PREDICTION AND VISUALIZATION OF MENTAL HEALTH PATTERNS USING MACHINE LEARNING TECHNIQUES*

Shwe Sin Ei[1] , Thet Thet Hlaing[2], Soe Mya Mya Aye[3]

Abstract

Mental health includes emotional, psychological, and social well-being. As the coronavirus pandemic has spread across the world, the public health crisis has brought with it considerable impact on social and economic. Besides physical health, the psychological impacts of COVID-19 also pose significant risks to mental-wellbeing as the greater levels of stress, depression and anxiety in people. In this paper, patterns of mental health were discovered using public data by machine learning approach. The main contribution of this paper is to provide awareness and nature of mental health to determine the factors that influence mental health across people's lifespans (e.g. genetics, cognition, demographics). This research use three machine learning classifiers: decision tree, random forest and k-nearest neighbour (KNN). The experimental results found that Random Forest classifier achieves the best accuracy.

Keywords: Mental Health, Machine Learning, Decision tree, Random forest, K-nearest neighbour

Introduction

Nowadays, the impact of modern age life style, mental illnesses have become more common in life. Depression, anxiety, stress and suicide are all more prevalent, year by year. Around 1-in-7 people globally (11% to 18%) have one or more mental or substance use disorders. Mental health is vital at every stage of life, not only from childhood to adulthood but also for old people, in every age.

Mental health can happen to all kinds of people from all kinds of works life, young or old, rich or poor. Good mental wellbeing doesn't mean a person is always happy or unaffected by external experiences. Mental health affects how people think, feel, and act. The impact of poor mental health creating tension, uncertainty, stress and sometimes significant changes in how people live their lives. Mental health awareness is so important that it can help you to understand your symptoms, get access treatment and perhaps you can overcome mental unhealthy conditions.

The information technologies offer a lot of opportunities for communication and social networking. In medical field, machine learning is widely used and it helps in mental health research and practice. With the help of advances technology, doctors can get a huge amount of information at a rapid rate. Machine learning in AI analyze these data and this allows data patterns are correctly discovered. Machine learning techniques aided quick and scalable analysis of complicated data as well as accurate predictions can be made.

In this paper, three machine learning classifiers: decision tree and random forest and k-nearest neighbors (KNN) were applied for research. Among these algorithms, decision tree and random forest are popular machine learning algorithms which belongs to the supervised learning technique. It can be used for both classification and regression problems in machine learning. Random Forest can predict the output with high accuracy and decision trees can perform classification without requiring much computation. K-nearest neighbors (KNN) is one of the simplest machine learning algorithms based on supervised learning technique. It can be used for classification as well as regression problems.

[1.] Department of Computer Studies, University of Yangon
[2.] Department of Computer Studies, University of Yangon
[3.] Department of Computer Studies, University of Yangon
* Best Paper Award Winning Paper in Computer Science (2022)

The paper is organized as follows. Section 2 describes the related works to clarify more on the approach. In Section 3, we describe the materials and method use in identifying the pattern of mental health. Section 4 describes the implementation of the system and finally, Section 5 describes results and future work.

Materials and Methods

Data source

This dataset is survey in 2014 from Kaggle which measures attitudes towards mental health and frequency of mental health disorders in the Tech workplace.

The original dataset has 1260 records and 27 variables. This research utilized 1260 records involving 8 selected variables. The sample dataset was as shown in table 1.

Table 1: Dataset of selected features of Mental Health

Sr. No.	Age	Gender	Self Employed	Family History	Work Interferes	Coworkers	Supervisor	Treatment
1	37	Female	Yes	No	Often	Some	Yes	Yes
2	44	Male	Yes	No	Rarely	No	No	No
3	32	Male	No	No	Rarely	Yes	Yes	No
4	31	Male	Yes	Yes	Often	Some	No	Yes
5	31	Male	No	No	Never	Some	Yes	No
6	33	Male	No	Yes	Sometimes	Yes	Yes	No
7	35	Female	No	Yes	Sometimes	Some	No	Yes
8	39	Male	No	No	Never	No	No	No
9	42	Female	No	Yes	Sometimes	Yes	Yes	Yes
10	23	Male	No	No	Never	Yes	Yes	No
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
1259	46	Female	No	No	Never	No	No	No
1260	25	Male	No	Yes	Sometimes	Some	No	Yes

Patterns in machine learning

A pattern is a series of data that repeats in a recognizable way. Pattern recognition is a process of finding regularities and similarities in the dataset. Machine learning technique classifies data based on statistical information or knowledge gained from patterns and their representation. In this study, the patterns of influencing factors on mental health are visualized by using matplotlib, data visualization and graphical plotting library of python.

Machine learning methods

Machine learning is a subfield of artificial intelligence. The concept of machine learning is that the computer program can learn and adapt to new data without human intervention. Machine learning involves a group of computational algorithms that can perform classification, pattern recognition and prediction. There are different types of machine learning Algorithms such as supervised learning, unsupervised learning, semi-supervised learning and reinforcement Learning.

Decision Tree Algorithm

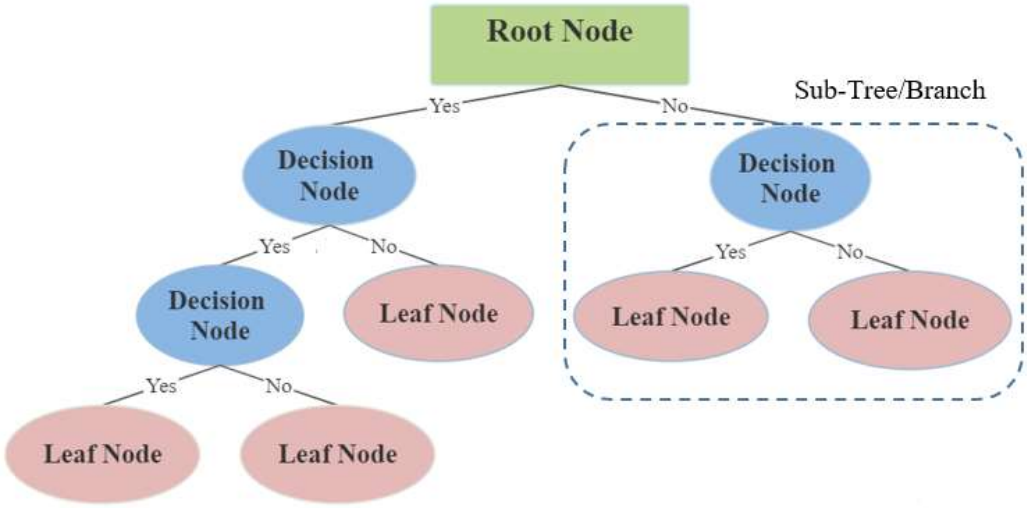

Figure 1: Decision Tree Structure

Decision Tree Algorithm which is also called 'CART' used for both classification and regression problems. It uses supervised machine learning technique. Because of its tree-structured classifier, it is called decision tree and it starts with the root node, which expands on further branches.

In a Decision tree, there are two nodes, decision node and leaf node. Decision nodes are used to make decision whereas Leaf nodes are the output of those decisions. A decision tree simply works, asks a question and based on the answer (yes/no), it is further split into branches and sub trees.

Random Forest Algorithm

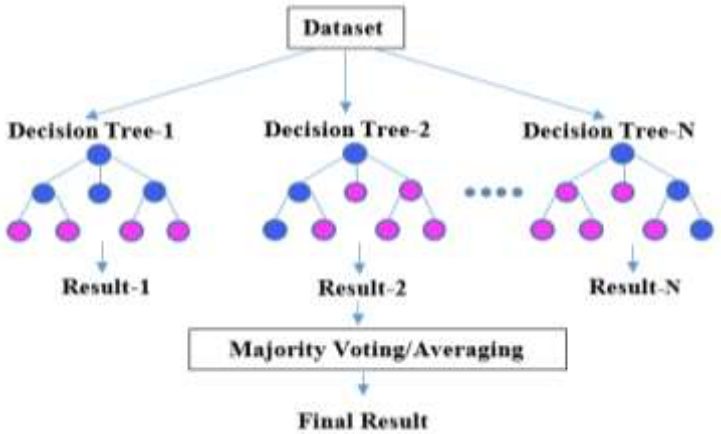

Figure 2: Random Forest Structure

Random Forest is a well-known machine learning algorithm that uses supervised learning techniques. Random Forest is a classifier that combines a number of decision trees on different subsets of a dataset to create a forest and feeds random features from the input dataset to them.

One of the most important features of the Random Forest Algorithm is that it can handle both categorical variables and continuous variables. The random forest uses the majority votes of predictions from each tree in the forest and it produce the final output of predictions. The greater the number of trees leads to the more accuracy of result can be produced and it prevents from over fitting.

K-Nearest Neighbour (KNN) Algorithm

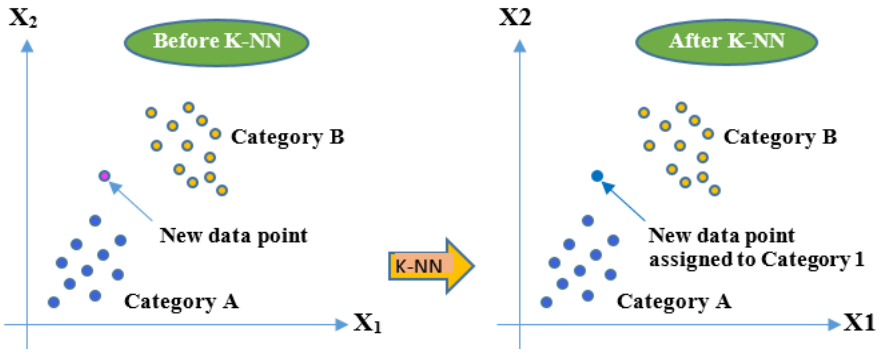

Figure 3: K-Nearest Neighbour Classifier

The K-Nearest Neighbors (KNN) algorithm is a simple supervised machine learning algorithm that can be used to solve both classification and regression problems. In pattern recognition, the KNN algorithm is a method for classifying objects based on closest training examples in the feature. Aleem Shumaila and Huda Noor ul [2022] said this rule simply retains the entire training set during learning and assigns to each query a class represented by the majority label of its K-Nearest Neighbors in the training set.

It is a major classical machine learning algorithm that focuses on the distance from new unclassified/ unlabeled data points to existing classified/labeled data points. For classification problems, a class label is assigned on the basis of a majority vote that means the label that is most frequently represented around a given data point is used.

Implementation and Results

In developing the system, Python programming language is used in Spyder Platform which is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python Language. The dataset that we used in our research is from Kaggle.com. In this study, we will implement the decision tree, random forest and K-Nearest Neighbors algorithms using Python's Scikit-Learn library.

Experimental Setup

The experiments were implemented on 8 selected attributes containing 1260 instances. The dataset was imported first from Kaggle and data cleaning i.e., removing duplicate records, incorrect and incomplete data in dataset was done. Then, the dataset was split into two separate datasets, train and test datasets for input dataset (features) and output dataset (target). The attributes description of datasets are as shown in table (2).

Table 2: Mental Health Dataset Description

No	Attribute Names	Description
1.	Age	18 ~ 70
2.	Gender	Male/Female/Queer
3.	Remote work	Do you work remotely (outside of an office) at least 50% of the time?
4.	Family history	Do you have a family history of mental illness?
5.	Work interfere	If you have a mental health condition, do you feel that it interferes with your work?
6.	Coworkers	Would you be willing to discuss a mental health issue with your coworkers?
7.	Supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
8.	Treatment	Have you sought treatment for a mental health condition?

Workflow Diagram of a Proposed System

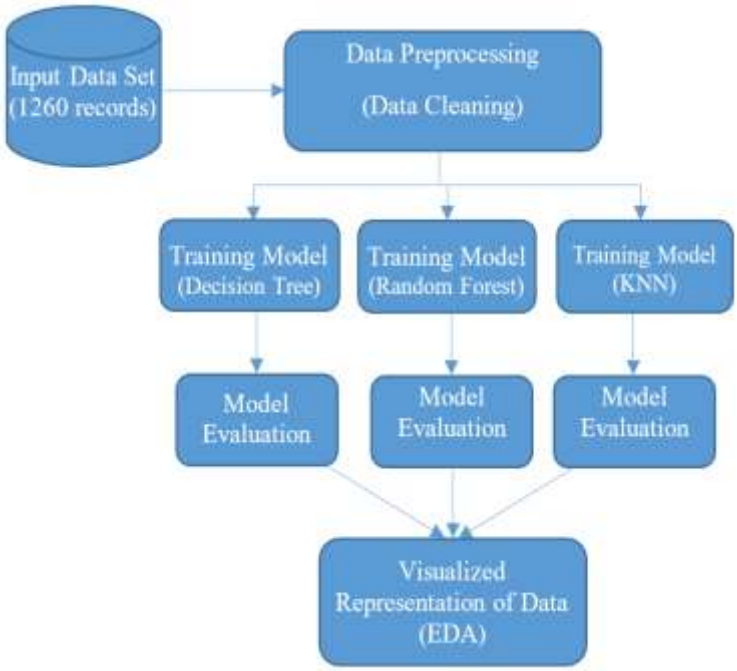

Figure 4: Workflow Diagram of a Proposed System

The work flow of the proposed system starts with input of dataset. The input data contains 8 attributes and 1260 employee records. In the first step, the data processing is done to clean the raw data. In the second step, three different classification algorithms in machine learning: Decision tree, Random forest and K-Nearest Neighbors are chosen and the model are trained. In the third step, models are evaluated to test the accuracy of each model. At the last stage, EDA data are displayed in the form of graphs and charts.

Building Models

In supervised learning, algorithms learn from labeled data. Classification techniques are used to determine which class is yes and no. In this study, three classification models are built using supervised machine learning techniques called decision tree classifier, random forest classifier and k-nearest neighbors classifier. So, these classifiers are imported from scikit-learn library which is the machine learning library in python as shown in figure (5). Decision trees were generated using CART algorithm. After decision tree generation, Random Forest was performed. Then, KNN classifier was imported from scikit-learn and create a model.

```
#Load Dataset
mh = pd.read_csv('survey.csv')
# Modeling
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, recall_score, plot_roc_curve
from sklearn.ensemble import RandomForestClassifier
# Splitting Data
from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score
# Tuning
mh.drop(columns=['Timestamp', 'Country', 'state', 'comments'], inplace = True)
mh.rename({'self_employed' : 'Self_Employed', 'family_history' : 'Family_History',
           'treatment' : 'Treatment', 'work_interfere' : 'Work_Interfere',
           'remote_work': 'Remote_Work','coworkers': 'Coworkers', 'supervisor': 'Supervisor',
           } , inplace = True , axis = 1)
mh['Age'].replace([mh['Age'][mh['Age'] < 15]], np.nan, inplace = True)
mh['Age'].replace([mh['Age'][mh['Age'] > 100]], np.nan, inplace = True)
```

Figure 5: Code Segments of Developing Classification Models

Afterwards, the models are trained on two separate datasets to train and test. The dataset is divided into training dataset and testing dataset, 70-80% of the dataset is used for the training and 20-30% is used for the testing. Evaluation of the models show that the best results are obtained if dataset is divided into 20% for testing, and the remaining 80% for training. The training dataset is used to train the machine learning models and the testing dataset is used to test the trained model. Our models will learn patterns on the data to make predictions whether an employee willingly to take treatment or not. The scores of different models are measured to evaluate and compare their accuracy. The k=5-fold Cross Validation was adopted for the training datasets and test datasets. The accuracy of the models were calculated using Confusion Matric and compared the performances of the models using recall, standard deviation, mean and cross validation scores as shown in figure (6).

```
def model_evaluation(model, metric):
    model_cv = cross_val_score(model, X_train, y_train, cv = StratifiedKFold(n_splits = 5), scoring = metric)
    return model_cv


tree_pipe_cv = model_evaluation(tree_pipe, 'recall')
knn_pipe_cv = model_evaluation(knn_pipe, 'recall')
rf_pipe_cv = model_evaluation(rf_pipe, 'recall')


for model in [tree_pipe, rf_pipe, knn_pipe]:
    model.fit(X_train, y_train)

score_cv = [tree_pipe_cv.round(5),
            rf_pipe_cv.round(5), knn_pipe_cv.round(5)]
score_mean = [tree_pipe_cv.mean(),  rf_pipe_cv.mean(), knn_pipe_cv.mean()]
score_std = [tree_pipe_cv.std(), rf_pipe_cv.std(), knn_pipe_cv.std()]
score_recall_score = [recall_score(y_test, tree_pipe.predict(X_test)),
            recall_score(y_test, rf_pipe.predict(X_test)),  recall_score(y_test, knn_pipe.predict(X_test))]
method_name = [ 'Decision Tree Classifier',  'Random Forest Classifier', 'K Nearest Neighbour']
cv_summary = pd.DataFrame({
    'method': method_name,
    'std score': score_std,
    'recall score': score_recall_score,
    'mean score': score_mean})
cv_summary
print(cv_summary)
```

Figure 6: Code Segments to calculate Accuracy and Performance of Models

Exploratory Data Analysis (EDA)

Visualization of the discovered patterns is important in order to communicate information efficiently using graphs, charts and tables. In this paper, different machine learning algorithms such as Decision Tree, Random Forest and K-Nearest Neighbors are used to identify the key features of mental health patterns that lead to mental health problems in working environment. Exploratory Data Analysis based on the selected features are shown in figure (7).

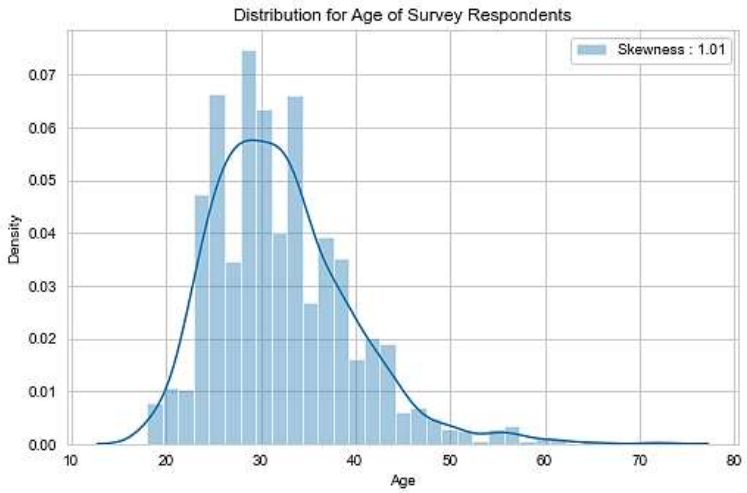

Figure 7: Density of respondents by age

From figure (7), age pattern indicated that most of the employees that fill the survey around the end 20s to early 40s.

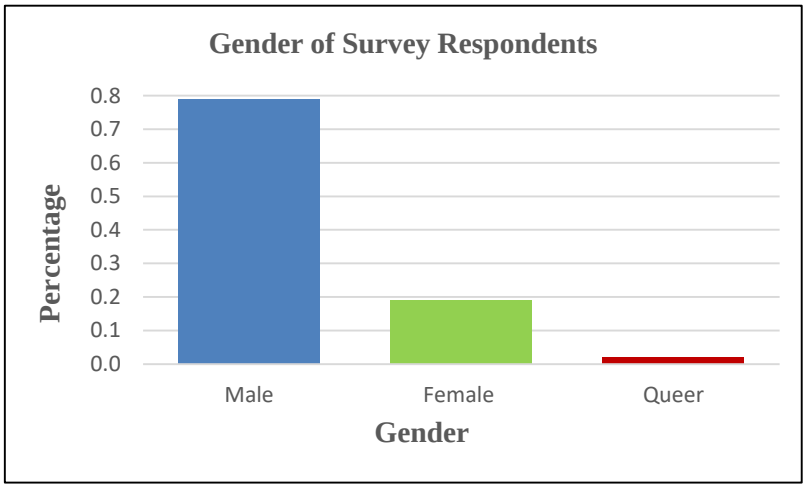

Figure 8: The respondents by Gender

According to gender pattern in figure (8), it was found that male are higher rate of respondents almost 79% of respondents are male.

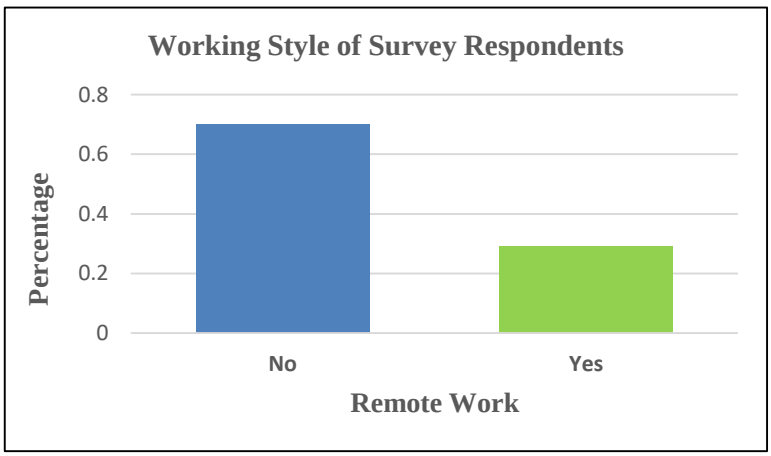

Figure 9: The working style of respondents

According to working style pattern in figure (9), around 70% of respondents don't work remotely, which means the biggest factor of mental health disorder came up triggered on the workplace.

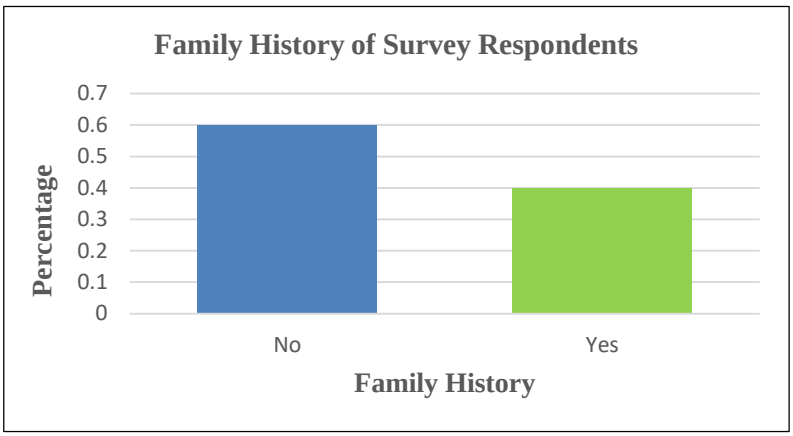

Figure 10: Family History of survey respondents

Figure (10) describes the 'family history of survey respondents by percentage'. According to the patterns, 40% of respondents who say that they have a family history of mental illness.

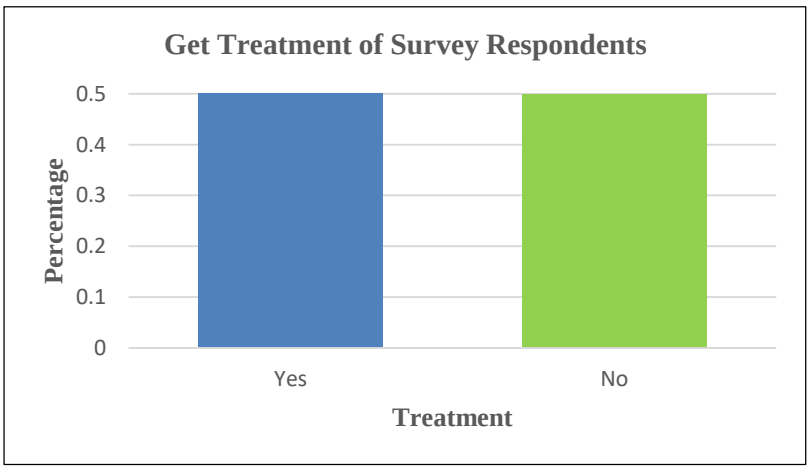

Figure 11: Percentage of getting treatments

Figure (11) describes the 'Percentage of getting treatments'. It was found that the percentage of respondents who want to get treatment is 50%.

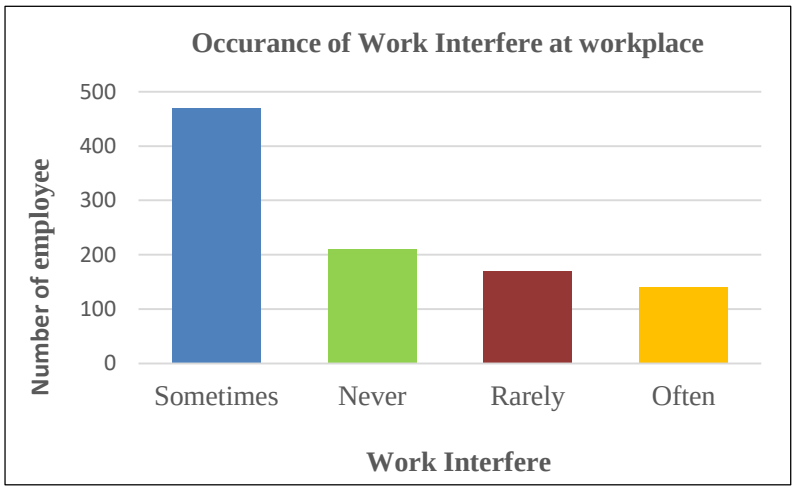

Figure 12: Mental Health that makes work interfere at workplace by percentage

In the above figure (12) shows that the 'occurrence of work Interfere at workplace by Percentage', about 78% of respondents have experienced interference at work with a ratio of rarely, sometimes, and frequently.

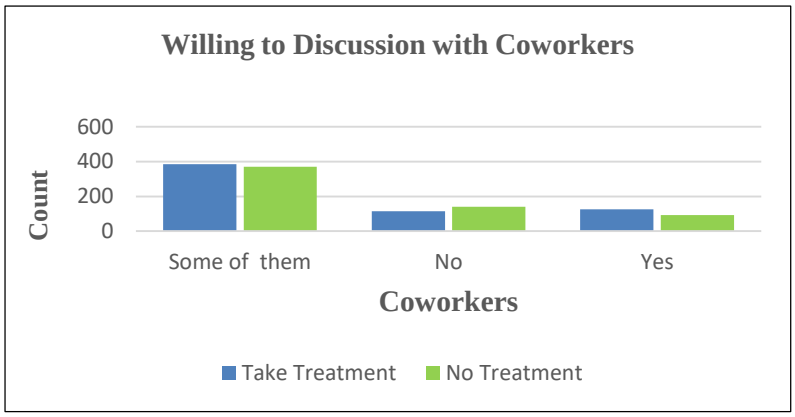

Figure 13: Respondents who are willing for discussion with coworkers.

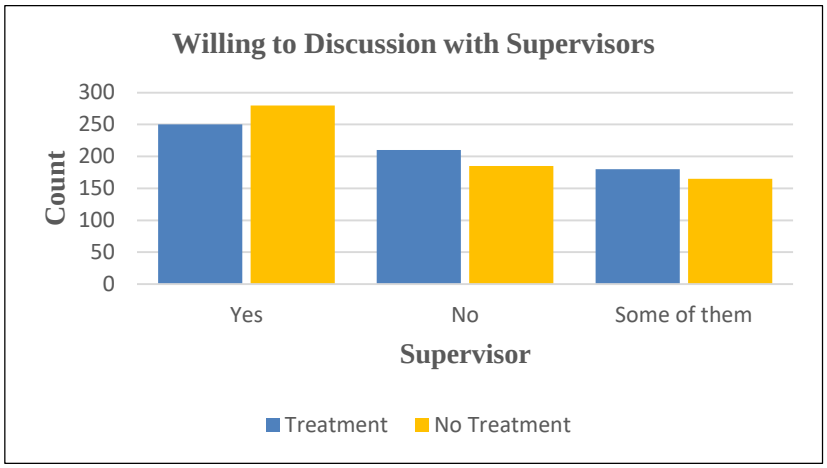

Figure 14: Respondents who are willing for discussion with supervisors.

The above figure (13) and (14) shows 'the number of poor mental health people who are willing to discuss with coworkers and supervisors'. According to a pattern released by exploratory data analysis, most people with mental health conditions are more inclined to make discussions with their superiors than coworkers.

Model Evaluation

Evaluating a model is the core part of creating an effective model. After building machine learning models by using Random forest, Decision tree and K-Nearest Neighbors algorithms, next accuracy of the models are measured to make improvements and continue until achieving a desirable accuracy.

The measurement scores described in table (3) are obtained by running the code of program of three different machine learning classifiers and results are as shown in figure (15).

```
In [3]: runfile('D:/mental_health/04Allscore_edit.py', wdir='D:/mental_health')
                    method  std score  recall score  mean score
0  Decision Tree Classifier   0.033664      0.617801    0.679451
1  Random Forest Classifier   0.018375      0.706806    0.762297
2       K Nearest Neighbour   0.027498      0.612565    0.589588
                    method                                cv score
0  Decision Tree Classifier  [0.64444, 0.64045, 0.69663, 0.73034, 0.68539]
1  Random Forest Classifier  [0.77778, 0.77528, 0.73034, 0.75281, 0.77528]
2       K Nearest Neighbour    [0.63333, 0.60674, 0.58427. 0.5618, 0.5618]
```

Figure 15: Standard deviation, Recall, Mean and CV scores of three Classifiers

Table 3: Accuracy Measurement Table for the model comparisons

Method	Standard score	Recall score	Mean score	CV score
Decision Tree	0.033664	0.617801	0.679451	[0.64444, 0.64045, 0.69663, 0.73034, 0.68539]
Random Forest	0.018375	0.706806	0.762297	[0.77778, 0.77528, 0.73034, 0.75281, 0.77528]
K-Nearest Neighbors	0.027498	0.612565	0.589588	[0.63333, 0.60674, 0.58427, 0.5618, 0.5618]

Standard score: Standard score is the number of standard deviations by which the value of a raw score (i.e., an observed value or data point) is above or below the mean value of what is being observed or measured.

Recall score: The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

Mean score: Mean score is to measure the accuracy of the model.

CV score: Cross-validation (CV) is used to estimate the skill of a machine learning model on unseen data.

Discussion

The objective of this study is to provide awareness of potential mental health and to discover the factors influencing on mental health among socio-demographics data and information on the occupations. The experimental results of three machine learning classifiers are shown in this paper and Random Forest classifier achieves the highest accuracy with 76% and K-Nearest Neighbors classifier is the lowest accuracy with 58% and that of Decision Tree classifier is 67%.

Based on Exploratory Data Analysis (EDA), data visualization makes better data-based decisions on the relation between mental health and associated factors. A histogram shows that most of the survey respondents are around 20s to early 40s and the distribution of ages indicates that to have younger employees. There is no statistically significant difference of ages between respondents that get treatment and no treatment. According to gender, almost 79% of respondents are male as the Tech work space and a queer is less than 2%. It was also found that 40% of respondents who say that they have a family history of mental illness, the plot shows that they significantly want to get treatment rather than without a family history. Relation with mental health and work interference, it was found that about 78% of respondents have experienced interference at work and mental health conditions sometimes become interfere at work is about

45%. The respondents result of question, 'Would you be willing to discuss a mental health issue with your coworkers and Supervisors', 18% of respondents who say yes to discuss it with coworkers, 60% of them want to get treatment and 40% of respondents who say yes to discuss with supervisor, only 55% of them want to get treatment. Based on survey respondents shown on graphs, generally mental health disorder came up triggered on the workplace, consequently it was found that mental health is closely related to the type of working style.

Conclusion

In this study, different machine learning algorithms such as Decision Tree, Random Forest and K-Nearest Neighbors were applied for classification, and identified factors associated with mental health patterns. This study found that Random Forest Model has the best accuracy among the three prediction models using Machine Learning Techniques. Exploratory Data Analysis (EDA) are used to analyze the data using visual techniques. Dataset for this study was from kaggle.com containing 1260 employee records with 8 selected attributes for this paper.

Acknowledgements

I would like to give my warmest thanks to Professor Dr. Thet Thet Hlaing, Department of Computer Studies, University of Yangon for guiding me through all the stages of doing my research. I would like to express my special thanks to Professor Dr. Soe Mya Mya Aye, Head of Department of Computer Studies, University of Yangon for her kind permission to carry out this research. My thanks for my gratitude to U Zaw Win Htun, Principal of Myanmar Mercantile Marine College for giving me a chance to do this research.

References

Aleem, S., Noor ul Huda, Rashid Amin, Samina Khalid, Sultan S. Alshamrani and Abdullah Alshehri, (2022) "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions", https://doi.org/10.3390/electronics11071111.

Das, J., Quy-Toan Do, Jed Friedman and David McKenzie, (2009) "Mental Health Patterns and Consequences: Results from Survey Data in Five Developing Countries", The World Bank Economic Review, Vol. 23, No. 1, pp 31-55.

Gold, A., Danny Gross, Abdul Latif Jameel, (2022) "Deploying Machine Learning to Improve Mental Health", MIT News, https://news.mit.edu.

Kaur, P., Ravinder Kumar and Munish Kumar, (2019) "A Healthcare Monitoring System Using Random Forest and Internet of Things (IoT)", Multimedia Tools Applications 78, 19905–19916.

Mind (2017) "Understanding mental health problems: Introduction of Mental Health Problems", Booklet, https://www.mid.org.uk.

Shafiee, S. M., Sofianita Mutalib, (2020) "Prediction of Mental Health Problems among Higher Education Student Using Machine Learning", International Journal of Education and Management Engineering, 10(6):1-9.

Thieme, A., Danielee Belgrave and Gavin Doherty, (2020) "Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems", ACM Trans. Human Computer Interaction. Vol. 27, No. 5, Article 34, pp 1-53.

Vaishnavi, K., U Nikhitha Kamath, B Ashwath Rao and N V Subba Reddy, (2022) "Predicting Mental Health Illness using Machine Learning Algorithms", Journal of Physics: Conference Series, AICECS 2021.