Final Project: The Prevalence of Depression among Women

Carmen Wang

12/18/2020

Introduction

The project aims to determine factors attributable to the prevalence of depression among women and use statistical learning models to make a prediction on the prevalence of depression.

The report will cover the issue background, information about data used in the analyses, methods and tools in the analyses, analysis results, and the discussion of the next step for future analysis.

Problem Statement

This is a regression problem that takes variables to make a prediction on the prevalence of depression among women. This project intends to determine whether depression is indeed more prevalent in women than in men, and if so, what factors affect women's mental health.

Background

Mental health disorders are more prevalent than one would expect. Researchers from Our World in Data in 2017 estimated that more than 10% of people globally lived with some sort of mental health disorders, and the existing statistics related to mental health are underreported and poorly understood (Ritchie and Roser, 2018). And according to the World Health Organization, depression is one of the leading causes of disability, and suicide is the second leading cause of death among young people.

Moreover, some studies show that most mental health disorders, such as depression, anxiety, bipolar and eating disorders, are more common in women than in men around the world(Ritchier, 2018). Mayo Clinic listed factors that would be associated with the risk of mental health disorders, especially with the risk of depression. According to Mayo Clinic, hormonal changes along with biological factors and cultural factors would contribute to depression in women. Specially, for example, Mayo Clinic states that "women are much more likely than men to live in poverty, causing concerns such as uncertainty about the future and decreased access to community and health care resources", and "women are more likely than men to experience sexual abuse." (Mayo Clinic, 2019) Instead of determining the mental health issue in general, this project will specifically focus on one type of mental health disorders: depression.

Data

Most indicators are reported by the World Bank, and other research institutes; however, as the focus on this project is social and political factors, I only downloaded relevant data, such as regime score, domestic violence, human rights score, etc from Our World in Data, as they organized the datasets by indicators.. In addition to the downloadable data, I also scraped the table indicating the share of women in national parliaments in the most recent election from Inter-Parliamentary Union. In addition, I scraped the list of UN member states from the UN's official site and the continent for each country from a table found on Statistics Time.

The unit of analysis is country-year. The dataset will cover UN member states from 2010 to 2017.

The outcome that would be analyzed and predicted is the prevalence of depression among women, since the focus is depression instead of all types of mental health disorders summed together. The data are measured as a percentage on 0-to-100 scale.

Independent variables include population density (measured by people per sq. km of land area), GDP per capita, human rights score (measured by how much the government protects and respects human rights, on -10-to-10 scale), regime type index(on a scale from -10 – full autocracy, to 10 – full democracy), deaths from terrorist attacks, the rate of violence by intimate partner, whether the country has a law prohibiting or invalidating child marriage, the share of women in parliaments (measured as a percentage on 0-100 scale), women's economic rights (an ordinal variable ranging from 0 to 3) and the average births given per women. While I intended to include human resources data in the mental health sectors, not every country reports them: for instance, the U.K and South Korea never reported it to the WHO, and Russia reported only some of the indicators. It is not reasonable to drop observations due to such missingness when other factors are complete, so I decided not to include the human resources data.

For these variables included in this project, there are still issues that need to be addressed. First, there contains missing data. To address this issue, if there are missing data in the dependent variable: the prevalence of depression, then these observations are not included; if the missingness exists in independent variables, then it is replaced with a value depending on the specific context. The detail will be addressed in steps of data wrangling.

For data wrangling, I first restricted the observations to only UN member states and from 2010 to 2017, because several countries' data were missing after the year of 2017. Then I combined datasets together with the merge function by country and year. As briefly mentioned earlier, I dropped observations that had missingness in the prevalence of depression in women, as it is the outcome that would be analyzed and predicted. For the regime type score, and women's economic rights, I filled the missing data with the country's average. For the deaths from terrorist attacks, since a terrorist attack is not a frequent event for many countries, I filled the missing data with 0. For the variable indicating whether there is a law prohibiting child marriage, the dataset did not cover every year, so in any of the reported years there was a value, then I replaced with 1 and otherwise with 0. Lastly, there is an exception when it comes to how I dealt with missing data: the share of women in parliaments. Since the election does not occur at the same frequency across the world, I used the most recent value that I scraped for all the observations. After the filling process, there were still missing values that were very difficult to address given my available information; thus, I dropped these observations, and dropped duplicate values that might have come at merges. I also created a variable which is the absolute difference between the prevalence of depression in women and the prevalence in men to observe the patterns.

Analysis

To answer the first part of my analytical question whether the prevalence of depression is indeed more common than in men, I calculated the statistics summary (which covers mean, median, standard deviation, minimum, maximum and 25th and 75th percentile) and used scatterplots to analyze the big picture of patterns in the prevalence. The scatterplots demonstrate the distribution of the prevalence of depression in women and men, and I created a visualization capturing the pattern in each continent.

Then to determine the factors contributing to depression among women, before building statistical learning models, I first included a correlation matrix and used a heatmap to visualize the matrix. The correlation matrix and heatmap would help us to determine the most and least correlated factors with the prevalence of depression. The result from the correlation matrix would be a reference to decide which variables would be included in the statistical learning models.

For statistically learning models, I included a linear regression model, K nearest neighbor, decision tree, bagging regressor, and random forest for predictions, and I used mean squared error and R-square as the criteria to choose the most predictive model since this is a regression problem.

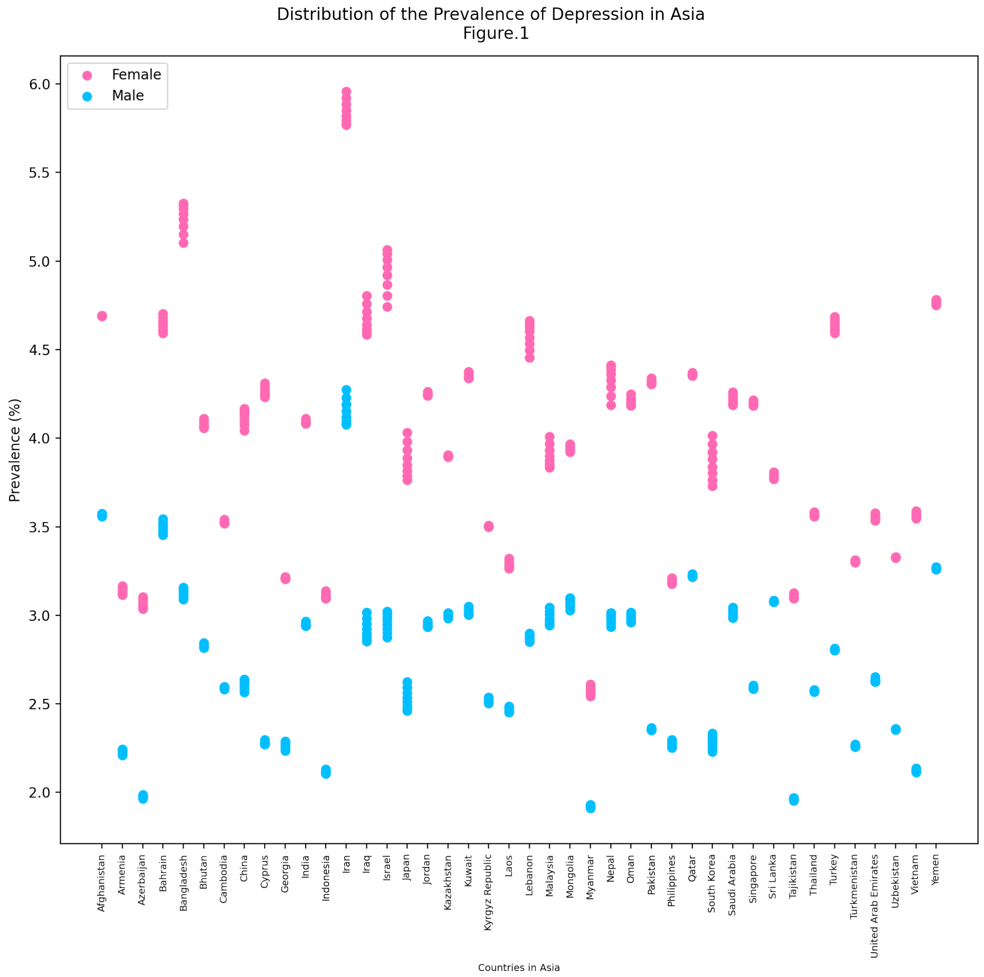
Linear regression is included because it is a classic and simple model that is easy to interpret the relationships between dependent variable and independent variables. We use OLS to estimate the values of the coefficients, and to interpret it, the more predictive the model is, the smaller the sum of all squared errors (calculated as the sum of squared distance between each data to the regressed fitted line) is.

K nearest neighbor (KNN) is useful for classification and regression, so I included it in this project for this regression problem. The KNN algorithm works under the assumption that similar things exist in certain proximity, meaning that similar things should be near each other. KNN is also a simple "model", as it just calculates the distance between data. In my machine learning process, I chose the 10, 15, 20, 25 and 30 nearest neighbors.

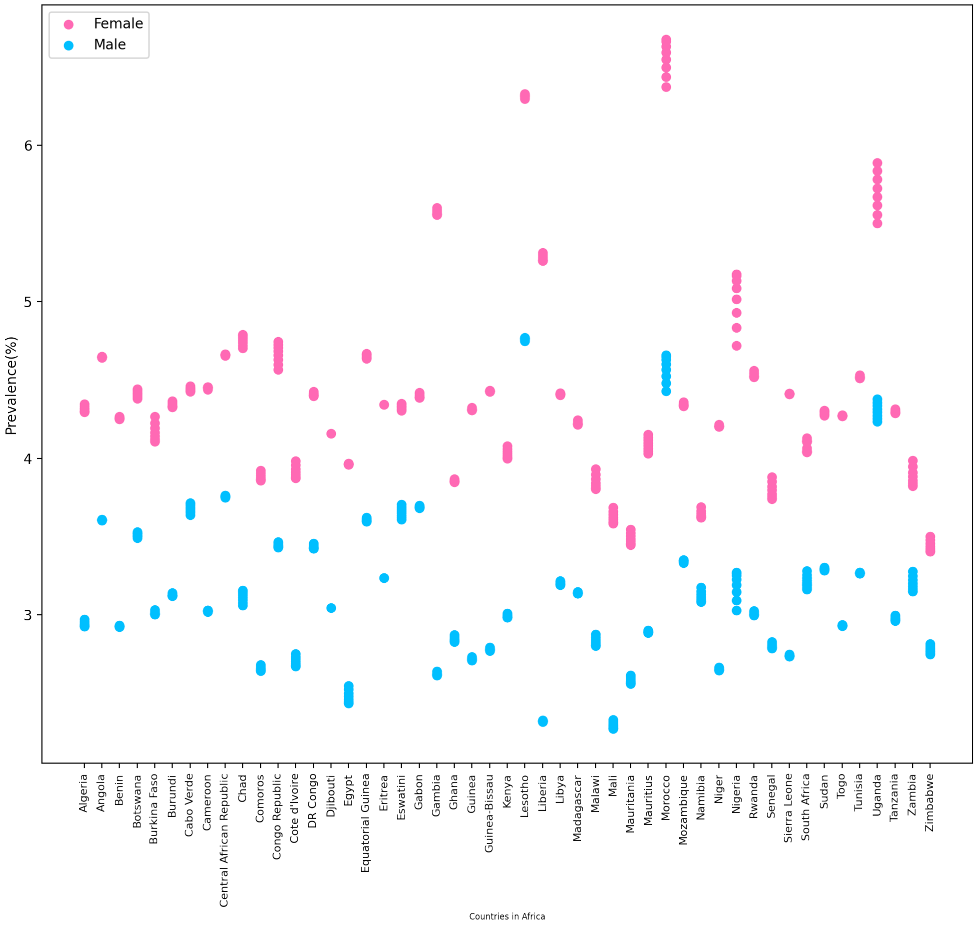
Then I included decision tree. Decision tree is a top-down approach that binarily splits the predictor step by step and it ultimately stops when a criterion is met. The model fit depends on the tree pruning. As the tree gets deeper, the model fit should be higher; nonetheless, deep trees might result in over-fitting. In this model, I chose a tree depth of 2, 3, 4, and 5 to control the maximum tree depth. Decision tree is also easy to interpret the result, but the model fit can be too low or too high.

Bagging as a procedure is building separate trees from a large number of subsets of data then averaging the prediction. By repeating the process and replacing with different subsets, it largely reduces the variance so that it generates an aggregate average result. The result might be predictive, but trees are correlated with each other.

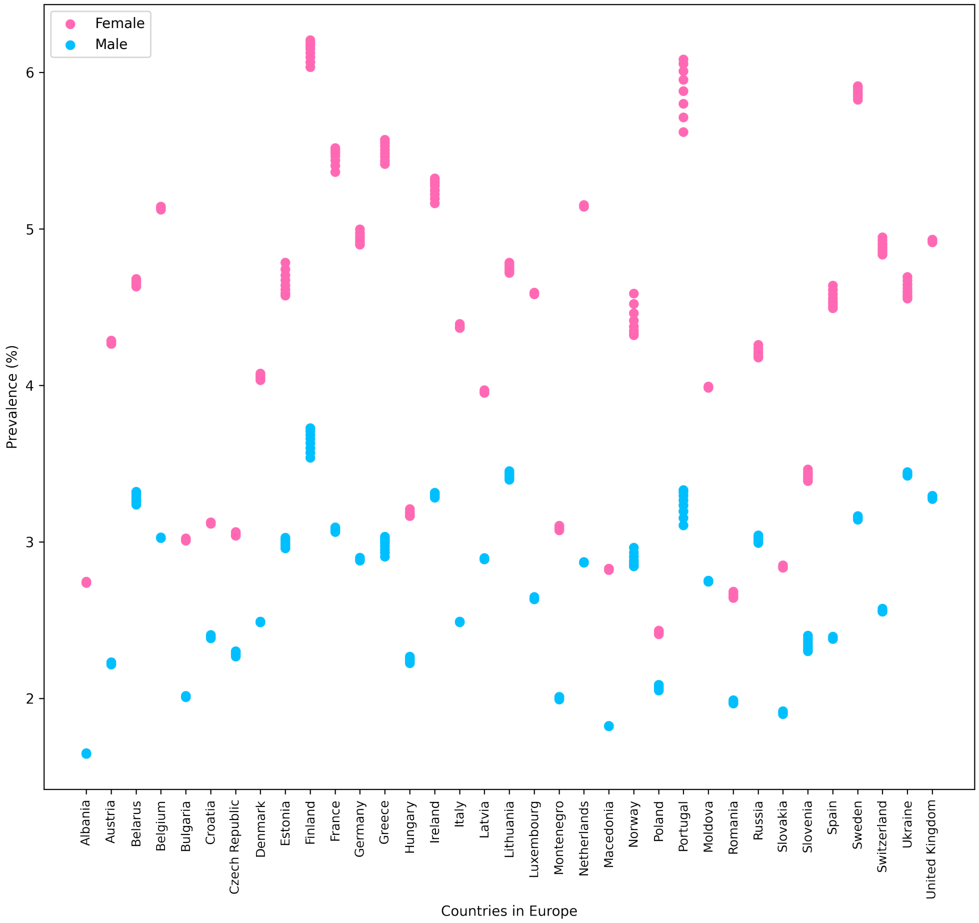
Given the flaws of bagging, random forest was included in this analysis. Similar to the idea of trees in bagging, the random forest approach takes a random sample at each split. I chose 2, 3, and 4 for the maximum tree depth, and 500, 600, 800 for the number of estimators.


Distribution of the Prevalence of Depression in Asia
Figure.1
Female
Male
Prevalence (%)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
Afghanistan
Armenia
Azerbaijan
Bahrain
Bangladesh
Bhutan
Cambodia
China
Cyprus
Georgia
India
Indonesia
Iran
Iraq
Israel
Japan
Jordan
Kazakhstan
Kuwait
Kyrgyz Republic
Laos
Lebanon
Malaysia
Mongolia
Myanmar
Nepal
Oman
Pakistan
Philippines
Qatar
South Korea
Saudi Arabia
Singapore
Sri Lanka
Tajikistan
Thailand
Turkey
Turkmenistan
United Arab Emirates
Uzbekistan
Vietnam
Yemen
Countries in Asia

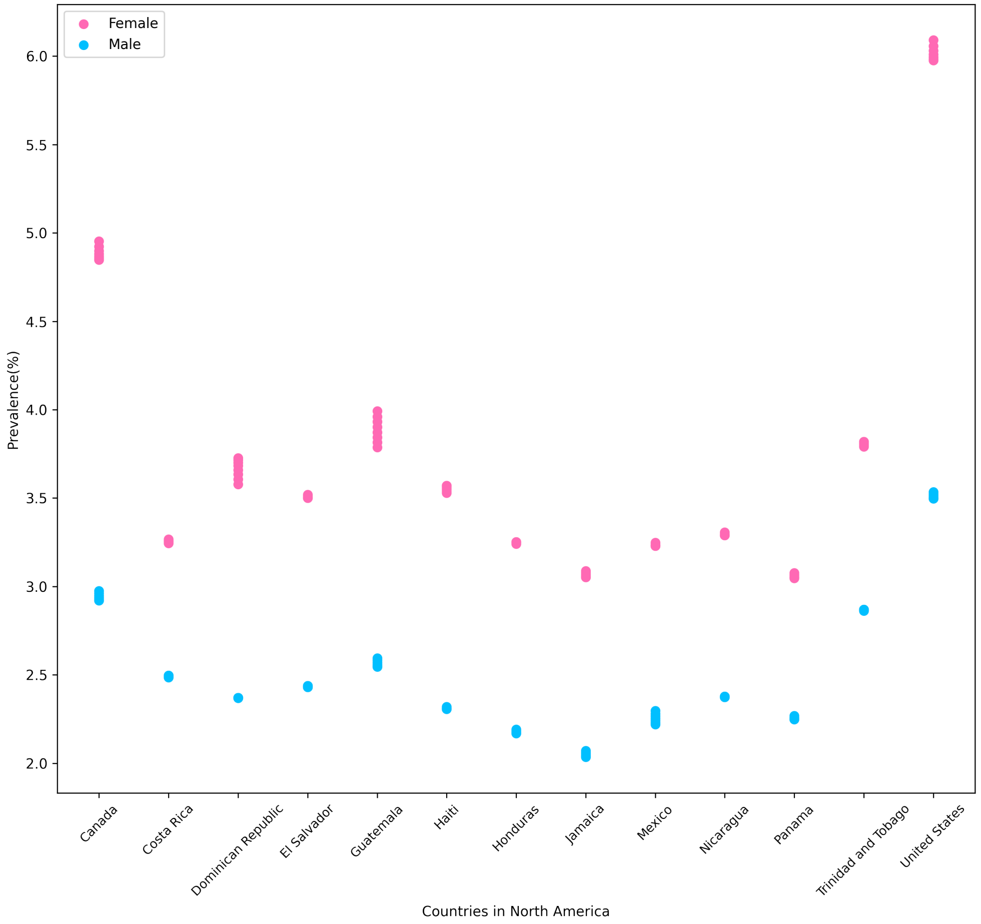
Distribution of the Prevalence of Depression in Africa
Figure.2



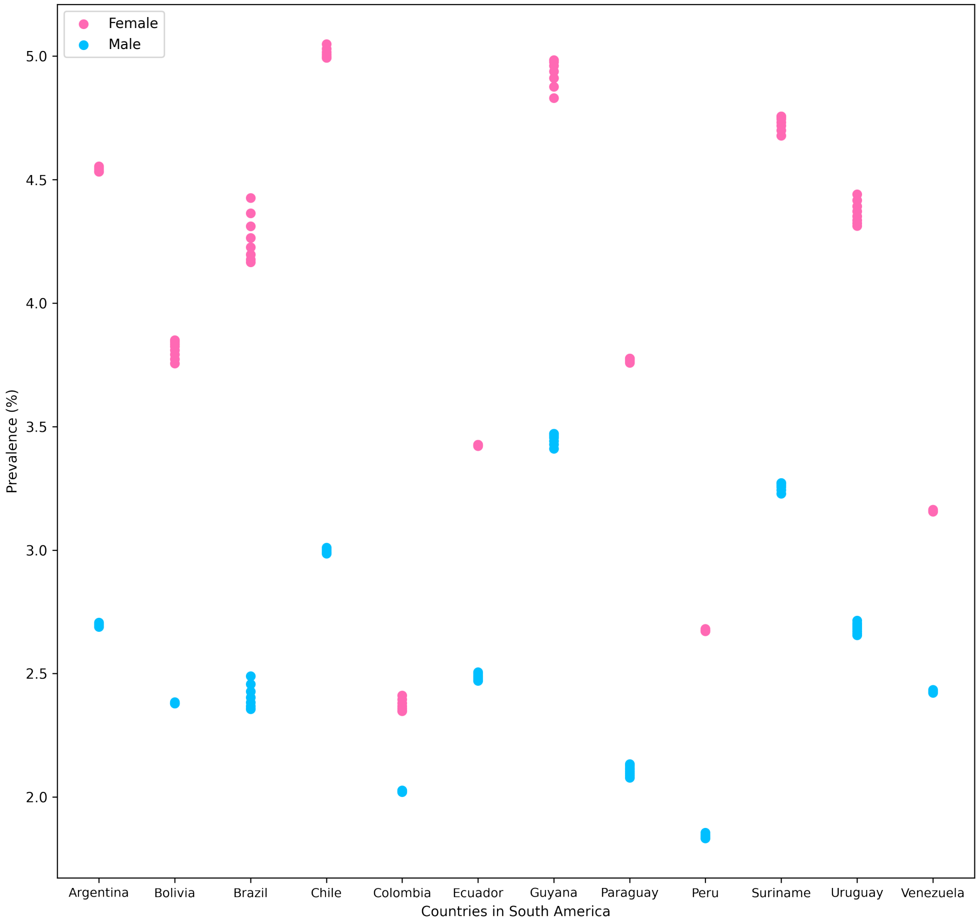
Distribution of the Prevalence of Depression in Europe
Figure.3



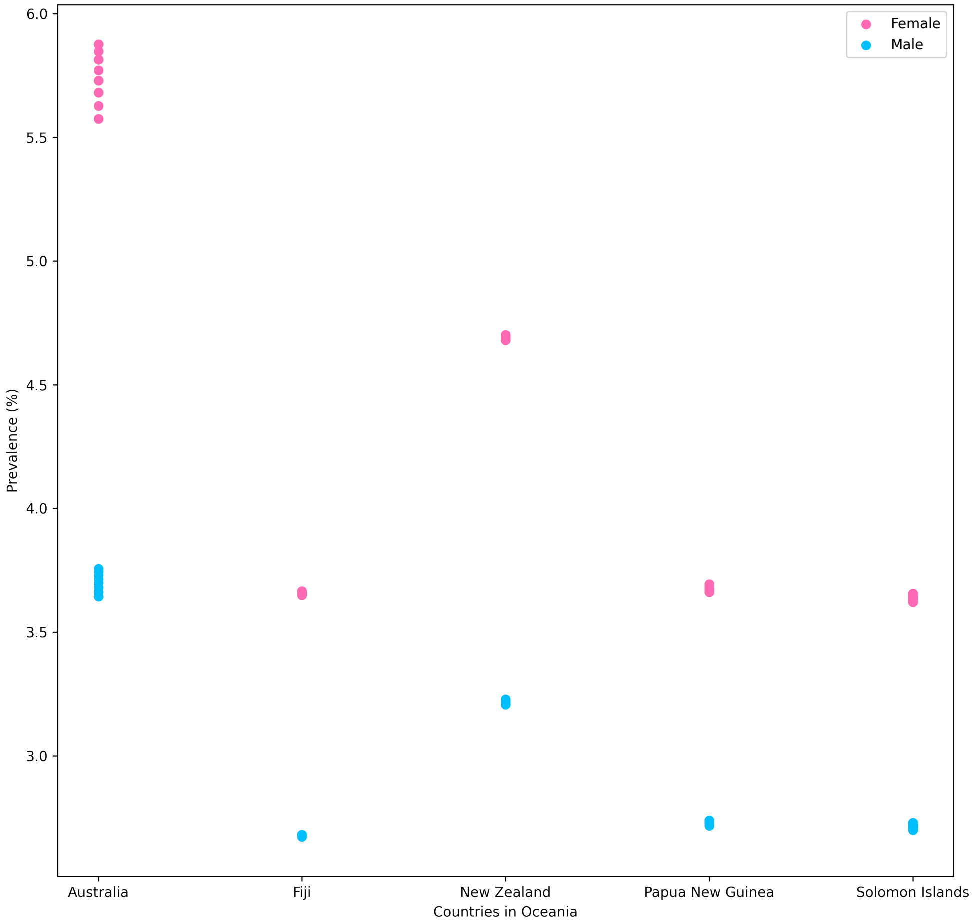
Distribution of the Prevalence of Depression in North America
Figure.4



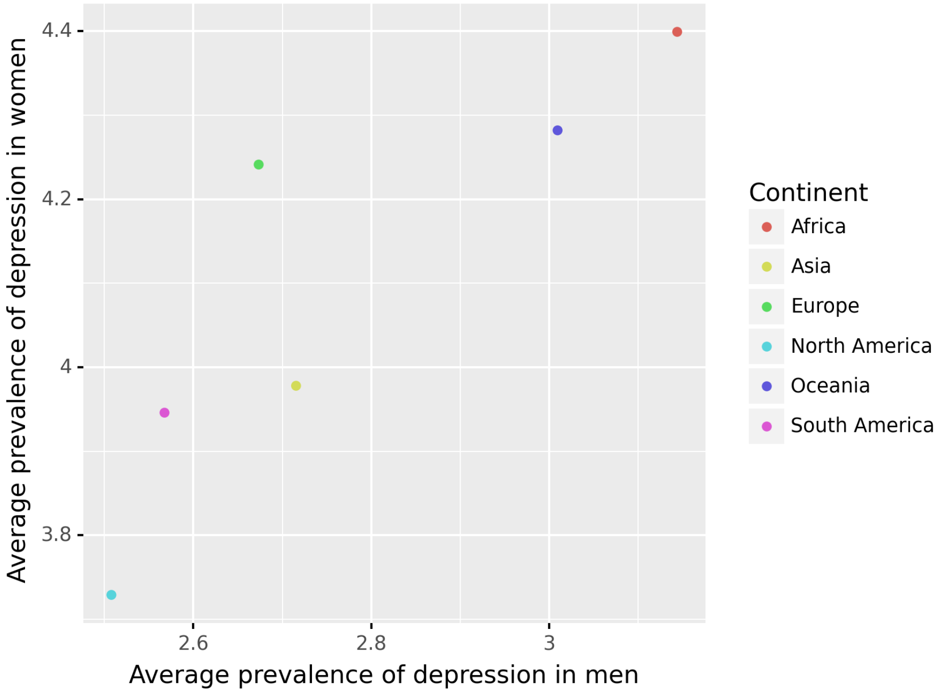
Distribution of the Prevalence of Depression in South America
Figure.5



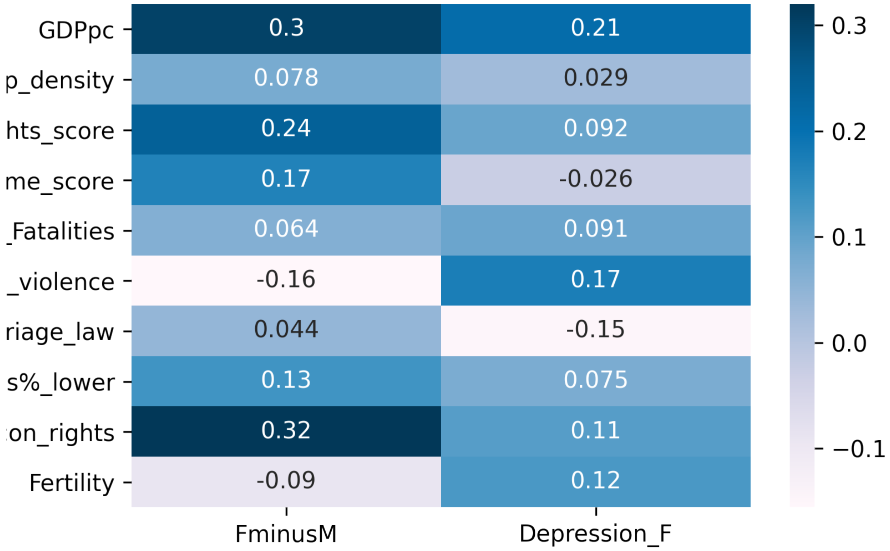
Distribution of the Prevalence of Depression in Oceania
Figure.6



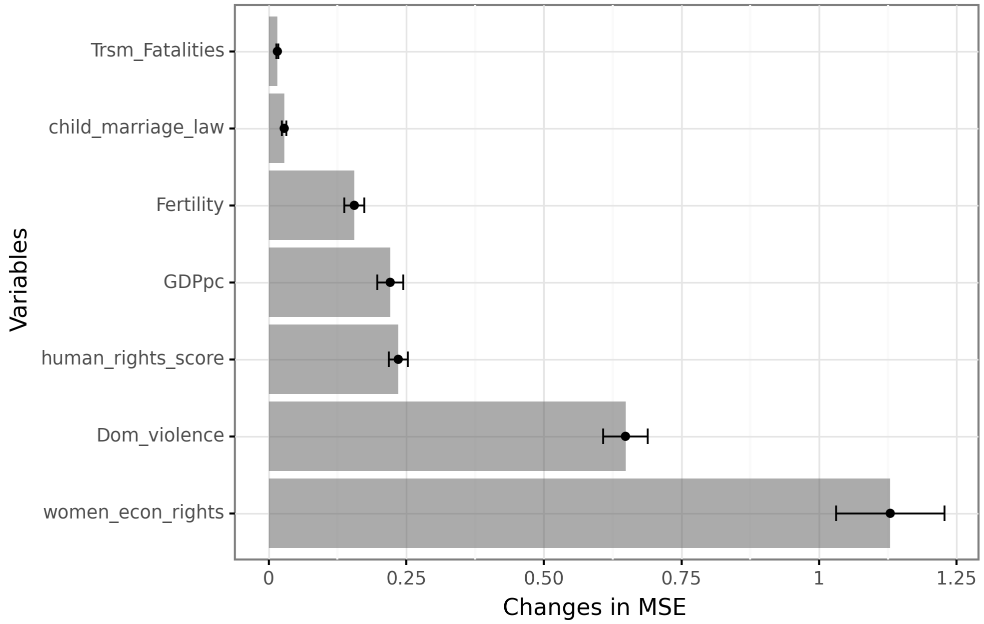
Average prevalence of depression across continents
Figure.7



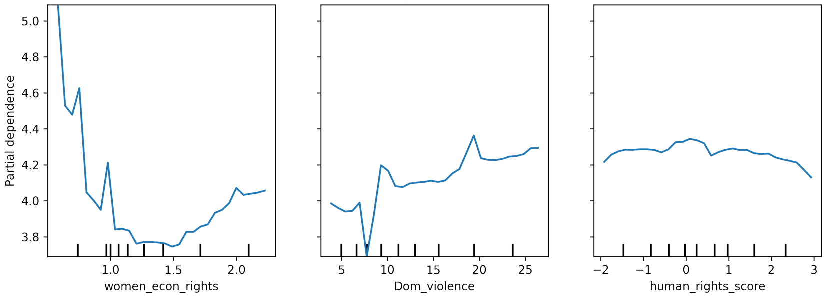
Correlation Matrix
Figure.8



Visualized Permutation Importance
Figure.9



Partial Dependency for 3 most important variables
Figure.10



Results

To answer the first part of the question, yes, depression is more prevalent in women than in men. Based on the statistics summary, across all 1243 observations that account for 156 countries, the prevalence of depression in women is higher than in men. The mean value of the prevalence of depression among women is 4.155% and the mean value of the prevalence of depression among men is 2.820%; the mean for women is around 47% higher than the mean for men. Out of curiosity, I queried the data and found that the highest prevalence rate of depression among women comes from Morocco in 2010, which is round 6.679%, and the lowest prevalence rate comes from Columbia in 2016, which is around 2.348%. On the other hand, among males, the highest prevalence of depression is about 4.77%, which is just slightly above the average value of prevalence among females, and the lowest is about 1.645%.

I visualized the distribution for each continent and referring to Figure.1 to Figure.5 we can observe a gap between genders in each continent, and by observing the overlapped points, the pattern remains pretty consistent over the years from 2010 to 2017. However, if the average prevalence rate is observed within each continent, the difference is not as significant as the statistics indicate, according to Figure.6.

With the appealing statistics and visualized results, it is reasonable to conclude that the prevalence of depression is more common in women than in men across the world, consistent with the previous research work and my assumptions.

Next, based on the correlation matrix, GDP per capita, rate of domestic violence, and fertility (average births given per woman) are the top three factors leading to the prevalence of depression in women. They have a correlation of 0.21, 0.17, and 0.12, respectively (Figure. 7). Whether the country has a law prohibiting or invalidating child marriage has a correlation of -0.15 with depression, which aligns with the intuition that in countries where girls are able (but are mostly forced) to get married the prevalence of depression is assumed to be higher. On the other hand, population density and the share of women in parliaments and regime score seem to be relatively unimportant in examining factors related to depression, based on a correlation of 0.029, 0.073 and -0.027. I don't have a legitimate explanation for the results on these three factors, but I did not expect the share of women in parliaments to be very predictive, because I filled the missing values with the most current ones. Therefore, the three least relevant factors are excluded from the statistical learning models.

According to the results from statistical learning models, the most predictive model is the bagging regressor, which generates a mean squared error of 0.069 for the out-of-sample data and 0.0088 for the in-sample data. With the test data, the mean squared error is 0.0416, and the R-squared is 0.9298. The R-squared is surprisingly high. However, according to the results from permutation importance, the model does have some highly predictive variables. As visualized in the graph (Figure.8), women's economic rights, rate of domestic violence, and human rights score are the three most important variables in the prediction model, and the results do not completely agree with the previously estimated correlation matrix. Deaths from terrorist attacks and law regarding child marriage are the least predictive variables. The results regarding the variable importance make sense to me as follows: based on the partial dependency graph (Figure.9), we can tell that there is almost a quadratic relationship between women's economic rights and the partial dependency, and it seems reasonable since one would expect that the marginal effect of women's economic rights would be higher when the rights are either very low or very high; the results of domestic violence are very intuitive, as just expected; and the marginal effect of human rights score is relatively even across the range. Also, the high R-squared might be resulted from the relatively steady change over the years since the numeric variables are very close to each other within each country. A solution to the problem would be only including one year of data for each country, but then the sample size would be too small to take useful insights from the results.

Discussion

I believe I achieved my objective set in the proposal that I am able to answer my analytical question. In my preliminary analysis, my results did not come out as expected, due to a different selection of sample. At first, I chose to use the prevalence of mental health disorders as my outcome; however, that variable is complicated composite of all types of mental health disorders, so it conceals the real patterns regarding the difference between genders. With additional research, I decided to focus on depression instead, because depression is one of the most common mental health disorders. Therefore, I was able to observe a clear pattern between genders and build a predictive statistical learning model. Although I have some doubts regarding the very high result, at the current stage, I have tried the best and I am satisfied with the outcome. One big challenge that could possibly confine the ability to generate a better result is the data: medical experts believe that global statistics on mental health are poorly understood and reported and mental health disorders are underdiagnosed, so there are chances that there exists huge measurement error in the datasets, potentially biasing the results. Additionally, reiterating what I mentioned earlier, I believe due to the country-year nature, data values within each country are too close to each other, so that the R-square is as high as 92.98%.

If more time is given, I would take a set of actions to expand analysis. First, I would explore more data related to health, such as government expenditure on health, exposure to mental health knowledge, and cost of healthcare, especially for mental health. Unfortunately, data related to health (both physical and mental health) are incomplete and somewhat outdated, so it would take a lot of time working on them; but it might be doable if more time is given. In addition, I would like to expand the dataset with more variables regarding the gender equality issue. Likewise, these data are incomplete and underreported, unless data are retrieved through individual country's sources. Third, I would look into other types of mental health disorders, such as bipolar and substance use, to determine if the conclusion still stand. And if the analysis covers other types of disorders, then more data ought to be retrieved, for instance, I would like to gather data about alcohol consumption, religion, access to drugs, median income, etc. In conclusion, I would like to explore additional data to conduct a more thorough analysis with more dimensions.

Word Count: 2638

Work Cited

Articles

Mayo Clinic. (2019). “Depression in women: Understanding the gender gap”. Published online at mayoclinic.org. Retrieved from https://www.mayoclinic.org/diseases-conditions/depression/in-depth/depression/art-20047725

Ritchie, H and Roser M. (2018). “Mental Health”. Published online at OurWorldInData.org. Retrieved from: ‘https://ourworldindata.org/mental-health’ [Online Resource]

Ritchie, H (2018). “Global mental health: five key insights which emerge from the data”. Published online at OurWorldInData.org. Retrieved from: ‘https://ourworldindata.org/mental-health’ [Online Resource]

Data

Cingranelli & Richards. (2014). WOMEN’S ECONOMIC RIGHTS - CIRI_WECON. The Quality of Government Institute. http://qog.pol.gu.se/data, http://www.humanrightsdata.com/

Gapminder(2017). Fertility Rate. GapMinder.

Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018. http://ghdx.healthdata.org/gbd-results-tool

FEMALE INTIMATE PARTNER VIOLENCE. Institute for Health Metrics and Evaluation (IHME). http://ghdx.healthdata.org/record/ihme-data/gbd-2017-health-related-sdgs-1990-2030

List of Countries by Continent. (2019). Statistics Time. Retrievd from http://statisticstimes.com/geography/countries-by-continents.php

Member States. United Nations. Retrieved from https://www.un.org/en/member-states/index.html

NUMBER OF TERRORIST INCIDENTS (GDT, 2018). Global Terrorism Database (GTD). University of Maryland. https://www.start.umd.edu/gtd/

POLITICAL REGIME (OWID BASED ON POLITY IV AND WIMMER & MIN). Our World in Data. Retrieved from https://ourworldindata.org/human-rights

Schnakenberg, K. E. & Fariss, C. J. (2014). Dynamic Patterns of Human Rights Practices. Political Science Research and Methods, 2(1), 1–31. doi:10.1017/psrm.2013.15 Fariss, C. J. (2019). Yes, Human Rights Practices Are Improving Over Time. American Political Science Review. Advance online publication. doi: 10.1017/S000305541900025X

“Women in national parliaments”. (2019). Inter-Parliamentary Union. Retrieved from http://archive.ipu.org/wmn-e/classif.htm

World Bank. (2017). Population Density. Food and Agriculture Organization and World Bank population estimates.http://data.worldbank.org/data-catalog/world-development-indicators

World Bank. (2017). GDP Per Capita. World Bank, International Comparison Program database. http://data.worldbank.org/data-catalog/world-development-indicators

World Bank. (2017). LAW PROHIBITS OR INVALIDATES CHILD OR EARLY MARRIAGE (1=YES; 0=NO). World Bank: Women, Business and the Law. http://data.worldbank.org/data-catalog/world-development-indicators

Packages

John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55 (publisher link)

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830 (2011) https://jmlr.org/papers/v12/pedregosa11a.html

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2. (Publisher link).

Waskom, M. et al., 2017. mwaskom/seaborn: v0.8.1 (September 2017), Zenodo. Available at: https://doi.org/10.5281/zenodo.883859

Wes McKinney. Data Structures for http://conference.scipy.org/proceedings/scipy2010/mckinney.htmlStatistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010) http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org