**CSCE 5215: MACHINE LEARNING**

**Project Report**

**Project Description:**

**Title:** Tech Well–Being Predictor

**Team:**

* Project Lead: Sujana Vadlamudi (Email: sujanavadlamudi@my.unt.edu, Role: Project Leader)
* Data Scientist: Prathyusha Papagari (Email: Prathyushapapagari@my.unt.edu, Role: Data Analysis) Machine Learning Engineer: Vineela Pamarthi (Email: Vineela pamarthi@my.unt.edu, Role: Model Development)
* Machine Learning Engineer: Aparna Patnala (Email: Aparna Patnala@my.unt.edu, Role: Model Development)

**Abstract:**

Mental illness is a significant public health concern, affecting millions of people worldwide. Despite the prevalence of mental illness, many individuals do not seek treatment due to various factors, including stigma, lack of access to care, and fear of negative consequences. Early identification and intervention are crucial for improving mental health outcomes. Machine learning (ML) has emerged as a powerful tool for analysing complex data and identifying patterns that may be difficult to detect using traditional methods. This project aims to utilize ML to predict the likelihood of an individual seeking treatment for mental illness based on a dataset of work-related and mental health-related variables.

The project employs a variety of ML algorithms, including logistic regression, K-Nearest Neighbours, decision trees, random forests, bagging, boosting, stacking, and neural networks, to predict treatment necessity. The performance of each model is evaluated using various metrics, including accuracy, precision, recall, and confusion matrix. The strongest predictors of mental health illness or certain attitudes towards mental health in the workplace are also identified.

The findings of this project have the potential to inform the development of interventions that promote early identification and treatment of mental illness. By identifying individuals who are at high risk of requiring treatment, healthcare providers can prioritize their care and ensure that they receive the support they need. Additionally, the findings can inform workplace policies and practices that create a more supportive environment for individuals with mental illness.

**Problem Specification:**

The supervised learning problem in this project is to predict whether a participant will seek treatment for a mental health condition based on the values of the data variables. This is a binary classification problem, as the outcome variable (treatment necessity) can only take on two values: 0 (not seeking treatment) or 1 (seeking treatment).

Upon comprehensive analysis and exploration of the dataset using Python, covariance/correlation matrices, and charts, key features were identified to gain a deeper understanding of their interrelationships. The selected features, including 'Age,' 'Gender,' 'family\_history,' 'benefits,' 'care\_options,' 'anonymity,' 'leave,' and 'work\_interfere,' were scrutinized for patterns and dependencies. This analysis aimed to unveil potential insights into how these factors collectively contribute to the participants' attitudes and experiences related to mental health. By focusing on these specific features, we aimed to process meaningful correlations and uncover any discernible trends within this multifaceted dataset.

**Data specification:**

The features in this dataset can be categorized into three main groups:

*Demographic features*: Age, gender, country, state (for US residents), self-employment status

*Work-related features*: Number of employees, remote work status, tech company status, mental health benefits, knowledge of mental health care options, employee wellness programs, resources for seeking help, anonymity protection, medical leave policies

*Mental health-related features*: Family history of mental illness, work interference due to mental health conditions, perceived consequences of discussing mental health with employer and co-workers, willingness to discuss mental health issues with co-workers and supervisors, attitudes towards mental health in the workplace, exposure to negative consequences for co-workers with mental illness

The goal of the project is to use these features to predict whether a participant will seek treatment for a mental health condition.

**Design and Milestones:**

There are multiple phases involved in building this project from importing required libraries to building and evaluating the model. Let us have a closer look the steps involved.

1. **Data Loading and Cleaning:**
   * The initial step is to load all the necessary libraries and the dataset from a CSV file.
   * Exploratory data analysis is performed, such as displaying data shape, statistics, and information to understand the data. In this step missing data is also identified.
   * Missing data is handled by dropping certain columns and filling missing values for others.
   * Libraries: **pandas** for data manipulation and analysis, **numpy** for numerical operations.
2. **Data Pre-processing:**
   * Cleaned the 'Gender' column by converting entries to lowercase and categorizing them into three groups: male, female, and trans.
   * Handled missing values in the 'Age' column by filling with the median and creating age ranges.
   * Converted categorical variables into numerical labels using Label Encoding.
   * Visualized the covariance matrix for feature variability.
   * Libraries: **sklearn** for pre-processing tasks like Label Encoding, **matplotlib** for visualizations.
3. **Exploratory Data Analysis (EDA):**
   * Various visualizations are created to explore relationships among different variables.
   * Distributions, bar plots, and histograms are used to analyse the impact of age, gender, family history, care options, benefits, and work interference on mental health.
   * Libraries: **seaborn** and **matplotlib** for creating various visualizations.
4. **Scaling and Fitting:**
   * Scaling the 'Age' column using Min-Max scaling. Here age the variable is scaled between 0 and 1, for simplifying the computations.
   * Splitting the dataset into training and testing sets. Same split of dataset has been used for all the models.
   * Libraries: **sklearn** for Min-Max scaling, train-test split, and model fitting.
5. **Modelling and Evaluation:**
   * Several machine learning models, including ensemble models are implemented and evaluated:
     + Logistic Regression
     + K-Nearest Neighbours Classifier
     + Decision Tree Classifier
     + Random Forests
     + Bagging
     + Boosting
     + Stacking
   * The models are tuned using techniques like GridSearchCV and RandomizedSearchCV.
   * The evaluation metrics include accuracy, confusion matrix, precision, AUC-ROC curve, and cross-validated AUC.
   * Libraries: **sklearn** for implementing machine learning models (Logistic Regression, KNeighbors Classifier, Decision Tree Classifier, Random Forests, Bagging, Boosting, Stacking), **sklearn.model\_selection** for hyperparameter tuning using GridSearchCV and RandomizedSearchCV, **sklearn.metrics** for evaluation metrics.
6. **Neural Network (TensorFlow):**
   * Features are converted into TensorFlow feature columns.
   * A neural network model is defined and trained using TensorFlow's Estimator API.
   * The model is evaluated on the test set.
   * Libraries: above mentioned libraries and **tensorflow**, which is used for building and training the neural network model.

**Resources and Related Projects:**

* "Machine Learning and Its Applications in Mental Health Care" by Torous and Resnik (2013): This article provides an overview of machine learning applications in mental health care. It discusses the potential of machine learning to improve diagnosis, treatment, and prevention of mental illness. [<https://pubmed.ncbi.nlm.nih.gov/30744717>]
* Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression, clustering, and dimensionality reduction algorithms. It's one of the most popular machine learning libraries in Python and the de-facto standard for supervised and unsupervised learning tasks. [ <https://scikit-learn.org> ]
* Keras is a high-level neural networks API for Python. It allows for easy and efficient building of deep learning models. It's capable of running on top of TensorFlow, CNTK, Theano, or MXNet as the backend. [<https://keras.io/api/> ]
* Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It's based on NumPy and provides a wide range of plot types, including line plots, bar charts, scatterplots, histograms, and more. Matplotlib is a powerful tool for data visualization and is widely used in scientific computing and data science. [ <https://matplotlib.org/> ]