

# Machine Learning Project

By James Ong

## Table of Content

### Table of Contents

<b>Introduction.....</b>	<b>2</b>
<b>Classification Dataset – Diabetes Outcome .....</b>	<b>2</b>
<b>Logistic Regression .....</b>	<b>3</b>
<b>Random Forest Classifier.....</b>	<b>4</b>
<b>Decision Tree Classifier .....</b>	<b>4</b>
<b>Conclusion .....</b>	<b>4</b>
<b>Regression Dataset – 2018 World Happiness Report .....</b>	<b>5</b>
<b>Lasso Regression.....</b>	<b>5</b>
<b>Ridge Regression .....</b>	<b>6</b>
<b>Multiple Linear Regression .....</b>	<b>6</b>
<b>Random Forest Regressor .....</b>	<b>6</b>
<b>K Neighbors Regressor .....</b>	<b>7</b>
<b>Conclusion .....</b>	<b>7</b>
<b>Unsupervised Learning – Mall customer segmentation dataset.....</b>	<b>8</b>

## Introduction

This report will be focused on using the different types of machine learning models on real world datasets. In this project, I will be using unsupervised learning models to find clusters of customers spending habits based on their demographics from the mall datasets. Regression models will be used to find factors affecting world happiness score on the 2018 world happiness dataset and classification models will be used to find out if a person will get diabetes based on their health factors. This project is done in python and will display graphs and models to support as evidence for our research question.

## Classification Dataset – Diabetes Outcome

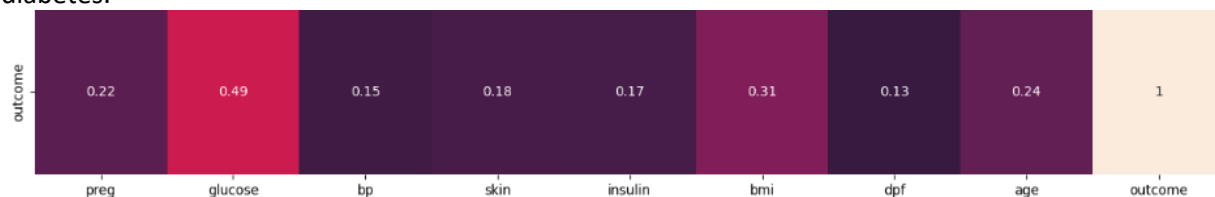
Diabetes is a rampant problem that is projected to only increase in the near future, therefore it is important to understand its causes and trends to find a prevention for future generations. Our research focuses solely on females with diabetes as it is shown that females are more prone to serious complications and at a higher risk of death. The dataset is obtained from National Institute of Diabetes and Digestive Kidney Diseases, and it consist of medical predictor variables and one dependent variable (outcome) from females aged 21 and above.

I start by identifying key research questions to focus on when building the machine learning models.

- What are the health factors that result in diabetes? Build a model to predict if a female patient will get diabetes with the given health factors.
- What are the top 3 highest correlated factors that result in diabetes?
- Does an older age patient affect the chance of diabetes?
- Does food and lifestyle choices play a part in the chances of getting diabetes?

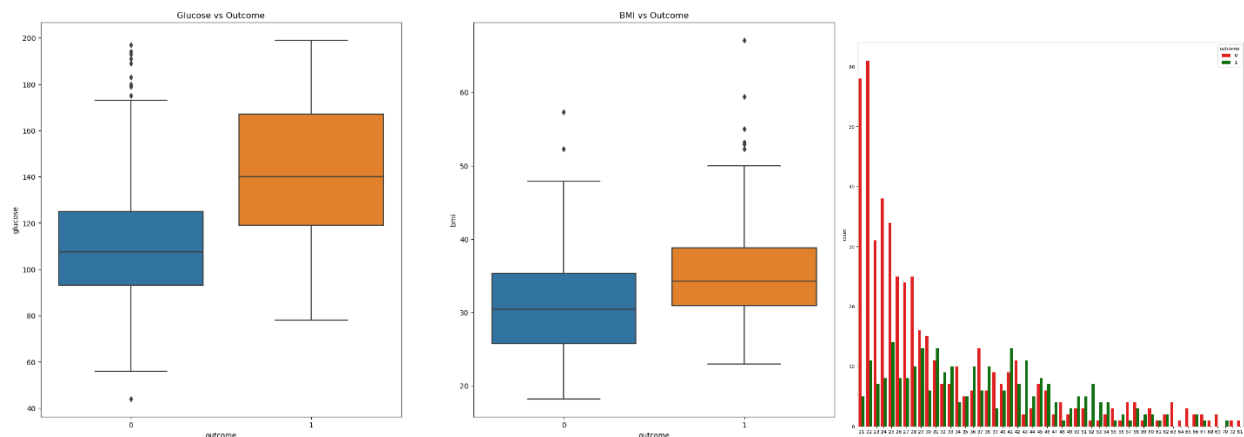
After importing the dataset and doing preprocessing of the data such as handling missing values and standardization, I proceed to check for multi-collinearity using VIF among the independent variables and It can be noted that variables such as glucose and bp had a higher VIF which could indicate a possibility of multi-collinearity existing, which I will use standard scaler to reduce the multi collinearity before splitting the data into training and testing sets. I also plotted a boxplot to note the outliers and replaced the outliers with the median of the variable, this method is also known as Median Absolute Deviation.

Lastly, I plot a heat map to understand the correlation of all the variables together. It can be noted that glucose, BMI and age of the patient are the top 3 variables that correlate with the outcome of getting diabetes.



Our statement can be further supported by visualizing the data using charts. As seen below in the boxplots, it can be noted that a higher glucose level and BMI would lead to diabetes. Research done by European Society of Cardiology proves similar findings, therefore preventive measure such as selecting healthier food options and partaking in exercise regularly can reduce the chance of diabetes. Age is also a factor that plays a part, but not as severe as the previous two factors. In the barchart below, it is noted that the diabetes (green bar) is more common among female patients after age 30 – 54, which is also known as middle age diabetes due to the fact that these group of patients are not taking care of their health by eating healthy and keeping an active lifestyle because of factors such as too copped up with work or pregnancies. When looking further into whether the number of pregnancies can risk the patient getting diabetes, it is observed from the boxplot that this is infact a possibility. This is also further supported by the medical reason that women is associated with a higher risk of developing gestational diabetes and type 2 diabetes due to factors such as weight gain, changes in hormones and an increased

in insulin resistance. Research done by National Medical Birth Register also suggest similar findings that can be use to support this argument.

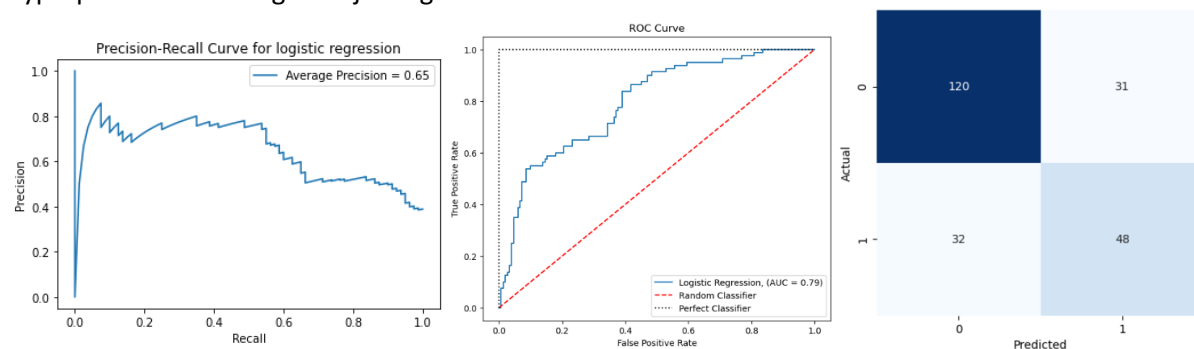


Moving on to create the machine learning models, I proceed to split the dataset into training and test dataset with proportions of 70% and 30% respectively. The dataset is noted to have a slight imbalance where 65% of patients has no diabetes but a slight imbalance should not be a major concern for the models. The model will be evaluated using F1 Score, AUC score and Accuracy (since the model is close to being balanced, accuracy is a useful scoring metric). The F1 score combines precision and recall using their harmonic mean and is used to compare performance of a binary classifier. The AUC score represents the area under the ROC curve, which tells us how efficient our model is at distinguishing between the positive and negative classes. Lastly, accuracy score measures the number of correct predictions made by a model over the total predictions made.

### Logistic Regression

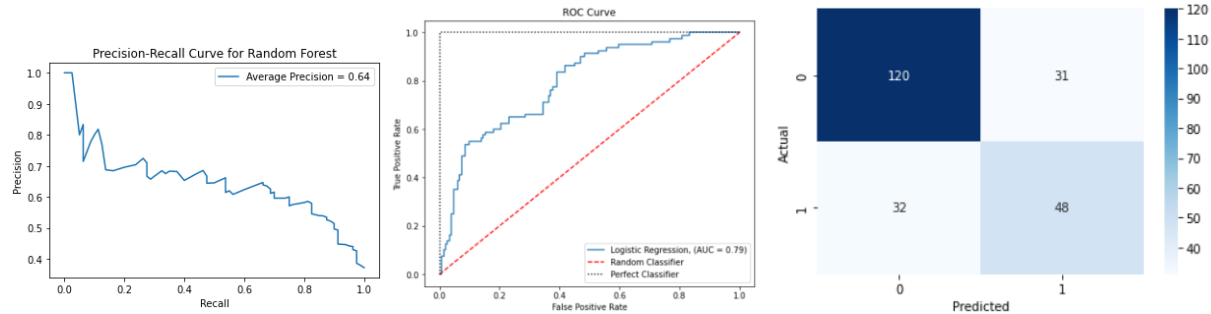
The first model I will be using is logistic regression as this model is a popular algorithm used for binary classification. It is based on a sigmoid function that transforms linear regression into a logistic regression that predicts the probability of patients that does not have diabetes. The formular for precision is  $tp/(tp+fp)$  which result in us getting 0.61, that means if the model predicts a patient has diabetes, it is correct 61% of the time.

The formular for recall =  $tp/(tp+fn)$  which result in us getting 0.60, which means the model correctly identifies 60% of all patients who has diabetes. The AUROC score of 0.79 suggest this model has good discriminatory ability such that a randomly selected patient with diabetes has a higher probability of having diabetes. The model also has an accuracy score of 0.72, which is ideal and realistic. Therefore, this model is a suitable model to be used but it could still be improved upon to increase its f1 score, such as hyperparameter tuning or adjusting the decision threshold.



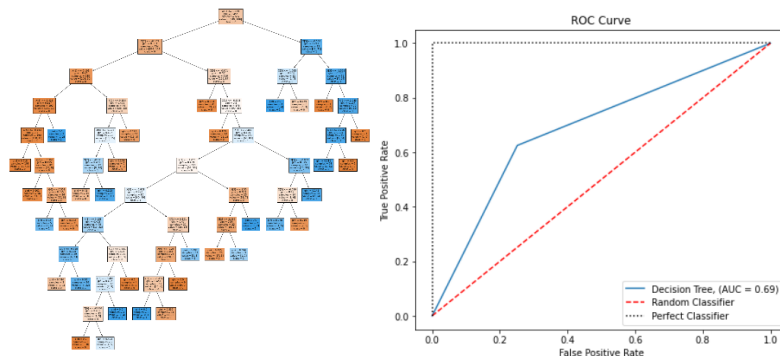
## Random Forest Classifier

The second model is random forest classifier. It is an ensemble learning method that contains multiple decision trees. Random forest overcomes the problem of overfitting that stems from a single decision tree by averaging the predictions of multiple decision trees which reduces the variance. The f1 score for random forest is 0.65 which is still considerably acceptable, along with a decent AUROC score of 0.79 and accuracy score of 0.75. This is by far the best model.



## Decision Tree Classifier

The next model is decision trees model, which works by splitting the dataset based on splitting criteria until a stopping condition is met. New data would be filtered through the root nodes and follow the branches based on its feature values and stops at the leaf node which provides the final output. It is an easy model to understand and interpret at a cost of being prone to overfitting. The model has a f1 score of 0.59 and AUROC score of 0.68667 and a decent accuracy score of 0.7, which makes it the poorest model in the bunch. Before visualizing the decision tree, I have to use the pruning technique to remove branches that use features of low importance, which reduces the chances of overfitting. Once that is done, the decision tree is shown below.



## Conclusion

Referring to the research questions at the beginning, we are able to find out that important factors that are correlated to diabetes are BMI, glucose level and number of pregnancies, with supporting evidence from external research. The models that I created are also able to predict the outcome of diabetes based on the variables, these models are decent but can be improved upon with methods such as hyper-parameter tuning and cross validation.

## Regression Dataset – 2018 World Happiness Report

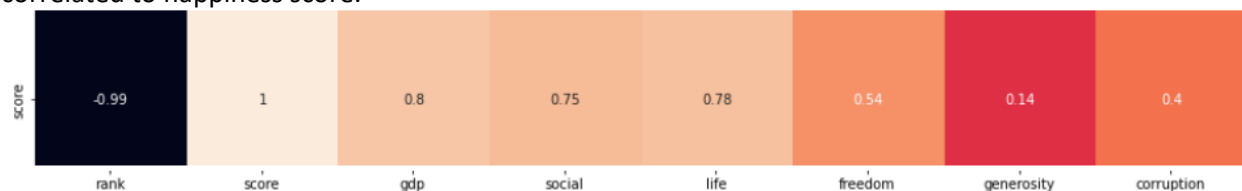
We as humans have always been interested in the pursuit of happiness. In recent years, data has been gathered about the world happiness report that scores countries on its happiness score based on social and economic factors of the country. This report aims to use machine learning and data analysis to find relationships and insights on how these variables can affect a nation's happiness, which can benefit by allowing government bodies to adjust and set policies to keep the country happy.

These are the research questions that I have proposed to have an idea for the development of the model.

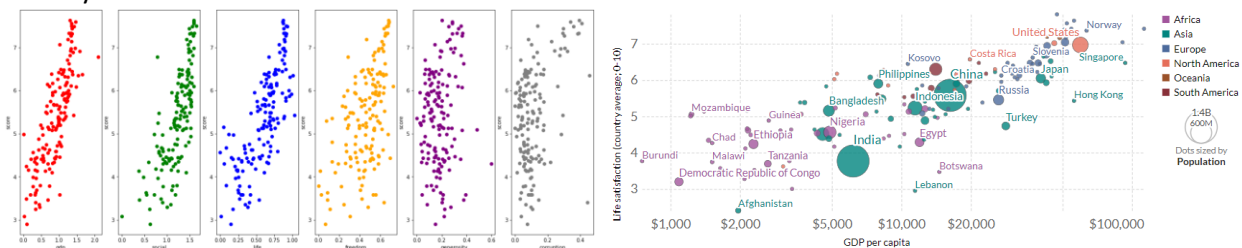
- What are the variables that have a high correlation to happiness score?

- Why are some countries happier than others? How can we improve the country's happiness?

I begin by importing the dataset and proceed to perform data preprocessing, cleaning up the data, checking and fixing outliers using the median absolute deviation method and checking for multi-collinearity. There are some missing values in the corruption column which are fixed by replacing the missing values with the mean of the column. A heatmap is plotted to identify which variables are highly correlated to happiness score.



It is seen that all these factors except generosity and corruption are highly correlated, this is further supported by analyzing each independent variable against score using a scatterplot, which all shows a linear trend. With the support of external research, it is seen that GDP has with national happiness, as a higher GDP tend to lead to citizens being wealthier, however social factors such as social support, healthy life expectancy and freedom of choice are also important factors that play a part besides wealth of a country.



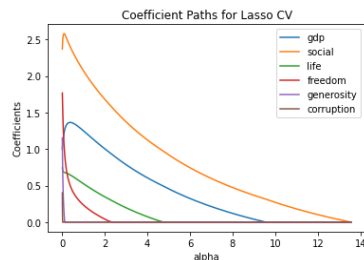
I proceed to standardize the data using standard scaler before splitting the data into training and test groups with a test size of 30%. The scoring metric that will be used for this regression dataset will be Mean Square Error (MSE) is the mean squared differences between the predicted and actual values. A lower MSE results in better performance, however it is sensitive to outliers. Root MSE (RMSE) is used by rooting square root of MSE, which is a better interpretation of MSE. Similar to MSE, a lower MSE indicates a better performance, but it is also sensitive to outliers and emphasizes large errors. R2 Score is a scoring metric that explains how well the data fit with the regression model.

## Lasso Regression

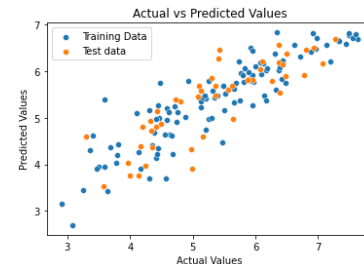
Lasso regression is a linear model with regularization techniques which uses a shrinkage method, The model uses the L1 regularization technique with parameter 'alpha' as the penalty term which is the amount of shrinkage implemented. Lasso regression shrinks the coefficients and helps to reduce model complexity and multi-collinearity.

The original model is decently scored, but after performing a 10-fold cross validation using GridsearchCV, there is a slight improvement in all scores except accuracy. It is also found that hyperparameter alpha is the best for the model when alpha = 0.01. A coefficient plot is to see the coefficient shrinking as alpha

changes, which indicates which features are important by observing which coefficients become zero first indicating less importance compared to features that remain non-zero for higher alpha values. We can observe that social and gdp are important features. Another plot used is actual vs predicted values, to visualize the model's performance by observing how much the predicted values differ from the training values. Despite having a few inaccurate points, generally predicted values are very similar to training values.



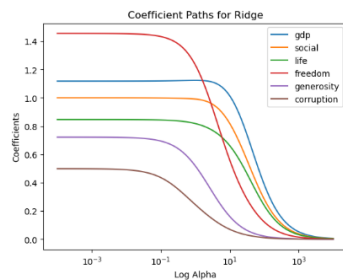
Model	R2 Score	MSE	RMSE	Accuracy
Lasso	0.776	0.278	0.527	0.713
Lasso CV	0.781	0.272	0.522	0.712



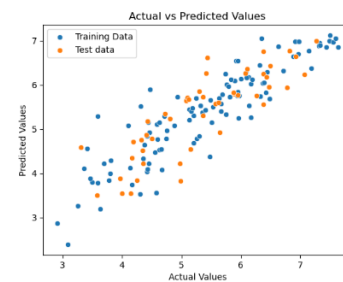
## Ridge Regression

Ridge regression is a linear model fitted with L2 regularization, which is used to reduce overfitting and reduce multi-collinearity. L2 Regularization works by adding a penalty term to the ordinary least squares and has the parameter alpha which is penalty term that controls the amount of shrinkage implemented on the model. The model has a r2 score of 0.785 indicating this model explain 78.5% of the variation in the target variable. Performing cross validation makes the results of the model slightly worse, which could be explained by model instability, such as noise in the dataset.

Looking at the coefficient path plot, it is observed with similar effect that gdp and social are the 2 most important feature, remaining non-zero for higher values of alpha. Similar to lasso regression, ridge regression actual vs predicted values are similar to each other, which indicates good performance.



Model	R2 Score	MSE	RMSE	Accuracy
Ridge	0.785	0.267	0.517	0.702
Ridge CV	0.782	0.270	0.519	0.712



## Multiple Linear Regression

Multiple Linear Regression is a linear model that explain the relationship between the independent variables and the dependent variable. This is a basic linear model that uses the ordinary least squares method, which minimizes the sum of squared differences of the actual and predicted values on dependent variable to obtain the best fit line. As seen, all the coefficients are positive, which indicates all of them are positively correlated to the happiness score , with social , life and freedom being the most impact on happiness score.

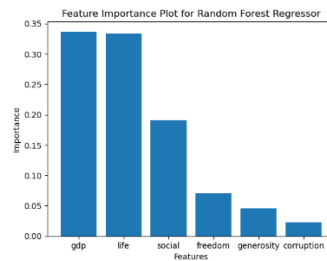
Model	R2 Score	MSE	RMSE	Accuracy
MLR	0.785	0.267	0.517	0.702

	Coefficient
gdp	0.986259
social	1.057343
life	1.119536
freedom	1.468661
generosity	0.892868
corruption	0.145474

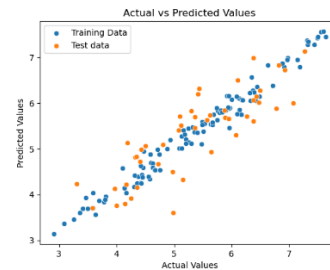
## Random Forest Regressor

A random forest regressor is an ensemble of multiple decision trees constructed together and averaging their predictions to create a more accurate model. This reduces the overfitting problem that occurs in single decision tree models. In comparison, this model has the best performance with a R2 Score of 0.908

and lowest MSE and RMSE score of 0.114 and 0.338 respectively. Upon performing 10-fold cross validation with parameters 'n\_estimator', we get an increase in accuracy and a lower MSE and RMSE, therefore. The feature importance plot allows us to visualize which features are important based on their impact on the model performance. As seen gdp and life are the most important features based on this model.

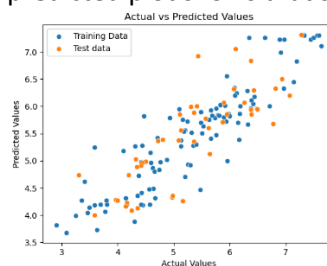


Model	R2 Score	MSE	RMSE	Accuracy
RFR	0.908	0.114	0.338	0.703
RFR CV	0.908	0.113	0.337	0.711



## K Neighbors Regressor

K neighbors regressor is based on the k nearest neighbors' algorithm. This model uses feature similarity to predict values of new data points. The model decides the class using K-nearest data points using the Euclidean distance, the center point for class is determined by the class with the highest number of datapoints in K-nearest data point. The model has a decent score and performing a 10-fold cross validation improves the r2 and accuracy score significantly by 0.052 and 0.11 respectively. The actual vs predicted plot shows that some of the test data are not similar and there are outlying data points.



Model	R2 Score	MSE	RMSE	Accuracy
KNN	0.777	0.276	0.525	0.658
KNN CV	0.829	0.211	0.460	0.768

## Conclusion

After analyzing the data set and examining the important features using different models, it is seen that GDP, social support and a healthy life expectancy are the top 3 variables that contributes to a country's happiness.

## Unsupervised Learning – Mall customer segmentation dataset

Understanding customers demographic and shopping patterns is an important process for retailers and business owners to stay ahead of competition. By doing so, retailers are able to create targeted marketing campaigns tailored for their customers and also improve overall customer shopping experience. This dataset contains the basic demographics of 200 customers such as gender, age, annual income and the spending score. In this report, I will be using various unsupervised learning models to group customers and understand their shopping habits and examine its relationship to their demographics. The clustering algorithms used are k means clustering and hierarchical clustering.