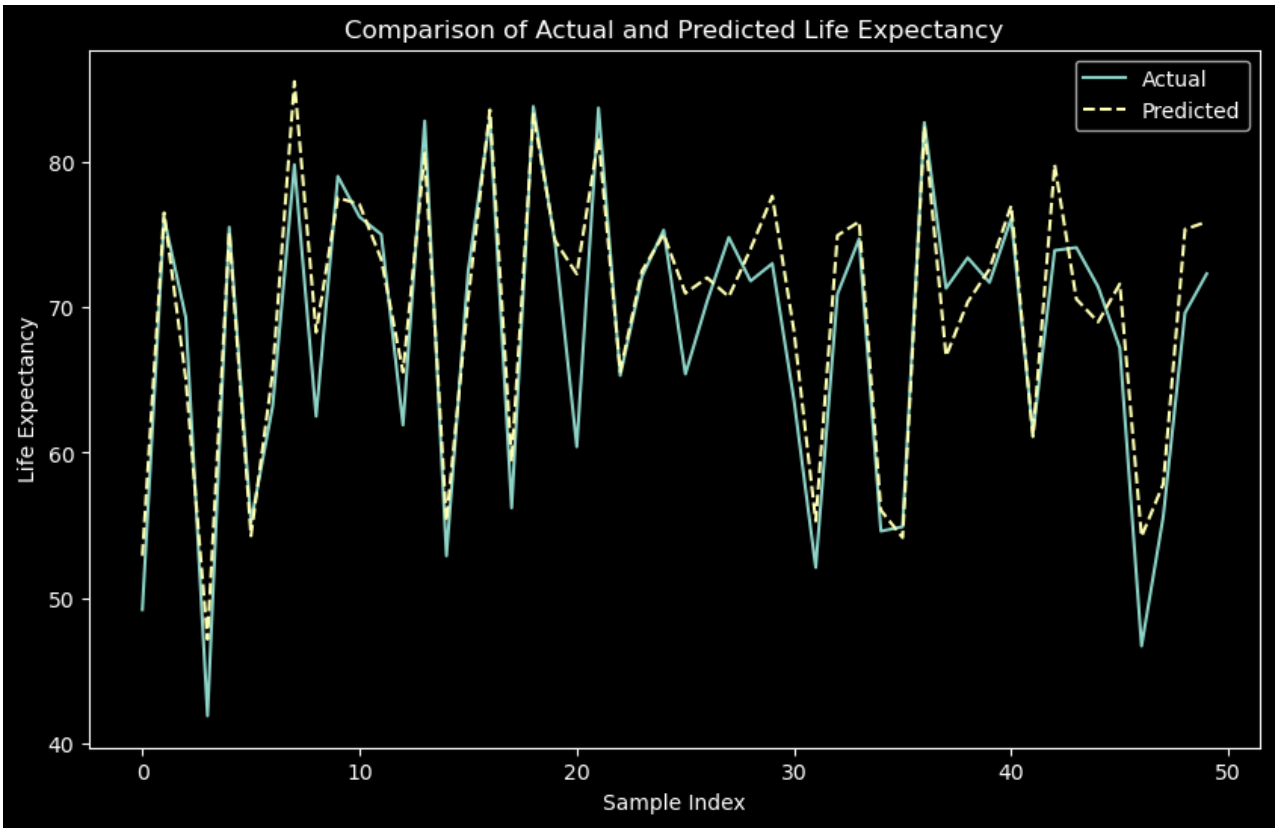


Figure 12: Hyperparameter predicted verses actual on 50 samples

Feature Selection

Feature Selection involves intentionally choosing a subset of relevant features for use in our model creation, or removing features as they are not relevant to the target. Based on our Exploratory Data Analysis (EDA) findings, I decided to focus on creating the best possible model using a maximum of three key features that were most impactful for predicting our target variable. This approach was exploratory in nature; despite it not being a primary focus as per Azadeh's guidance, I was keen on examining all aspects of Machine Learning, and given the obvious advantages of such a simple and small model such as only having to collect 3 data types, and the computational efficiency.

From my EDA I selected the following features as they had some of the highest coefficients to the target. The three I picked were: ['HIV-AIDS', 'AdultMortality', 'IncomeCompositionOfResources']. I was able to achieve an unscaled R^2 score of 0.631 and a scaled of 0.707 which is quite good considering the simplicity of the model, and the small amount of computation required to run this model.

Model Conclusions, and Evaluation

Feature Preprocessing: Scaling large features like population and normalising data to near Gaussian distributions significantly improved model scores. Additionally, capping outliers in all datasets contributed to score enhancements.

Hypertuned Polynomial Function: The hypertuned polynomial function exhibited further improvements in model performance, potentially overfitting compared to regularisation models which are supposedly more Generalisable.

Exploration of Complementary Methods: Further exploration will involve complementary methods such as tree-based models, like Decision Trees, Random Forests, and ensemble methods like gradient boosting to enhance hyperparameter tuning.

Improved Data Preprocessing Strategies: In future iterations, I will focus on refining data preprocessing strategies by addressing mismatched values, handling zero outliers, and prioritising complementary methods for better generalizability and accuracy.

Surprising Performance of Multivariate Model: The Baseline Multivariate model's simplicity and inclusion of all features yielded surprisingly effective results, defying initial expectations of poorer performance due to the dataset's complexity and uncorrelated data.

Interest in Similar Health Datasets: The project's success, and fun nature has sparked interest in exploring similar health-related datasets. I have used before being the Kaggle WHO dataset, so I had some prior knowledge / familiarity with this dataset.

References

1. Azadeh Alavi, Pubudu Sanjeevani, and Rumin Chu. CANVAS RMIT for Machine Learning COSC2673 Particularly Labs, Lectures, and Tutorials.
2. Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for Machine Learning. Cambridge University Press. Retrieved from https://mml-book.github.io/book/mml-book.pdf.
3. Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press. Retrieved from https://ebookcentral.proquest.com/lib/rmit/detail.action?docID=6246595&pq-origsite=primo.
4. Kumarajarshi. (2018). WHO Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data.
5. Matplotlib Documentation https://matplotlib.org/
6. Plotly Express https://plotly.com/python/plotly-express/
7. Sklearn Documentation https://scikit-learn.org/stable/
8. World Data Bank 2002-2017 records. Retrieved from https://data.worldbank.org/.

Appendices

Appendix 1: README_ass1COSC2673_s3952320_Oisin_Aeonn.txt

Appendix 2: Notebook_ass1COSC2673_s3952320_Oisin_Aeonn.ipynb

Appendix 3: Notebook_ass1COSC2673_s3952320_Oisin_Aeonn.pdf

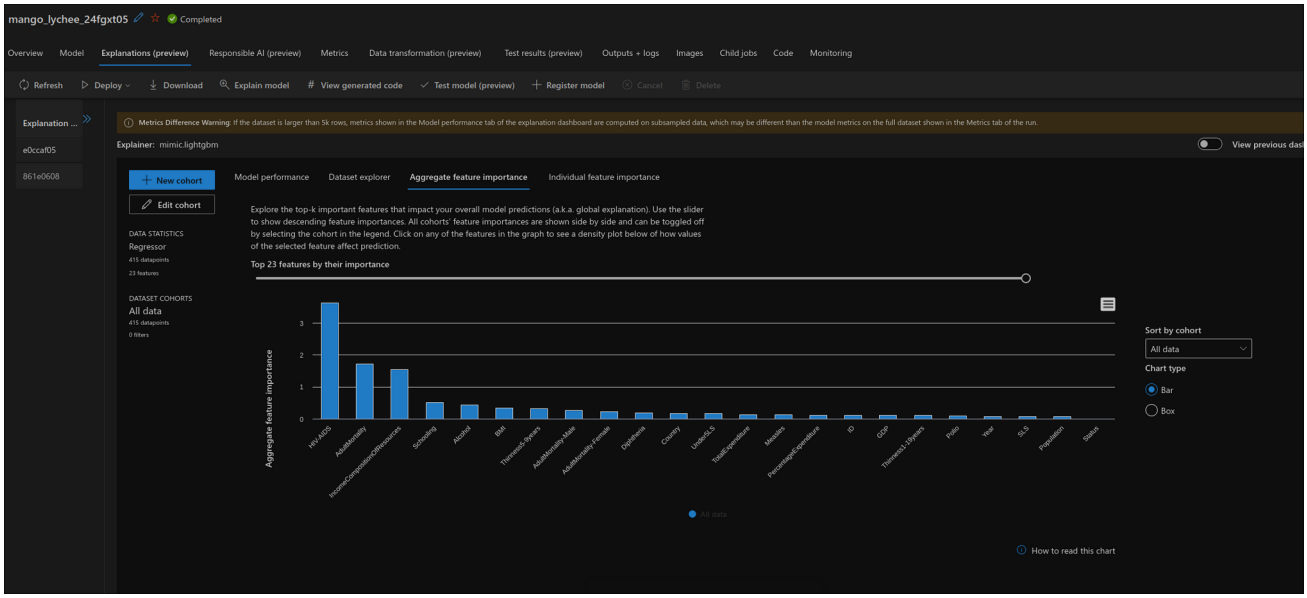
Appendix 4: Report_Ass1COSC2673_s3952320_Oisin_Aeonn.pdf (this file)

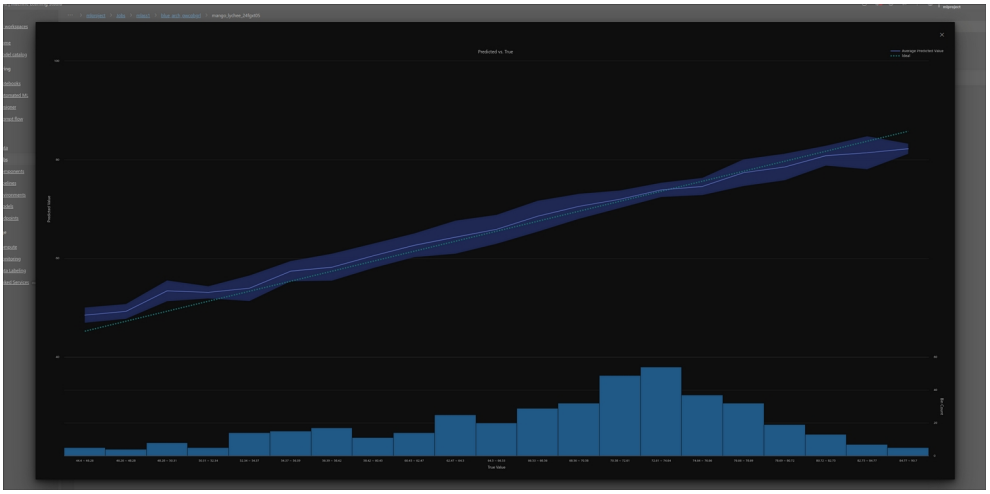
Appendix 5: Predictions_ass1COSC2673_s3952320_Oisin_Aeonn.csv

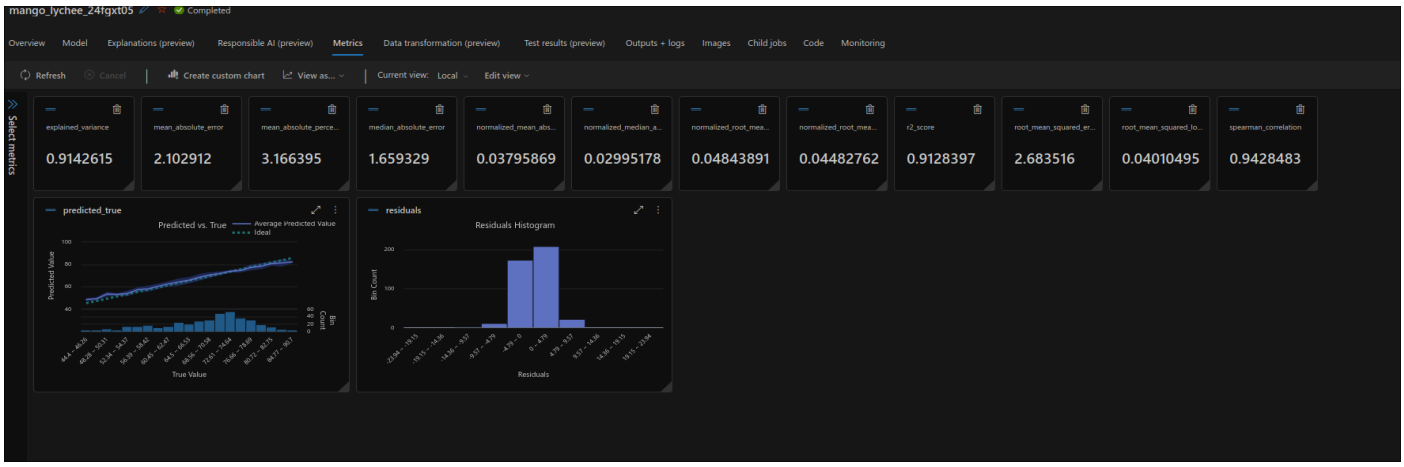
Appendix 6: Datasets, and other required materials in this .zip folder

Azure AutoML

Azure AutoML: I employed Azure AutoML to enhance and implement advanced machine learning techniques, which will be elaborated a bit in the video.







Complementary Methods

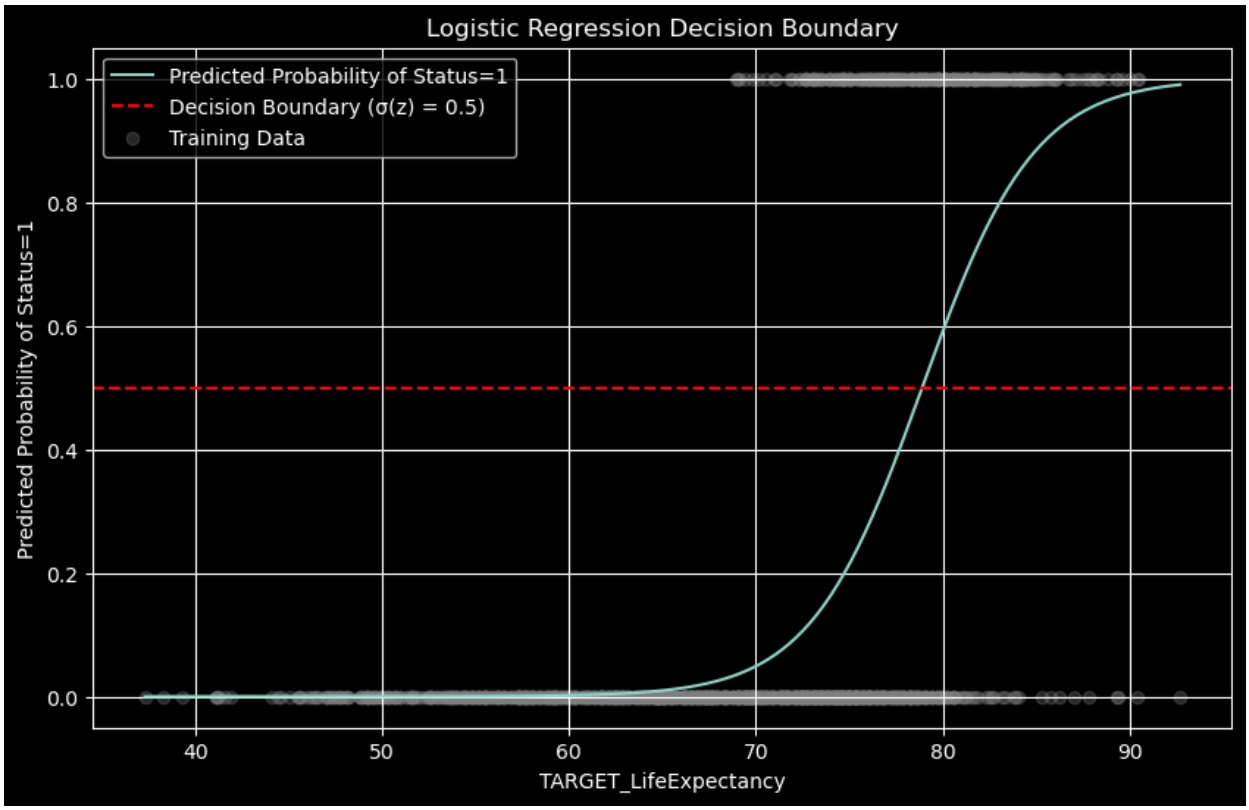
While the methods employed extend beyond the scope of weeks 1-4, they primarily focus on hyperparameter tuning, a critical aspect contributing to 15 marks on the assignment. These were also covered in week 5, so I thought that it would be fun to get a head start.

The exploration beyond the assignment requirements was driven by a desire to deepen my understanding and expertise in machine learning techniques, reflecting a commitment to continuous learning and improvement.

Introduction of Voting Ensemble Method: To further enhance model performance, a Voting Ensemble Method will be employed, using gradient boosting to leverage the collective wisdom of multiple models to make more accurate predictions.

Logistic Regression

Limitations of Logistic Regression: Logistic Regression is unsuitable for predicting numerical values such as TARGET_LifeExpectancy. Therefore, a quick and simple prediction of the binary variable "Status" was chosen instead to have fun, and gain more experience doing a binary classification.

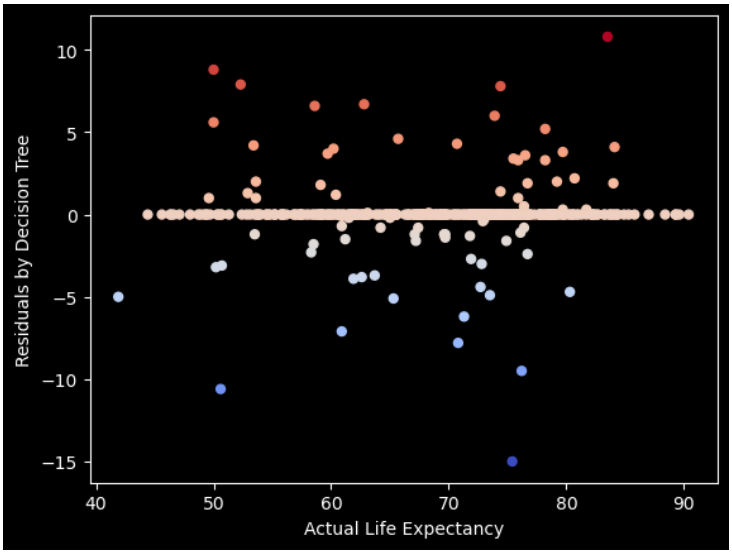


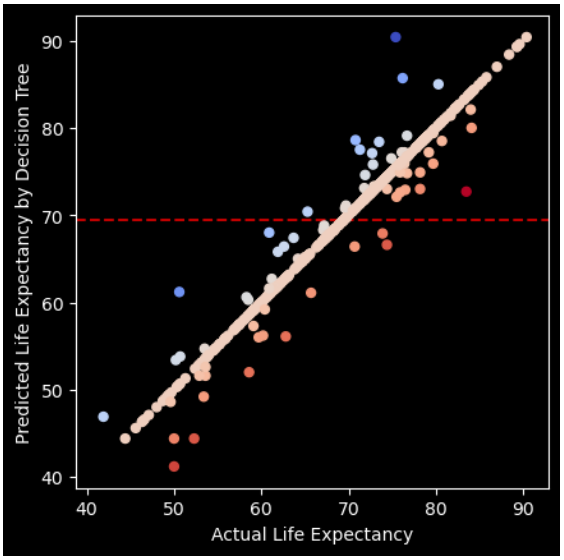
Decision Tree

Decision Tree benefits:

- Clear interpretability.
- Minimal data prep.
- Robust to outliers.

Depth of the Decision Tree: 27
Number of leaves: 1592
RMSE for Decision Tree: 1.9042311796476206
R^2 score for Decision Tree: 0.9610376028562522



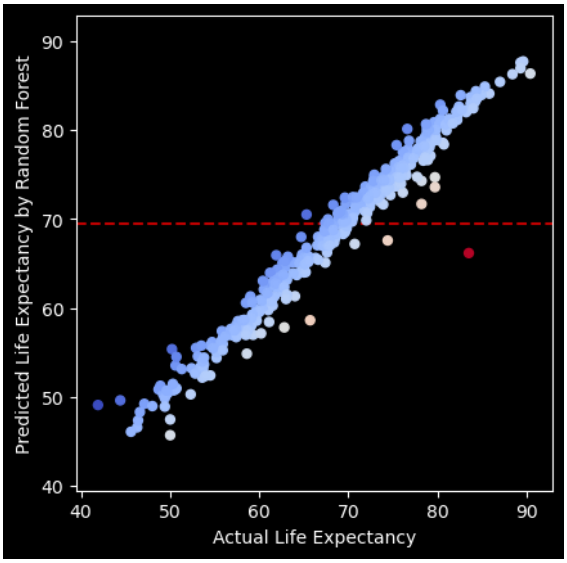


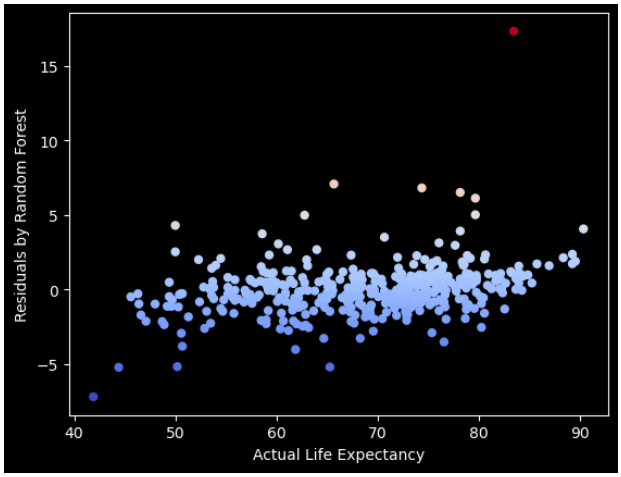
Random Forest

Random Forest benefits:

- High accuracy.
- Reduces overfitting.
- Feature importance insights.
-

Number of estimators in the Random Forest: 100
RMSE for Random Forest: 1.789243906410334
R^2 score for Random Forest: 0.9656010316061836





Ensemble Voting

Ensemble benefits:

- Corrects errors iteratively.
- Flexible optimization.
- Handles mixed data.

Number of estimators in Gradient Boosting: 100
RMSE for Gradient Boosting: 2.616638122564423
R^2 score for Gradient Boosting: 0.9264312031177964