# Abstract:

Mental health is rapidly becoming a major concern of people and thus requires research that will enhance the understanding of variables affecting the same. The dataset is hosted on Kaggle and contains data of surveys out of which a set of established for patients’ mental health conditions, work characteristics and socio-demographics. It is designed to address the patterns that might exist between mental health and the organisational culture with the view of establishing research that can guide the possible changes for future policies and intercessions.

Features that are included in the dataset include age, gender, family history of mental illness, current mental health status, provisions at the workplace, and prior experience with mental health treatment. It also measures organization related factors such as availability and accessibility of resources from employers, employer’s perception to mental health issues, and employer’s perception towards how mental health may affect performance at workplace. People, and more specifically researchers, data scientists and Mental Health advocates can use this kind of data set to create awareness programs to encourage people’s mental health, launch education campaigns and champion for policies that actively encourage good Mental Health at work places.

# Introduction

Today in a world increasingly fast and increasingly demanding, mental health has become one of the most significant issues troubling people and societies worldwide. To address those challenges, we need to better understand what determines mental health, and how prevalent these factors are across demographic and occupational groups. The Mental Health Dataset is a rich source which contains information about the mental health of people with variables like gender, occupation, family history, coping methods, etc. This dataset provides a basis from which to uncover patterns, identify risk factors, and create interventions to achieve mental health. After pre-processing the dataset contains 284,858 unique records with 16 columns to capture various attributes of mental health. Demographics like gender and country, occupational settings, personal and familial mental health histories, behavioural patterns, stress levels and coping struggles are among these.

Also important to the dataset is the presence of indicators such as "treatment," "growing stress" and "mental health interview" that let us know both what challenges are present and what support systems are available. The breadth of information the dataset offers make it a valuable source to gain knowledge of mental health on different levels. This project addresses the first major problem of identifying what stresses mental health, and how often do people struggle with their mental wellbeing. In particular, this aims to solve a multi class classification problem, where we want to predict what categories, a person is placed into ( mood swings or not, as per a dataset used). We solve the problem by comparing different machine learning models and evaluate their ability to solve the classification problem.

**The Challenging Part**

Mental health research is inherently complex due to its multifaceted nature, and working with this dataset introduces several unique challenges:

1. **Diversity of Variables:** Categorical and ordinal such as Days Indoors and Growing Stress needs to be carefully encoded and analysed for meaningful insights from the dataset. This diversity presents a challenge of balancing it while keeping interpretability.
2. **Imbalanced Data:** The dataset is highly skewed with an 82 percent of respondents being male. If not properly dealt with this imbalance can create bias in model which in turn means that you can't generalize the findings.
3. **Interconnected Factors:** There are a combination of personal, social, and occupational factors to mental health. Advanced statistical and machine learning techniques are necessary to isolate the effects of individual variables or to identify causal relationships from correlations.
4. **Data Cleaning and Pre-processing:** The dataset already had missing values, duplicates, and columns which were not useful, like timestamps. Basically, these issues were handled during pre-processing in such a way that no critical information was lost, which was a major consideration for assuring data quality.
5. **Ethical Considerations:** It is a sensitive topic of mental health so any analysis has to be done ethically. Interpretations and conclusions about mental health challenges should not stigmatize or oversimplify.

# Solution:

* **Feature Selection and Assumptions :** It is determined that the models are built with the help of pipelines which use the MinMaxScaler to normalize the features so that input data stays within a consistent range which is necessary for models that are sensitive to the magnitude of features like logistic regression. Here we assume that all features are equally contributing to the prediction task, so normalization is a necessary pre-processing step.

Nevertheless, the method could be extended by applying feature selection techniques (e.g., recursive feature elimination, RFE or feature importance ranking) to select and retain the most important predictors. In other words, the models are effective generalizers, and we assume the training and test datasets representatively belonging to the larger population.

* **Techniques for Model Building and Parameter Tuning:** The solution evaluates four models: We do select methods that include Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Every model is enclosed within a pipeline that makes it extremely easy to incorporate pre-processing and model training into one step.

The patterns are learned and tested with 10-fold cross validation that is a technique of reducing over fitting by dividing the data into different training and testing sets. This makes the measure more reliable and not inclined to a particular split of data depending on how the dataset is parsed. Although none of the code specifies how the hyperparameters are tuned, default values are set.

Additional steps could be added to the pipelines themselves by using methods that are filters such as a grid search or randomised search in order to optimise the hyperparameters that are used in models such as how deep a tree needs to be or how many estimators should be used in ensemble algorithms.

* **Handling Data Skewness and Inconsistencies:** Some problems such as data inconsistency or data skewness are resolved by the pre-processing step of scaling. The possible problem of class imbalance, more characteristic of mental health data, can affect the measurement of accuracy greatly.

It has been done to reduce the above problem by including precision, recall, and F1-score into the evaluation of models which gives an insight of the performance of models on each class. For instance, an indication of high recall meaning the ability to identify the target condition reflected by “Mood Swings”. To take it a little further, sampling techniques could also tackle skewness directly.

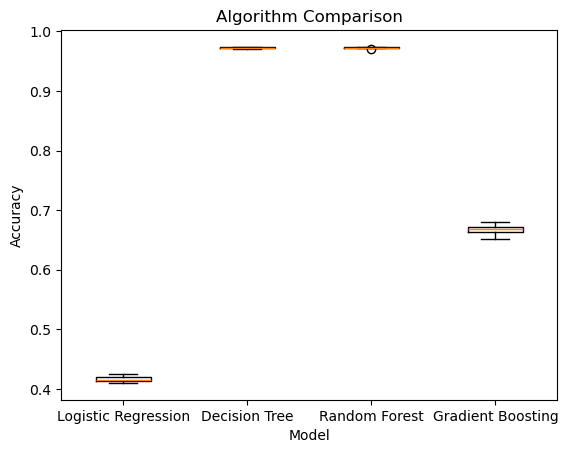
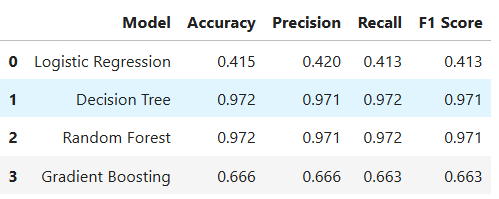
It was also possible to oversample the minority class or under sample the majority group and get datasets of equal sizes for training. Random forest kind of methods also help to reduce some of the above imbalance as the result coming from the individual trees are combined.

* **Model Comparisons and Evaluation:** The output is then presented through box plots in order to compare the percent accuracy of cross-validation among models, highlighting the Random Forest and Decision Tree as the models with the highest percent accuracy estimations of approximately 97.2 %.

This result indicates that Gradient Boosting is just moderately good while the Logistic Regression is struggling due to its inability to handle non-linear relationships as seen in the data. The confusion matrices give much information on how specific models are performing in classification terms. These matrices depict true positives, false positives, true negatives and false negatives, so that it is easy to know where models go wrong.

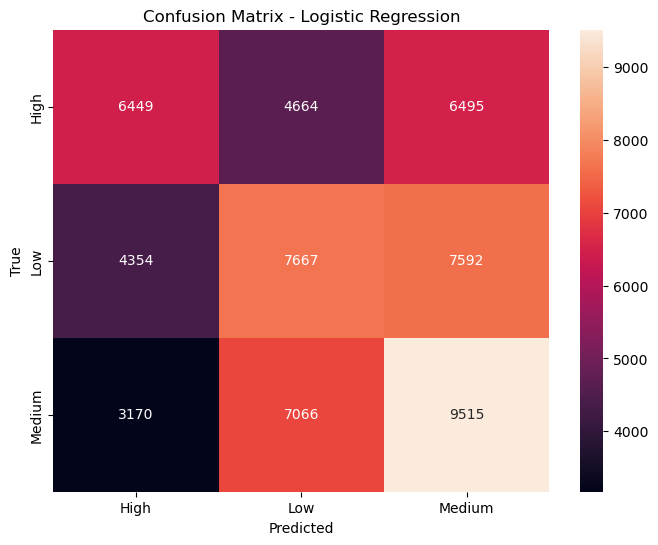
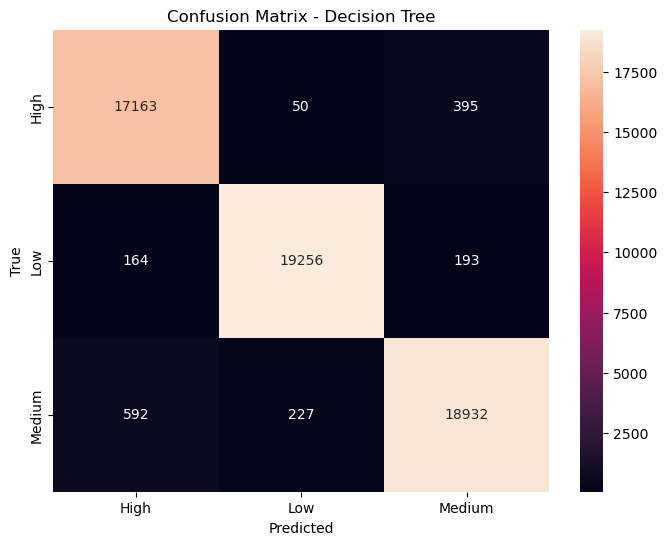
# Empirical Experiments

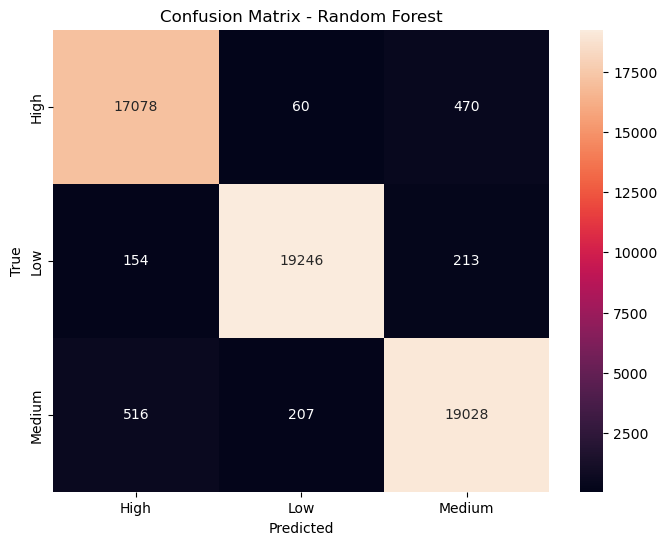
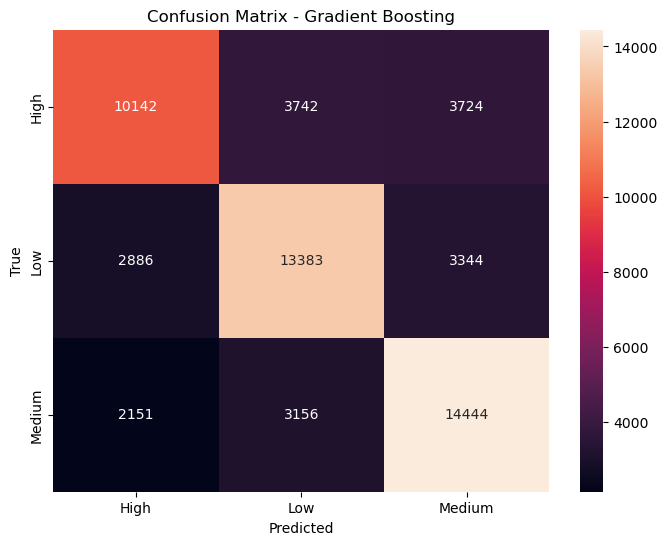
In the experiments that were conducted, the four models of machine learning namely Logistic Regression, Decision Tree, Random Forest and Gradient boosting algorithm was tested for effectiveness in predicting the “Mood Swings” from the data set that was given. The experiments were carried out with cross validation with ten folds of data for better and accurate results. Logistic Regression was used as the baseline analysis because it is simple and can identify linear associations. By comparing with seemingly more sophisticated techniques such as Random Forest and Gradient Boosting, enhanced capability in dealing with possibly non-linear correlations was demonstrated.

**Figure: Model performance metrics comparision**

The results of the evaluation present a clear picture of disparities in model performance. For Logistic Regression, the mean accuracy of about 41.6% shows how effective a linear model can be on a dataset which contains non-linear patterns most likely. Random Forest and Decision Tree had an average accuracy of about 97.2 percent making the two models powerful in defining hierarchical decision boundaries. Using Gradient Boosting, we obtained a slightly above average cross-validation score of 66.7% that signified its usefulness as an ensemble model whilst pointing out the possibility of there being specific parameters that could be tuned to increase the model’s performance. Below are the confusion matrix plots for all models we developed.

**Figure: Confusion matrix for Logistic regression , Decision tree, Random forest and Gradient boosting classifiers**

Additional findings that arose involved compared models based on something other than accuracy alone. Confusion matrices offered breakdown of prediction results which showed that both Random Forest and Decision Tree models had very low rates of false positive and false negatives High precision, recall, and F1 scores were achieved by the two models . These results substantiate that these models are well applicable to cases with imbalanced data or cases with subtle feature interaction. As for Gradient Boosting, it indicated a conduct worse than Random Forest algorithm but given that the algorithm integrates several hyperparameters such as learning rate or number of boosting iterations, it is possible to improve it.

# Discussion

Nevertheless, the approach used in this experiment is quite smooth with slight drawbacks. First, one of the issues is that most of the models, accordingly, use default or minimally tuned hyperparameters. For example, detecting accuracy Random Forests and Gradient Boosting provide but they may not be optimized enough since the hyperparameters tuning was not done much throughout the process. More general hyperparameters that could have been also tuned include the number of estimators, maximum depth or learning rate for which other methods like the grid search or Bayesian optimization could have been applied to explore.

Effectively, another drawback is that imbalances or noise in the data are not addressed directly in the present approach. While measures such as accuracy and F1 scores mean high performance of model, they do not translate all possible biases in class predictions. For example, all the models might choose the oversampled class, say, Class 2 (e.g., “No Mood Swings”), which will not be very accurate.

Last but not least, the very structure of the experiment might be generalized to study different forms of modelling or even other types of architectures. All these models, including Random Forest and Logistic Regression, were helpful, but deep learning methods like neural networks also may be helpful depending on the relationship between the variables since the non-linear model could be useful for dataset to some extent in the case of using methods like dropout to prevent overfitting.

# References:

* A Machine Learning-supported Analysis of University Students’ Mental Health. (2021). *Machine Learning Theory and Practice*, *2*(3). https://doi.org/10.38007/ml.2021.020302
* BH, S., K, N., C, D., P, B., Duraipandian, K., Rajan, S., & Sathiyanarayanan, M. (2022). Mental Health Analysis of Employees using Machine Learning Techniques. *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*. https://doi.org/10.1109/comsnets53615.2022.9668526
* Kim, H., Kim, M., Park, K., Kim, J., Yoon, D., Kim, W., & Park, C. H. (2023). Machine learning-based classification analysis of knowledge worker mental stress. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1302794
* Oyebode, O., Alqahtani, F., & Orji, R. (2020). Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews. *IEEE Access*, *8*, 111141–111158. https://doi.org/10.1109/access.2020.3002176