Dataset Name: LIFE EXPECTANCY

Group Name: Zzz-02

On Campus/cloud: On Campus

Student ID	Student name	Individual Contribution
218047883	Amber Jain	5
218043304	Cedric Quenette	5
218302635	Isaac Gleeson	5
218188385	Miles Danswan	5

* **5** – **Contributed significantly, attended all meetings**
4 – **Partial contribution, attended all meetings**
3 – **Partial contribution, attended few meetings**
1 – **No contribution, attended few meetings**
0 – **No contribution, did not attended any meetings**

NOTE: IF ANY OF THE CELLS IN INDIVIDUAL CONTRIBUTION MARK IS EMPTY ALL STUDENTS WOULD GET 3 MARK BY DEFAULT

Section 1: Brief Summary & ML Problem Formulation (max 2 pages)

Summary of Group Assignment 1

Our analysis of the dataset heavily revolved around the key attribute of *Life Expectancy*, what attributes affect life expectancy and perhaps could be used to produce an accurate forecast.

An analysis into *Living conditions* and the attributes that impact it, was an investigation to determine what factors lead to one country having a higher life expectancy to the next country.
Our analysis when considering the attributes of various disease data(Hepatitis B, Measles, Polio, HIV/AIDS and Diphtheria), Schooling and Body Mass Index let us draw a concise conclusion; that was, countries with higher levels of education provide a higher life expectancy in all ages. Although specifically with children, it was found that schooling increased all health related attributes such as thinness (through schooling that implements/provides dietary options) and increased Body Mass Indices. This study also showed substantially lower rates of diseases, likely to be a cause of school mandated immunisation programs.

From data gathered and investigated, our conclusion was that schooling was a prominent factor in increasing life expectancy by starting with the youngest population to ensure their future is as healthy as possible.
Acknowledging that schooling is an important value in life expectancy, could this attribute be used to predict how long we could live?

How can we apply Machine Learning?

There are many ways we could effectively utilise machine learning techniques to prove our hypothesis or create a beneficial utility to predict or correct data.

In early stages of planning development, an idea was shared that interested the whole team; this was, can we classify developed and developing countries with the data we have and then can we predict the timeline where developing countries could become a developed country and vice-versa, are there countries backtracking in health progress.

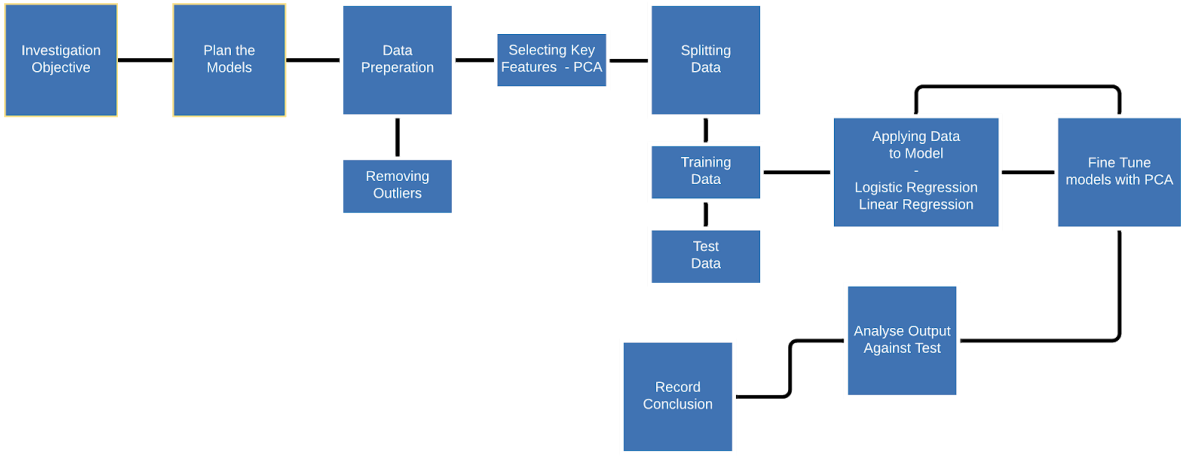
So as a team, our plan is to classify the data into countries that are developing and countries that are developed and then to use forecasting methods to predict when a country might move between the two classifying categories.

By the means of:

- A PCA model, to find where the main variance in the dataset lies, and how many dimensions can be removed, whilst maintaining high levels of accuracy (used to improve model execution).
- Logistic Regression, Using the data given to us (from PCA) to build our own model to determine whether the country more accurately fits in the *Developing* field or the *Developed* field

- Linear Regression, using the key components highlighted by the PCA; We can predict life expectancy.

Machine Learning Flowchart



Section 2: Results and Discussion (max 7 pages)

Feature Selection

Feature selection is the process of selecting the best or highest contributing factors of a dataset, so we can (a) visualise a dataset in high dimensions [4+] and (b) improve the efficiency of our Machine Learning algorithms.

We can decide which features to extrapolate from the dataset with several techniques, allowing us to quantify how each feature contributes to the dataset. The Contribution of each feature is slightly different between techniques, however, the majority of techniques revolve around Variance or Quality/Quantity of Information.

The technique that will be discussed here is Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

One way of selecting features is to create a set of values, representing a range of Maximum Variance throughout the Dataset. PCA applies this idea to prioritise and highlight maximal variation in a dataset, without losing significant information.

PCA uses the Covariance Matrix to calculate the Eigenvalues and Eigenvectors of the dataset. Eigenvalues and Eigenvectors represent a matrix, in a decomposed form. This process is known as Eigen-decomposition, and is one of several matrix decomposition techniques.

Using the Life Expectancy Dataset, we conducted PCA to find subject variance, which allowed us to remove unnecessary or ineffective attributes. The purpose of this is to streamline the relevant attributes for further analysis / prediction.

The results of PCA on the Life Expectancy Dataset show what features account for what proportion of the variance in the dataset, as well as to what degree, each attribute contributes to each Principal Component.

This data can be used in place (as is returned by the PCA algorithm), however, it is more intuitive to represent the data in terms of the original feature space, the apply modelling techniques.

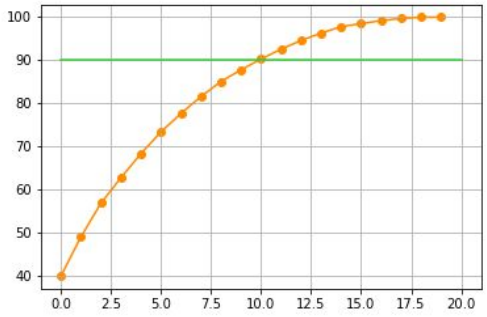


Figure .1 shows the cumulative explained variance for each of the Principle Components. The explained variance is an Eigenvalue, representing the degree of variance for a component.

The explained variance meets our **90%** criteria at the **10th Principle Component**. This is a good result, allowing us to **halve** our Feature space.

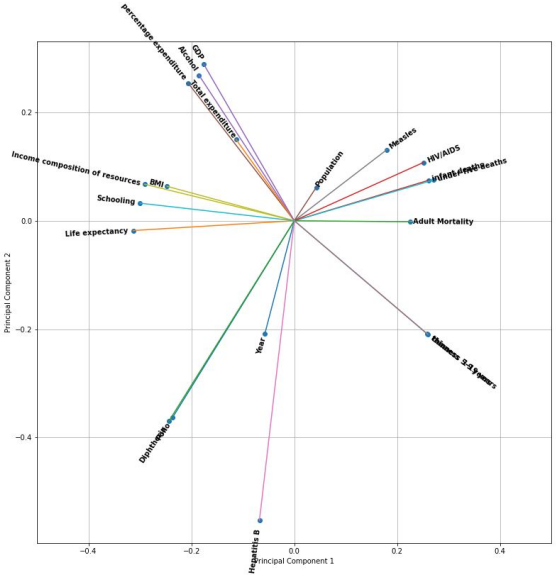


Figure 2. shows how each attribute contributes to the first and second Principle Components (PCs). **Important degrees** of contribution can be found in the first, second and fourth quadrant. For example *Measles*, *HIV/AIDS*, I*nfant & Under Five Deaths*, *GDP*, and *Thinness* all positively contribute significant variance to their respective quadrants.

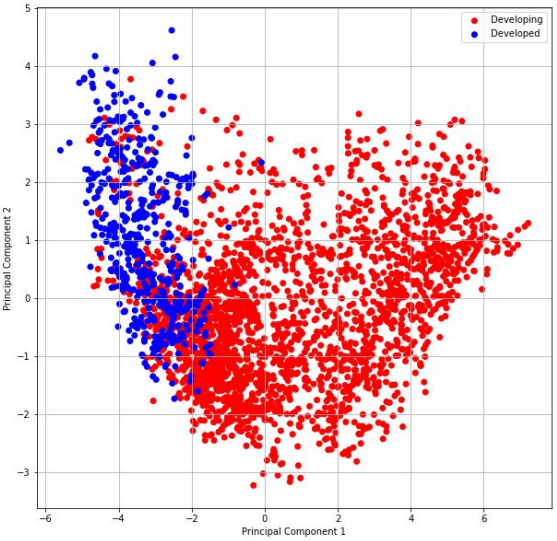


Figure 3. compared the degree of contribution of Developing vs Developed countries in the First and Second Principle Component. We can see that Developed Countries have strong variance surrounding the second PC, however, Developing Countries have both strong or no variance.

These diagrams allow us to see the dataset in different forms. These new forms show how the features of the dataset affect and contribute to the overall variance, which we can remove to achieve the best results for our AI and ML models.

This is a general overview of PCA and it's application to the continuous features of the Life Expectancy dataset. However, the following problems will apply PCA with a subset of the Life Expectancy dataset, to achieve the desired results.

Logistic Regression for Country Status

We have applied a logistic regression[n] model. We have performed PCA on the dataset and used the results as features to train our model. The label is Status of the countries(developing or developed). The goal of the regression was to predict whether a country was classified as developed or developing based on the aforementioned features. This can enable us to insert custom components into the model and predict whether that country is developed or developing. This model can be used practically to make informed decisions about which factors to optimise for to ensure a country gets a developed status.

Preparing the Data

We have dropped columns which were nominal such as Country, irrelevant such as Year or have several nan values such as GDP and created a reduced dataset. We then encoded Status and assigned the reduced dataset as our features and the Status as our label. We then standardized the data which was then fitted into our PCA model. The PCA indicated that out of 12 features, the first 7 accounted for over 90% of the variance in the dataset, thus, we created a dataframe of the transformed first 7 components and used that as our features to train our model. We then split this dataframe into 75% training set and 25% test set.

Training the Model

We fit the features (first 7 transformed PCA components) and labels from the training set in the model and used the testing set to predict the labels based on the test features.

Testing and Evaluating the Model

We created a confusion matrix[n] between the training labels and the test labels.

True Positives Predicted developed, actually developed	111
True Negatives Predicted developing, actually developing	565
False Positives Predicted developed, actually developing	32
False Negatives Predicted developing, actually developed	27

The results are good. Because most countries in the dataset are developing, it is normal to see a high number of true negatives (predicted developing, actually developing). To show the quality of our model, we have calculated its accuracy using the predicted labels against the test labels. We got an **accuracy score of 92%** which shows the good accuracy of our model.

Validation Using K-Fold Cross Validation

To confirm the accuracy of our model, we have also applied K-Fold Cross Validation. this divides the entire dataset into specific slots, and then iterates through each slot while keeping it as the test and the rest of them as a training data set. The mean accuracy score from this technique is also **92%**. This confirms the quality of our model.

Linear Regression to predict Life expectancy

While exploring the life expectancy dataset, we found strong correlation of the data with following variables based on Pearson's Correlation coefficients:

- Schooling
- Under-five deaths
- Percentage expenditure
- Income composition of resources

From finding the correlation of variables with Life expectancy, we moved on to predicting it in this study.

Simple Linear Regression

Life expectancy of Australia for the next 15 years

We started with predicting Australia's Life expectancy which has been on rise since the 1990s based on the WHO data. This was mainly done to kick start our prediction model and follow the key observations. The following graph was the result with the equation of the straight line as :

$$y = 0.12x - 172$$

where 'y' is the dependent variable (life expectancy) and 'x' is the independent variable (year). The graphs display this information precisely.

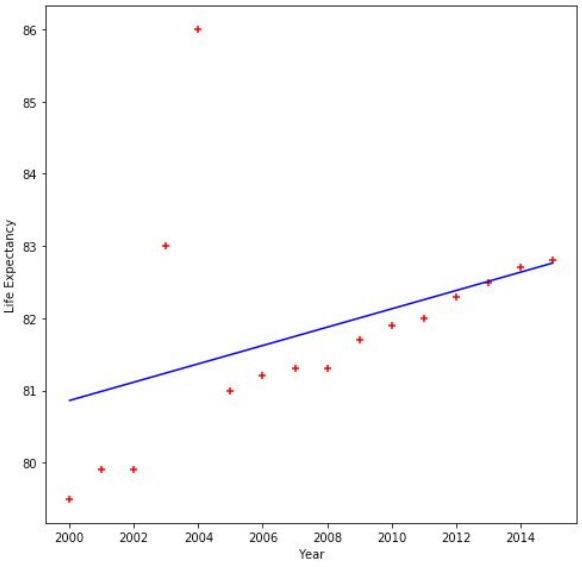


Figure 1: Training (2000-2014)

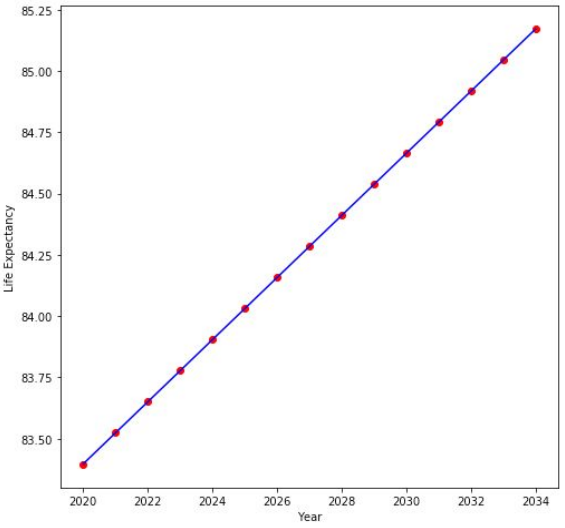


Fig 2: Prediction (2020-2034)

Figure 1 represents how the data was trained from the present to derive the line equation and the life expectancy prediction for Australia is shown in Figure 2.

Multivariate Linear Regression involving Nominal and Continuous data types

The features were selected based on how strongly they correlate with Life Expectancy and from the PCA analysis performed in earlier sections. Following features have been used to predict life expectancy of the world:

- Schooling (Continuous)
- Under-five deaths (Continuous)
- Percentage expenditure (Continuous)
- Income composition of resources (Continuous)
- Status (**Nominal**)

The following equation is being used to train the model and predict life expectancy:
Life Expectancy = (m1 * schooling) + (m2 * Percentage Expenditure) + (m3 * under5deaths) + (m4 * income_comp) + (m5 * Status) + b

The continuous values can be directly used to train models and predict values, which can be done using the 'Sklearn' library's 'LinearRegression()' functionality.

For nominal values like 'Status' of a country, i.e., developed or developing, some amount of pre-processing is required as a nominal value cannot be compared with a continuous value.

This preprocessing is called 'Encoding' which assigns 'developing' and 'developed' as 1 and 0 respectively and helps with training the data for prediction. The data can then be transformed using the 'LabelEncoder()' function to generate a new column with 0s and 1s representing the country status. The next step is using the 'OneHotEncoder()' Library to encode the values and then finally predict the values using Linear Regression.

Sample number	Status	Schooling	Under 5 deaths	Percentage Expenditure	Income Composition of Resources	Predicted Life Expectancy
1	1	10.1	61	71.3	0.25	53.45
2	0	10.1	61	71.3	0.25	55.47
3	0	10.1	61	11.3	0.25	55.37
4	0	10.1	61	11.3	0.50	62.33

The above table contains a few test samples performed with following observations:

- Changing the status from developing country to developed country, keeping all other parameters same increased the Predicted life expectancy by 2 years.
- Percentage expenditure when reduced to 11.3 from 71.3 had almost no effect on life expectancy, which reduced by 0.1 years.
- Income Composition of Resources has the largest impact, when doubled, predicted life expectancy increases by 10 years.

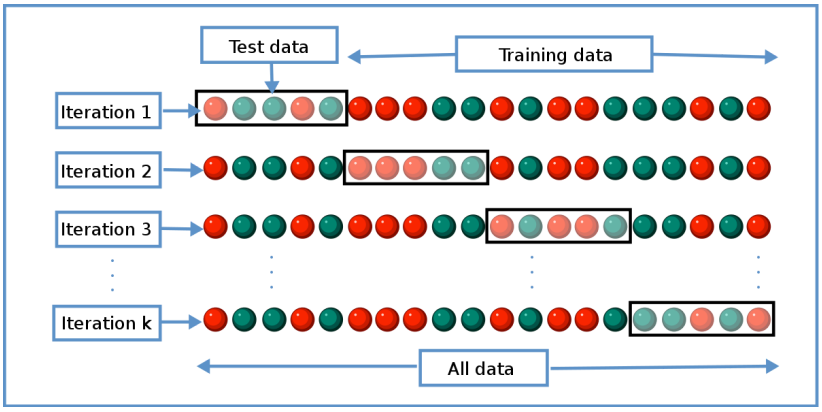
The sample size used here is small just to get a brief idea from the model. To have a substantial test size, divide the data between test data and training data while randomly picking both these subsets.

70% of data is randomly picked for training data and the rest of it for testing the data. So our sample size for testing the data is about 900 samples.

After applying the functions from Sklearn Library, we get an output of an array with predicted life expectancy values. To validate these values, one approach is to compare these predicted values to the actual values which can be done using the accuracy score. We received an accuracy score of 65% for this model.

Though this accuracy score is solely based on one fold, it can be improved by using K fold cross validation.

Validating results using K-fold Cross Validation



The figure displays how the K fold analysis works. We have about 3000 data points where 2100 are the training data and 900 the test data. We have performed a six fold analysis, which means that the

training and test data has been picked up 6 times separately in each iteration to calculate the accuracy score for each fold. The accuracy of the resultant life expectancy has been about **67.18 %.**

Section 3: Conclusions (max 1 page)

PCA was used to evaluate the need for all of the features in the Life Expectancy dataset. We found that 90% of variance throughout the dataset was represented by only 10 components (half the original feature set).

For logistic regression, we have found that the dataset can be used to predict whether a country is developed or developing. Additionally, new data can be used in the model to predict the status of that country. This has practical implications with regards to finding which factors to optimise for to ensure the country is classified as developed or developing.

Furthermore, by looking at the confusion matrix, accuracy score and accuracy score after using K-Fold Cross Validation, we have confirmed the accuracy of our model where it is accurately predicting the status of a country more than 90% of the time.

Similarly, we have predicted the life expectancy of our dataset using Linear Regression.
Based on Status, Schooling, Under five deaths, percentage expenditure and Income composition of resources. This has given us insight into what factors will lead to an increase in life expectancy.

Developed countries are still more likely to have a higher life expectancy and will continue to do so. Schooling has a huge impact on increasing life expectancy in the near future. Under five deaths is still the biggest cause which brings down the life expectancy prediction by a long margin. Although percentage expenditure has negligible impact on life expectancy, spending more income on healthcare has a significant impact on increasing it. Using K-fold Cross Validation, we made sure our model is accurate with a 70% accuracy rate.

We could improve the model by either selecting better features to increase our accuracy score or by using fewer features and trying to maintain our accuracy score. The benefit of using fewer features is that it is easier to make predictions based on new data as fewer data points are needed to get a prediction from the model.

Section 4: References

[1] Datatofish.com. 2020. *Example Of Logistic Regression In Python - Data To Fish*. [online] Available at: https://datatofish.com/logistic-regression-python
[Accessed 24 May 2020].

[2] Medium. 2020. *Confusion Matrix Visualization*. [online] Available at: https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea
[Accessed 24 May 2020].

[3] analyticsvidhya.com. 2016. PCA: A Practical Guide to Principal Component Analysis in R & Python. [online] Available at: https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/
[Accessed 25 May 2020]

[4] support.minitab.com. 2019. Interpret the key results for Principal Components Analysis. [online] Available at: https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/interpret-the-results/key-results/
[Accessed 25 May 2020]

[5] cs.otago.ac.nz. 2002. A tutorial on Principal Components Analysis. [online] Available at: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf [Accessed 25 May 2020]

[6] stats.stackexchange.com. 2010. What are principal component scores?. [online] Available at: https://stats.stackexchange.com/questions/222/what-are-principal-component-scores
[Accessed 25 May 2020]

[7] Scikit-learn.org. 2020. *3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.23.1 Documentation*. [online] Available at: https://scikit-learn.org/stable/modules/cross_validation.html
[Accessed 28 May 2020].

[8] Brownlee, J., 2020. *Principal Component Analysis For Dimensionality Reduction In Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction-in-python/
[Accessed 29 May 2020].

[9] Forum, a., 2020. *What Exactly Is Called "Principal Component" In PCA?*. [online] Cross Validated. Available at: https://stats.stackexchange.com/questions/88118/what-exactly-is-called-principal-component-in-pca
[Accessed 29 May 2020].

Pareek, M., 2020. *Understanding Principal Component Analysis (PCA)*. [online] Riskprep.com. Available at:
https://www.riskprep.com/all-tutorials/36-exam-22/132-understanding-principal-component-analysis-pca
[Accessed 29 May 2020].