ST3189 Machine Learning

3 April 2023

Prepared for:

ST3189 Coursework

Prepared by:

CHEN, PIN-SYUE

Student ID: 200618629

Table of Contents

Unsupervised Learning (World Happiness Report)

Introduction & Background

Happiness is a fundamental human goal, and its importance cannot be overstated. Happy individuals tend to lead healthier, more positive, and well contribute to the society. At the same time, the countries with high levels of happiness are more prosperous, and innovative. All of these further highlighted the need for individuals, businesses, and governments to prioritise happiness and well-being.

In order to measure and understand happiness, we use the data published by the United Nations Sustainable Development Solutions Network, measuring various indicators that contribute to overall happiness, including GDP, life expectancy, family, health, freedom, trust in government, and generosity. Aim to provide policymakers with insights to improve the well-being of their citizens.

Despite the importance of happiness, the substantive issue is that significant inequality exists in happiness levels between developed and emerging countries. Addressing these gaps and promoting happiness for all individuals, regardless of background or socio-economic status, is crucial for creating a more equitable and prosperous world. In this report, we will learn about the distribution of happiness in the world, the relationship between the GDP (economy) and happiness, and the impact of various variables on happiness. Finally, machine learning will cluster countries to gain deeper insights.

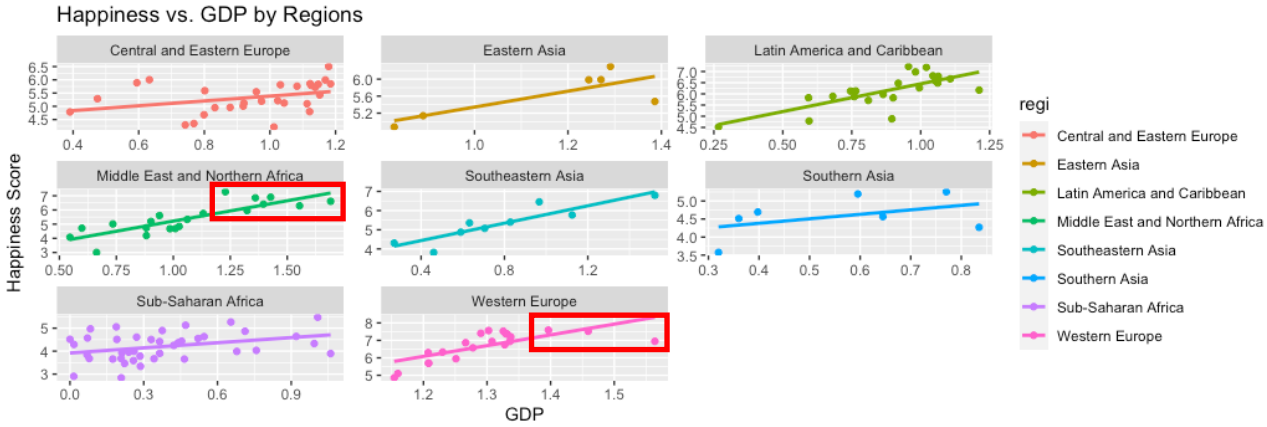
Exploratory Data Analysis

The box plot and world map visualise happiness levels and distribution across different regions. One striking observation is that developed regions have higher happiness than developing countries. To begin with, Oceania, the Americas, and Western Europe, which are highly developed, rank among the top four happiest regions in the box plot and are marked in green on the map. On the other hand, the Middle East and Eastern Europe exhibit mixed happiness levels due to the presence of both developed and developing countries, as shown by the brown areas on the map. At the same time, the box plot reveals that the “Middle East and North Africa” exhibit a wide distribution of happiness levels, representing a wide disparity in happiness levels within that area. In contrast, predominantly developing or undeveloped, Southern Asia and sub-Saharan Africa own the lowest happiness scores. Notably, almost all the orange and red areas on the map are located in sub-Saharan Africa.

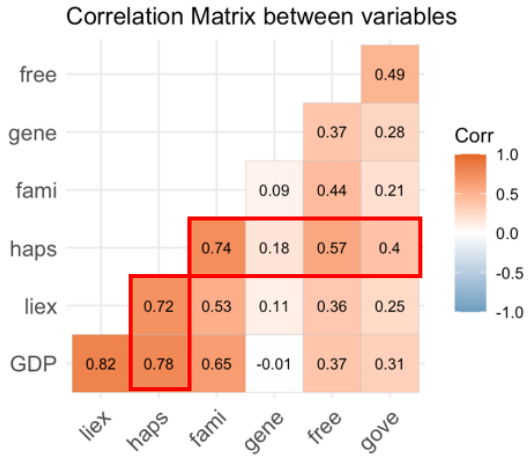
According to Lykken and Tellegen (1996), happiness is a stochastic phenomenon that is determined by a genetic set-point. They stated that people in developed countries have a higher probability of obtaining happiness because they have more resources, like education, wealth, and advanced medical services, to improve their overall well-being. This further verifies our result.

However, for those countries with lower happiness scores, the governments need to review the reasons why the people are unhappy. International organizations can also further investigate the causes of unhappiness and improve them, such as diseases, wars, etc.

Next, correlation analysis will examine the relationship between GDP and happiness across different regions. However, GDP is a measure of a country's economic output, which somewhat relates to its income. To illustrate the relationship between these two, we plotted the GDP score on the x-axis and the happiness score on the y-axis. The higher the score on the y-axis, the happier the people in that country are. Here, we will exclude North America and Oceania due to the limited number of countries. Notably, all lines in the figure slope from the bottom left to the top right, revealing a clear pattern: Countries with higher GDP tend to have higher happiness.



One of the interesting findings corresponds to the theory of, the Easterlin Paradox (Easterlin, 1974), mentioning there is indeed a direct positive correlation between happiness and income among countries, but after it reaches a certain point, the correlation will often no longer be significant, and even higher income will not continue increasing happiness. We can see this clearly from “Middle East and North Africa” and “Western Europe” which have extremely high happiness. When they reach a certain happiness value, they tend to be stable. The rest do not have this phenomenon because they have not yet reached a certain level.



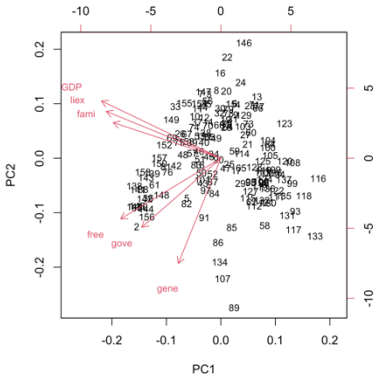
The matrix on the right gives us the information that which are the key indicators affecting happiness. It presents the correlation between variables. A positive correlation value indicates that two variables change in the same direction, increasing and decreasing together.

Furthermore, we found some positive correlations (0.78, 0.74, 0.72, 0.57) between family, GDP, life expectancy, freedom, and happiness. These four variables are the most influential factors in determining happiness levels. In contrast, trust in government, and generosity have lower correlation values with happiness, indicating weaker impacts on overall happiness levels.

Data Modeling (PCA, K-means Clustering, Hierarchical Clustering)

Principal Component Analysis (PCA) is a data reduction technique that helps identify potential correlations and patterns in complex data. The results obtained by PCA will be used as the basis for the K-means clustering later. To begin with, we get the result in R-code, stating PC1 and PC2 are the principal components that capture the 70.03% variation in the data. In other words, these

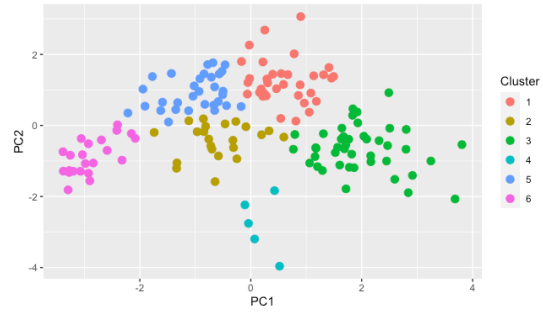
two contain most information from those six variables, which have different internal proportions inside each principal component. Nevertheless, in the biplot, the direction of the arrow indicates the correlation, while the length indicates its strength. Here, the direction of the arrows shows that GDP, life expectancy, and family are positively correlated with PC2, while freedom, trust in government, and generosity are negatively correlated. These six variables will have different strengths of correlations with each principal component. PC1 also works in the same way. After getting the PC1 and PC2, let's move on to K-means clustering.



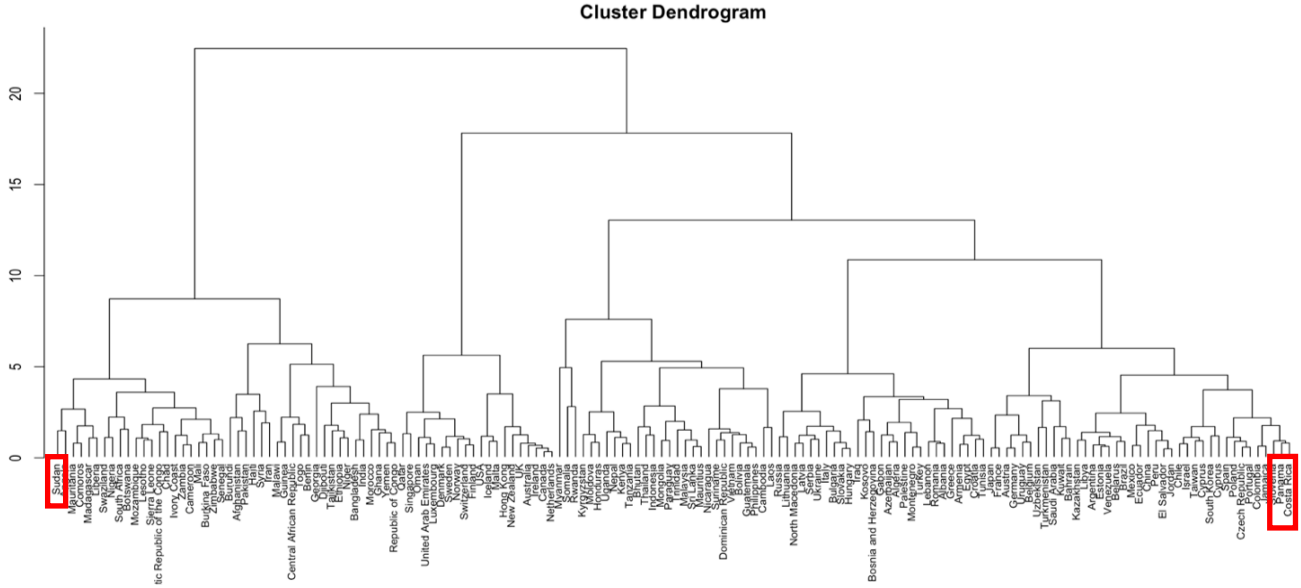
Here, K-means clustering classifies all the countries into six clusters based on their similarities in PC1 and PC2. Each point on the scatter plot below represents a country, and different colours represent different groups. Unlike exploratory data analysis (EDA), which only visualises surface data, K-means clustering uses machine learning to classify similar countries more detailedly.

The countries covered by the six clusters are listed below. They have similar backgrounds and usually lead to similar levels of happiness. The picture on the right is the result of K-means clustering. From the distance and colour, we can observe how the country is affected by the principal components. The significant distance between these clusters reflects the different backgrounds that led to their assignment into different groups. For instance, Cambodia, Laos, Myanmar, Somalia, and Rwanda are all in cluster 4, meaning they are similar in GDP, life expectancy, family, health, freedom, trust in government, and generosity. Moreover, the United Nations also describes them as "least developed countries" (LDCs), verifying they are indeed in a similar condition that belongs to backward countries.

```
Cluster 1: Serbia, Ukraine, Hungary, Montenegro, Armenia, Romania, Russia, Azerbaijan, Bulgaria, Kosovo, Lithuania,
Bosnia and Herzegovina, Croatia, Moldova, North Macedonia, Latvia, Albania, China, Peru, El Salvador, Honduras, Mor
occo, Iraq, Jordan, Tunisia, Lebanon, Palestine, Egypt, Algeria, Turkey, South Africa, Botswana, Gabon, Greece

Cluster 2: Uzbekistan, Georgia, Turkmenistan, Kyrgyzstan, Mongolia, Uruguay, Nicaragua, Suriname, Trinidad, Bolivi
a, Paraguay, Dominican Republic, Guatemala, Chile, Thailand, Malaysia, Indonesia, Philippines, Vietnam, Bhutan, Sri
Lanka, Mauritius

Cluster 3: Tajikistan, Haiti, Syria, Yemen, Iran, Bangladesh, Afghanistan, Pakistan, Nepal, India, Zambia, Democrat
ic Republic of the Congo, Swaziland, Nigeria, Sudan, Zimbabwe, Cameroon, Chad, Lesotho, Mauritania, Mozambique, Dji
bouti, Ghana, Kenya, Burundi, Malawi, Togo, Tanzania, Madagascar, Ethiopia, Sierra Leone, Angola, Mali, Republic of
Congo, Comoros, Uganda, Senegal, Niger, Guinea, Ivory Coast, Central African Republic, Benin, Burkina Faso, Liberia

Cluster 4: Cambodia, Laos, Myanmar, Somalia, Rwanda

Cluster 5: Slovakia, Czech Republic, Poland, Estonia, Belarus, Kazakhstan, Slovenia, Japan, South Korea, Taiwan, Co
lombia, Panama, Mexico, Argentina, Costa Rica, Jamaica, Brazil, Ecuador, Venezuela, Israel, Kuwait, Libya, Saudi Ar
abia, Bahrain, Italy, France, Cyprus, Belgium, Portugal, Spain, Cyprus

Cluster 6: Australia, New Zealand, Hong Kong, United Arab Emirates, Qatar, Oman, Canada, USA, Singapore, Switzerlan
d, Iceland, Denmark, Norway, Ireland, Finland, Netherlands, Sweden, Malta, Austria, Luxembourg, UK, Germany
```



Hierarchical clustering is a method that groups similar objects or data points based on the similarity of variables that affect happiness. This function is like K-means clustering, but it offers an advantage because there is no need to set the number of clusters. It allows for a clearer visualisation of the relationships between countries. The figure below shows the results of hierarchical clustering on this data.



In hierarchical clustering, the x-axis represents countries, while the branch heights on the y-axis represent the distances or differences between clusters. The closer the countries are on the graph, the more similar their backgrounds. For example, Sudan on the left of the image has a different profile than Costa Rica on the right. Costa Rica is next to Panama, meaning both have similar backgrounds, corresponding to what we did in K-means clustering, that these two countries belong to the same cluster.

In conclusion, this research report has a high reference value. It clearly shows the happiness inequality in the world. It also points out that developing countries can increase their happiness by driving citizens' income through economic growth. As for developed countries, they can try to break through the limitations of the Easterlin Paradox to achieve higher happiness by improving other factors that affect the happiness score, like family, freedom or even life expectancy. In machine learning, we classified all the countries into six groups. Among them, countries can emulate each other, and the government can also formulate policies and improve the existing environment to increase the happiness of citizens.

Regression (Life Expectancy)

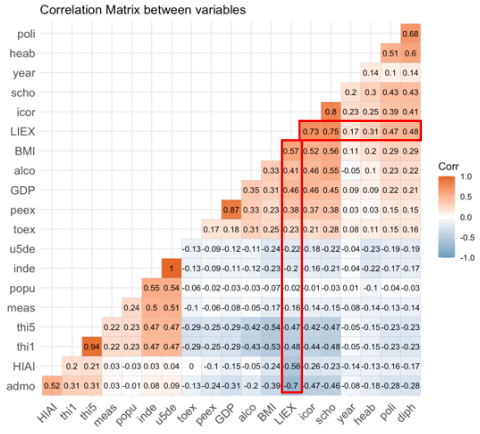
Introduction & Background

Life expectancy indicates people's overall health and well-being in a certain region. A higher life expectancy is usually accompanied by a healthier and longer lifespan, having better conditions to pursue a higher level of spiritual achievements, such as the happiness mentioned in the previous unsupervised learning section. Furthermore, life expectancy is deeply affected by the environment, economy, and society, that’s why it is also a goal and standard that the national government, policymakers, and influencers must refer to and improve.

We used data from the Global Health Observatory (GHO) under World Health Organization (WHO) for a more in-depth assessment and understanding of life expectancy. It includes 22 variables, such as national, social, economic, medical and health-related factors. Having indicators for different aspects also gives us more information to predict lifespan accurately.

Here, the substantive issue is the serious gap in life expectancy in the world, especially the inequality between developed and developing nations. Next, we will also discuss the global life expectancy trend and its relationship with education, GDP, disease, income, and life expectancy. Lastly, a machine learning regression model will be built to use these variables to predict life expectancy. We will also compare the fit between different models.

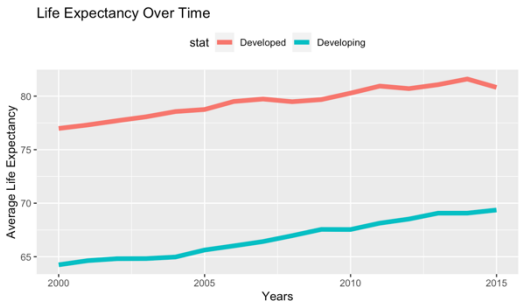
Exploratory Data Analysis



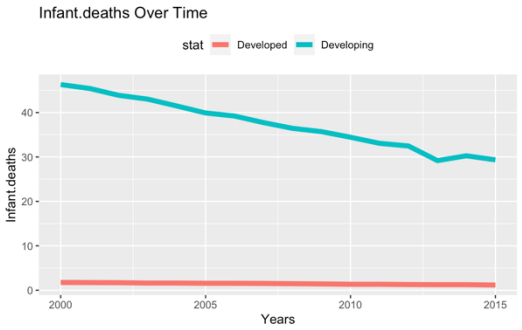
The correlation matrix shows most variables are weakly correlated. For this report, we will focus on the variables that are closely related to life expectancy: Schooling (scho), Income composition of resources (icor), 15~60 Adults Mortality (admo), HIAI, and BMI, as they all have acceptable correlations with life expectancy. However, we must first declare that the high correlation between the two does not necessarily represent a causal relationship. Although years of education and life expectancy are highly correlated, we

cannot guarantee an increase in education will definitely lead to an increase in life expectancy. It may just happen that countries with higher years of education have a longer life expectancy, leading to a high correlation. Thus, we can draw good insights from this report, but the exact cause and effect still need further research.

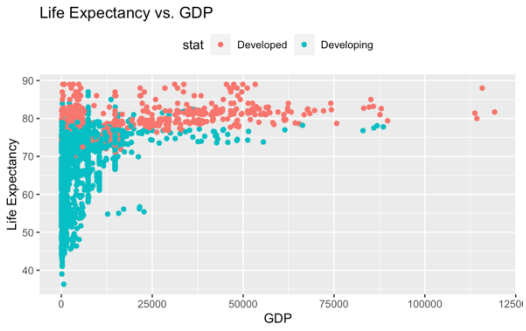
The figure on the right illustrates that the average life expectancy in developed countries is approximately 80 years. In comparison, developing countries have an average life expectancy between 65 and 70 years, indicating a significant gap of almost 15 years. However, there is still a positive growth trend in overall human life expectancy year by year.



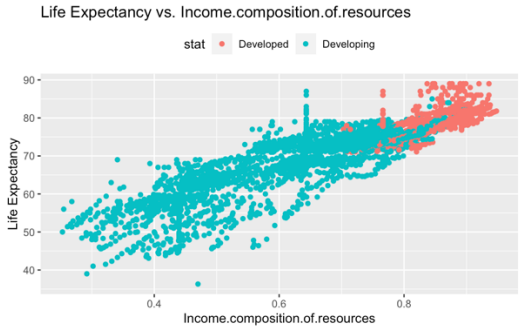
Next, we explore infant death over time, calculated as the number of infant deaths per 1000 population. The line representing developed countries is relatively stable and close to zero. In contrast, the line for developing countries shows a decreasing trend but still with a higher rate of 30 cases per 1000 population. It is necessary to further investigate the high infant death rate and identify the causes, such as insufficient medical resources or lack of vaccine access.



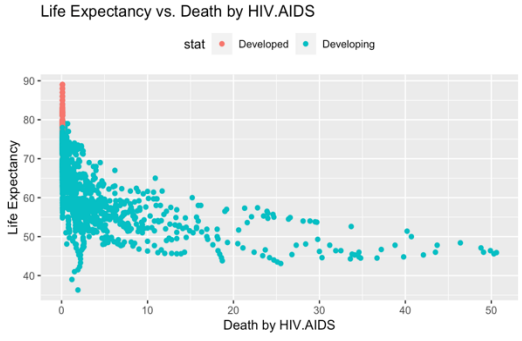
From the life expectancy vs GDP scatter plot, most of the developing countries have a lower GDP and lower life expectancy because the blue dots are distributed on the lower left. However, there are still many red dots with low GDP that have a high life expectancy. In conclusion, GDP may not be the key factor affecting the life expectancy. The scatter plot does not show an obvious relationship or trend between these two.



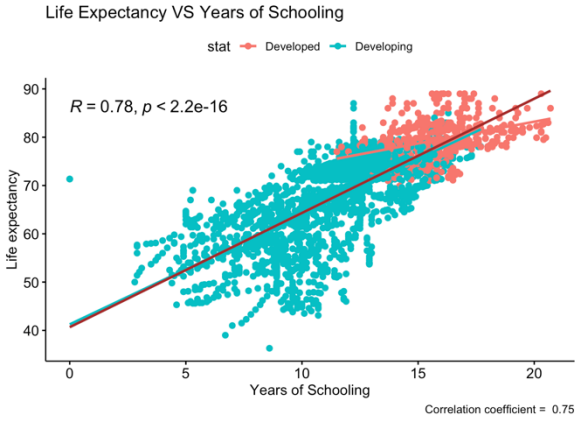
Income composition of resources (ICR) is a subcomponent of the Human Development Index (HDI) that measures the diversity and distribution of a country's income sources. A higher ICR represents a more balanced and diversified industrial structure, making the economy more resilient. The graph on the right shows that ICR has a strong positive correlation with life expectancy, suggesting that improving ICR can higher the life expectancy in a country.



The graph illustrates the deaths per 1000 due to HIV/AIDS (0-4 years). The developed countries have negligible deaths, whereas developing countries have a wide range of deaths from 0 to 50. Therefore, it is imperative for these countries to take effective measures to combat the spread of HIV/AIDS. Further analysis of the environmental factors, lifestyle, medical facilities, and



education level can help devise appropriate strategies to reduce the infection rate and enhance the lifespan and quality of life.



The scatter plot depicts the relationship between the years of education and life expectancy. The graph demonstrates that most developed countries have a high average of 15-20 years of education, whereas developing countries have less than 15 years of education. The brown regression line highlights the correlation between the years of education and life expectancy, indicating that higher education leads to an increase in life expectancy. In fact, from daily life, we can realize that in today's competitive society, it is more difficult for people with low education to win good jobs and live a good life. Moreover, lacking education has a direct impact on perceptions of health, such as belief in traditional remedies rather than modern medical treatment. Therefore, education plays a crucial role in enhancing the overall lifespan of a country or region. Governments and international institutions need to provide the population with the necessary education and knowledge to promote healthy lifestyles, which can lead to an improvement in the life expectancy of individuals.

While there is no single theory that directly focuses on the impact of social and economic factors on life expectancy, many studies have identified valuable conclusions. First, individuals with higher incomes tend to have longer life expectancies due to better access to resources and a higher quality of life. Second, higher levels of education are associated with longer life expectancies because they lead to higher income and healthier lifestyles. Third, access to quality healthcare is a crucial factor in determining life expectancy, as preventative care and advanced treatments can significantly improve health outcomes. The above points have fully verified our EDA results.

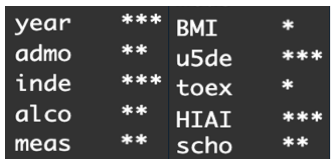
Data Modeling (Linear, Ridge, Lasso, Random Forest, CART)

Machine learning is a powerful tool that uses algorithms and statistical models to uncover associations or hidden patterns from data, which can then be used to make predictions. This report uses five different regression techniques to predict life expectancy based on 21 independent variables. Each model has its advantages and disadvantages, and we will evaluate their accuracy by comparing their RMSE and R-squared values.

Briefly explain these five models, linear regression is a simple algorithm that works well when there is a linear relationship between variables, while ridge and lasso regression is a regularisation technique that helps prevent overfitting. Overfitting means the model performs well on the training data but poorly on the new data. Random forest and CART are decision-tree-based algorithms that recursively divide the data into smaller subsets based on the most important variables.

From the perspective of RMSE, random forest is the best choice among the five models because it makes accurate predictions and minimises the difference between predicted and actual values. In contrast, linear regression is preferable in R-squared because it shows how well the model fits the data and explains the relationship between variables.

	model	rmse	r2
4	Random Forest	1.948867	0.9596562
2	Ridge Regression	2.146315	0.9496403
1	Linear Regression	2.281884	0.9663384
3	Lasso Regression	3.743000	0.8450462
5	CART	3.745520	0.8466512



In the summary of the linear regression model, we have identified several important variables that have a direct impact on life expectancy. The statistical significance level of each variable is indicated by the number of stars on the right-hand side of the summary table, with more stars indicating a stronger impact on life expectancy. Our analysis shows that infant mortality rate (inde), alcohol consumption (alco), measles (meas), prevalence of HIV/AIDS, and years of schooling (scho) all have significant impacts on life expectancy. Furthermore, body mass index (BMI) and government expenditure on health (toex) also have some influence on life expectancy.

These findings demonstrate that life expectancy is primarily influenced by people's health. Therefore, medical institutions should prioritise improving basic healthcare for the population, while governments can increase national health budgets to promote better health outcomes. For instance, providing free regular health checkups can help improve the health status of the population and increase life expectancy.

This report can also help in decision-making, as companies, international agencies, and governments can use machine learning predictions to formulate and evaluate strategies. For example, knowing the importance of education and health, the government can create free educational programs for basic educational levels or provide free education for healthcare program graduates as it is essential to promote better health service and of course contribute to higher life expectancy. Additionally, understanding which factors are important for lifespan also provides a basis for further exploration to extend life expectancy.

Overall, this research has significant practical applications and valuable insight. It sheds light on key factors influencing life expectancy and provides practical guidance for policymakers to pursue strategies that promote greater lifespan and healthier living in their respective countries. Developing countries can reduce serious disparities by using these indicators to increase life expectancy.

Classification (Mental Health)

Introduction & Background

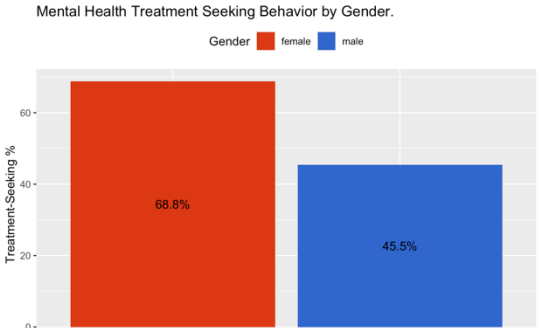
Mental issue is a common cause of suffering in nowadays society. It leads to emotional, behavioural, cognitive, and even thoughts of suicide in severe cases. Mental health problems can be caused by genetic inheritance, environmental influences, and various stressors such as work pressure, interpersonal relationships, and societal pressures. However, mental problems can be effectively treated with appropriate medical intervention to reduce symptoms.

For this study, we used data sourced from Kaggle, which includes 27 variables related to potential causes of mental health issues. It contains information on personal details, daily life, work environment, physical health, and other factors. Our objective is to use this data to build machine-learning classification modes predicting whether an individual should seek medical help for their mental health issues.

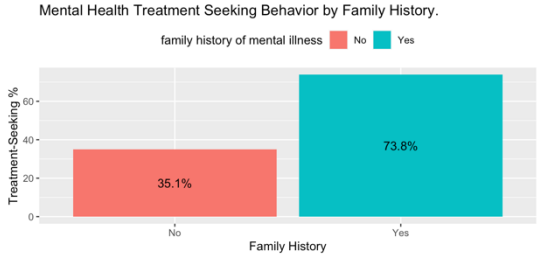
The substantive issue here is a lack of moderate treatment options in the workspace and correct cognition and knowledge about psychological problems. Many people struggle to find the right channels to access treatment or are hindered by a lack of resources in this area. In this report, we explore the distribution of mental health treatment-seeking behaviour across gender, family history,

tech industry affiliation, and age groups. Additionally, we examine how frequently mental health issues interfere with work productivity. Finally, we will compare the performance of different machine learning models in predicting whether individuals should seek treatment for their mental health issues.

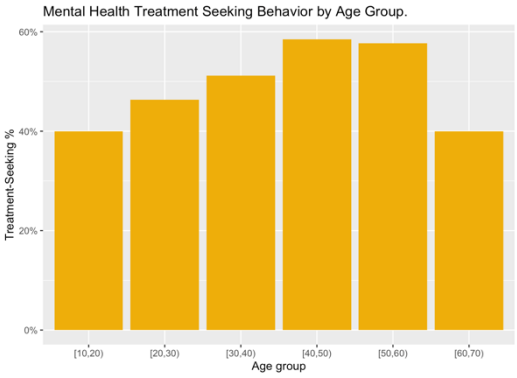
Exploratory Data Analysis



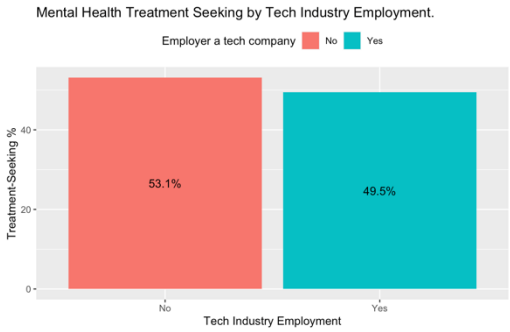
In the three charts below, one notable finding was that 66% of females seek treatment compared to only 45.5% of males. One possible explanation could be the social stigma surrounding men seeking counselling, which may make them less willing to seek help.



Furthermore, nearly 73.8% of people with a family history of psychosis sought treatment. People without a family history may be relatively unfamiliar with this mental issue. The government can make psychological counselling more accessible and educational institutes can promote the importance of psychological treatment.



Next, let's explore the relationship between age and seeking mental health treatment. In the chart on the right, those aged between 40 and 50 on the highest percentages, 59%, who sought treatments. Furthermore, the data also reveals that more than half of people will seek mental treatment between the age of 30 to 60. Enterprises and governments should prioritize addressing mental health concerns for this demographic. Governments can play a role by implementing policies that protect the rights of workers and alleviate pressure, while enterprises can offer mental health resources and support to their employees. By doing so, we can create a healthier and more productive workforce.



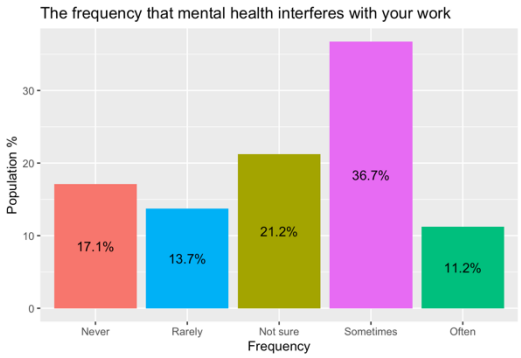
We continue to delve deeper into mental issues related to the workplace. It is often said that employees working in the technology industry are more likely to suffer from psychological problems because the technology industry work environment is relatively high pressure, long hours, solitary work, and goal orientation. These several reasons make stress easy to cause physical and mental problems. However, we generate the chart on the right dispels the myth that the tech sector has higher rates of mental issues, as employees in this industry seek psychological counselling even at a lower rate than in other industries.

Moreover, this data also includes the difficulty of asking for leave due to mental issues. Even after filtering out the "Don't know", it is evident that this difficulty factor still represents a lot. This finding indicates that there is a significant lack of support for mental health in the workplace. Both the

government and companies have ample room for improvement in this area, and measures must be taken to address this critical issue.

Don't know	Somewhat difficult	Somewhat easy	Very difficult	Very easy
557	121	263	95	203

Next, we look into the frequency with that mental health interferes with work. For this analysis, we have classified missing data as "Not sure". The chart shows only less than one in five stated that mental health had never interfered with their work. Nevertheless, we all know addressing mental health concerns in the workplace can improve productivity and well-being. However, there is still a serious portion of the respondents (nearly 50%) who feel that their mental health is sometimes or often affected by work. This highlights the importance of addressing mental health issues in the workplace and providing resources and support for employees to manage their mental health. It also suggests that there may be a need for employers to prioritize mental health initiatives and create a workplace culture that promotes mental wellness.



Data Modeling (Logit, CART, LDA, Random Forest, Neural Network, KNN)

This report has done unsupervised learning (world happiness index), and regression (life expectancy). Finally, we're going to use "classification" to predict whether a person should seek psychiatric help based on the available 22 variables. Since this report is for readers without background knowledge of machine learning, we here briefly describe these three different methods. Firstly, regression predicts continuous values, in the previous part, we predict the average life expectancy among many countries. Secondly, classification predicts discrete values. In this case, based on this person's condition suggest whether he needs to go to the doctor or not. Lastly, unsupervised learning involves finding patterns or relationships in data without any predefined labels. Clustering is a common unsupervised learning technique that groups similar data points. In our first part, we group countries with similar natures and conditions to understand the happiness level.

Here we briefly explain the classification machine learning models we use. Firstly, Logit uses a function to model the probability of a binary response. LDA tries to maximize the separation between classes. Random Forest uses decision trees to classify instances, and the predictions of many trees are combined to make a final prediction. KNN assigns an instance to the class most common among its nearest neighbours. Neural Networks models complex relationships between inputs and outputs by simulating the structure of the human brain. Again, each model has its advantages and disadvantages, and we evaluate their performance by comparing their accuracy and AUC values.

Here we can see that CART and Random Forest perform best. Cart's accuracy is as high as 0.82%, which means it correctly predicts 82% of the samples, while Random Forest has an AUC of 0.87, which is the most effective model for distinguishing between positive and negative examples. We do not discuss the rest of the models, because they are not as suitable as these two. Now that we have a good model,

	model	Accuracy	AUC
3	CART	0.8243243	0.84
2	Random Forest	0.8108108	0.87
1	Logit Regression	0.7513514	0.82
6	LDA	0.7486486	0.82
4	Neural Network	0.7270270	0.74
5	KNN	0.7162162	0.79

we only need to input variables in the future to know whether we need to participate in treatment. This is an excellent example of how to measure a person's psychological state and also as a reference for researchers. Furthermore, the summary of the logit regression model highlights several indicators that have significant statistical significance levels, which play a vital role in determining whether an individual need to seek for mental treatment. Key variables such as family history, work influence by mental health, and benefits provided by the workplace. Therefore, these factors are more important than others.

This mental health report can be used by governments, educational institutions, and companies to shape policies and improve the well-being of their citizens or employees. For instance, our chart on the impact of mental health conditions on work highlights the importance of addressing the mental health of employees to enhance productivity. Additionally, the report indicates that still have some people find it challenging to take medical leave for mental health conditions, which underscores the need for improved access to mental health services. Governments and medical institutions can play a vital role in promoting the necessary mental treatment in the workspace and reducing the stigma surrounding mental health issues in society.

Reference

1. Lykken, D. T., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. Psychological Science, 7(3), 186-189. https://doi.org/10.1111/j.1467-9280.1996.tb00355.x
2. Easterlin, R. A. (1974). Does economic growth improve the human lot? Some empirical evidence. In P. A. David & M. W. Reder (Eds.), Nations and households in economic growth: Essays in honour of Moses Abramovitz (pp. 89-125). Academic Press.
3. McLoughlin, K., & Gough, B. (2021). Exploring the use of mobile health interventions to support weight management in adults: A comprehensive review of the literature. International Journal of Environmental Research and Public Health, 18(5), 2455. https://doi.org/10.3390/ijerph18052455
4. World Health Organization. (2021). Prevalence of obesity among adults, BMI ≥ 30, age-standardized (%). Global Health Observatory data repository. Indicator metadata registry. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3131
5. World Health Organization. (2021, March). Mental health in the workplace. Fact sheet. https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work
6. United Nations Development Programme. (2021). Human Development Index. UNDP Data Center. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI