iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 Issue: VI Month of publication: June 2022

DOI: https://doi.org/10.22214/ijraset.2022.45134

Prediction of Mental Health (Depression) Using Data Science Technique

Yuvasri K[1], Kannan R[2]
[1]*Student Department of Computer Science and Engineering*
[2]*Assistant Professor in Computer Science and Engineering*
GOJAN School of Business and Technology,Redhills, Chennai-52.

Abstract: Early diagnosis of mental health problems helps the professionals to treat it at an earlier stage and improves the patients quality of life. Depression is one of the leading causes of disability worldwide. This article provides an overview of AI and current applications in healthcare, a review of recent original research on AI specific to mental health, and a discussion of how AI can supplement clinical practice while considering its current limitations, areas needing additional research, and ethical implications regarding AI technology. So, there is an urgent need to treat basic mental health problems that prevail among children which may lead to complicated problems, if not treated at an early stage. Machine learning Techniques are currently well suited for analyzing medical data and diagnosing the problem. The attributes have been reduced by applying Feature Selection algorithms over the full attribute data set. The accuracy over the full attribute set and selected attribute set on various machine learning algorithms have been compared. However, caution is necessary in order to avoid over-interpreting preliminary results, and more work is required to bridge the gap between AI in mental health research and clinical care.

Keywords: Machine Learning, Artificial Intelligence, Natural Language Processing, Visualization, Deployment

I. INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.

The term “data science” was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market. Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

II. PROPOSED SYSTEM

A. Data Wrangling

In this section of the report will load in the data, check for cleanliness, and then trim and clean given data set for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

B. Data Collection

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done.

C. Building The Classification Model

The mental health predicting, ML algorithms prediction model is effective because of the following reasons: It provides better results in classification problem.

- It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.
- It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

D. Construction of a Predictive Model

Machine learning needs data gathering have lot of past data's. Data gathering have sufficient historical data and raw data. Before data pre-processing, raw data can't be used directly. It's used to pre-process then, what kind of algorithm with model. Training and testing this model working and predicting correctly with minimum errors. Tuned model involved by tuned time to time with improving the accuracy.

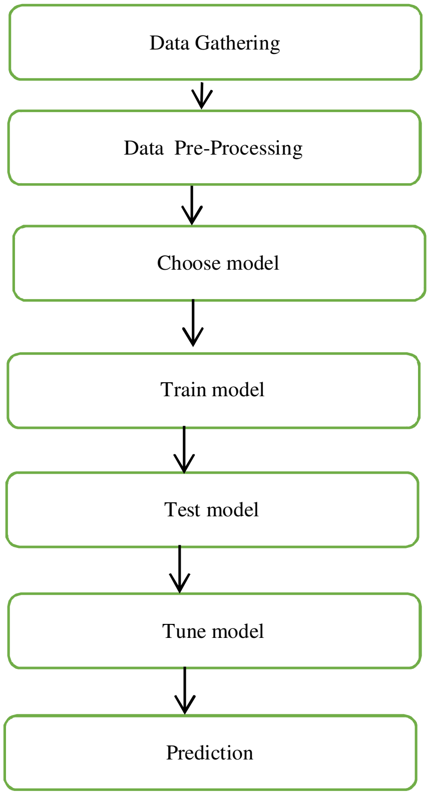

Fig. 1 Process Of Data flow Diagram

III. MODULES DESCRIPTION

A. List of Modules

- Data Pre-processing
- Data Analysis of Visualization
- Comparing Algorithm with prediction in the form of best accuracy result
- Deployment Using Flask

B. Data Pre-Processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the data set. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given data set. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training data set while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation data set is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters.

Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different data cleaning tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

C. Exploration Data Analysis Of Visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a data set and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

D. Comparing Algorithm With Prediction In The Form Of Best Accuracy Result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new data set, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracy.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below 4 different algorithms are compared:

- Logistic Regression
- Random Forest
- Decision Tree Classifier
- Naive Bayes

E. Deployment

Flask (Web Frame Work)

Flask is a micro web framework written in Python.

It is classified as a micro-framework because it does not require particular tools or libraries.

It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself.

Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Flask was created by Armin Ronacher of Pocoo, an international group of Python enthusiasts formed in 2004. According to Ronacher, the idea was originally an April Fool's joke that was popular enough to make into a serious application. The name is a play on the earlier Bottle framework.

When Ronacher and Georg Brand created a bulletin board system written in Python, the Pocoo projects Werkzeug and Jinja were developed. In April 2016, the Pocoo team was disbanded and development of Flask and related libraries passed to the newly formed Pallets project. Flask has become popular among Python enthusiasts. As of October 2020, it has second most stars on GitHub among Python web-development frameworks, only slightly behind Django, and was voted the most popular web framework in the Python Developers Survey 2018. The micro-framework Flask is part of the Pallets Projects, and based on several others of them.

Flask is based on Werkzeug, Jinja2 and inspired by Sinatra Ruby framework, available under BSD licence. It was developed at pocoo by Armin Ronacher. Although Flask is rather young compared to most Python frameworks, it holds a great promise and has already gained popularity among Python web developers. Let's take a closer look into Flask, so-called "micro" framework for Python.

IV. CONCLUSIONS

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of mental health.

A. Future Work

- Mental health prediction to connect with cloud model.
- To optimize the work to implement in Artificial Intelligence environment.

REFERENCES

[1] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, 2013.

[2] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations.," In Proceedings of the 5th Annual ACM Web Science Conference, 2013.

[3] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from twitter activity," In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems,2015.

[4] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in twitter," In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[5] Y. Tausczik and J. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," Journal of Language and Social Psychology, 2010.

[6] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From adhd to sad: analyzing the language of mental health on twitter through self-reported diagnoses," In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology, 2015.

[7] M. Trotzek, S. Koitka, and C. Friedrich, "Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression," Proceedings of the 8th International Conference of the CLEF Association,CLEF 2017, Dublin, Ireland, 2017.

[8] M. Trotzek, S. Koitka, and C. Friedrich, "Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia," Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, 2018.

[9] N. Liu, Z. Zhou, K. Xin, and F. Ren, "Tua1 at erisk 2018," Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, 2018.

[10] D. Losada, F. Crestani, and J. Parapar, "Overview of erisk 2018: Early risk prediction on the internet (extended lab overview)," Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018,Avignon, France, 2018.

[11] E. A. R´ıssola, M. Aliannejadi, and F. Crestani, "Beyond modelling:Understanding mental disorders in online social media," Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, 2020.

[12] S. Burdisso, M. Errecalde, and M. Montes-y Go´mez, "A text classification framework for simple and effective early depression detection over social media streams," Expert Systems With Applications, Vol. 133,2019.