Utrecht University

Predicting Mental Health using the BRFSS – CRISP-DM Analysis

Report Portfolio Assignment for Applied Data Science Profile

Anja Tolpekina

5537584

Abstract

Mental health problems can have an enormous impact on individuals' wellbeing and society as a whole. It is important that these problems are prevented or treated as soon as they arise, but this can be difficult to achieve for several reasons. It would be helpful to be able to identify target groups that at a higher risk of mental health problems in a low cost and easy way. A possible solution for this problem, is the using the Behaviours Risk Factor Surveillance System (BRFSS). The BRFSS is the largest health survey in the world, and it is conducted annually. This makes it an ideal candidate dataset to build a model on that can monitor changes in the need of mental health care and prevention. In order to achieve this goal, a decision tree and a random forest model were build on the data from the BRFSS that can predict mental health status. The decision tree model performed slightly better and identified the following factors as being strongly related to the outcome: prior diagnosis of depression or anxiety and the amount of days the individual has recently felt down, depressed, or hopeless. With this study we hope to have laid the groundwork to make an efficient monitoring tool that could enable monitoring the state of mental health across different populations in the Unites States.

Keywords- *Machine Learning, Mental Health, Decision Tree, Random Forest, Data Mining*

Contents

1. Introduction

1.1. Motivation

Over the past few decades there has been increasingly more attention for the prevention and treatment of mental health problems. With good reason, as poor mental health has large implications for individuals' wellbeing and their functioning in society. Depression alone is a leading cause of disability worldwide [1] and more than 50% of people will be diagnosed with a mental illness at some point in their life [2]. The prevalence of mental health problems has large economic implications on a global scale. A lot of money is spent on treating mental health disorders, making prevention and treatment not only an ethical goal but also a financial one. The World Health Organisation (WHO) estimates that poor mental health costs the world economy 2.5 trillion USD a year and this number is predicted to rise to 6 trillion USD by 2030 [3]. These enormous sums are both the result of treatment costs and of reduced productivity of affected individuals.

It is important to give people the correct treatment when mental health issues arise. These problems often get worse, and therefore more expensive to treat, over time. For example, one of the largest outcome factors for people with schizophrenia is how quickly they receive treatment for their first psychotic episode [4]. The longer it takes an individual to get effective treatment, the worse their outcome will be. A similar effect occurs with individuals suffering from depression. People generally recover from a major depressive episode within a few months or one to two years. This occurs even if they do not receive treatment, however, without proper treatment there is a larger risk of relapse, and this can make further treatment more difficult [5]. A large problem is that not everyone is able or willing to receive treatment. Therefore, it would be beneficial to be able to identify individuals with mental health issues. This would enable targeting groups that are at higher risk and increase awareness about mental health among them, set prevention measures in place, and ensure that there are good treatment options available for them.

While identifying such groups of people can be challenging, there is a good candidate method. Data mining is an emerging field within health care and can be used to provide a prognosis and increased understanding of disease classification [6]. The most common use of data mining in mental health, is the use of classification algorithms that aim to select the correct diagnosis for each patient [7]. However, it is not only useful within patient populations, but data mining can also be used to predict mental health status in the normal population [8]. Therefore, the first goal of this project is to use data mining to build a model that can predict an individual's mental health status based on other characteristics. The second goal is to examine this model to determine which factors contribute most strongly to the prediction of poor mental health. The findings could be used for several purposes. For example, governments can allocate funding and focus interventions to target populations that are at higher risk for developing mental health illnesses. In addition, greater understanding of the causes of poor mental health, can help improve treatment.

1.2. Background

The Behaviours Risk Factor Surveillance System (BRFSS) is the largest ongoing telephone health-survey in the world [9]. As the BRFSS is both expansive in the number of respondents and the different mental and physical health related topics it covers, it is a good candidate to build our data mining model on. This survey was established by the Centres for Disease Control and Prevention (CDC), a federal organization in the United States, tasked with the role of national public health agency. The organisation was founded in 1946 and since then its main goal has been to protect public health and safety [10]. To achieve this objective, the CDC endeavours to control and prevent disease, injury, and disability. Their focus is on infectious disease, food borne pathogens, and environmental health. The organisation studies non-infectious diseases, are as well, such as chronic diseases caused

by obesity, smoking, sedentary lifestyles, and other risk factors [11]. In more recent years, the CDC has expanded their scope of health to include mental health in addition to physical health.

The CDC has always been a science-based organization and with the emergence of modern technologies they are becoming more data-driven as well. They have established, maintained, and shared several large datasets and survey systems, the BRFSS being one of them. Findings from the BRFSS have been used for targeting and promoting healthy behaviour and it has proven to be a powerful tool [12]. Thus far, it has mostly been utilized in the prevention and understanding of physical disease, but it could also be used to further the understanding of mental health. Some interesting results have been published already. Women caregivers have been found to have significantly worse mental health than their male counterparts [13] and another study found that respondents with serious psychological distress were ten times as likely to receive treatment for mental health problems[14].

However, no comprehensive study has been published using the BRFSS to examine the risk factors for poor mental health in general as of this date. The objective of this project is to gain insight into the factors that are associated with mental health disorders by studying this dataset. Most studies that have been published on this subject, have had a priori hypotheses about risk factors or subpopulations at risk. While these uses certainly have their merits, the BRFSS data gives an opportunity to create a better understanding of which behavioural risk factors, socio-demographic factors and other possible factors most strongly influence mental health and which do not. Therefore, in this project, no assumptions will be made beforehand, to get an unbiased overview of the relationship between mental health and other information available in the BRFSS.

1.3. Research Objectives

In this project, the aim is to build a comprehensive model to help understand the risk factors for poor mental health and to provide possible targets to help improve and prevent mental health problems among the population. In order to achieve this goal, the relevant risk factors in the BRFSS will be utilized to build a model that predicts the respondent's mental health status. The outcome variable that will be used to quantify mental health status, is whether an individual is receiving medicine or treatment from a health professional for an emotional problem. Such a model has the potential to give us insight into which factors contribute strongly to mental health and which do not. Ideally, to achieve this goal, a model that can be easily interpreted should be chosen.

While many different studies have been published that attempt to predict mental health status of individuals, only few could be found that use the BRFSS. Those that do, only look within certain subpopulations such as caretakers [13]. This project aims to broaden the scope and look at the entire population to get a birds-eye overview of mental health status. As the BRFSS is already carried out every year in every state of the United States, a prediction model based on it, could be an ideal tool for monitoring the state of mental health across different populations. This model could be used on both national or state level to enable government to identify troubling developments and intervene as necessary.

Previous studies that used data mining to predict mental health status on different data sets, have found models with accuracy scores ranging between 72-82%[15]–[17]. The goal of this project is to find a model that can perform at least as well. Therefore, the first success criterium is to create a model that can predict whether an individual is receiving mental health treatment with an accuracy of at least 75%. The second success criterium is to identify the factors that have the strongest relationship with the need for mental health treatment.

1.4. Research Methodology

The goal of the project is to gain more understanding of the risk factors for poor mental health. This will be achieved by building a model that can predict whether a person is receiving

treatment for mental health conditions based on other factors. Furthermore, this model can give us insight into which risk factors are associated with the need for treatment. There are several factors to consider when selecting a model. Firstly, to achieve the goal of this project it is important that the model is interpretable and not a 'black box' model. While support vector machine (SVM) models and neural network (NN) models can be highly valuable and have been used in previous studies to predict mental health status [16], [18], they can be difficult to interpret and will therefore not be used in this project. In addition, it is important to consider the assumptions of models. Another type of model that has been used in previous research is the naïve bayes classifier [15], [17]. This is an interpretable model that performs well when the assumption of independent predictors is not violated, however, this can be very difficult to achieve. A model that is interpretable and would not have its assumptions violated in this case, is the random forest classifier. It has been used before to predict mental health problems in adolescents and to build a monitoring system for physical and mental health[16], [19].

A random forest classifier is suitable for this project as the goals is to predict class assignment (whether an individual is receiving treatment, or they are not). The only disadvantage is that random forest classifier can be biased when dealing with categorical variables and those make up a large part the dataset[20]. Decision trees, however, don't have this problem. They are well equipped to handle both numerical and categorical data, do not require normally distributed data, and are very intuitive and easy to interpret. A random forest classifier fits multiple decision trees on different sub-samples of the dataset and uses the results thereof to build a more accurate model with less over-fitting. As both these types of models seem to have advantages and disadvantages as pertains to this project, the decision was made to build both models and compare their accuracy to determine which one is most suited to this particular case.

To achieve the project goal and build the model, the steps of the Cross Industry Standard Process for Data Mining (CRISP-DM) model are followed. This is an open standard process model that provides a structured methodology for data mining projects [21]. The CRISP-DM consists of the following steps: business understanding, data understanding, data preparation, modelling, evaluation and deployment. All these steps are carried out for this project; however, they are not described in this sequence in this paper. The following steps are taken for this project. First, the data is extracted from the source and loaded into Jupyter Notebook 6.4.8. Following, the data is checked for possible quality issues and cleaned. Then, the data is prepared for modelling. As mentioned prior, two different classification models are made, and their predictive performance is compared to ascertain which one is the best fit. Both a decision tree and random forest model are built on a part of the data and afterwards tested on a different part to ensure reproducibility and prevent an overfitted model. The model outcomes are evaluated using different accuracy measures and the best one is selected to be interpreted and the outcome discussed.

2. Methods

2.1. Materials

The dataset used in this project is the BRFSS dataset from 2015 which was originally published by the CDC [9]. There are annual BRFSS datasets available on the site of the CDC as recent as 2020. Unfortunately, due to technical difficulties pertaining the format of the files, they could not be used directly for this project. Instead, the data set was sourced from Kaggle, where a CSV-formatted version of the survey from 2015 was available[22]. In addition, a text-file containing information on each variable in the files was extracted from the site of the CDC as well [23].

The dataset consists of 330 columns and 441,456 entries. Every entry represents one respondent and 375,059 (85%) of them have completed the interview. The majority of the columns has integer values and the remaining seven are character values. Most of the integer columns represent coded data and are therefore either ordinal or nominal data, but there are also a few interval and ratio variables such as weight, height, and number of children. Four of the variables that

have character values, hold information about the date of the interview and could easily be transformed into integer values. The remaining three character variables have ordinal values.
For further description of the dataset, the codebook published by the CDC can be consulted[23]. Data exploration was performed by visually inspecting the data, creating figures such as histograms and scatter plots and using functions to examine the individual values of variables. During the initial data exploration, it became clear that there was a large percentage of missing values. While this seemed at odds with the fact that 85% of respondents had completed their interview, this is because the BRFSS interview consists of both core modules, which are included in every interview across different states, in addition to more specific modules that are included in certain states but not in others.

2.2. Data Preparation

2.2.1. Data Selection

The first step in the data preparation was to select the data that would be used to build the model. Decisions within this process were made based on the relevance of the variables to the research question and on the data quality. From the 330 variables and 441.456 records a subset was made consisting of 64 variables and 5957 records. In the following section, this process will be described in more detail.

Firstly, the BRFSS interview is not the same in every state. As is mentioned in the previous section, there are certain core modules, which are included in every interview, in addition to more specific modules that are included in certain states but not in others. One of these modules is the 'Anxiety and Depression' module which consists of questions such as: 'Over the last 2 weeks, how many days have you had little interest or pleasure in doing things?'. As the goal of this project is to get a better understanding of the risk factors for poor mental health, the variables corresponding to this module were all included in the data selection. This particular module was only included in the interviews conducted in the state of West Virginia. In total, there were 5,957 records obtained in this state and all records from other states were excluded as they did not contain the 'Anxiety and Depression' module. This decision prompted the removal of the 'State' variable as well, as this became uninformative.

Furthermore, there were 12 variables which only contained the value 'HIDDEN'. For different reasons the data within these columns was not displayed in the published dataset. These variables where all excluded, in addition to variables that were judged to hold little meaningful information regarding the data mining goal. Examples are variables regarding the date of the interview and questions such as 'Is this the correct phone number?'. Variables with a high number of missing values were excluded for practical reasons. These variables were generally questions that belonged to specialized modules that were not assessed in West Virginia or follow-up questions such as 'Are you currently taking medicine for your high blood pressure?' following the question 'Have you ever been told that you have high blood pressure?'. Furthermore, there was a large number of calculated variables which held practically the same information as other variables in slightly different forms ('Reported weight in pounds' and 'Reported weight in kilograms'). In such cases, the variable that was the most informative or practical was chosen and the others were excluded to avoid multicollinearity. These steps collectively resulted in a large decrease in the data volume. While, this is often undesirable, in this case it had the benefit of less technical constraints, as computing a model on the original dataset would have been very time consuming.

2.2.2. Data Cleaning

To clean the dataset, the following factors were considered: missing data, duplicate data, outliers, structure issues. Firstly, all records with missing data were excluded. Instances containing values which code for 'Refused' or 'Don't know' were excluded as well, because they were not

informative in relation to the research question. This decision decreased the 5,957 records to 1766. Instances of missing data were checked whether they could be reasonably imputed, however, upon inspection this was not expected to produce valid results. As the number of records was still sufficient for modelling, the decision was made to exclude all instances containing missing values and to not attempt imputing any. In addition, the data was checked for duplicate instances but there appeared not to be any.

Furthermore, the data was checked for outliers and values that did not belong in the dataset. A large part of the selected variables was either categorical or ordinal. These were visually inspected for incorrect values or strange distributions with histograms and by printing all the unique instances for every variable. To examine the data in the interval variables, boxplots were made and inspected. A few variables had some visible outliers but upon closer inspection, these were assessed to be not out of the scope of realistic situations (such as being 209 cm tall or eating seven pieces of fruit a day). Therefore, they were kept in the dataset for analysis.

Lastly, there were two structural issues that had to be fixed. In twelve columns that appeared to be interval scale, the number 88 was used to code for zero. For example, 'How many children less than 18 years of age live in your household?' would have the possible values 1 – 87: number of children, 88: None. In these cases, the value 88 was replaced by 0 to enable valid values for modelling. In addition, 'Yes/No' variables were coded as Yes: 1, No: 2. This was changed to No: 0, Yes: 1 for the same reason.

2.2.3. Data Construction

There was little data construction necessary as the dataset was very comprehensive and already provided some calculated variables that were helpful, such as BMI and whether the interviewed individual met the criteria to be a binge drinker. The only data construction performed, was One-Hot Encoding for the categorical and ordinal variables as this is required by most modelling techniques to yield accurate results. Dummy variables were created for the categorical and ordinal variables that had more than two possible values (as these were already encoded as 0/1).
After performing all the steps described above, the result was a clean dataset formatted in a Pandas dataframe and ready for modelling. The variables of the selected subset of data are shown in the Appendix (Table 2) and more details about these variables and all the other ones of the original dataset, can be found in the codebook pdf-file.

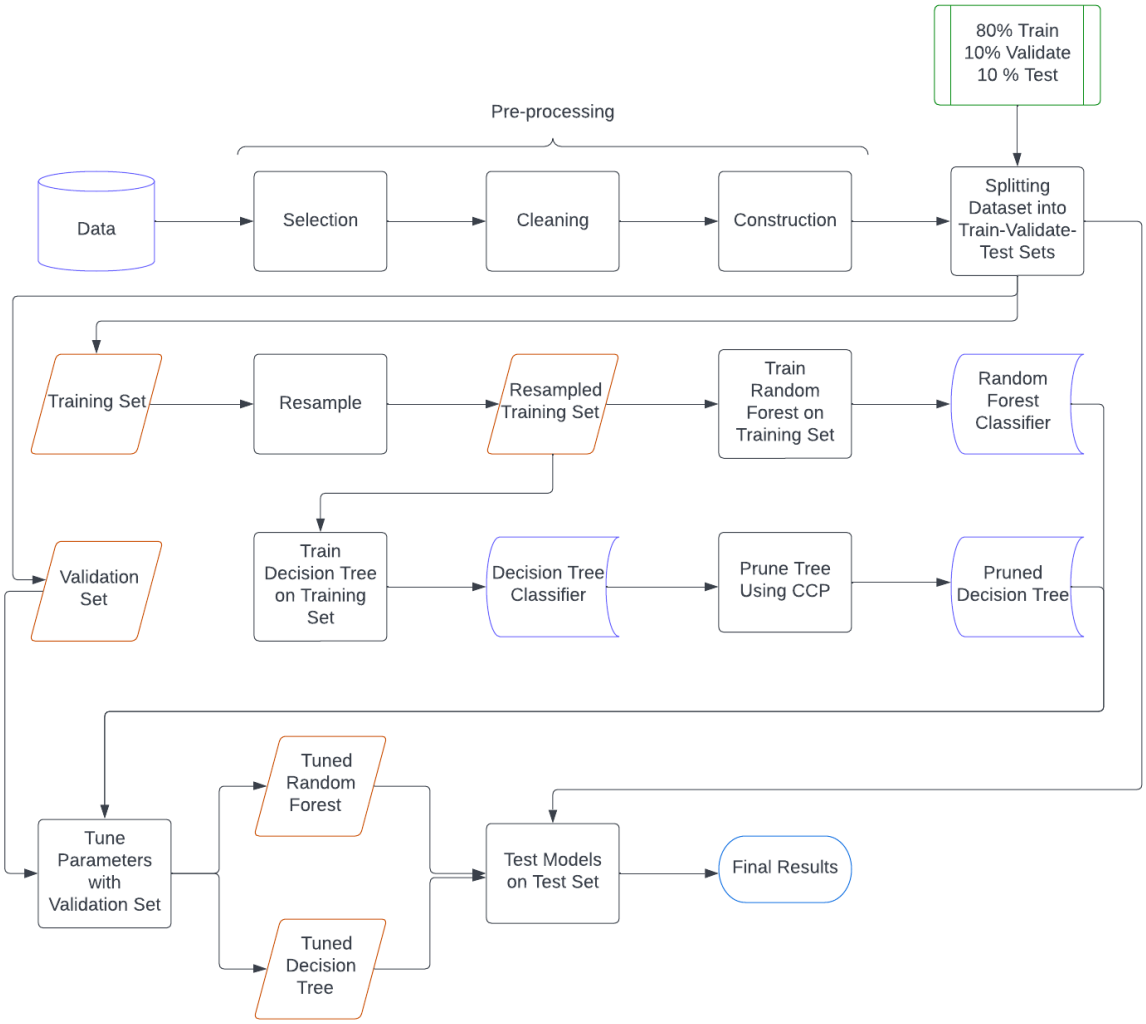
2.3. Modelling

2.3.1.Test Design

The data will be split into a training, validation, and testing set to avoid overfitting (respectively 80%, 10%, and 10%). Furthermore, the training sample will be resampled because the target variable is not entirely balanced, as more people do not receive treatment for mental health problems than those who do. After fitting the initial decision tree on the resampled training data, the tree will be pruned. First, cost complexity pruning will be used to get the effective alpha values of subtrees. Then cross validation will be used to find the best value for alpha. This alpha value will then be used to create the definitive decision tree and the accuracy of this tree will be tested with the validation subset.
In addition, a random forest classifier will be trained on the training sample. The parameters of both models are then tuned using the validation subset. Lastly, the testing subset is used to evaluate both models and select the one with the best fit. A schematic overview of the test design, in addition to the data pre-processing, is given in Figure 1.

Figure 1

Flowchart of the pre-processing and modelling steps



Note. The initial dataset is pre-processed by performing data selection, data cleaning, and data construction. The cleaned data is subsequently split into training, validation, and testing subsets. The training set is resampled, after which it is used to train a decision tree classifier and a random forest classifier. The decision tree is then pruned using cost complexity pruning. The parameters of the pruned decision tree and random forest are tuned using the validation set. Lastly, the tuned models are tested on the testing set.

2.3.2. Model Description

To build the models the following parameter settings were used. Firstly, the maximum value for alpha was excluded from the list of potential alpha values that were generated by cost complexity pruning, as this would result in a decision tree with no leaves and only a root. Furthermore, when selecting the optimal value for alpha, the values below 0.01 were excluded as well. While these lower alpha values gave very high within training sample accuracy, they were overfitted and did not perform well on the testing data. For the cross validation, five folds were used because increasing the number of folds did not improve the results and slowed down the processing time. Lastly, for the random forest the value for n_estimators (number of trees to build) was set to 300, as a higher value did not yield a significantly more accurate model.

3. Results

To assess the models, confusion matrices were made, and four different performance values were calculated. These performance measures include:

- Sensitivity: is also referred to as true positive rate, recall or hit rate. Sensitivity reflects the ratio of truly positive instances among all positive results.

$$Sensitivity = \frac{TP}{P}$$

- Specificity: is also referred to as selectivity or true negative rate. Specificity reflects the ration of truly negative instances among all negative results, making it the counterpart to sensitivity.

$$Specificity = \frac{TN}{N}$$

- Accuracy: is the ration of correct predictions (both true positives and true negatives) among the total number of instances. It combines both sensitivity and specificity to give a better impression of a model's overall performance.

$$Accuracy = \frac{TP + TN}{P + N}$$

- Balanced Accuracy: while accuracy may be a useful measure, it can give skewed result when dealing with class imbalance. In such cases, balanced accuracy is a better measure of performance. It is the arithmetic mean of sensitivity and specificity.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

Another measure that was considered was the F1-score. This measure combines precision and recall into one metric by calculating the mean between those two scores. However, F1-score does not reflect how many true negatives are being classified. In this instance, a performance measure that reflects both the true negatives and the true positives is preferable. The goal is to create a model that can both accurately identify whether an individual is receiving treatment and when they are not. As there is class imbalance in this dependent variable of our dataset, the main focus will be on how well the models perform in relation to balanced accuracy. Sensitivity, specificity, and normal accuracy are published as well to create a better understanding of the differences between the decision tree model and the random forest model. The results can be seen in Table 1.

Table 1

Performance measures of the decision tree and random forest models.

Performance measure	Decision Tree	Random Forest
Sensitivity	0.981	0.750
Specificity	0.736	0.888
Accuracy	0.808	0.847
Balanced Accuracy	0.858	0.819

While the decision tree yields a much better sensitivity and balanced accuracy, the random forest classifier resulted in a higher specificity and accuracy. Both models performed well and met the data mining goal. However, for this particular project the decision tree is more suitable because the main measurement of performance is the balanced accuracy.

Due to the better performance, the decision tree model is selected and further examined. The model decided upon the class label based on the following three factors:

1. Whether the individual had ever been diagnosed with depression.
2. Whether the individual had ever been diagnosed with anxiety.
3. How many days over the last 2 weeks the individual has felt down, depressed, or hopeless.

4. Discussion

The goal of this project was to build a model to help understand the risk factors for poor mental health and to provide possible targets to help improve and prevent mental health problems among the population. More specifically, the research objectives for this goal were to create a model that can predict whether an individual is receiving mental health treatment with an accuracy of at least 72% and to identify the factors that have the strongest relationship with the need for mental health treatment. To achieve this result, the BRFSS dataset was extracted, examined, and cleaned. Following which, a decision tree and random forest model were created.

Both models had an accuracy of 80% or higher, thus the first success criterium has been met. The decision tree was selected as the best model for this project, as it had a better balanced accuracy than the random forest model. The random forest model had a better score for the normal accuracy measure, but the balanced accuracy is more fitting as performance measure in this case, as we are dealing with class imbalance. Nevertheless, it is quite unusual that a decision tree performs better than a random forest model. To ensure this was not a random occurrence, five different random state values were tried for the sub-setting of the data into training, validation, and testing data. All five times the decision tree had a better balanced accuracy. This is surprising, as a random forest classifier fits multiple decision trees on different sub-samples of the dataset and uses the results thereof to build a more accurate model. This generally results in less over-fitting; however, decision trees can be biased when dealing with categorical variables, which made up a large portion of the dataset[20]. Possibly, this bias caused the random forest to perform worse than the decision tree.

The performance on the second research objective is less straightforward. While the decision tree did provide the variables which were used to predict whether an individual received mental health treatment these are not very insightful upon closer inspection. The variables with the biggest influence on the decision tree where whether an individual had ever been diagnosed with depression or anxiety. This contrasts with other studies that aimed to find risk factors for mental illness. These studies link mental illness most strongly to factors such as gender, physical health, employment, and socio-economic characteristics [8], [24], [25]. Furthermore, it is not surprising that individuals with a prior diagnosis of depression or anxiety are receiving medication or other treatment for emotional problems, as there is a large chance that the current treatment is for the prior diagnosis. This finding might be in support for the internal validity of the model, as the findings sounds ration. In addition, it might indicate that after initial recovery, the individuals that suffered from depression or anxiety in the past, are at a higher risk of needing mental health treatment for relapse or a new condition. This would suggest that it is advisable to monitor them long-term even after they have recovered. However, to know this for sure, we have to know whether the current treatment is in fact the result of the prior diagnosis or not. In addition, these findings do not help identify new target groups for mental health improvement. Another issue with the model is that it tests whether an individual is receiving treatment and not whether they are in need of it. There could be different reasons for someone in need of treatment, to be unable or unwilling to receive it. This is somewhat accounted for in the model as access to insurance is incorporated, however, it would have been better to use a different target value if it had been available in the dataset.

5. Conclusion

The aim of this project was to build a model to help understand the risk factors for poor mental health and to provide possible targets to help improve and prevent mental health problems among the population. To achieve this goal, the BRFSS survey data from 2015 was used to build a model that can predict whether an individual is receiving mental health treatment. Initially two models were made, a decision tree model and a random forest model. Against expectation, the decision tree model outperformed the random forest model in regard to balanced accuracy. This might be explained by the fact that the dataset contained a large number of categorical variables, which can cause bias in

random forest models. The decision tree classified instances based on the following three variables: whether the individual had ever been diagnosed with depression, whether the individual had ever been diagnosed with anxiety, and how many days over the last 2 weeks the individual has felt down, depressed, or hopeless. The research objective of building a model that can accurately predict mental health status was achieved. However, this model was less informative than was hoped, as it did not provide new potential targets for mental health prevention and treatment.

There were several limitations to this study. Firstly, the goal was to make a nationwide prediction model, however, this was not possible as the predicted variable was part of a module of the survey that was only conducted in West Virginia. In addition, the target variable captured whether an individual was receiving medication or treatment for mental health issues, not whether they were in need of them. This could have strongly impacted the model as it misses individuals who are in need of treatment but are unable or unwilling to receive it. Lastly, due to technical reasons and time constraints all instances with missing data (this includes answers such as 'don't know/not sure' and 'refused') were excluded from analysis. This led to a large decrease of instances in the dataset and might have caused a bias, as people with mental health problems are more likely to experience confusion, inattention, or other executive function problems that would make these answers more likely [26].

For these reasons, we recommend that further studies find a dataset with a smaller percentage of missing values and/or try to impute them so these instances can still be incorporated in the model. Following, we recommend that another output variable is chosen for a model to predict mental health status. This is a difficult variable to get an accurate measure of when dealing with self-reported data instead of professional assessment, but there are better options possible than the one used in this project. Lastly, we recommend that the BRFSS incorporate the mental health module into the core modules, as the societal and economic problems surrounding the issue are expected to grow rapidly over the next few years [3]. In conclusion, we build a classification model that can predict mental health status with a high accuracy. There are some problems with the model that have to improved in further research, however, with this project the groundwork was laid to make an efficient monitoring tool that could enable monitoring the state of mental health across different populations in the Unites States.

References

[1] "Depression." https://www.who.int/news-room/fact-sheets/detail/depression (accessed Jul. 15, 2022).

[2] R. C. KESSLER *et al.*, "Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative," *World Psychiatry*, vol. 6, no. 3, p. 168, Oct. 2007, Accessed: Jul. 02, 2022. [Online]. Available: /pmc/articles/PMC2174588/

[3] The Lancet Global Health, "Mental health matters," *The Lancet Global Health*, vol. 8, no. 11, p. e1352, Nov. 2020, doi: 10.1016/S2214-109X(20)30432-0.

[4] N. S. Tirupati, T. Rangaswamy, and P. Raman, "Duration of Untreated Psychosis and Treatment Outcome in Schizophrenia Patients Untreated for Many Years:," *http://dx.doi.org/10.1080/j.1440-1614.2004.01361.x*, vol. 38, no. 5, pp. 339–343, Jun. 2016, doi: 10.1080/J.1440-1614.2004.01361.X.

[5] L. Ghio *et al.*, "Duration of untreated depression influences clinical outcomes and disability," *Journal of Affective Disorders*, vol. 175, pp. 224–228, Apr. 2015, doi: 10.1016/J.JAD.2015.01.014.

[6] S. G. Alonso *et al.*, "Data Mining Algorithms and Techniques in Mental Health: A Systematic Review," *Journal of Medical Systems 2018 42:9*, vol. 42, no. 9, pp. 1–15, Jul. 2018, doi: 10.1007/S10916-018-1018-2.

[7] G. Azar, C. Gloster, N. El-Bathy, S. Yu, R. H. Neela, and I. Alothman, "Intelligent data mining and machine learning for mental health diagnosis using genetic algorithm," *IEEE International Conference on Electro Information Technology*, vol. 2015-June, pp. 201–206, Jun. 2015, doi: 10.1109/EIT.2015.7293425.

[8] P. A. Idowu *et al.*, "Predictive Model for the Risk of Mental Illness in Nigeria Using Data Mining," *International Journal of Immunology*, vol. 6, no. 1, pp. 5–16, 2018, doi: 10.11648/j.iji.20180601.12.

[9] "CDC - BRFSS." https://www.cdc.gov/BRFSS/ (accessed May 27, 2022).

[10] "Centers for Disease Control and Prevention | HHS.gov." https://web.archive.org/web/20200410150453/https://www.hhs.gov/about/budget/fy-2020-cdc-contingency-staffing-plan/index.html (accessed May 27, 2022).

[11] "Mission, Role and Pledge | About | CDC." https://www.cdc.gov/about/organization/mission.htm (accessed May 27, 2022).

[12] "CDC - About BRFSS." https://www.cdc.gov/brfss/about/index.htm (accessed May 27, 2022).

[13] V. J. Edwards, L. A. Anderson, W. W. Thompson, and A. J. Deokar, "Mental health differences between men and women caregivers, BRFSS 2009," *http://dx.doi.org/10.1080/08952841.2016.1223916*, vol. 29, no. 5, pp. 385–391, Sep. 2016, doi: 10.1080/08952841.2016.1223916.

[14] S. S. Dhingra, M. M. Zack, T. W. Strine, B. G. Druss, J. T. Berry, and L. S. Balluz, "Psychological distress severity of adults reporting receipt of treatment for mental

health problems in the BRFSS," *Psychiatric Services*, vol. 62, no. 4, pp. 396–403, 2011, doi: 10.1176/PS.62.4.PSS6204_0396/ASSET/IMAGES/LARGE/PSS6204_0396_FIG004.JPEG.

[15] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, "Classification Algorithms based Mental Health Prediction using Data Mining," pp. 1174–1178, Jul. 2020, doi: 10.1109/ICCES48766.2020.9137856.

[16] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, "Predicting mental health problems in adolescence using machine learning techniques," *PLOS ONE*, vol. 15, no. 4, p. e0230389, 2020, doi: 10.1371/JOURNAL.PONE.0230389.

[17] B. Hao, L. Li, A. Li, and T. Zhu, "Predicting Mental Health Status on Social Media A Preliminary Study on Microblog," *LNCS*, vol. 8024, pp. 101–110, 2013, Accessed: Jul. 03, 2022. [Online]. Available: http://ccpl.psych.ac.cn:10002

[18] M. Sumathi and B. Poorna, "Prediction of Mental Health Problems Among Children Using Machine Learning Techniques," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, 2016, Accessed: Jul. 16, 2022. [Online]. Available: www.ijacsa.thesai.org

[19] P. Kaur, R. Kumar, and M. Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)," *Multimedia Tools and Applications 2019 78:14*, vol. 78, no. 14, pp. 19905–19916, Feb. 2019, doi: 10.1007/S11042-019-7327-8.

[20] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, pp. 1–21, Jan. 2007, doi: 10.1186/1471-2105-8-25/FIGURES/11.

[21] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining".

[22] "Behavioral Risk Factor Surveillance System | Kaggle." https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?resource=download (accessed May 27, 2022).

[23] "CDC - 2015 BRFSS Survey Data and Documentation." https://www.cdc.gov/brfss/annual_data/annual_2015.html (accessed Jun. 20, 2022).

[24] O. E. Amoran, T. O. Lawoyin, and O. O. Oni, "Risk factors associated with mental illness in Oyo State, Nigeria.: A community based study," *Annals of General Psychiatry*, vol. 4, no. 1, pp. 1–6, Dec. 2005, doi: 10.1186/1744-859X-4-19/TABLES/4.

[25] F. I. Matheson, K. L. W. Smith, G. S. Fazli, R. Moineddin, J. R. Dunn, and R. H. Glazier, "Physical health and gender as risk factors for usage of services for mental illness," *J Epidemiol Community Health*, vol. 68, no. 10, pp. 971–978, Oct. 2014, doi: 10.1136/JECH-2014-203844.

[26] S. F. Logue and T. J. Gould, "The neural and genetic basis of executive function: Attention, cognitive flexibility, and response inhibition," *Pharmacology Biochemistry and Behavior*, vol. 123, pp. 45–54, Aug. 2014, doi: 10.1016/J.PBB.2013.08.007.

Appendix

Table 2

The column names that were included in the selected dataset on which the model was build and their description.

Column Name	Description
GENHLTH	General Health
PHYSHLTH	Number of Days Physical Health Not Good
MENTHLTH	Number of Days Mental Health Not Good
POORHLTH	Poor Physical or Mental Health
HLTHPLN1	Have any health care coverage
PERSDOC2	Multiple Health Care Professionals
MEDCOST	Could Not See Doctor Because of Cost
CHECKUP1	Length of time since last routine checkup
BPHIGH4	Ever Told Blood Pressure High
TOLDHI2	Ever Told Blood Cholesterol High
CVDINFR4	Ever Diagnosed with Heart Attack
CVDCRHD4	Ever Diagnosed with Angina or Coronary Heart Disease
CVDSTRK3	Ever Diagnosed with a Stroke
CHCSCNCR	Ever told you had skin cancer?
CHCOCNCR	Ever told you had any other types of cancer?
CHCCOPD1	Ever told you have chronic obstructive pulmonary disease, emphysema or chronic bronchitis?
HAVARTH3	Told Have Arthritis
ADDEPEV2	Ever told you had a depressive disorder
CHCKIDNY	Ever told) you have kidney disease?
DIABETE3	Ever told) you have diabetes
SEX	Respondents Sex
MARITAL	Marital Status
EDUCA	Education Level
RENTHOM1	Own or Rent Home
VETERAN3	Are You A Veteran
EMPLOY1	Employment Status
CHILDREN	Number of Children in Household
INCOME2	Income Level
INTERNET	Internet use in the past 30 days?
QLACTLM2	Activity Limitation Due to Health Problems
USEEQUIP	Health Problems Requiring Special Equipment
BLIND	Blind or Difficulty seeing
DECIDE	Difficulty Concentrating or Remembering
DIFFWALK	Difficulty Walking or Climbing Stairs
DIFFDRES	Difficulty Dressing or Bathing
DIFFALON	Difficulty Doing Errands Alone
HIVTST6	Ever tested HIV
SXORIENT	Sexual orientation or gender identity
TRNSGNDR	Do you consider yourself to be transgender?
ADPLEASR	Days had little pleasure doing things
ADDOWN	Days felt down, depressed or hopeless
ADSLEEP	Days had trouble with sleep
ADENERGY	Days were tired or had little energy
ADEAT1	Days ate too little or too much

ADFAIL	Days felt like failure or let family down
ADTHINK	Days had trouble concentrating
ADMOVE	Days talked to move slower or faster than usual
MISTMNT	Receiving medicine or treatment from health pro for emotional problem
ADANXEV	Ever told you had an anxiety disorder
_ASTHMS1	Computed Asthma Status
_DRDXAR1	Respondents diagnosed with arthritis
_ RACE	Computed Race-Ethnicity grouping
_ AGE80	Imputed Age value collapsed above 80
HTM4	Computed Height in Meters
WTKG3	Computed Weight in Kilograms
_BMI5	Computed body mass index
_SMOKER3	Computed Smoking Status
_RFBING5	Binge Drinking Calculated Variable
_DRNKWEK	Computed number of drinks of alcohol beverages per week
_FRUTSUM	Total fruits consumed per day
_VEGESUM	Total vegetables consumed per day
STRFREQ_	Strength Activity Frequency per Week
_PACAT1	Physical Activity Categories
_PAINDX1	Physical Activity Index

Note. For more details about these variables, see the codebook pdf-file.