Review of Mental Health Applications using Data Science with ML Techniques

Manjunath R.

Professor,
Department of CSE,
R. R. Institute of Technology, Bengaluru,
Karnataka, India

Rajashwari C.

UG Student,
R.R. Institute of Technology,
Visvesvaraya Technological University,
Bangalore, India

ABSTRACT

According to the World Health Organization (WHO), amongst 1.3 billion of the country's population, about 90 million Indians i.e., 7.5 % of them endure one or the other kind of mental disorder. The WHO also predicted that, the population suffering from mental illness would be around 20% by 2020 without foreseeing the corona virus pandemic. That counts over 200 million Indians who may be affected mentally and the number would even more raise and worsen the instances caused due to the other effects of the pandemic such as the caged feeling due to lockdown, loneliness, financial distress, etc. This paper aims at considering few Machine Learning algorithms such as Random Forest, SVM, K-NN, ID3, Naïve Bayes and C4.5 and find the best suitable algorithm for detecting and predicting mental illness accurately. Also, a few SDLC frameworks are considered to provide the integrated results of the algorithms in Mental Health. This survey of algorithms mainly focuses on classifying the emotional states and detecting mental illness in more accurate form. As a result of which, C4.5 algorithm was found to be more accurate.

Keywords: Random Forest, SVM, K-NN, ID3, Naïve Bayes, C4.5, LDTM, Data Analytical Framework.

INTRODUCTION:

Mental Health can be coined as the well-being state of an individual in terms of emotions and the other psychological parameters. The abnormal variations in a state of an individual in regard to thoughts, emotions and social functioning leads to mental illness. According to the World Health Organization (WHO), amongst 1.3 billion of the country's population, about 90 million Indians i.e., 7.5 % of them endure one or the other kind of mental disorder. The WHO alsopredicted that, the population suffering from mental illness would be around 20% by 2020 without foreseeing the corona virus pandemic. That counts over 200 million Indians who may be affected mentally and the number wouldeven more raise and worsen the instances caused due to the other effects of the pandemic such as the caged feeling due to lockdown, loneliness, financial distress, etc. The top 5 most common mental illnesses are the depression, anxiety, bipolar affective disorder, Schizophrenia and other Psychoses, Dementia. The traditional and modern ways are the broader sectors of medicinal practices in India. The traditional practices are inclusive of the various categories ofhealers such as the Ayurveda practitioners, folk healers, spiritual healers and Allopathic healers. The modern methods of curing involve the usage of Psychotropic drugs that are preferred by most of the psychiatrists and also theelectroconvulsive therapy (ECT) or other psycho therapies. Since the modern psychiatric facilities in India, areavailable only in the urban areas, there is a paucity of the modern amenities for the people in the rural areas. This is the main reason why 80% of the population depends on the traditional methods such as Ayurvedic and Unani systems of medicine, religious practices such as the prayers, fasting, magical rituals that are considered to be indigenous in nature.

RELATED WORK:

Ela Gore et al. [1], aimed at providing customary algorithms with their characteristics and efficiencies which would result in predicting appropriate models. The possibility of filling a gap between patient and therapists in disclosing exact problems to the therapists using machine learning. Recognizing the problems and choosing the features based on the EEG signal data & then designing a particular solution using machine learning algorithms regarding the mental health issues. It mainly focused on giving ML algorithms using which the emotional states were classified and for discovering the mentally occurring illnesses. Kunawut Boonkwang et al. [2] aimed at the research of multiple attempts of suicide byan individual using Data mining techniques such as ID3, C4.5 and naïve Bayes, mainly comparing the efficiencies of these techniques for classification & ensemble models. As a part of which the researchers used synthetic minority over- sampling technique (SMOTE) to maintain stability over the ideation of suicide reattempt and those not attempting categories. 57 ascribes in regard to self-harm surveillance report (RP.506S) of Khon Kaen Rajanagarindra Psychiatric hospital were used by the researchers. The tested results showed that C4.5 was an optimistic model for categorizing accurately at the rate of 90.69%. The ensemble model presented the finest results of 91.26. The Characteristics analysed using C4.5 were 474 patterns. They concluded that amongst 500 people, some people have the suicide ideation with 42 patterns each and over those 42 patterns, only 22 patterns gave the accuracy of 90%. Shahidul Islam Khan et al. [3] aimed at testing various classification algorithm in order to forecast the mental disorder during which 466 mental health datasets of Bangladesh were analysed to find the correlativity amongst diagnosis and attributes. Machine learning algorithms in particularly Random forest, SVM, K-nearest neighbor were applied and estimations were made to gauge the performances of these algorithms to spot the mental related issues. As a part of this, Random forest algorithm stood out as one of the best algorithms in finding such accuracy. Bakhtiyar Ahmed et al. [4] aimed and proposed one of the software development methodologies named Lean Design Thinking Methodology (LDTM) for guiding the modern data projects development. LDTM comprises of advantages of CRISP-DM along with which innovatory Design Thinking and Lean Startup strategies in order to introduce an approach split up into 3 stages consisting of 7 steps. Finally, they concluded that there is no one particular method or approach, but a collective togetherness of features of diverse approaches may lead to proper guidance for upcoming data projects. EdmonBegoli et al. [5] focused on the challenges faced by ML and AI in recent years addressing the insufficient volumes of data to be provided for training which would affect on the accuracies of the outcome of mental health sector. Hence, introduced a new framework known as SynthNotes which would generate a free text notes that is derived from the statistics permitting the natural language depictions of a mentally affected individuals and characteristics of such patients. SynthNotes is an adaptable and ascendable which incorporates various fundamental models and addresses diverse conditions. They reported on usageand competencies of SynthNotes and application of the same over the ongoing work adding to it the content forethought and training the real data with generative deep-learning methods. Charith Silva et al. [6] aimed at motivating mental health professionals to use the data science approach to solve the challenges they encounter in mental health. Before adapting data science into such field, it is also important to have fine data analysis framework with clear instructions. Hence, the authors have given the new analytical framework to implement data science approach in the field of mental health which not only find solutions for problems but also prophesying time and resource needed.

METHODOLOGIES:

This section describes the analysis methodologies and approaches explored during this survey. The perspicacity of such study with different perspectives would bring out innovation in the field of machine learning & data science in mental health applications. Here, we have contemplated the existing Machine Learning algorithms i.e., Random Forest, SVM, K-NN, ID3, Naïve Bayes, C4.5 and also, frameworks including SynthNotes, Lean Design Thinking Methodology (LDTM), Data science models in a quick footnote.

Random Forest:

Random forest classification method that can also be called as ensemble classification method is one of the supervised learning methods. The ensemble method makes use of the predictor attributes to calculate the average outof the result obtained from individual predictions. The Sampling subsets of training data are randomly contained in the bagging method through which the Random forest can be trained. Using the tree bagging method, each of the trees in random forest are fitted and aggregation of all those trees is calculated and result is produced. Random forestis efficient when compared to Support Vector Machine (SVM) and K-Nearest Neighbor(K-NN) algorithms.

Support Vector Machine (SVM):
SVM is one of the unique supervised learning models of machine learning which is the most popular way when it comes to Classifications and Regression problems usually when there is huge amount of dataset. This algorithmtakes as input a large dataset that comprises of noisier data and makes decision effectively with respect to the datasets reducing the noise data and constructing hyperplane by estimating the best attainable margin that is comparison among hyperplane and support vector. These super vectors split up the dataset into a high dimension space. Larger the dataset, larger is the performance of SVM.

K-Nearest Neighbor (K-NN):
The K-NN algorithm is one of the non-parameterized methods to identify patterns during Classifications and Regression. It was proposed by Thomas Cover. It takes as input the training data and on the basis of resemblance, it plots the whole of datasets into n-number of dimensions finding the best suited neighbors with regard to value of k.

C4.5:
C4.5 accepts categorical data, continuous data and also miscellaneous values. C4.5 considers training data as itsinput and produces a single tree as its output. It is a better classifier method that backtracks to the tree so that the nodes can be eliminated or internal structure of the tree can be modified. C4.5 is more efficient when compared to ID3 and Naïve Bayes algorithms.

Iterative Dichotomiser 3 (ID3):
ID3, one of the classification algorithms was invented by Rose Quinlan in order to produce a decision tree from given dataset. It is also predecessor of C4.5. This algorithm is mainly used in the fields of machine learning and Natural language processing. It takes as input the dataset considering it as the root node and performs iteration upon unused attributes and estimates entropy and information gains. After which it takes the data with minimal entropy and make it a subset or sub node of the root node forming a decision tree.

Naïve Bayes:
Naïve Bayes is one of the probabilistic algorithms that is a successor of Bayes theorem. These algorithms are highly scalable since the probabilities calculated at the earlier stages can be used to analyse the upcoming cases. It follows the Bayes theorem whose equation is:

$$\frac{P(A|C)\ P(C)P(C|A)}{P(A)}$$

Where P(C|A)- Posterior Probability with Class C and Attribute A, P(A|C)- Probability that training data consists of Class C and Attribute A, P(C)- Prior Probability of C, P(A)- Prior Probability of A.

Random Forest vs C4.5:
There isn't a single tree that can be drawn and analyzed in the method of Random forest whereas, in C4.5 there is a single tree and to guard against the over fit, pondering must be done in case of pruning. Investigation of generalized errors is very less required since there is no over fitting in Random forest. Hence, Random forest is unbalanced whereas C4.5 has a better balance w.r.t training errors and generalized errors. Therefore, amongst Random forest andC4.5, C4.5 stands out to be accurate in terms of prediction.

Synth Notes::
SynthNotes focuses on the challenges faced by ML and AI in recent years addressing the insufficient volumes ofdata to be provided for training which would affect on the accuracies of the outcome of mental health sector. Hence, introduced a new framework known as SynthNotes which would generate a free text notes that is derived from the statistics permitting the natural language depictions of a mentally affected individuals and characteristics of such patients. SynthNotes is an adaptable and ascendable which incorporates various fundamental models and addresses diverse conditions.

Lean Design Thinking Methodology:
Lean Design Thinking Methodology (LDTM) guides the modern data projects development. LDTM comprises of advantages of CRISP-DM along with which innovatory Design Thinking and Lean Startup strategies in order to introduce an approach split up into 3 stages consisting of 7 steps.

By bringing together these three approaches, the intention is to allow development teams the flexibility and opportunity to routinely improve a data model by iteratively and incrementally acting upon accuracy statistics and user feedback. Therefore, to advance the implementation of the already existing and upcoming data projects LDTM can be used.

LDTM is categorized into various stages such as Business, Data and Product which consists of 7 steps.These steps are:

- Work Discovery: Defining the problem statement, setting objectives and defining requirements.
- Analytical Approach: Analysing the methodologies, strategies in reference to project development.
- Data Resources: Selected approaches with data in particular formats are specified.
- Data Preparation: Collection of data, data cleansing, and visualizing the content of the data.
- Building Minimum Viable Product (MVP): Simple solution for the dataset to get intentional data.
- Measure Value: Implementation of MVP, assessment of features, measuring the solutions.
- Learn & Update: Learn and update using the assessment result and enhance the project.

The LDTM is depicted in below Figure 1 with above steps:

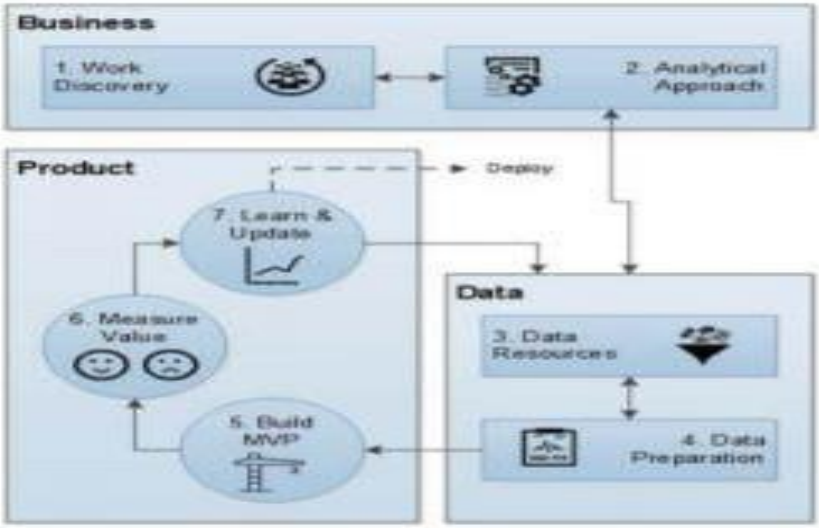

Figure 1. Lean Design Thinking Methodology

Data Science approach of Development of Mental Health projects:

Motivating mental health professionals to use the data science approach to solve the challenges they encounter in mental health. Before adapting data science into such field, it is also important to have fine data analysis framework with clear instructions. Hence, the new analytical framework to implement data science approach in the field of mental health which not only find solutions for problems but also prophesying time and resource needed.

This is a new framework using the data science which can be used in the development of Mental Health estimation projects using data science approach. Below fig 2. shows the flow of SDLC of data science approach which consists of problem definition, Requirement gathering, Data Acquisition phase, Analysis and Visualization of data attributes, data fusion, filtering and pre-processing, Data cleansing, feature selection, Data partitioning, Predictive modelling using various domains of data science, visual data exploration for better understanding, evaluation of predictive model and knowledge extraction through which the performance of model increases analyzing the new cases.

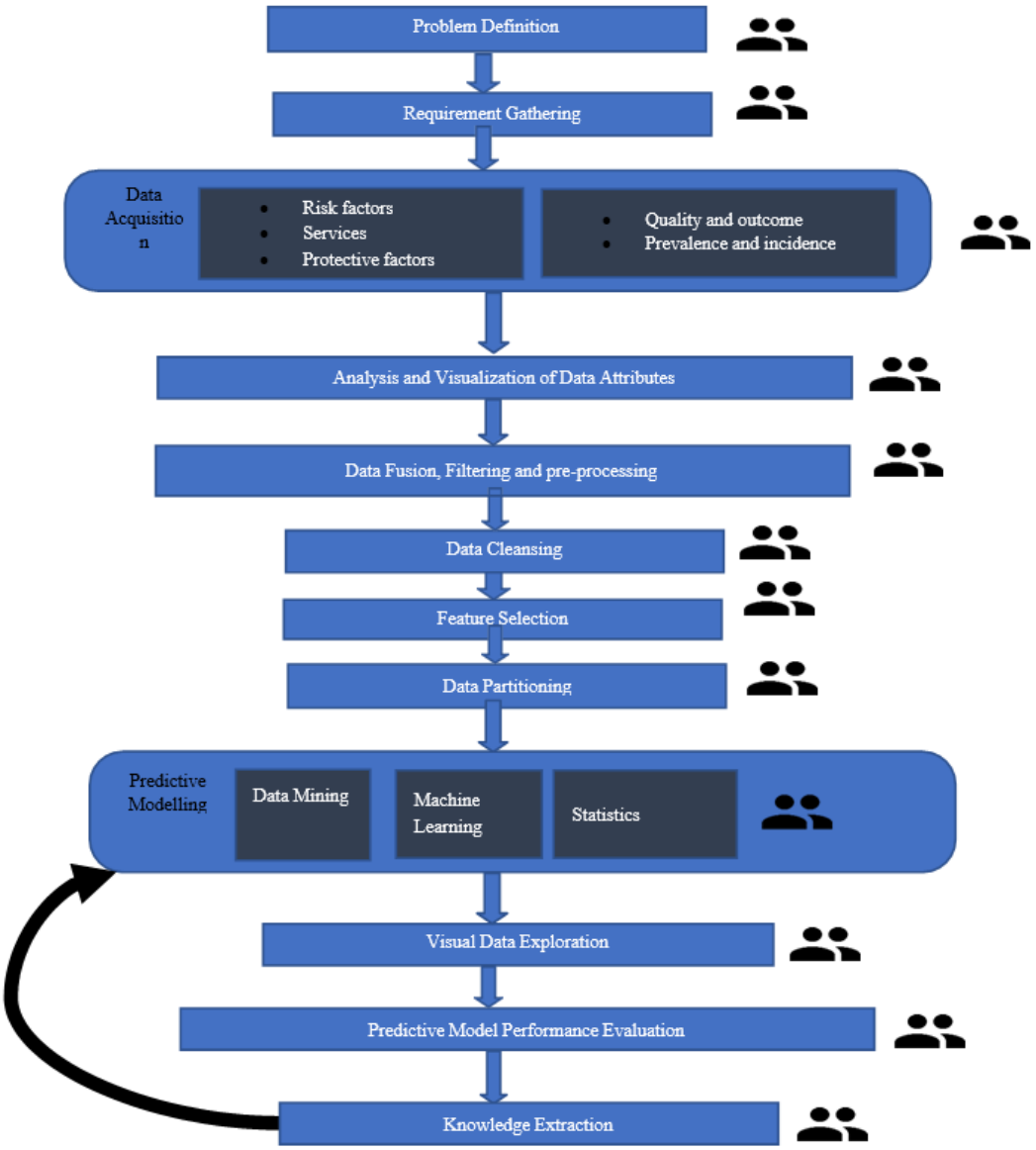

Figure 2.Framework for Software Development Life cycle (SDLC) using data science approach

RESULTS AND DISCUSSIONS:

This section describes the performances of machine learning algorithms and the analysis of their accuracy in terms ofmental health predictions.

- Performance analysis of Algorithms such as Random Forest, SVM and KNN:

Table 1: Comparison of algorithms in terms of original dataset & Synthetic dataset

Data	RF	SVM	KNN
Original dataset(466)	0.851	0.787	0.787
Synthetic dataset(2600)	0.859	0.844	0.851

- Performance of C4.5 :C4.5 has an accuracy 90.69%.

Table 2: Comparison of algorithms in terms of Single Model & Ensemble Model

Type	Method	Accuracy (%)	Preci-sion	Recall	F-Meas-ure
Single Model	Naïve Bayes	70.40%	0.708	0.704	0.703
	ID3	90.60%	0.908	0.907	0.906
	C4.5	**90.69%**	**0.908**	**0.907**	**0.907**
Ensemble Model	Bagging	90.82%	0.910	0.908	0.908
	Random Forest	91.17%	0.913	0.913	0.912
	Voting (C4.5,Naïve Bayes, Random Forest)	**91.26%**	**0.914**	**0.913**	**0.913**

- Comparison of Performances of Random Forest and C4.5

C4.5 is more accurate than Random Forest as shown in below table:

Table 3. Comparison of Random Forest and C4.5

Algorithms	Accuracy
Random Forest	85.1%
C4.5	90.69%

CONCLUSION:

In around 1.3 billion of the country's population, around 7.5% of them suffer from mental health disorder. The main focus ofpeople in today's world is on the physical health. They know what kind of workout and diet they need to follow and how to maintain their body. That is good, but equal importance must be given to mental health also.Here comparison of various Machine Learning algorithms is done and finally C4.5 algorithm is identified as the best algorithm so far to detect and predict the mental health. The usage of frameworks is also considered so that the integrated results of using the algorithm and the frameworks can provide more accuracy.People must also consider few simple but effective measures like Meditation in their daily routine, sharing their feelings openly with their closed ones and socializing.Government or the institutes or the organizations also must take measures in establishing rehabilitation centres in even the remote localities and create awareness about the mental health and conduct workshops in this aspect.In future, other algorithms can also be compared that provide more accuracy and many datasets can be considered that improves the prediction accuracy and any app can be developed with more and more benefits in this regard. Any other frameworks can also be developed to improve the detection and prediction of mental health disorders and also curing the mental illness.

REFERENCES:

B. Ahmed, T. Dannhauser and N. Philip, "A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects," 2018 10th Computer Science and Electronic Engineering (CEEC), Colchester, United Kingdom, 2018, pp. 11-14, doi: 10.1109/CEEC.2018.8674234.

B. Severtson, L. Franks and G. Ericson, "What is the Team Data Science Process?", Microsoft Azure, 2017. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-datascience-process/overview

C. Silva, M. Saraee and M. Saraee, "Data Science in Public Mental Health: A New Analytic Framework," 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, 2019, pp. 1123-1128, doi: 10.1109/ISCC47284.2019.8969723.

Diederich, J., Al-Ajmi, A., &Yellowlees, P. (2007). Ex-ray: Data mining and mental health. Applied Soft Computing, 7(3), 923-928. doi:10.1016/j.asoc.2006.04.007

Dooshima, M. P. (2018). A Predictive Model for the Risk of Mental Illness in Nigeria Using Data Mining. International Journal of Immunology, 6(1), 5. doi:10.11648/j.iji.20180601.12

E. Begoli, K. Brown, S. Srinivas and S. Tamang, "SynthNotes: A Generator Framework for High-volume, High-

fidelity Synthetic Mental Health Notes," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 951-958, doi: 10.1109/BigData.2018.8621981.

E. Gore and S. Rathi, "Surveying Machine Learning Algorithms On Eeg Signals Data For Mental Health Assessment," 2019 IEEE Pune Section International Conference (PuneCon), Pune, India, 2019, pp. 1-6, doi: 10.1109/PuneCon46936.2019.9105749.

K. Boonkwang, S. Kasemvilas, S. Kaewhao and O. Youdkang, "A Comparison of Data Mining Techniques for Suicide Attempt Characteristics Mapping and Prediction," 2018 International Seminar on Application for Technology of Information and Communication, Semarang, 2018, pp. 488-493, doi: 10.1109/ISEMANTIC.2018.8549835.

K. Thoring and R. Mueller, "Design Thinking vs. Lean Startup: A comparison of two user-driven innovation strategies", in 2012 International Design Management Research Conference, Boston, Massachusetts, 2012, pp. 151-161.

S. I. Khan, A. Islam, A. Hossen, T. I. Zahangir and A. S. M. LatifulHoque, "Supporting the Treatment of Mental Diseases using Data Mining," 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2018, pp. 339-344, doi: 10.1109/ICISET.2018.8745591.
