INJ INTERNATIONAL NEUROUROLOGY JOURNAL

Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods

[1]**Varsha Naik,** [2]**Ahbaz Memon,** [3]**Snehalraj Chugh,** [4]**Yashada Nikam,** [5]**Tanaya Patole,** [6]**Debajyoti Mukhopadhyay**

[1,2,3,4,5]Dr. Vishwanath Karad MIT World Peace University, Survey No: 124, Paud Rd, Kothrud, Pune, 411038, Maharashtra, India.
[6]WIDiCoReL Research Lab, Mumbai 400601, Maharashtra, India.

Abstract

Open Sourcing Mental Illness (OSMI), conducts surveys and discussions with IT companies to raise awareness about mental illnesses, ascertaining that 51% of IT employees face mental health difficulties. As part of our research, we applied a variety of machine learning techniques and approaches to the OSMI 2016 dataset.

The first emphasis was on cleaning up rogue data, adjusting odd values, and testing null values. Through encoding and normalising the data, we used supervised and unsupervised learning models, with combinations of serial and parallel approaches by changing "hyper-parameters" to determine the optimal accuracy. Primary results were obtained by selecting, extracting, and eliminating features from higher-dimensional data sets. The Random Forest, an ensemble model, gave the best accuracy of 97%. far better than earlier studies. We hope that through this study, researchers and mental health professionals will be better aware of mental health disorders and thereby minimise the stigma associated with them.

Key Words: Mental Health; Linear Regression (LR); K Nearest Neighbor (KNN); Decision Tree (DT); Random Forest (RF); Support Vector Machine (SVM); Logistic Regression (LogR).

Introduction

IT industry has flourished as a result of the contributions of these exceptional people. Always praised in terms of productivity and inventiveness, there is a darker side to it all. The stakes in the digital world have never been higher than they are now. Tech start-up entrepreneurs confront ever-increasing competition, and they labour tirelessly to turn their ideas into viable business plans. This fast-paced and innovative industry puts immense pressure on them to run an efficient and successful firm and keep their position at the top of the heap. Hazardous working conditions, such as long hours and hard deadlines, are common in this industry.

More than one in five IT and computer workers suffer from mental health concerns, according to OSMI. In recent studies, an entrepreneur's chance of having a mental health condition has increased by 50% (Fofana et al., 2020) . Open Sourcing Mental Illness (OSMI) is a non-profit organisation that strives to promote public awareness about mental health issues in the IT sector and eradicate the stigma associated with them. There is a huge financial cost associated with mental illness, and these professional areas must be made aware of this. Organizations benefit from OSMI's assistance in identifying and addressing these challenges (Muruganandam et al., 2020).

To maximise production and ensure the health and happiness of workers, creating a pressure-free work environment is a top priority. Counselling, professional coaching, developing coping workshops, and healthcare education and awareness are just a few options for helping professionals deal with stress and maintain their mental health. The likelihood of these kinds of initiatives proving effective increases when workers who will benefit from them are identified early on.

ISSN:2093-4777 | E-ISSN:2093-6931
DOI: 10.5123/inj.2023.4.in87
Vol. 27 Iss. 4 (2023)

We seek to simplify this procedure by utilising ML techniques to construct a model to predict the stress faced by the employees, which was accomplished by OSMI but had poor required output, delivering low accuracy and erroneous forecasts about the psychological wellbeing of individuals (Brouwers, 2020). In order to improve the

accuracy, we take procedures like preprocessing and adjusting hyperparameters into consideration. In addition to helping HR managers get a better understanding of their workers, a model like this may be used to make proactive efforts to reduce the likelihood of a staff member departing or functioning below expectations. By using advanced prediction techniques, a person's need for mental health care may be predicted well in advance.

Data Preprocessing

We drew on the OSMI 2016 dataset, which is freely accessible and contains information on IT and tech employees. For each of the 63 questions in this dataset, there are criteria on which each tech worker was tested. The survey also includes 1433 employee answers, which encompass both professional and personal aspects of the workers, and so give insights into the work environment experienced in IT firms. Further analysis of the data revealed that the original dataset had 2196 null and duplicate items, which can be seen in the Figure 1 as stage 1. We started by removing the irrelevant observations, null entries, and duplicate items from each column (18 columns). Following a thorough examination of the dataset, we discovered that there were 5682 null entries, as shown in the figure, which represents the first step of the analysis outlined here. Most of these variables were categorical, making it impossible to infer the values that were missing. Since we didn't want to change the values of the columns until we did our exploratory data analysis, we kept these variables.

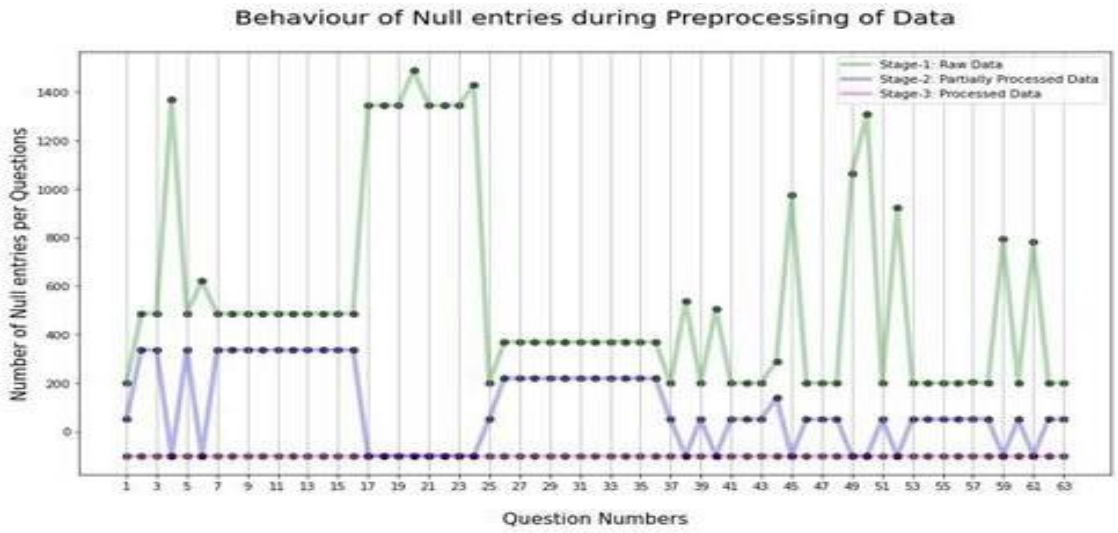
Before preprocessing, we had to remove all NaN values from a dataset in order to make our algorithm powerful and to make our model stand out (Islam et al., 2019; Papadopoulos et al., 2021). Therefore, we employed zero and one allocation to assist. Because of this, we renamed our workers' genders as 0 and 1, which is to say, they were renamed as 0 and 1. Yes and No were likewise substituted with 0 and 1 replies. In addition, "no" and "not much" are substituted by 0 in several cases. The class-dependent method is the name given to this particular replacement method.

In addition, we used the class priority approach to distribute the odd reply among classes that already had comparable responses. All of the "neutral" replies were placed in "class 0," and then the "yes/no" questions were gathered and placed in "class 1." Class 1 now contains all of the responses. Class 2 was formed by merging all the comments that included the word "maybe" together. Additionally, we classified all of the replies that were identical to "no" into classes 3 through 4, as well as "class-1" and "class-2" for the responses that were the most difficult to categorise.

As a result, several nations were assigned to "The United States" since they had a relatively low number of tech employees in comparison to other countries. We focused on the respondents' gender preferences while composing the survey. As a result of the answers, each gender has had 49 different submissions sent to the contestants. In order to Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods 5 eliminate any remaining outliers, we've separated these numbers into three subcategories: "man," "female," and "queer." It was discovered that men were the most popular of the three categories.

According to the generalised ideology, several of the ages reported by the workers were incorrect. As a result, based on publicly available information, we estimated the youngest worker was 18 years old and the oldest retiree was 65 years old. The number 30 was assigned to responses with a score of more than 65 but less than 18 points. Due to the fact that employees under the age of 30 constituted a large majority of the workforce and, thus, the dataset, adjusting their ages to the median would have a little impact on the dataset's correctness.

Figure 1 Null items were deleted from each column in phases so that the data was not corrupted.



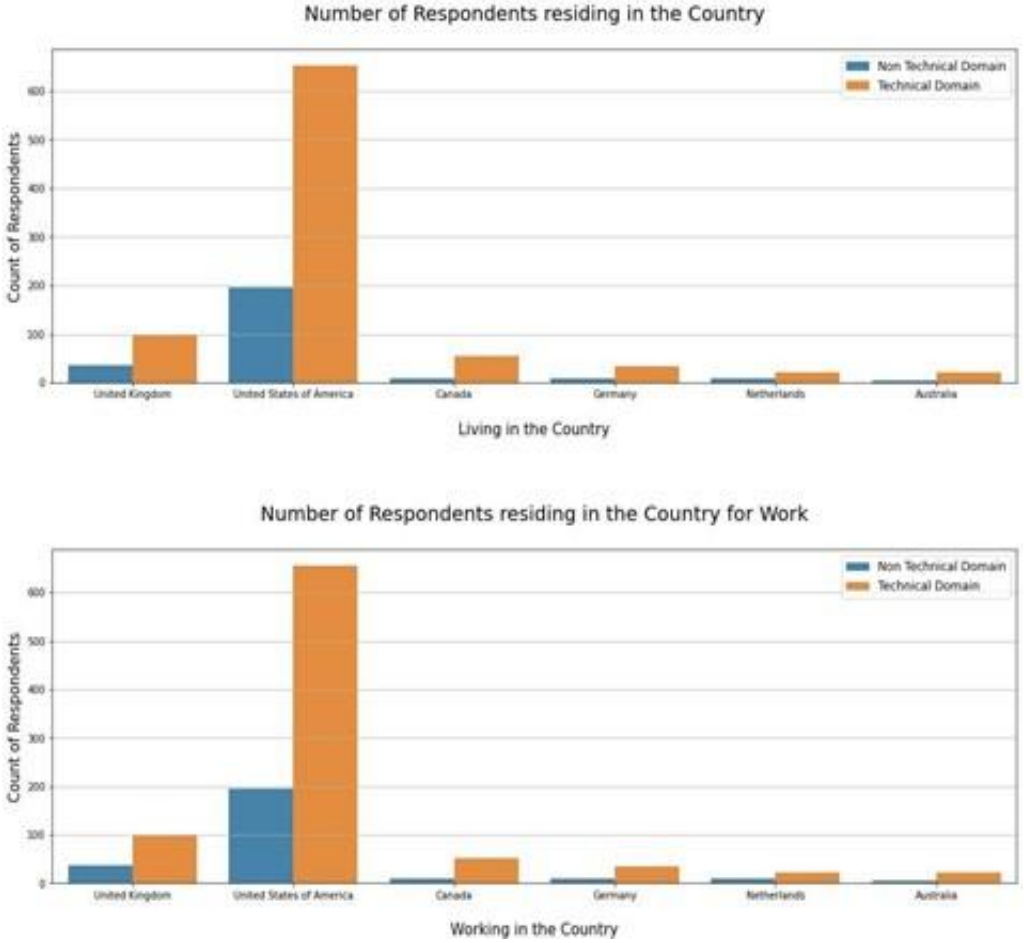
There was a tremendous quantity of data gathered for each aspect since this dataset includes respondents from all around the globe. It doesn't matter where workers work or reside; the same rules apply. That's how it was determined that most people lived and worked in these five countries: United States, Canada (Canada), Australia (Australia), UK (UK), Netherlands (NL). The United States was the most common nation to respond to the survey. It was possible to categorise the number of workers in each organisation, with values ranging from 1 to over 5000 (Laijawala et al., 2020). Thus, in order to make the dataset more efficient, we used the replacement by values approach to transform the categorical values to integer values (Reddy et al., 2018). Each company's workforce is now represented by figures such as 5, 25, 100, 500, 1000, and 5000.

Data Encoding

Data is encoded using libraries like One Hot Encoder, Transaction Encoder, and Label Binarizer in accordance with the prescribed format for safe transmission through encoding.

Using the unique values of each feature, one hot encoder creates the categories. As shown in the picture, we converted the categorical information to statistical information for the nations wherein the workers reside. Each employee's position of work was encoded using the Activity Encryption approach. This feature was stripped of its distinct values and handled as a set of separate qualities, thus portraying them in binary code (0 & 1). The Label Binarizer algorithm, which accepts categorical data as input and returns binary labels, was also applied to the replies from each nation where workers are employed in the column shown in Figure 2. This kind of preprocessing will help the model perform better.

Figure 2 The following bar graph illustrates the number of employees working or residing in "The United Kingdom," "The United States of America," "Canada," "Germany," "The Netherlands," and "Australia" in IT workplaces, regardless of whether they are in technical or non-technical jobs.



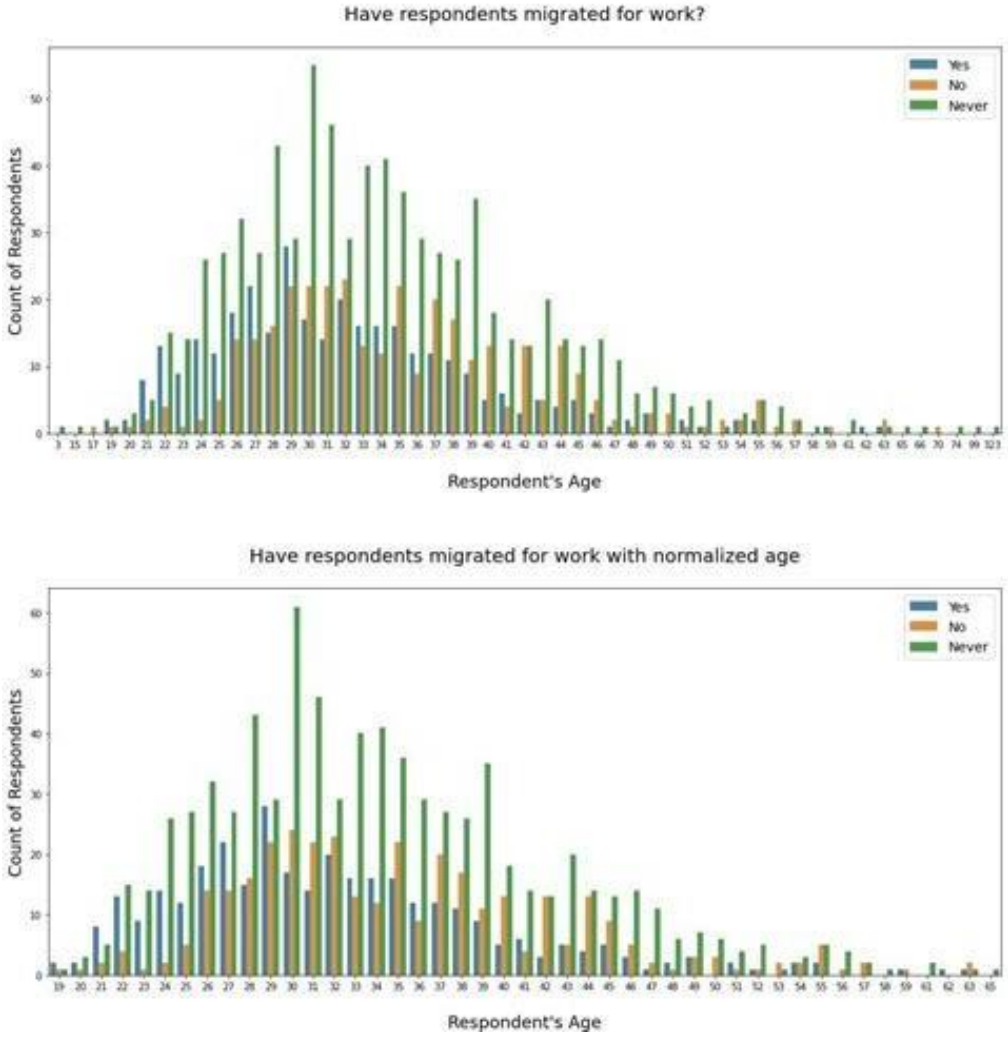
Data Normalization

Data normalisation is the process of converting the dataset's numeric columns to a standard scale so that the dataset's wide range of values isn't deformed or its contents obscured (Jacobson et al., 2020). For each organization, we utilised the Standard Scaler Normalization and MinMax Scaler Normalization for the number of workers and the age of employees (Sano et al., 2018). To help with the execution, each of them was hand-picked.

Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods 7 Normalization by Standard Scaler The pace of convergence throughout the optimization process increases with conventional scaler normalisation. The Standard Scaler reduces the variation and mean of each column to one unit. For each feature, it assumes that our data is normally distributed with a standard deviation of 1, and then scales the data such that the distribution is centred around 0. Before performing a StandardScaler Transformation, this function is used to normalise the null values, which is exactly what it does (Zebin et al., 2019). The column representing the total number of workers at each company was normalised using this technique is represented in the Figure 3.

DOI: 10.5123/inj.2023.4.in87
ISSN:2093-4777 | E-ISSN:2093-6931
Vol. 27 Iss. 4 (2023)

Figure 3 This graph displays the percentage of workers that answered "Yes," "No," or "Never" when asked whether they had relocated for employment and worked remotely at an IT company.



Normalization by Minmax Scaler

We used the Minmax Scaler Figure object in the Machine Learning library to normalise the column based on the age of workers. Using this estimator, each feature was altered and scaled to fit inside a certain range in the training set (Saha and Sharma, 2020). Finally, our dataset has been cleaned, encoded, and normalised to include no null values at this point, Seen in stage 3 of Figure 1.

Unsupervised Learning on Data Using Flat Clustering

When it comes to artificial intelligence, heuristic algorithms are used to help machines learn through data. There are several applications for clustering as a data analysis technique and as an unsupervised technique (Rueda and Krishnan, 2018). Clustering splits data into subgroups in which comparable examples are grouped, while instances that vary belong to separate groups. Because all clusters in a flat clustering technique are the same, it is better for big datasets when performance is critical. For our model, we used the K Means and the K Modes approaches.

K Means

A Euclidean representational area is assumed in the K Means approach, making it the simplest method. Inertia is usually shown as a result of this (Yuan and Yang, 2019; Yeasmin et al., 2020). The elbow approach and the silhouette technique are both used to determine the optimal number of clusters, refer Figure 4.

Elbow method

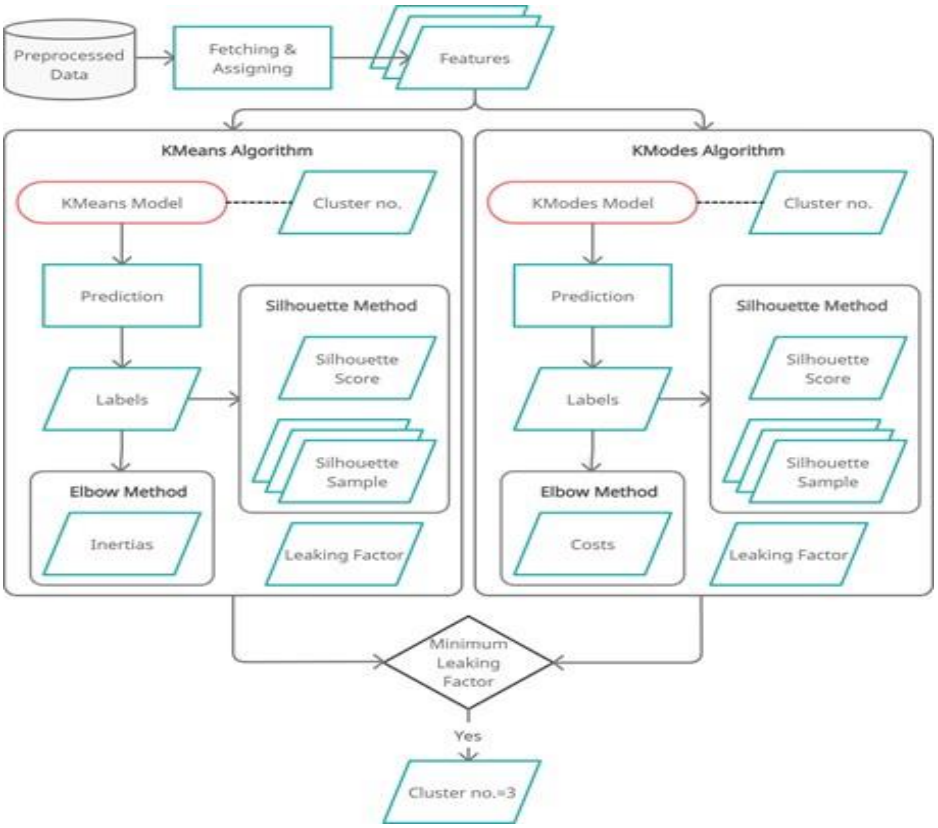
Clustering methods such as K-Means need the user to pick the number of clusters. One of the most effective methods for calculating the ideal value of k is the elbow approach. As the number of clusters increases, the proportion of variation explained increases as well. The scatter plot shows how much variation can be explained by a given number of clusters (Srividya et al., 2018). If our model is working well, it will keep finding k until it reaches the "elbow point," at which point the rate of reduction suddenly switches drastically shown in Figure 5. We were unable to choose a specific K since we found several local optimum points. As a consequence, we switched to the Silhouette Method in order to get the best outcomes possible.

Silhouette method

Using the silhouette approach, one may find the distances between any two points in a cluster. In the -1 to +1 range, the values may be found. It shows that the sample is far away from its neighbouring cluster and is quite near to the one it is allocated to. For similar reasons, the point's value of -1 shows that it's closer to its neighbouring cluster than its actual cluster (Srividya et al., 2018). A value of 0 indicates that the object is located on the cusp of the distance between the two clusters. For each sample in a separate cluster, the silhouette scores are derived using aspects of the data. For each cluster, a silhouette score is calculated to calculate the average of the silhouette coefficients for each sample.

Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods 9

Figure 4 This figure shows flat clustering, in which models and their methods were applied to the dataset parallel, and depending upon the criteria, highlighted models and methods were chosen for our model.



There was a correlation between the points and the 3rd cluster, which we display in Figure 6. As a result, we may conclude that K means cluster number 3 is the ideal point. By analysing silhouette plots, we can determine the ideal K value in our situation, which is 3, for K-means clustering.

K Modes

The K Modes clustering algorithm is an extension of the K-means clustering algorithm (Sangodiah et al., 2021). It defines clusters based on the number of matching categorical values between data objects (Grané et al., 2020).

Figure 5 The X-Axis shows the number of clusters every iteration, while the Y-Axis shows the normalised costs and interias. It was found that KMeans was superior to KModes in terms of accuracy since it looked to be distorted.

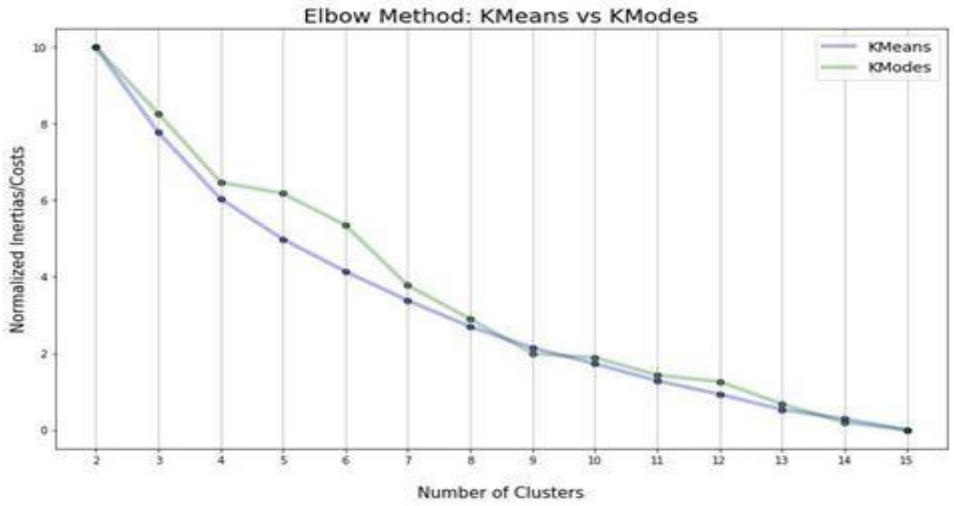
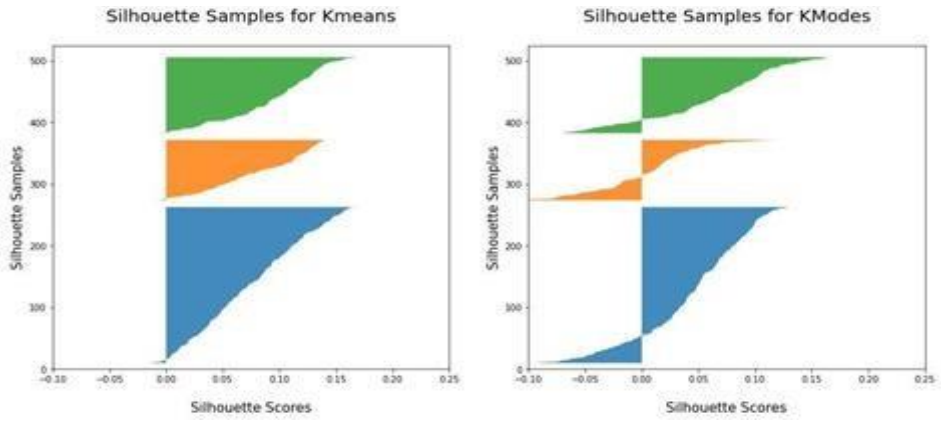


Figure 6 This figure shows the Silhouette Samples for KMeans and KModes determining the leaking factor and the difference between them.



Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods 11

Elbow method

The elbow method was adopted for the selection of the pre-assumed number of clusters. We couldn't settle on a particular K, because more than one local optimal point was acquired (Dewia and Dwidsamaraa, 2301). This gives its outcome in cost; hence, to get optimal results, we shifted to the Silhouette Method.

Silhouette method

We employed the silhouette method again while checking for the cluster number using K Modes. The results we were getting weren't up to the mark. The points coincided a lot, and hence, at this point, we concluded that the silhouette method using K Modes wasn't working for our dataset (Papachristou et al., 2018).

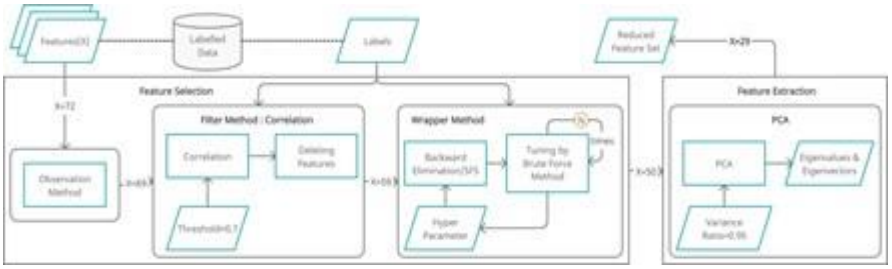
DOI: 10.5123/inj.2023.4.in87
ISSN:2093-4777 | E-ISSN:2093-6931
Vol. 27 Iss. 4 (2023)

The values of the outcomes we get from KMeans and Kmodes are normalized, i.e., cost and inertia, and thus used to compare, which can be seen in the graph. Here we can see that the elbow method for KMeans performs better than Kmodes. We then used the silhouette method to get more accuracy for the best number of classes for both methods as seen in Figure 6. Silhouette generates silhouette scores as well as samples for both the methods and, thus, is used to get the number of classes. For our model, we created a particular formula called the leaking factor, which helps in deciding a common ground on which class number and method is better. This helps in understanding the outcome and working out which setting performs better for our model. As a result, the model performs better when the leaking factor is lower.

Feature Engineering

Finally, we applied feature selection techniques after adopting unsupervised learning on the dataset to increase accuracy and save training time (Chancellor and Choudhury, 2020) . Due to the fact that the number of characteristics increases inexorably, this method was used. As a result, we were able to exclude data that was contributing to our model's decreasing accuracy due to being irrelevant, deceptive, and sparse. A total of 72 features, seen in Figure 7 , are gathered via this process of data extraction, which includes removing characteristics that aren't significant and analysing the data.

Figure 7 Models and techniques were applied to the dataset in a mix of series and parallel, and based on the criteria, models and methods shown in this figure were selected for feature reduction using Feature Selection and Extraction (FSE). We eventually reduced the feature set to 27.



Filter Method

Filter approaches to aid in analysing the relevance of predictions made outside of predictive models and then modelling just those that are projected to meet certain conditions. We also need the classes in order to get the filter function. This is accomplished by obtaining the correlation matrix, which enables us to discern the connections between the columns (Abd- Alrazaq et al., 2019).

Correlation

Data analysis at this point contained 69 characteristics, refer Figure 7, and we utilised the correlation approach to remove those that were not essential. Maps that demonstrate correlations between variables are called heat maps (Alonso et al., 2018). To discover the pairwise correlations among all the variables in the data frame, we used heat-map correlations to analyse the data frame (Braquehais et al., 2020). There were only 59 characteristics remaining after the correlation algorithm was used, and all the non-existent values were immediately eliminated. In addition, characteristics that included irrelevant data were eliminated. Thus, we were able to eliminate the uncertainty by using correlation.

Wrapper Method

A wrapper approach, which offers a variety of elimination strategies, evaluates many models by adding and removing predictors to identify the best combination that maximises model performance. On the other hand (Yitayih et al., 2021), the backward elimination method is used in our model, refer Figure 7.

Backward Elimination by Sequential Forward Selection Method

The technique begins with a "universal set," which is a term for the whole collection of characteristics. It takes out the poorest aspects of the dataset one by one. Selecting the optimal combination of characteristics using a criteria function and permutations is the goal of sequential forward selection. Once a certain number of characteristics have been identified, the procedure repeats itself. For each iteration, the Y-axis shows the combination of characteristics that were added (the initial iteration is at the bottom). The X-axis displays the classification accuracy (percentage). We may regulate the number of processors by n tasks using a random forest classifier as an SFS estimator. When we use "Floating" and "Scoring," we get more accurate predictions by subtracting and squaring the new forecast, respectively. We ran over the whole collection of features in order to

ISSN:2093-4777 | E-ISSN:2093-6931
DOI: 10.5123/inj.2023.4.in87
Vol. 27 Iss. 4 (2023)

get the greatest possible mix of accuracy and features, seen in Figure 8. A feature that didn't belong in this combination was deleted, and we trained the remaining features using a loop, which we monitored for each loop's performance to see how well they were doing. The best combination of features was scored at 92% using the SFS and hyper-parameters tweaked to get an ideal number of 50 elements. At this point, we have 50 columns remaining, shown in Figure 7.

Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods 13

Figure 8 This figure shows the combination of the best feature set depending on the positive and negative deviation of accuracy. The reduced feature set, ie., 59, which we got from Filter Method had been backward eliminated on the basis of their deviations in wrapper method.

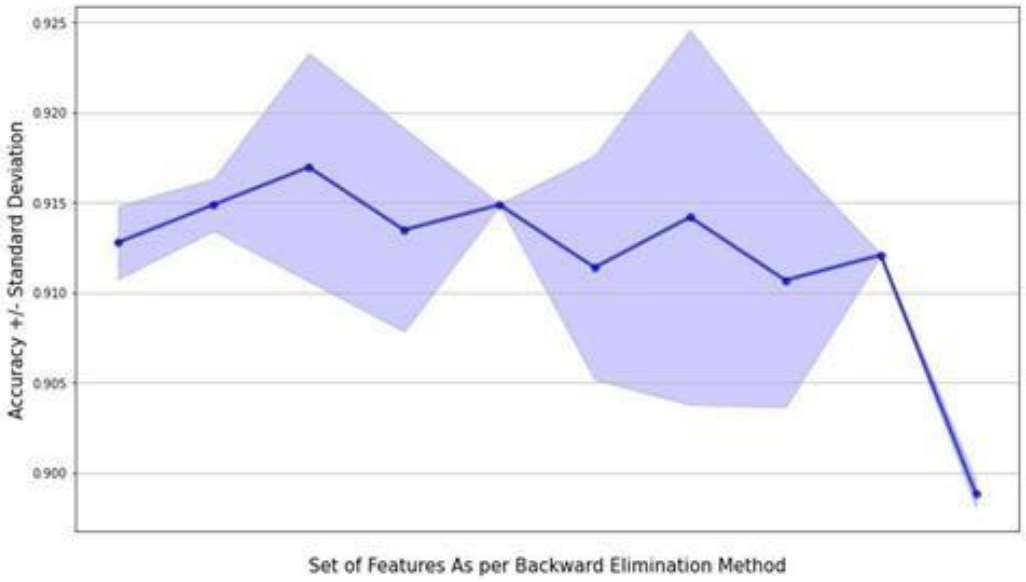
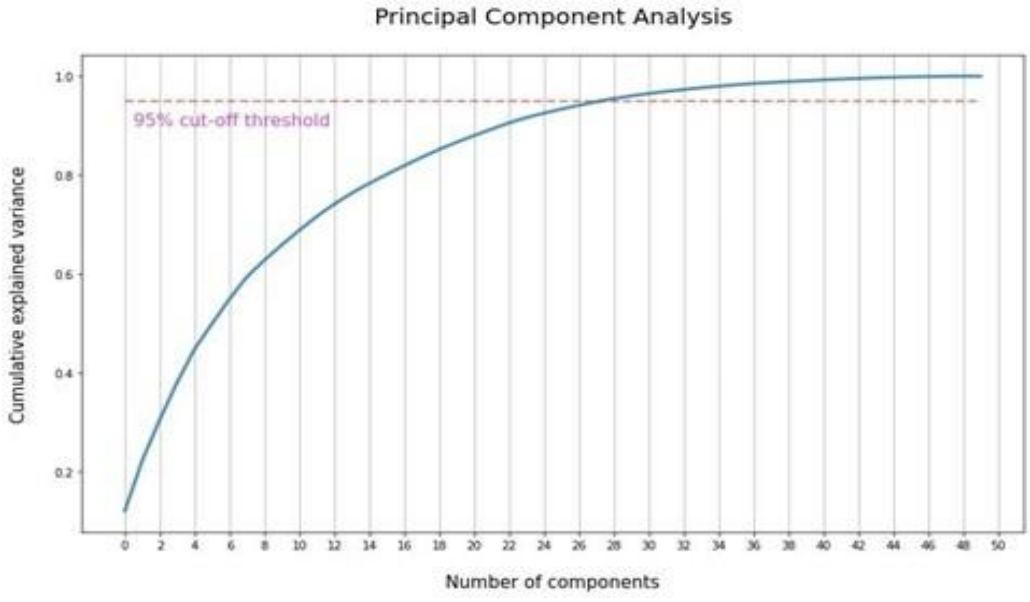


Figure 9 The reduced feature set, ie, 50, which we got from the Wrapper Method, had been reduced to a lower dimension without losing any data with the help of PCA. The optimal threshold of 95% had been extracted.



4.3 Principal Component Analysis

By using the SFS approach, we were able to delete 50 features that were no longer useful to us. While 50 features are a time-consuming process, it is possible. There are 50 distinct characteristics that make up the feature pattern. We need to reduce the dimensions of this data such that we don't lose any overall information. It is possible to minimise the dimensions of features by using Eigenvalues and Eigenvectors to calculate Eigenvalues and Eigenvectors (Suryavanshi et al., 2020; Clara et al., 2019). Using Principal Component Analysis, we can extract the most useful features from a dataset. It has been previously shown that achieving a variance of 95% is an excellent choice shown in Figure 9. Thus, the total number of features in the dataset was decreased to only 27 refer to Figure 7

Supervised Learning on Data

When a machine learns from prior acts and improves without being explicitly instructed, it is a subset of artificial intelligence (AI). Depending on the learning process or the input data, there are two basic forms of machine learning. The input and result of a supervised learning experiment are both known in advance since the data is labelled. A key difference between the two approaches is that in self-paced learning, we are given labels for the independent variables, but we are not given labels for the ones that are being learned. Only supervised learning methods will be employed in this case since the dataset is tagged, which is represented in Figure 10.

Linear Regression

Due to the fact that the splitter's training data has greater dimensions, linear regression is an effective method for data analysis. This means that the 27 characteristics from the splitter may be reduced to lower dimensions without losing any of the present data. In order to reduce the number of characteristics to just one, we employed principal component analysis (PCA) once again (Suryavanshi et al., 2020; Clara et al., 2019). In order to create the coefficient and intercept, we must first train a linear regression model on single-dimensional data. The seed for the random number generator is the random state. The Random State class is thus be used. To determine our loss, we'll use the cost function, which takes into account the actual and anticipated classes (Jacob et al., 2021; González-Sanguino et al., 2020). It will then attempt to optimise and identify the curve's lowest point. The coefficient and intercept will be controlled when the data is iterated through. Coefficients and intercepts have been fine-tuned and the total cost has been decreased up till now.

Logistic Regression

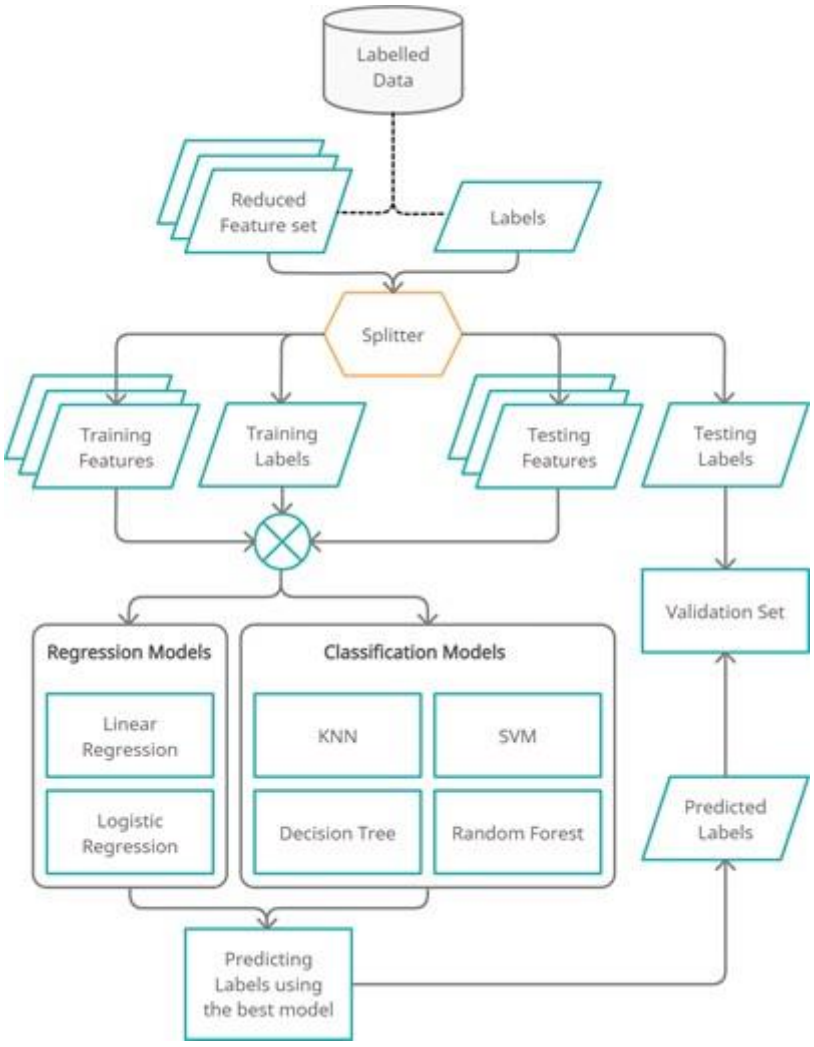
According to our observations, the supervised dataset was not linear. Hence, we used Logistic Regression on non-linear separable boundaries. In cases where the logistic regression could be used for both regression and classification, we chose the coefficients and direction at random. Because of this, we need to use the Logistics Cost function to determine our logistic loss. One or more dependent variables are used to predict the dependent variable Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods15 using the sigmoid function. Binary numbers between 0 and 1 are the most common form of the dependent variable. The sigmoid activation function is shown in Equation (1).

$$\sigma(\text{value}) = 1/(1 + e^{-\text{value}}) \qquad (1)$$

The logistic regression can be represented as followed by Equation (1).

$$\text{Ln}, f_0((p(y=1)/(1-p(y=1))= w_0+ w_1 x_1+\cdots+ w_n x_n \qquad (2)$$

Figure 10 The figure shows the flow of reduction of features using Feature Selection and Extraction, in which models and their methods were applied to the dataset in a combination of series and parallel, and depending upon the criteria, highlighted models/methods were chosen.



As a result of the non-linearity curve's many local minima, the elastic-net penalty criteria and regularisation by a factor of 10 assist us to attain global minima. Finally, we try to determine the Curve's global minima by optimising our model. Coefficients and intercept will be controlled as the data is iterated through. There is a reduction in total cost as a result of these optimizations.

K Nearest Neighbor

KNN is a lazy classifier since it keeps the data points during training and then runs through all of the pieces of data, sorting them, and identifying classes as it is predicting. A method known as the KNN classifier may be used to label data (Haines-Delmont et al., 2020). First, the KNN procedure is divided down into three parts: Minkowski Distance, Euclidean Distance, Manhattan Distance, Canberra Distance. Finally, he makes a forecast and focuses on gaining the trust of his neighbours. K closest neighbours are used to determine which class a data point will fall into when categorising it. The formula for Euclidean distance, which is the closest distance, is used in these cases.

Data points are then allocated to classes based on their perceived likelihood. Classification is based on how similar the independent variables are to comparable instances in the known data. KNN uses distances between all of the training locations and a single testing point to make predictions.

Depending on the situation, the kind of distances might change. We need to arrange the distances in decreasing order once we have calculated them. A K distance of odd is chosen first because the worst-case situation is that the forecast will be subjective if K is even and the distances are equal. The best bet for the testing point is to rely

ISSN:2093-4777 | E-ISSN:2093-6931
DOI: 10.5123/inj.2023.4.in87
Vol. 27 Iss. 4 (2023)

on the majority of the class as per Equation (1).

(x,y)= ∑_(j=1)^k▒√((x_j- y_j)^2) (3)

Decision Tree

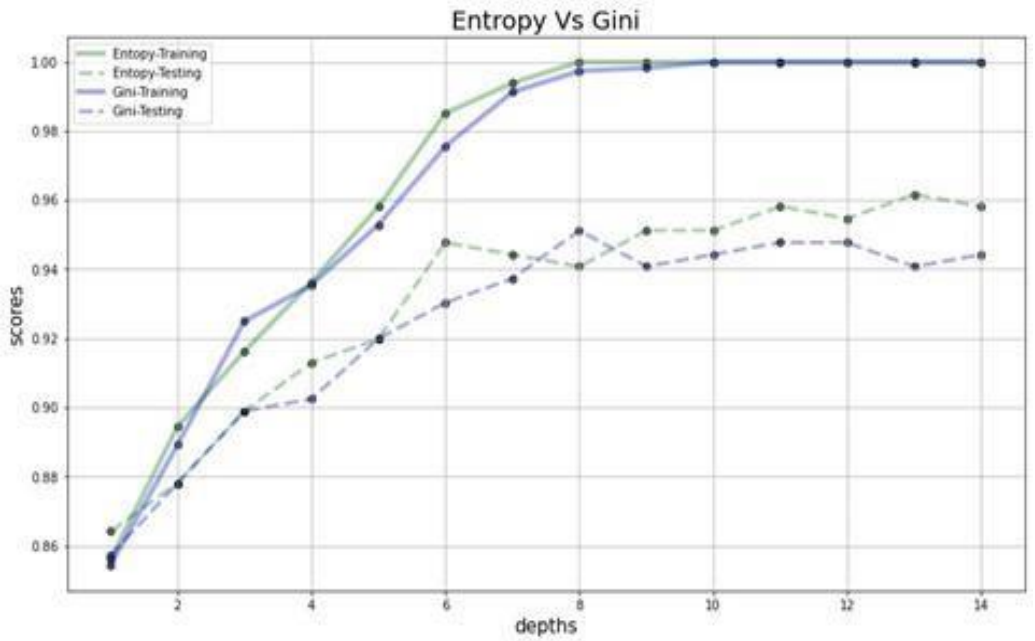
Our data contains a large number of characteristics and entries, which makes it a time- consuming task to train and forecast for each iteration (Alonso et al., 2018). One of the most memory-efficient data structures is the tree data structure. Feature and splitting criteria are used to form a tree in the Decision Tree.
Overall, the points of computation in the Decision Tree are:

If a node is pure, the output is only the respective class.
If no Feature is Left to split upon, Out-put is the majority.
Finding the best feature to split upon in every iteration of the depth of a tree.
The recursive call to the self-function and analysing the objective metrics which are Accuracy Score, Information Gain, Gain Ratio, Gini Index.

When we split the features, we have to analyse which feature is going to give the best classification.

Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods 17 Criteria such as Entropy, Gini, and Max-Depth are derived from hyperparameters (Patel and Prajapati, 2018; Mienye et al., 2019) . There is an issue with Decision Trees since they readily overfit the data. However, we were able to fix the issue by fine-tuning the hyperparameter. With 11 as a Max-Depth, we've chosen Entropy as a criterion in order to prevent overfitting by keeping the training and testing set's points near enough together, seen in Figure 11.

Figure 11 Test results were used to compare Entropy and Gini's performance. The depth of the Decision Tree was used to examine the separation of training and testing lines.



Random Forest

Ensemble models utilise a variety of machine learning methods in order to improve their performance over other models. We've utilised enumerators in Random Forest to replicate the N-Decision Tree or a collection of N-Difference Trees. Using the N decision tree, it creates random batches of data and trains it over them. Weight is applied to a decision tree to determine the priority of a certain branch and to train the model (Speiser et al., 2019). We use metrics such as entropy, Gini, and Max-Depth as a basis for selecting hyperparameters (Probst et al., 2019). The random forest has an issue with overfitting the data. However, we were able to fix the issue by fine-tuning the hyperparameter. It was decided that entropy and the maximum depth of 10 would ensure that the training and testing sets' score points were near enough to prevent overfitting. We were able to overcome the difficulty of over-fitting since our findings were so broad.

ISSN:2093-4777 | E-ISSN:2093-6931
DOI: 10.5123/inj.2023.4.in87
Vol. 27 Iss. 4 (2023)

Support Vector Machine

An SVM is a supervised machine learning technique that may be used for classification. After the analysis, the data is non-linear, more complex, and more susceptible to being over- fitting. In order to differentiate between the classes, we have employed the margin lines. When it comes to kernels, they may be linear, poly, RBF, sigmoid, or precomputed (Huang et al., 2018). The margins, on the other hand, can be soft or hard. Its two margin limits have changeable lengths between them, which may be adjusted up by learning from data. Graphically Due to margins, the binary class system region is split into two distinct portions. Misclassified data in N-dimensional features may be handled by these two components. We've experimented with a variety of optimization parameters. Afterwards, the classes are categorised based on the optimal distance between the margin, coefficients, and intercept. To optimise performance, we've experimented with a variety of kernels. For our dataset, the RBF kernel performed well in terms of precision.

Results and Accuracy

After the data had been cleaned up, there were no undesirable values or negative deviation points left in the data set. An important part of pre-preparation was making sure the data had been normalised and encoded in a categorised format. KMeans of hyperparameter 3 were utilised to form the clusters using the silhouette technique, seen in Table 1. We used correlation (with a threshold of 0.7), then SFS (with an estimator of 50), and finally PCA (with a variance threshold of 0.95) to extract the features, Figures 8 and 9. A Random Forest, an Entropy Criteria, and a maximum depth of ten were used to make class predictions. The Figure 11 shows a 97 per cent accuracy rate. Based on a variety of criteria and accuracy measures, the best possible routes were chosen at each phase of the conversion process for OSMI 2016 Survey data, as seen in the Figure 12.

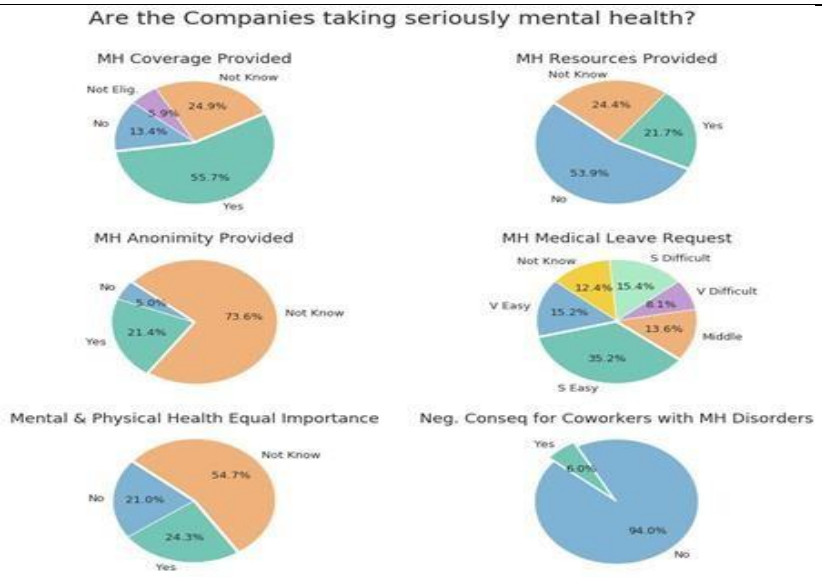
Table 1 Table shows the results of Flat Clustering models, their methods, and Hyper Parameters in respective columns. Which are analyzed using an accuracy matrix, Inertia, Cost and Silhouette Scores. Eventually, Number of Cluster was selected as 3 for the model KMeans using the Silhouette Method.

Flat Clustering	Method	Inertia/Cost/Silhouette Score	Hyper Parameters
			Cluster Number
K Means	Elbow Method	50000	5
	Silhouette Method	0.075	3
K Modes	Elbow Method	27000	3
	Silhouette Method	0.036	3
K Means for cluster number 3 is decided by Silhouette Method.			

DOI: 10.5123/inj.2023.4.in87

ISSN:2093-4777 | E-ISSN:2093-6931
Vol. 27 Iss. 4 (2023)

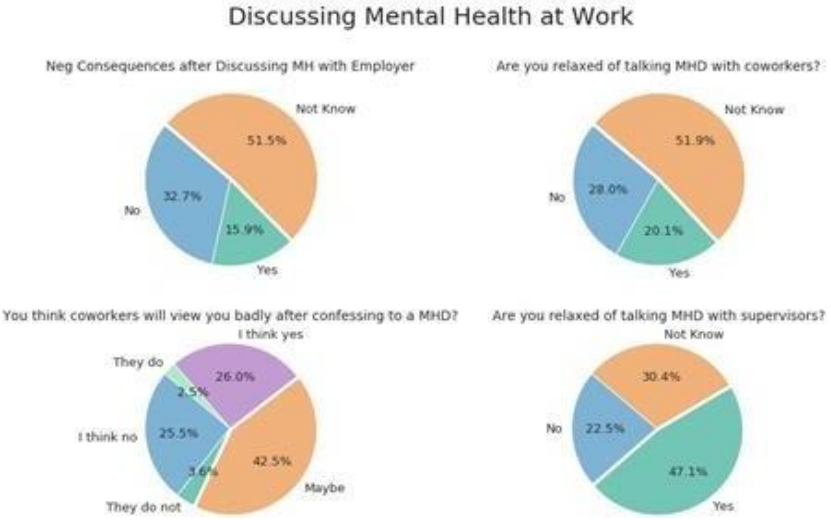
Figure 12 The number and proportion of IT workplaces offering Mental health coverage, resources, anonymity providence, and leave are shown according to various conditions. 'Yes', 'No', 'Not known', etc are polls of employees working in respective companies, as the answers of these respective questions



A single model, the decision tree, has already been used to determine the relative relevance of the 14 qualities under consideration. For prediction purposes, the supervised model was trained using several machine learning techniques. Boosting had a maximum accuracy of 75%, while bagging had a maximum accuracy of 69%. The KNN classifier had the highest proportion of false positives in the given example, indicating that it was very inaccurate. The accuracy of the remaining models, on the other hand, was about 70%, all of this shown in Figure 14.

Feature Selection and Extraction, on the other hand, was the approach we used in our research, in which models and procedures were applied sequentially to the dataset in both series and parallel, with the most relevant models and methods being picked based on the criteria. Finally, we were able to pare down the features to a manageable 27. Random Forest was the most accurate model at 97%, while the remainder of the models averaged 88% accuracy, Figure 14. When all of the models and techniques were applied simultaneously to the datasets, we performed flat clustering, shown in Table 2.

Figure 13 The number and proportion of the consequences faced by the employees in IT workplaces after appealing to their employers are communication gaps, losing opportunities, etc. are shown here basically showcasing their mental health



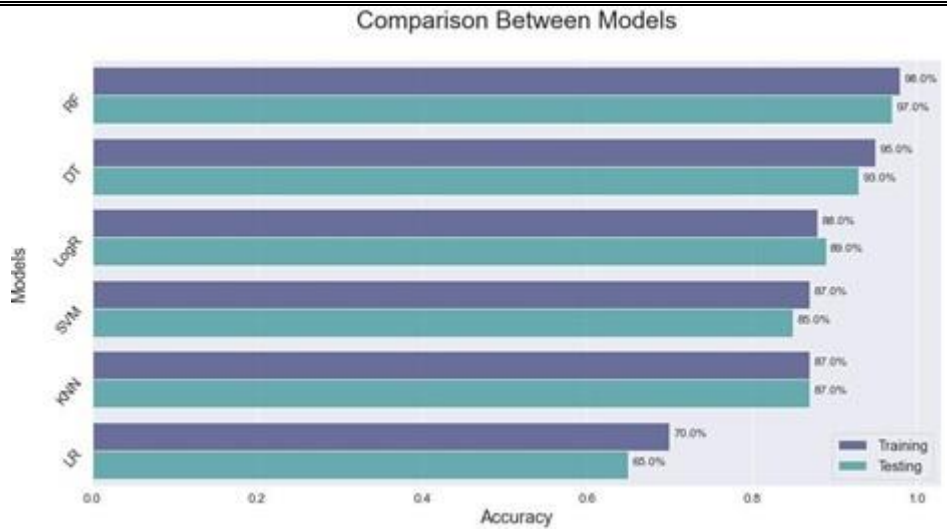


Figure 14 Accuracy Percentage of different algorithms shown in respective bars, where Linear Regression has the worst accuracy and Random Forest has the highest accuracy of 98%

Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods 21

Table 2 Table shows the results of Supervised Learning models, their methods, algorithms & Hyper Parameters, ie., Criterion, Penalty, Regularization Factor (C), Iteration (N), Max-Depth in Decision Tree and Gamma in SVM, etc in respective columns. Which were analyzed using an accuracy matrix; F1 Scores for individual classes and overall scores for all classes

Supervised Learning	Algorithms	Hyper Parameter		F1 Scores						Overall Scores	
		Criterion/ Penalty	C/N/Max depth/gamma	Class 0		Class 1		Class 2		All Classes	
				Training	Testing	Training	Testing	Training	Testing	Training	Testing
Regression	Linear Regression	Gradient Descent	1	0.069	0.065	0.071	0.066	0.069	0.067	0.70	0.65
	Logistic Regression	Elasticnet	10	0.9	0.9	0.86	0.85	0.87	0.8	0.88	0.89
Classification	K Nearest Neighbor	Mean	3	0.85	0.84	0.88	0.87	0.87	0.87	0.87	0.85
	Decision Tree	Entropy	11	0.98	0.97	0.94	0.92	0.945	0.93	0.95	0.93
		Gini	11	0.97	0.96	0.95	0.92	0.96	0.92	0.97	0.92
	Random Forest	Entropy	10	1	0.97	0.99	0.98	0.98	0.965	0.98	0.97
		Gini	11	0.99	0.975	0.95	0.94	0.965	0.94	0.97	0.96
	Support Vector Machine	Linear	0.1/1/10/100	0.85	0.86	0.84	0.85	0.84	0.83	0.85	0.855
		RBF	1	0.87	0.87	0.85	0.855	0.86	0.85	0.87	0.87
		Poly	0.1/1/10/100	0.86	0.86	0.85	0.85	0.83	0.84	0.86	0.85

Conclusion

To better understand the mental health issues faced by IT workers, this research was conducted. As a result of the long hours necessary to stay on top of the newest technical advancements, employees in this area are certain to get psychologically exhausted. Our models aim to anticipate the consequences of mental health illnesses in overworked employees in order to identify the organisations most likely to be impacted. By conducting surveys and workshops with IT companies, Open Sourcing Mental Illness (OSMI) hopes to promote awareness of the mental illness and provide support.

We observed that 76% of IT employees were male as the sector grew. Many employees in the IT business want mental health coverage, confidentiality, leave acceptance, and other benefits. A communication gap, missed opportunities, HR politics, and more are all consequences for IT employees who go to their managers for help. With regards to high-tech companies, the number of people in countries such as the UK, U.S. and Canada is working or living in countries such as Germany and Australia in technical or non-technical roles, and this is

ISSN:2093-4777 | E-ISSN:2093-6931
DOI: 10.5123/inj.2023.4.in87
Vol. 27 Iss. 4 (2023)

increasing.

There was a 75 percent accuracy rate for boosters and a 69 percent accuracy rate for baggers in earlier research. A total of 27 characteristics were achieved in our investigation using a variety of models and approaches. We improved Random Forest's accuracy to 97 percent, while the other models averaged 88 percent accuracy.

References

1. Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., and Househ,
2. M. (2019). 'An overview of the features of chatbots in mental health: A scoping review'.
3. International Journal of Medical Informatics, Vol 132, pp. 103978–103978.
4. Alonso, S. G., Torre-Díez, I. D. L., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., and Franco, M. (2018). 'Data mining algorithms and techniques in mental health: a systematic review'. Journal of medical systems, Vol 42, No 9, pp. 1–15.
5. Braquehais, M. D., Vargas-Cáceres, S., Gómez-Durán, E., Nieva, G., Valero, S., Casas, M., & Bruguera, E. (2020). The impact of the COVID-19 pandemic on the mental health of healthcare professionals. pp. 613-617.
6. Brouwers, E. P. (2020). 'Social stigma is an underestimated contributing factor to unemployment in people with mental illness or mental health issues: position paper and future directions'. BMC psychology, Vol 8, No 1, pp. 1–7.
7. Chancellor, S. and Choudhury, M. D. (2020). 'Methods in predictive techniques for mental health status on social media: a critical review'. NPJ digital medicine, Vol 3, No 1, pp. 1–11.
8. Clara, R. A., Simon, D., Noelia, G., and Barbara, A. (2019). 'Critical elements in accessible tourism for destination competitiveness and comparison: Principal component analysis from Oceania and South America'. Tourism Management, Vol 75, pp. 169–185.
9. Dewia, N. P. M. N. and Dwidsamaraa, I. B. (2301). 'Implementation of K-Modes Algorithm for Clustering of Stress Causes in University Students'. Jurnal Elektronik Ilmu Komputer Udayana p-ISSN, pages 5373–5373.
10. Fofana, N. K., Latif, F., Sarfraz, S., Bashir, M. F., and Komal, B. (2020). 'Fear and agony of the pandemic leading to stress and mental illness: An emerging crisis in the novel coronavirus (COVID-19) outbreak'. Psychiatry Research, pages 113230–113230.
11. González-Sanguino, C., Ausín, B., Castellanos, M. A., Saiz, J., López-Gómez, A., Ugidos, C., and Muñoz, M. (2020). 'Mental health consequences during the initial stage of the 2020 Coronavirus pandemic (COVID-19) in Spain'. Brain, behavior, and immunity, Vol 87, pp. 172–176.
12. Grané, A., Albarrán, I., and Lumley, R. (2020). 'Visualizing inequality in health and socioeconomic wellbeing in the EU: Findings from the share survey'. International Journal of Environmental Research and Public Health, Vol 17, No 21, pp. 7747–7747.
13. Haines-Delmont, A., Chahal, G., Bruen, A. J., Wall, A., Khan, C. T., Sadashiv, R., and Fearnley, D. (2020). 'Testing suicide risk prediction algorithms using phone measurements with patients in acute mental health settings: feasibility study'. JMIR mHealth and uHealth, Vol 8, No 6, pp. 15901–15901.
14. Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., and Xu, W. (2018). 'Applications of support vector machine (SVM) learning in cancer genomics'. Cancer genomics & proteomics, Vol 15, No 1, pp. 41–51.
15. Analysis and Prediction of Factors Influencing the Mental Health of IT Professionals Using Machine Learning Methods23
16. Islam, M. R., Miah, S. J., Kamal, A. R. M., & Burmeister, O. (2019). A design construct of developing approaches to measure mental health conditions. Australasian journal of information systems, Vol 23.
17. Jacob, L., Smith, L., Armstrong, N. C., Yakkundi, A., Barnett, Y., Butler, L., ... & Tully, M. A. (2021). Alcohol use and mental health during COVID-19 lockdown: A cross-sectional study in a sample of UK adults. Drug and alcohol dependence, Vol 219, pp. 108488.
18. Jacobson, N. C., Lekkas, D., Price, G., Heinz, M. V., Song, M., O'malley, A. J., and Barr, P. J. (2020). 'Flattening the mental health curve: COVID-19 stay-at-home orders are associated with alterations in mental health search behavior in the United States'. JMIR mental health, Vol 7, No 6, pp. 19347–19347.
19. Laijawala, V., Aachaliya, A., Jatta, H., and Pinjarkar, V. (2020). 'Classification algorithms based mental health prediction using data mining'. 2020 5th International Conference on Communication and Electronics Systems (ICCES), pages 1174–1178.
20. Mienye, I. D., Sun, Y., and Wang, Z. (2019). 'Prediction performance of improved decision tree-based algorithms: a review'. Procedia Manufacturing, Vol 35, pp. 698–703.
21. Muruganandam, P., Neelamegam, S., Menon, V., Alexander, J., and Chaturvedi, S. K. (2020). 'COVID-19 and severe mental illness: impact on patients and its relation with their awareness about COVID-19'. Psychiatry research, Vol 291, pp. 113265–113265.

ISSN:2093-4777 | E-ISSN:2093-6931
DOI: 10.5123/inj.2023.4.in87
Vol. 27 Iss. 4 (2023)

22. Papachristou, N., Barnaghi, P., Cooper, B. A., Hu, X., Maguire, R., Apostolidis, K., Miaskowski, . ., and C (2018). 'Congruence between latent class and K-modes analyses in the identification of oncology patients with distinct symptom experiences'. Journal of pain and symptom management, Vol 55, No 2, pp. 318–333.
23. Papadopoulos, P., Abramson, W., Hall, A. J., Pitropakis, N., and Buchanan, W. J. (2021). 'Privacy and trust redefined in federated machine learning'. Machine Learning and Knowledge Extraction, Vol 3, No 2, pp. 333–356.
24. Patel, H. H. and Prajapati, P. (2018). 'Study and analysis of decision tree based classification algorithms'. International Journal of Computer Sciences and Engineering, Vol 6, No 10, pp. 74–78.
25. Probst, P., Wright, M. N., and Boulesteix, A. L. (2019). 'Hyperparameters and tuning strategies for random forest'. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, Vol 9, No 3, pp. 1301–1301.
26. Reddy, U. S., Thota, A. V., and Dharun, A. (2018). 'Machine learning techniques for stress prediction in working employees'. 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pages 1–4.
27. Rueda, A. and Krishnan, S. (2018). 'Clustering Parkinson's and age-related voice impairment signal features for unsupervised learning'. Advances in Data Science and Adaptive Analysis, Vol 10, pp. 1840007–1840007.
28. Saha, K. and Sharma, A. (2020). 'Causal factors of effective psychosocial outcomes in online mental health communities'. Proceedings of the International AAAI Conference on Web and Social Media, Vol 14, pp. 590–601.
29. V. Naik et al Sangodiah, A., Spr, C. R., Jalil, N. A., Nee, A. Y. H., and Subramaniam, S. (2021). 'Investigation on Mental Health Well-Being for Students Learning from Home Arrangements Using Clustering Technique'. In Congress of the International Ergonomics Association, pp. pages 113–122. Springer
30. Sano, A., Taylor, S., Mchill, A. W., Phillips, A. J., Barger, L. K., Klerman, E., and Picard,
31. R. (2018). 'Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study'. Journal of medical Internet research, Vol 20, No 6, pp. 9410–9410.
32. Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, Vol 134, pp. 93-101.
33. Srividya, M., Mohanavalli, S., and Bhalaji, N. (2018). 'Behavioral modeling for mental health using machine learning algorithms'. Journal of medical systems, Vol 42, No 5, pp. 1–12.
34. Suryavanshi, N., Kadam, A., Dhumal, G., Nimkar, S., Mave, V., Gupta, A., Gupte, . ., and N (2020). 'Mental health and quality of life among healthcare professionals during the COVID-19 pandemic in India'. Brain and behavior, Vol 10, No 11, pp. 1837–1837.
35. Yeasmin, S., Banik, R., Hossain, S., Hossain, M. N., Mahumud, R., Salma, N., & Hossain, M. M. (2020). Impact of COVID-19 pandemic on the mental health of children in Bangladesh: A cross-sectional study. Children and youth services review, Vol 117, pp. 105277.
36. Yitayih, Y., Mekonen, S., Zeynudin, A., Mengistie, E., and Ambelu, A. (2021). 'Mental health of healthcare professionals during the early stage of the COVID-19 pandemic in Ethiopia'. BJPsych Open, Vol 7, No 1,.
37. Yuan, C. and Yang, H. (2019). 'Research on K-value selection method of K-means clustering algorithm'. J, Vol 2, No 2, pp. 226–235.
38. Zebin, T., Peek, N., and Casson, A. J. (2019). 'Physical activity based classification of serious mental illness group participants in the UK Biobank using ensemble dense neural networks'. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1251–1254.