



Coursework- Artificial Intelligence

Darsheta Dinesh Babu

32126858

MSc. Artificial Intelligence

Task 1: Proposal

Scope:

The scope of this project is to successfully predict whether an individual working in the tech industry has sought mental health treatment or not. With technology growing rapidly, the tech industry is facing immense pressure to update to the new advancements. This fast-paced technological advancement can have adverse effects on human mental health as they put in all the efforts to make AI and its technology huge. The primary objective of the model is to incorporate Machine Learning techniques to build a predictive model which can find individuals in the tech sector, with the help of a survey which is collected from various people working in the tech sector to see whether they have sought treatment or not.

Data:

The Kaggle dataset which is used in this analysis has a diverse set of features collected from the individuals working in the tech sector through a survey(<https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey/data>) The answers from the survey respond to a lot of diverse questions which is related to predicting whether an individual needs treatment or not. The dataset contains data such as age, gender and educational background. Job-related factors like working hours and company size, information about the individual's previous mental health records, and also whether they have been in treatment currently or before.

The emphasis of this dataset is its potential to explore the complicated relationship between the tech industry and mental health. By analysing this we can provide a lot of insights which will improve the quality of work in the workplace. Though it answers a lot of questions, the efficiency of the model can only be achieved if the data is treated properly. Before predicting, the missing values should be handled and also the data should be encoded in a way that both categorical and numerical features can be treated properly.

Baseline:

The baseline of the project includes developing a well-defined and simple model. This includes a lot of steps like

- **Data pre-processing-** To develop the model, we have to import the necessary packages and the dataset. Once the dataset is imported, the structure of the dataset should be analysed and the patterns of the data should be known. We need to check for missing values and perform techniques which will handle the missing values.
- **Feature selection-** to build a model with maximum accuracy, the features should be selected in such a way that the model can perform well. Only the relevant features with which the model accuracy will be high should be given for the predicting model.
- **Train-test model-** The analysis can't be done for the whole model. To maximise the efficiency and the accuracy of the model, the dataset should be split into training and testing datasets, the training set should be 80% of the whole dataset and the testing dataset should consist of 20% of the whole dataset.

- **Model training-** The algorithms we are going to use to build this project are all classification algorithms like Random Forest, Logistic Regression and Support Vector Machine. The training data is given to these algorithms and the accuracy, precision, recall and F1 score are calculated to know the performance of different models. Three algorithms are compared to know which one performs very well.

System Evaluation:

To say that the system is performing well, we need to keep in mind certain factors.

- **Accuracy:** accuracy is considered the most important factor that tells how well a model is performing. It provides an overview of how well the model is accurate.
- **Precision and Recall:** Recall evaluates the model's capacity to catch every positive event, whereas precision examines the accuracy of positive predictions. Maintaining a balance between recall and accuracy is essential to preventing false negatives and positives.
- **F1 score:** The F1 score provides a thorough assessment that takes both false positives and false negatives. It is calculated as the mean of accuracy and recall. When there is a disparity between the features, it is very handy.

There are many challenges when it comes to evaluating the system. The challenges are checking whether the model is built ethically, ensuring the quality of data and the effect of false positives and false negatives.

Task 3: Final Report

Introduction:

With technology growing rapidly, there is a high demand for innovation and inventions. This seems like a very good thing in general but not to forget this also leads to people working in tech jobs having heavy deadlines and stress. Taking care of our mental health is as important as working for a better future. This project aims to find out if there are patterns that can tell us if someone has looked for mental health support with the help of algorithms. By using this AI model as a lens, we hope to close the gap between mental health advocacy and technological innovation, creating a work environment that supports the holistic well-being of those driving the technology sector while simultaneously thriving on cutting-edge breakthroughs.

Dataset and Pre-Processing:

The dataset, sourced from Kaggle, comprises diverse features collected through a survey in the tech industry. It predominantly features categorical attributes alongside a single numerical variable. Given the dataset's origins in employee responses, instances of nonsensical values necessitate scrutiny. Additionally, an expected prevalence of missing values requires strategic handling. Proper encoding and careful preprocessing are essential for meaningful analysis and reliable insights.

Exploratory data Analysis:

The process of examining and condensing a dataset's primary features to comprehend its organisation, trends, and most important discoveries is known as Exploratory Data Analysis. Throughout EDA, data analysts examine the data, spot trends, find outliers, and get a basic grasp of the correlations between variables using a variety of statistical and visual aids. EDA is an essential phase in the data analysis process since it directs further investigations and influences the choice of suitable modelling approaches.

When analysing the data, we can see that many features have inappropriate values which do not make sense. Hence, we go through every feature and handle the feature either by dropping the unnecessary values or finding the value which occurs most and filling the missing data with that.

- **Missing values**

The presence of missing values within a dataset, particularly in the features "state," "self-employed," "work interfere," and notably "comments," necessitates careful consideration due to its potential impact on the accuracy and performance of machine learning models. Upon examination, it was observed that the "comments" feature exhibited the highest incidence of missing values. To mitigate this issue, a systematic approach was adopted, wherein missing values in the "state," "self-employed," and "work interfere" features were addressed through imputation. The mode of each respective feature was computed and utilized to fill in the missing values. However, given that the "comments" feature contained more than 80% missing values, a strategic decision was made to exclude this particular feature from further analysis.

This comprehensive approach ensures the integrity of the dataset while aligning with best practices in handling missing values in the context of machine learning.

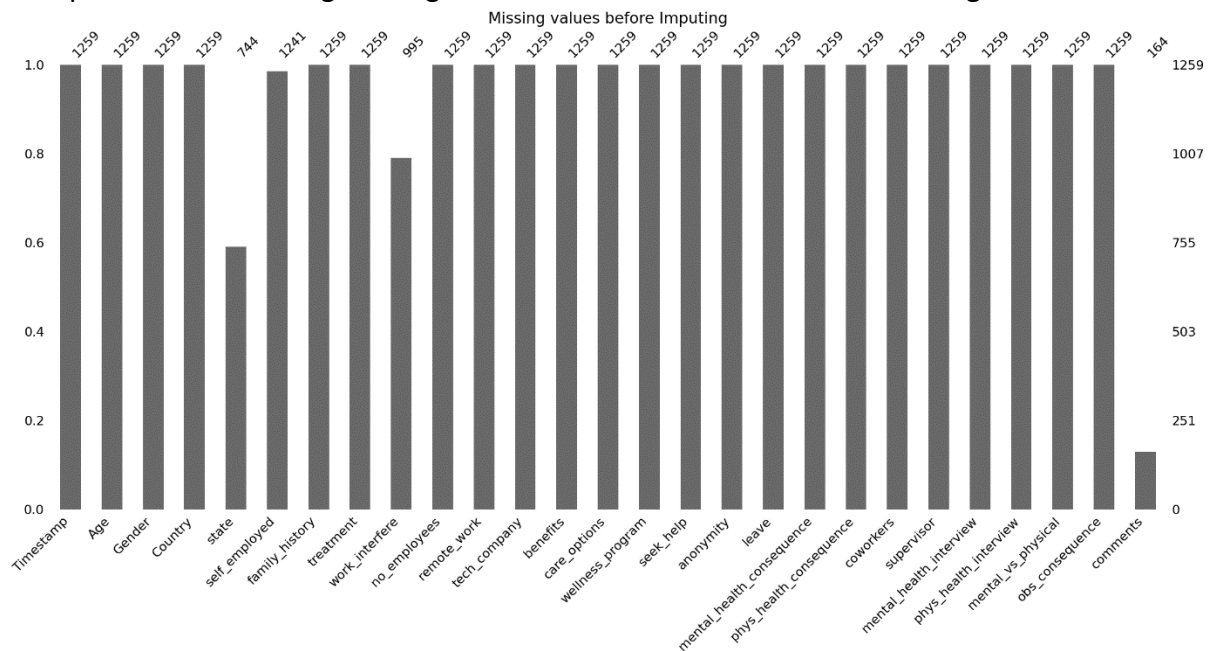


Fig 1 Missing Values before imputing

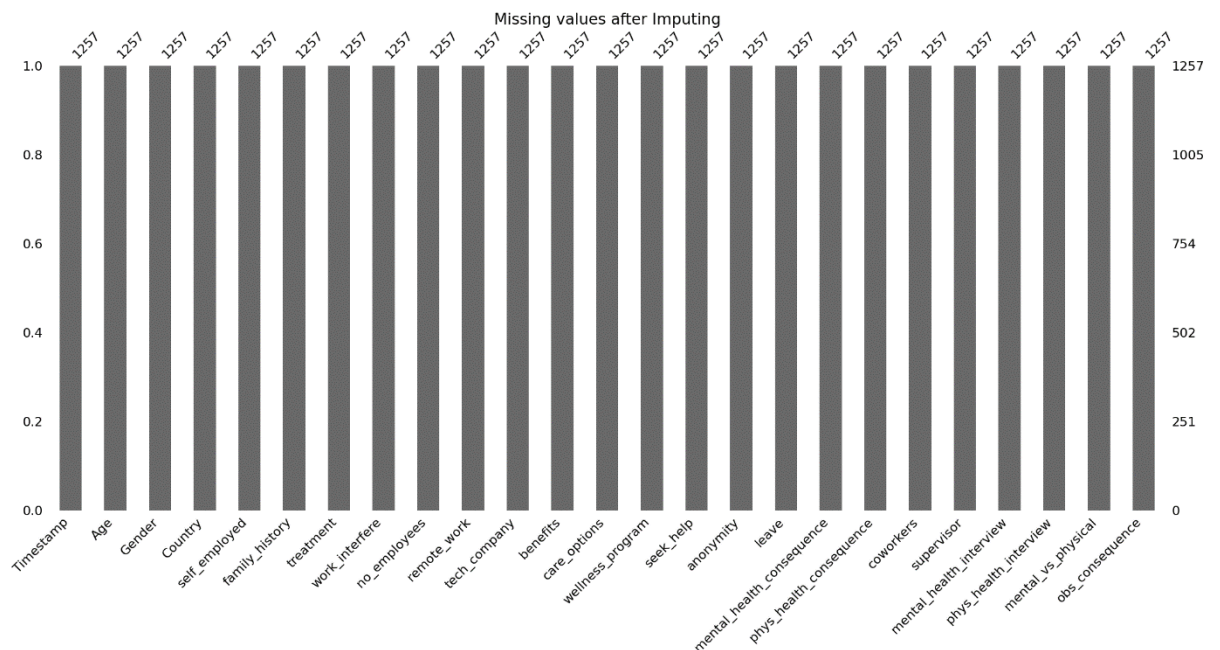


Fig 2 Missing values after Imputing

- **Duplicate values**

Since the information is collected from the employees through a survey, the answers to many of the features are not accurate which should be handled as they lead to model failure. The feature age had many inappropriate values so to handle that, the values less than 18 and greater than 75 are replaced by the value which occurred the most in the feature. Likewise in the gender feature, there were values other than male and female so three variables were created- male, female and other, and grouped every unique feature to fall under one of the free variables.

Visualisation

The process of graphically representing the data is called visualisation. Visualisations can be done in various ways using graphs, charts, maps and dashboards. The goal of visualisation is to understand the values even more clearly and to find out the patterns in the dataset. Visualisation makes the dataset easier for technical and non-technical users to understand.

In this project, the target feature which is the treatment feature is visualised just to know the distribution of the values. To make the model perform well, we need to know the distribution of the target variable so that we know the distribution of other features along with the target.

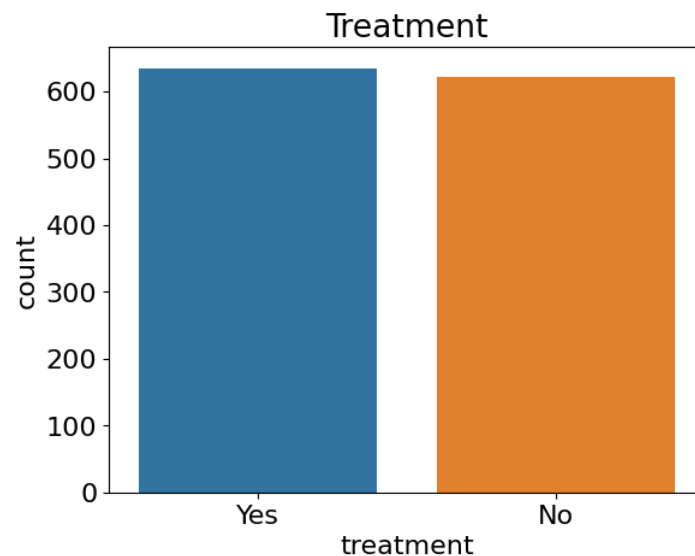
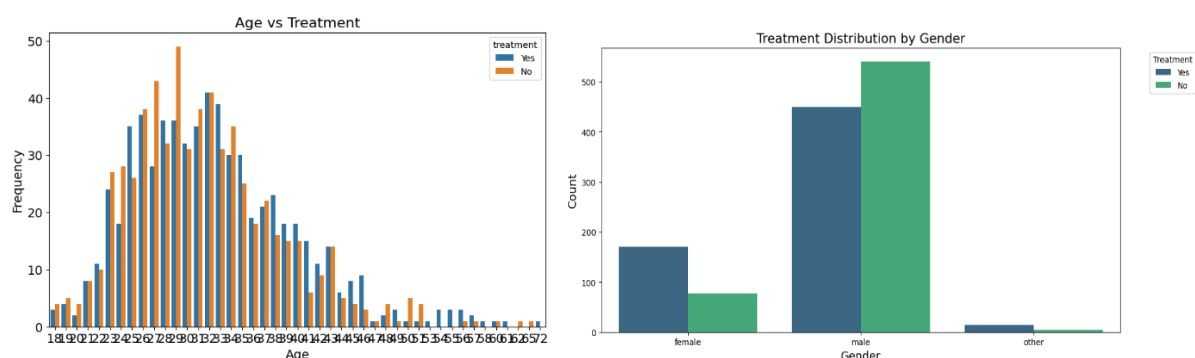


Fig 3 Target feature

Each of the features is visualised to know the distribution of the feature. The target feature that is, the "treatment" feature is compared with all other features and visualised to know the pattern. This is done to understand the data even more clearly and also



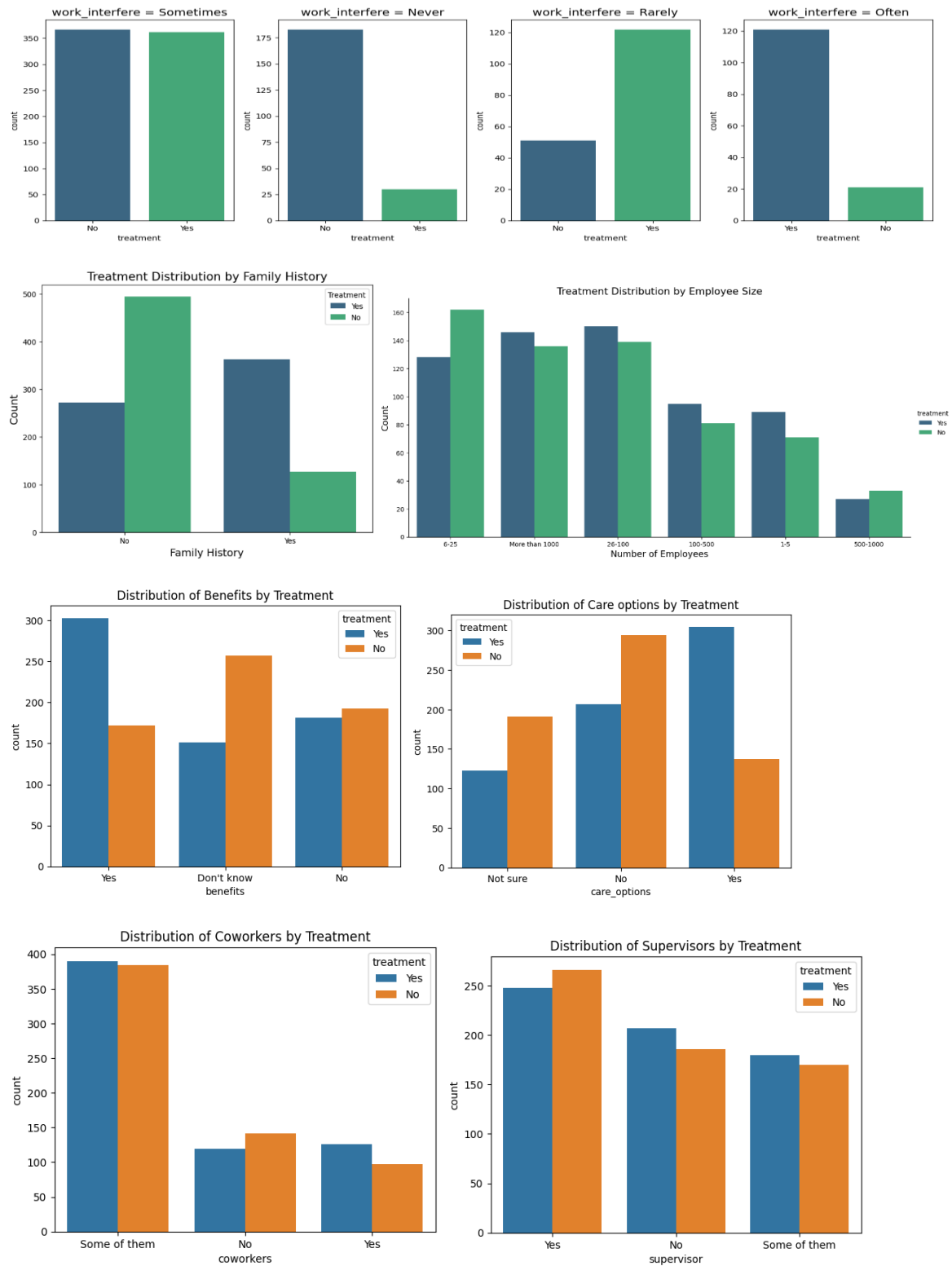


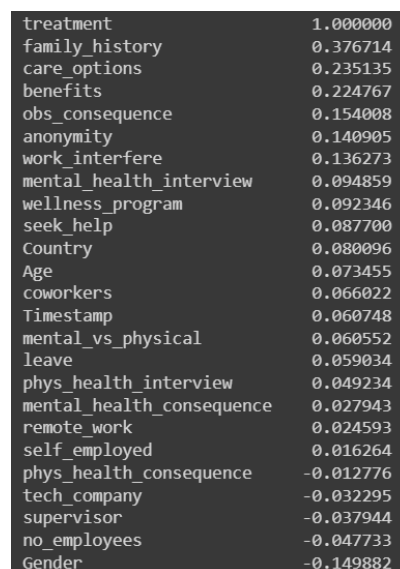
Fig 4. Visualisation

Pearson Correlation

A statistical metric used to quantify the linear relationship between two continuous variables is Pearson's correlation coefficient (r) (Statistics Solutions, 2021). In essence, it

provides you with the amount that one variable often changes in reaction to changes in the other, particularly in a linear fashion (Statistics Solutions, 2021)

In this project, the correlation values are displayed in stack format and the redundant relationships are removed. The function calculates the Pearson correlation coefficients between all the features in the input data frame that is “df_copy”. This code is used to analyse and visualise the strength of correlations between different features in the dataset excluding redundant relationships and correlations with the target variable that is ICU.



treatment	1.000000
family_history	0.376714
care_options	0.235135
benefits	0.224767
obs_consequence	0.154008
anonymity	0.140905
work_interfere	0.136273
mental_health_interview	0.094859
wellness_program	0.092346
seek_help	0.087700
Country	0.080096
Age	0.073455
coworkers	0.066022
Timestamp	0.060748
mental_vs_physical	0.060552
leave	0.059034
phys_health_interview	0.049234
mental_health_consequence	0.027943
remote_work	0.024593
self_employed	0.016264
phys_health_consequence	-0.012776
tech_company	-0.032295
supervisor	-0.037944
no_employees	-0.047733
Gender	-0.149882

Fig 5. Correlations

Feature Engineering

Feature engineering is the art of converting raw data into features that are more suitable for use in machine learning models. It's an important step in the machine learning pipeline, as it can significantly impact the performance of your model (Ali Shahzad et al., 2019)

Categorical values can't be fed into the model as the algorithms are primarily designed to operate on numerical data. The model uses mathematical operations and techniques that rely on numbers to make predictions. So instead of giving the categorical values, we use encoding techniques to convert the categorical values to numerical values. The one which is used here is the label encoder where a unique integer is assigned to each category within a feature.

Model building

Model building involves creating algorithms that analyse and discover patterns from data and can make predictions on the data. To build a model, we need to understand the specific problem that the AI wants to solve.

- **Split the data to train and test-** For unbiased evaluation, the data frame is split into train and test. The data frame is split into four data frames x_train, y_train, x_test and y_test. The x_train has all the features except the target feature which is the “TREATMENT” feature and y_train has the target feature.

- **Feature Importance-** A random forest model is built for finding the importance of the features. Instead of performing the analysis for the model, we can select the features which will be useful for the analysis and the important features are displayed. In this, the top 25 features are displayed.

➤ Random Forest

Random forest is a learning technique in machine learning where it is operated by creating several decision trees during training and the mode of the classification is given as the output. The random forest model is built with the important features and the accuracy is checked. The random forest gives an accuracy of 75%. We can understand more from the classification report.

Classification Report for Random forest:					
	precision	recall	f1-score	support	
0	0.75	0.74	0.74	120	
1	0.77	0.77	0.77	132	
accuracy			0.76	252	
macro avg	0.76	0.76	0.76	252	
weighted avg	0.76	0.76	0.76	252	
Confusion Matrix:					
[[89 31]					
[30 102]]					

Fig 6. Classification report for Random forest

Precision: Precision is used in binary and multiclass classification to estimate the accuracy of the model.

$$\text{Precision} = \frac{\text{True Positives} + \text{False Positives}}{\text{True Positives}}$$

Recall: Recall is used in classification to calculate the capability of a model to correctly find all relevant instances from the total instances.

$$\text{Recall} = \frac{\text{True Positives} + \text{False Negatives}}{\text{True Positives}}$$

F1 score: F1 score is used in classification which combines precision and recall into a single value which can be very useful in cases where there is an imbalance between the classes.

$$\text{F1 score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Macro avg: Macro avg stands for macro average which is a method of calculating the average performance across multiple classes in a multi-class classification problem

Confusion matrix:

A table used in classification to evaluate a machine learning model's performance is called a confusion matrix. By contrasting the model's predictions with the actual real values, it summarises the outcomes. With the help of the confusion matrix, we can compute the precision, recall and f1 score

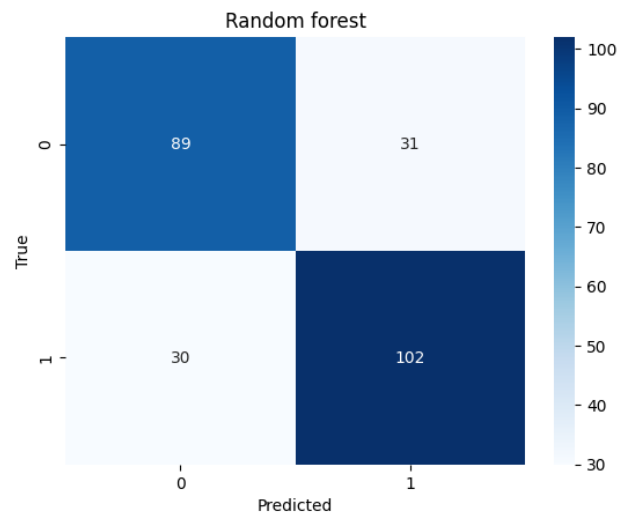


Fig 7. Confusion Matrix for Random Forest

This matrix shows the confusion matrix heatmap where true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

True Positive (TP): When a medical test correctly identifies a patient as having the condition, the model accurately predicts the positive class.

True Negative (TN): When a medical test correctly identifies a person as not having the condition, the model predicts the negative class.

False Positive (FP): When a medical test mistakenly identifies a patient as having a mental health illness, the model forecasts the positive class wrongly. Another name for this is a Type I mistake.

False Negative (FN): When a medical test misses a patient who truly has the condition, the model has predicted the negative class inaccurately. Another name for this is a Type II mistake.

➤ **Logistic Regression**

Logistic regression is a statistical method that is used to analyse data and make predictions about the probability of a binary outcome. For example, it can be used to predict whether a customer will buy a product, whether an email is spam, or whether a patient has a disease (Gaurav Chauhan, 2018). The logistic regression model accuracy came up to 71%

Classification Report for Logistic Regression:				
	precision	recall	f1-score	support
0	0.71	0.70	0.70	120
1	0.73	0.73	0.73	132
accuracy			0.72	252
macro avg	0.72	0.72	0.72	252
weighted avg	0.72	0.72	0.72	252
Confusion Matrix for Logistic Regression:				
[[84 36]				
[35 97]]				

Fig 8 Classification Report for Logistic Regression

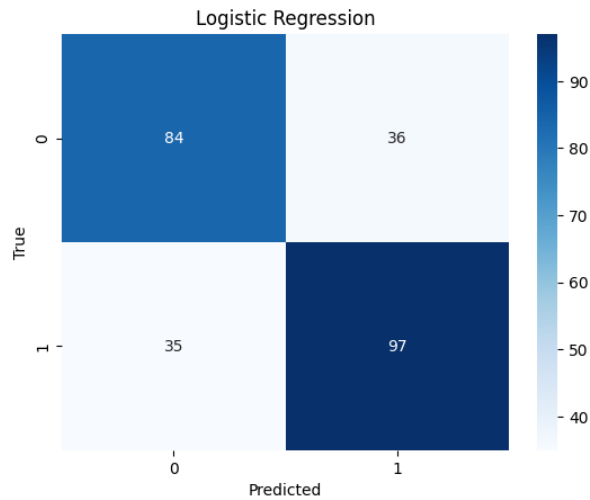


Fig 9. Confusion matrix for Logistic Regression

➤ Support Vector Machine

Support Vector Machines (SVMs) is a supervised learning algorithm widely used for classification and regression tasks. They operate by identifying a hyperplane that separates the data points belonging to their respective classes. This hyperplane maximizes the margin between the classes, leading to a robust and generalizable model. The accuracy of the model is 52%

```

Classification Report for SVM:
              precision    recall  f1-score   \

0               0.50        0.84        0.63
1               0.62        0.23        0.34

 accuracy          0.52
 macro avg         0.56        0.54        0.48
 weighted avg      0.56        0.52        0.48

Confusion Matrix for SVM:
[[101  19]
 [101  31]]

```

Fig 10. Classification Report for SVM

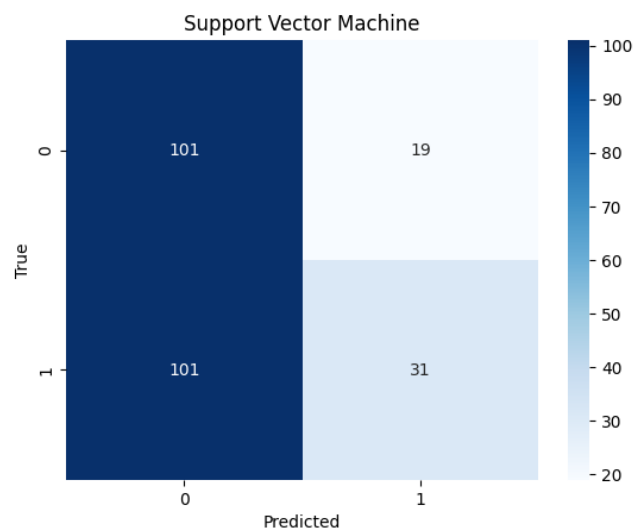


Fig 11. Confusion Matrix for SVM

Comparison of three models:

Models	Accuracy	Precision	Recall	F1 score
Random Forest	0.76	0.75	0.74	0.74
Linear Regression	0.72	0.71	0.70	0.70
Support Vector Machine	0.52	0.50	0.84	0.63

Upon analysis, it is evident that the Random Forest model outperformed both Logistic Regression and SVM with respect to overall accuracy, precision, recall, and F1-score. With an accuracy of 76%, Random Forest demonstrated a balanced performance in predicting mental health conditions based on the survey data. The precision and recall scores of 0.75 and 0.74, respectively, indicate that the model achieved a good balance between identifying positive cases and avoiding false positives.

Conclusion

In conclusion, for this particular mental health tech survey dataset, the Random Forest model emerged as a robust and effective choice for classification tasks, showcasing superior performance compared to alternative models. It is recommended to consider Random Forest as a promising algorithm for predicting mental health conditions in similar datasets, providing valuable insights for the development of mental health tech applications.

Reference list

Ali Shahzad, M., Noor, R., Ahmad, S., Mian, A. and Shafait, F. (2019). *Feature Engineering Meets Deep Learning: A Case Study on Table Detection in Documents* | *IEEE Conference Publication* | *IEEE Xplore*. [online] [ieeexplore.ieee.org](https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8945929). Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8945929> [Accessed 12 Jan. 2024].

Gaurav Chauhan (2018). *All about Logistic regression*. [online] Medium. Available at: <https://towardsdatascience.com/logistic-regression-b0af09cdb8ad>.

Madhavi, P. and Satyanarayana, S.V. (2023). *Big Data Analytics for Electrical Systems using Machine Learning Algorithms* | *IEEE Conference Publication* | *IEEE Xplore*. [online] [ieeexplore.ieee.org](https://ieeexplore.ieee.org/document/10149919). Available at: <https://ieeexplore.ieee.org/document/10149919> [Accessed 18 Jan. 2024].

Statistics Solutions (2021). *Pearson's Correlation Coefficient*. [online] Statistics Solutions. Available at: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/>.