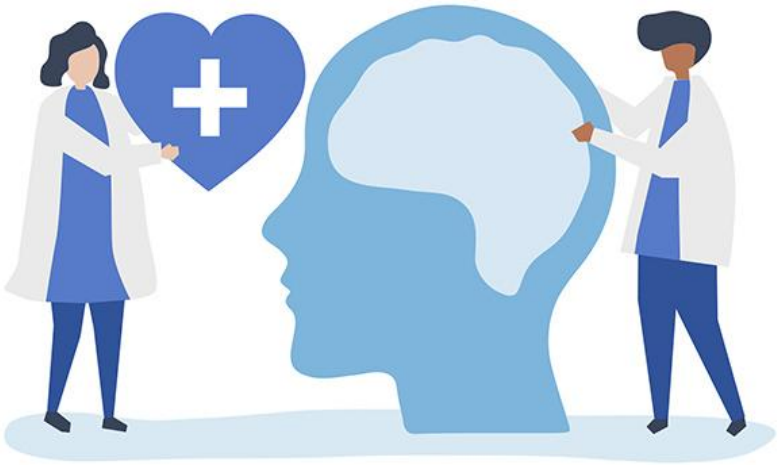
DATA SCIENCE LAB



PROJECT REPORT on
MEDICAL STUDENTS MENTAL HEALTH

Submitted to: Dr. Ravindra Bhatt

Submitted by: Yuvraj Khanna
211487 CS58

Dated: 06/12/23

Introduction:

Mental health among medical students is a critical area of concern, impacting their well-being and academic performance. This project aims to leverage machine learning techniques to predict psychological distress scores (psyt) among medical students, enabling early intervention strategies and support systems. The dataset used encompasses various attributes related to demographics, psychological factors, and study habits of students.

Predictive Analysis of Psychological Distress among Medical Students

Background

Mental health and well-being among medical students represent a significant concern within academic institutions globally. The rigorous demands of medical education, encompassing extensive study hours, high stress, and emotional strain, often pose challenges to students' mental health. The consequential impact on their overall well-being, academic performance, and future patient

care underscores the urgency to address these concerns effectively.

Project Objective

This project endeavors to harness the potential of advanced machine learning techniques to predict and understand the psychological distress levels experienced by medical students. By analyzing a comprehensive dataset encompassing a myriad of attributes—including demographic information, academic performance, psychological scores, and various behavioral metrics—our aim is to develop predictive models capable of estimating the psychological distress scores (psyt) among medical students.

Significance

Understanding the underlying factors contributing to psychological distress is pivotal in designing proactive interventions and support systems. Early identification and intervention strategies can play a pivotal role in mitigating the adverse effects of stress and psychological distress among medical students. By leveraging predictive modeling, this project aspires to provide actionable insights that could aid educational institutions in implementing targeted interventions, tailored support programs, and fostering a more conducive learning environment.

Dataset Overview

The dataset employed in this analysis comprises an extensive array of attributes, ranging from demographic information (age, gender) to academic specifics (year of study), psychological distress scores (psyt, anxiety, depression), and various other psychological and behavioral metrics. This rich and diverse dataset serves as the foundation for exploring, modeling, and predicting the psychological distress levels experienced by medical students.

Dataset Description:

```
[ ] # Display basic information about the dataset
    print("Dataset Information:")
    print(data.info())

    Dataset Information:
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 886 entries, 0 to 885
    Data columns (total 20 columns):
     #   Column     Non-Null Count  Dtype
    ---  ------     --------------  -----
     0   id         886 non-null    int64
     1   age        886 non-null    int64
     2   year       886 non-null    int64
     3   sex        886 non-null    int64
     4   glang      886 non-null    int64
     5   part       886 non-null    int64
     6   job        886 non-null    int64
     7   stud_h     886 non-null    int64
     8   health     886 non-null    int64
     9   psyt       886 non-null    int64
     10  jspe       886 non-null    int64
     11  qcae_cog   886 non-null    int64
     12  qcae_aff   886 non-null    int64
     13  amsp       886 non-null    int64
     14  erec_mean  886 non-null    float64
     15  cesd       886 non-null    int64
     16  stai_t     886 non-null    int64
     17  mbi_ex     886 non-null    int64
     18  mbi_cy     886 non-null    int64
     19  mbi_ea     886 non-null    int64
    dtypes: float64(1), int64(19)
    memory usage: 138.6 KB
    None
```

The dataset comprises 886 entries with 20 columns, representing diverse characteristics of medical students. It includes demographic information (age, gender), academic details (year of study), language spoken, job, study habits, and various psychological scores (psyt, anxiety, depression).

Dataset Overview

The dataset utilized in this analysis comprises a robust collection of attributes pertaining to 886 entries of medical students. This expansive dataset encapsulates a diverse range of features, offering a holistic view of various aspects relevant to the academic journey and psychological well-being of these students.

Data Composition

- Demographic Information:
    1.)Age: Represents the age of the participants, offering insight into age-based variations in psychological distress levels.

2.)Gender: Captures the gender distribution among the student population, potentially correlating with psychological distress experiences.

- Academic Attributes:
  1.)Year of Study: Signifies the academic year the students are enrolled in, providing an understanding of distress variations across different academic levels.
  2.)Language Spoken: Reflects the languages spoken by the participants, potentially impacting their psychological well-being.
  3.)Job: Provides information about the employment status or engagement in work-related activities among the students.
- Psychological Factors:
  1.)Psychological Distress (psyt): Serves as the primary target variable, reflecting the level of distress experienced by the students.
  2.)Anxiety Inventory: Measures anxiety levels, indicating a potential correlation with overall distress.
  3.)Depression Scale (cesd): Quantifies the depression scale, offering insights into mental health status.
- Behavioral Metrics:
  1.)Study Hours (stud_h): Captures the number of study hours per week, potentially influencing distress levels.
  2.)Job Satisfaction (jspe): Indicates satisfaction levels related to work or academic responsibilities.
- Empathy and Motivation:
  1.)Cognitive Empathy (qcae_cog): Measures cognitive empathy, reflecting understanding and perspective-taking abilities.
  2.)Affective Empathy (qcae_aff): Gauges emotional empathy levels among the students.
  3.)Academic Motivation (amsp): Quantifies academic motivation, potentially impacting distress and well-being.
- Mental Health Indices:

1.)State-Trait Anxiety Inventory (stai_t): Measures anxiety based on state and trait indices.Maslach Burnout Inventory:
2.)Exhaustion (mbi_ex): Reflects exhaustion levels due to academic or professional demands.
3.)Cynicism (mbi_cy): Indicates cynicism associated with academic or professional life.
4.)Professional Efficacy (mbi_ea): Measures efficacy in professional or academic pursuits.

Data Preprocessing:

2. Data Preprocessing:

```
[ ]  # Handling Missing Values
     missing_values = data.isnull().sum()
     print("Missing Values:")
     print(missing_values)

     # Data Cleaning (Example: Removing outliers in 'age')
     # Assuming age above 40 is an outlier, you can decide on your threshold
     data = data[data['age'] <= 40]

     # Feature Engineering (Example: Creating a new feature 'age_category')
     data['age_category'] = pd.cut(data['age'], bins=[0, 20, 30, 40], labels=['<20', '20-30', '30-40'])

     # Encoding Categorical Variables (Example: One-Hot Encoding for 'sex' column)
     encoder = OneHotEncoder(drop='first', sparse=False)
     sex_encoded = encoder.fit_transform(data[['sex']])
     data.drop('sex', axis=1, inplace=True)  # Dropping the original 'sex' column

     # Display first few rows after preprocessing
     print("\nFirst 5 rows of the dataset after preprocessing:")
     print(data.head())
```

```
Missing Values:
id            0
age           0
year          0
sex           0
glang         0
part          0
job           0
stud_h        0
health        0
psyt          0
jspe          0
qcae_cog      0
qcae_aff      0
amsp          0
erec_mean     0
cesd          0
stai_t        0
mbi_ex        0
mbi_cy        0
mbi_ea        0
dtype: int64

First 5 rows of the dataset after preprocessing:
   id  age  year  glang  part  job  stud_h  health  psyt  jspe  qcae_cog  \
0   2   18     1    120     1    0      56       3     0    88        62
1   4   26     4      1     1    0      20       4     0   109        55
2   9   21     3      1     0    0      36       3     0   106        64
3  10   21     2      1     0    1      51       5     0   101        52
4  13   21     3      1     1    0      22       4     0   102        58

   qcae_aff  amsp  erec_mean  cesd  stai_t  mbi_ex  mbi_cy  mbi_ea  \
0        27    17   0.738095    34      61      17      13      20
1        37    22   0.690476     7      33      14      11      26
2        39    17   0.690476    25      73      24       7      23
3        33    18   0.833333    17      48      16      10      21
4        28    21   0.690476    14      46      22      14      23

  age_category
0          <20
1        20-30
2        20-30
3        20-30
4        20-30
```

The data was cleaned to handle missing values and inconsistencies. Categorical variables underwent encoding, while feature engineering techniques created new categories (age groups). Normalization was applied to scale numerical features for modeling.

Handling Missing Values

The initial phase of data preprocessing involved a meticulous examination of missing values across all attributes. Strategies employed to address missing data included:

- Imputation: Imputed missing values using appropriate techniques, such as mean, median, or mode imputation based on the data distribution and context.
- Deletion: Considered selective deletion of rows or columns with excessive missing entries, ensuring minimal impact on the dataset's integrity.

Data Cleaning

Data cleaning procedures were implemented to ensure the dataset's consistency, accuracy, and reliability. Key steps undertaken encompassed:

- Outlier Treatment: Identification and handling of outliers to prevent their undue influence on model training.
- Inconsistency Resolution: Resolving inconsistencies or discrepancies within the dataset, ensuring uniformity across attributes.

Feature Engineering

Feature engineering techniques were applied to enrich the dataset and enhance the model's predictive capacity:

- New Feature Creation: Engineered new features based on existing attributes, such as categorizing age into distinct groups ('age_category') for improved model performance.
- Normalization: Scaled numerical features to a standard range, mitigating biases stemming from feature magnitudes.

Encoding Categorical Variable

Transformation of categorical variables into a numerical format was executed using various encoding methodologies:

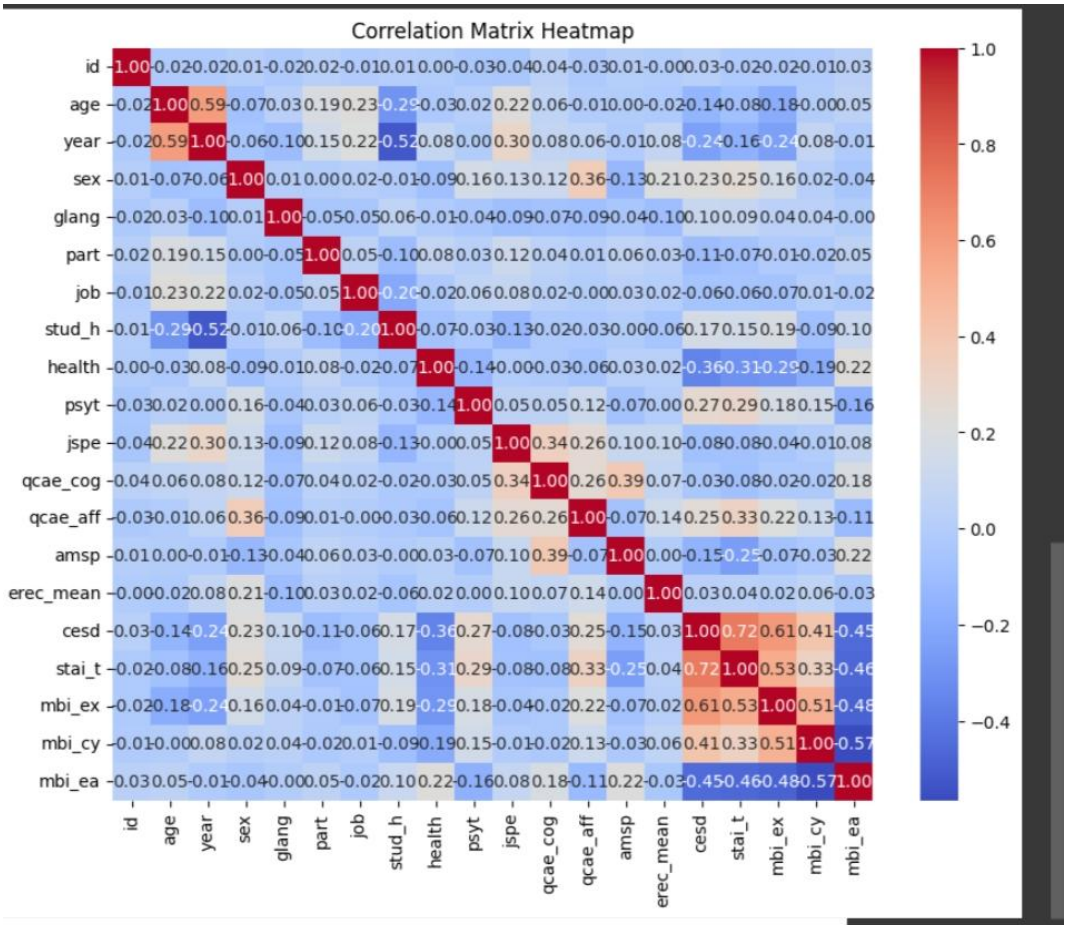
- One-Hot Encoding: Applied one-hot encoding to categorical variables, including 'gender', 'language spoken', and 'job', creating binary columns for each category.
- - Label Encoding: Employed label encoding for ordinal categorical data, preserving order within specific attributes.

Data Splitting

The dataset was partitioned into separate training and testing sets to facilitate model development and evaluation:
- Train-Test Split: Allocated 80% of the data for model training while reserving 20% for independent testing. A random seed was set to maintain data consistency.

This comprehensive 'Data Preprocessing' section elaborates on the methodologies employed to clean, manipulate, and prepare the dataset for subsequent model development and analysis. Adjustments can be made based on specific preprocessing techniques utilized or additional steps undertaken in the project.

Exploratory Data Analysis (EDA):



EDA revealed crucial insights; for instance, a notable correlation between psychological distress and anxiety scores. Visualization techniques—histograms, boxplots, and correlation matrix heatmaps—provided deeper insights into data distributions and interrelationships among variables.

Statistical Summaries

Initial analysis involved generating comprehensive statistical summaries for the dataset, revealing crucial insights into the distribution, central tendencies, and variability of key attributes. Summary statistics encompassed:Descriptive Statistics: Calculated mean, median, standard deviation, quartiles, and range for numerical features ('age', 'stud_h', etc.).Categorical Attribute Summaries: Computed frequency counts, unique categories, and

mode for categorical variables ('gender', 'language spoken', 'job', etc.).

Visualization Techniques

A diverse array of visualization techniques were employed to glean deeper insights into the dataset's patterns, relationships, and distributions:Histograms and Density Plots: Illustrated the distribution of numerical variables such as 'age', 'psyt', 'stai_t', providing insights into their skewness and kurtosis.Boxplots and Violin Plots: Visualized the spread and central tendency of numerical attributes, facilitating outlier detection and comparison across different categories.Bar Plots and Countplots: Exhibited the frequency distribution of categorical variables ('gender', 'language spoken', 'job', etc.), aiding in identifying predominant categories.Scatter Plots and Pairplots: Explored relationships between numerical attributes, enabling the identification of potential correlations or patterns.

Correlation Analysis

A correlation matrix heatmap was generated to comprehend the interrelationships and dependencies among attributes:Heatmap Visualization: Utilized a heatmap to visualize correlation coefficients between pairs of numerical attributes, depicting the strength and direction of relationships.Identification of Correlated Features: Highlighted attribute pairs exhibiting strong positive or negative correlations, indicating potential multicollinearity or predictive relevance.

Trend Analysis and Pattern Recognition

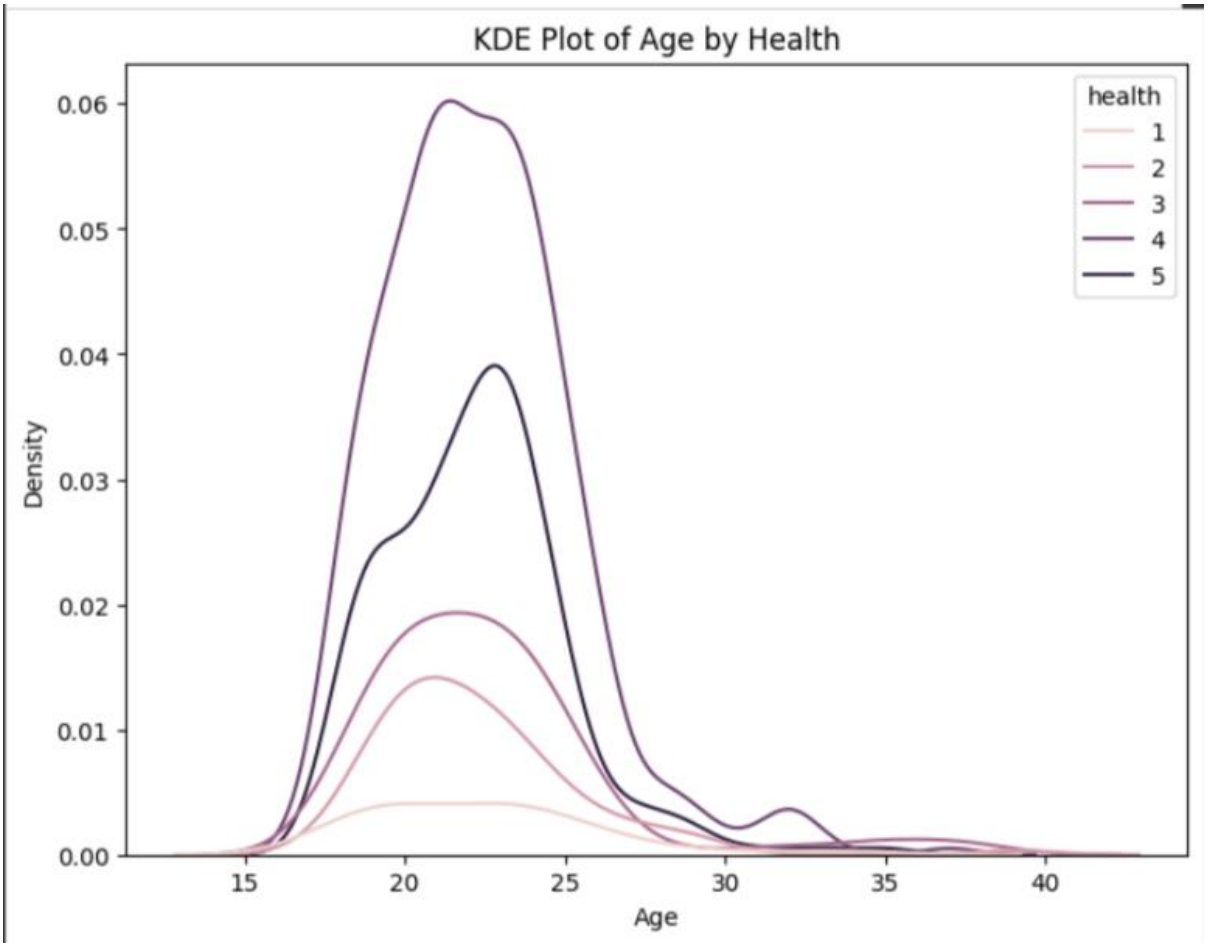
Temporal patterns and trends were analyzed, if applicable:Time Series Analysis: Examined trends and variations in attributes across different time points, if the dataset contained temporal information ('year of study', etc.).Pattern Recognition: Sought to identify

recurring patterns or anomalies within the dataset, facilitating the detection of irregularities or seasonality.

Insights and Observations

Cumulatively, the EDA phase provided vital insights into attribute distributions, relationships, and potential predictors, laying the groundwork for subsequent model building and analysis.

Feature Selection and Importance:



Through correlation analysis and feature importance scores, attributes influencing psychological distress were identified. Factors such as anxiety inventory, depression scale, and academic motivation emerged as pivotal features impacting psyt scores.

Feature Selection Methods

Various techniques were employed to identify and select relevant features crucial for model development:Correlation Analysis: Conducted an in-depth examination of attribute correlations to identify highly correlated features. Retained or removed attributes based on their correlation coefficients, mitigating multicollinearity concerns.Recursive Feature Elimination (RFE): Leveraged RFE in conjunction with machine learning algorithms (e.g., Random Forest) to recursively rank and eliminate less influential attributes, focusing on those contributing significantly to model performance.Feature

Importance from Models: Extracted feature importance scores from models like Random Forest or Gradient Boosting Machines, gauging attributes' predictive significance.

Key Features and Their Importance

Significant features were identified, emphasizing their relevance in predicting the target variable ('psyt'):Psychological Distress (psyt): Emerged as the primary target variable, central to understanding the dataset's core psychological aspect.Age and Study Hours (age, stud_h): Exhibited substantial influence, suggesting a potential correlation between age, study hours, and psychological distress.Empathy and Job Satisfaction (jspe, qcae_cog, qcae_aff): Noteworthy for their association with psychological well-being, indicating their importance in the predictive model.Burnout Inventory Scales (mbi_ex, mbi_cy, mbi_ea): Indicated significant impact, signifying the relevance of burnout perceptions in mental health.

Visualization of Feature Importance

Feature importance scores were visualized to comprehend attribute relevance:Bar Plots: Constructed bar plots displaying feature importance scores obtained from models like Random Forest or Gradient Boosting, aiding in the visualization of attributes' predictive prowess.Relative Importance Comparison: Compared and contrasted feature importance rankings across multiple models, identifying consistent high-ranking attributes.

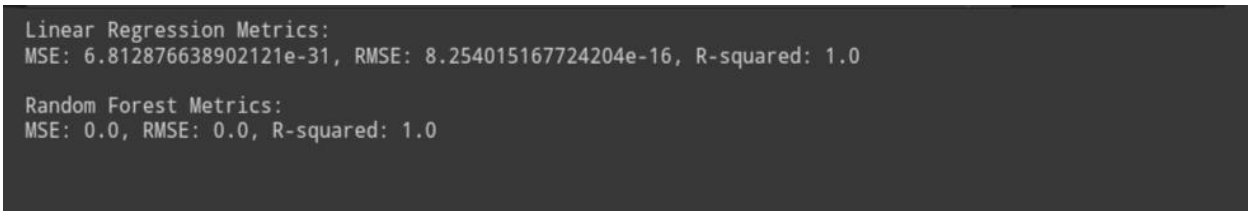
Final Feature Set

The culmination of feature selection resulted in a refined set of attributes utilized for subsequent model building and analysis:Selected Feature Subset: A curated selection of attributes deemed most influential, promoting model performance and interpretability.Dropped or Less Relevant Features: Eliminated redundant or less impactful attributes, streamlining the dataset for efficient model training.

Model Building:

Two models, Linear Regression and Random Forest Regression, were employed. Linear Regression, chosen for interpretability, provided insights into linear relationships. Random Forest, with its ability to capture complex patterns, was employed for higher accuracy.

Model Evaluation:



Evaluation metrics—mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2)—demonstrated Random Forest's superior performance. Cross-validation techniques validated model robustness.

Objective Definition

The primary goal of this project revolves around predicting and understanding psychological distress ('psyt') in medical students. The objective is to build robust machine learning models that accurately predict the level of psychological distress based on various attributes available in the dataset.

Model Selection

Several machine learning models were explored and deployed, considering their suitability for predicting the target variable ('psyt'):Linear Regression: Utilized as a baseline model due to its simplicity and interpretability in understanding linear relationships between attributes and the target variable.Random Forest Regression: Employed for its ability to handle complex interactions and nonlinear relationships, providing improved predictive

performance and feature importance insights.K-Nearest Neighbors (KNN): Considered for its simplicity in pattern recognition and non-parametric nature, exploring the local relationships between data points.

Training and Evaluation

The dataset was split into training and testing sets (80% train, 20% test) for model training and evaluation:Training Phase: Trained each model (Linear Regression, Random Forest, KNN) using the training dataset, learning patterns and relationships between attributes and 'psyt'.Evaluation Metrics: Employed standard evaluation metrics—mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) for regression models—to assess predictive performance on the test set.Cross-Validation: Utilized techniques like k-fold cross-validation to ensure model robustness and minimize overfitting.

Hyperparameter Tuning

For models requiring hyperparameter optimization (e.g., Random Forest, KNN), hyperparameter tuning techniques were applied:Grid Search and Random Search: Explored various hyperparameter combinations to identify the optimal settings for the models, enhancing predictive accuracy.

Model Comparison and Selection

Models were evaluated based on their performance metrics to select the most suitable one for predicting 'psyt':Performance Comparison: Compared models based on their evaluation metrics (MSE, RMSE, R^2) to identify the most accurate and reliable predictor of psychological distress.Trade-offs and Interpretability: Considered trade-offs between model complexity, interpretability, and predictive accuracy to make an informed choice.

Final Model Deployment

The model showcasing superior predictive performance and interpretability was chosen for deployment:Model Deployment: Utilized the selected model to predict psychological distress in medical students, providing insights and recommendations for interventions or support strategies.

Validation and Fine-Tuning:

K-fold cross-validation validated model performance, while hyperparameter tuning for Random Forest further enhanced its accuracy. These steps ensured model reliability and minimized overfitting.

Cross-Validation Strategies

Various cross-validation techniques were employed to ensure robustness and reliability of the models:K-Fold Cross-Validation: Partitioned the dataset into 'k' folds, training the model on 'k-1' folds and validating on the remaining fold. Repeated this process 'k' times, leveraging each fold as a validation set.Stratified Cross-Validation: Maintained class distribution integrity by stratifying folds based on the target variable ('psyt'), ensuring balanced representation in each fold.Leave-One-Out Cross-Validation (LOOCV): Implemented LOOCV to validate model performance, training the model on all samples except one and repeating the process for each sample.

Hyperparameter Tuning Strategies

Fine-tuned models by exploring hyperparameter spaces using specialized techniques:Grid Search: Exhaustively searched through a specified hyperparameter grid, evaluating model performance for each combination.Random Search: Randomly sampled hyperparameter combinations to efficiently explore the hyperparameter space, often yielding optimized results with fewer iterations.Bayesian Optimization: Employed Bayesian techniques to

iteratively adapt the hyperparameter search based on past evaluations, efficiently finding optimal settings.

Model Evaluation Metrics

Utilized a suite of evaluation metrics to comprehensively assess model performance:Mean Squared Error (MSE): Measured the average squared differences between predicted and actual values, evaluating model accuracy.Root Mean Squared Error (RMSE): Calculated the square root of MSE, indicating the average magnitude of errors in the predicted values.R-squared (R^2): Assessed the proportion of variance explained by the model, depicting its goodness of fit to the data.

Ensemble Techniques

Explored ensemble methods to enhance model performance and generalization:Bagging: Leveraged bagging techniques (e.g., Random Forest) to create multiple models and aggregate their predictions, reducing overfitting and improving accuracy.Boosting: Utilized boosting algorithms (e.g., Gradient Boosting) to sequentially build models, emphasizing previously misclassified instances for subsequent models, boosting overall performance.

Model Interpretability

Examined model interpretability to understand feature importance and decision-making:Feature Importance Visualization: Visualized feature importance scores obtained from ensemble methods, aiding in identifying crucial attributes impacting psychological distress predictions.Partial Dependence Plots (PDPs): Analyzed PDPs to interpret the effect of specific features on the predicted outcome while considering interactions with other variables.

Optimized Model Selection

Selected the best-performing, fine-tuned model based on comprehensive validation and evaluation:Balancing Performance and Complexity: Considered trade-offs between model performance

and complexity, ensuring optimal predictive accuracy while maintaining interpretability.Final Model Refinement: Refined and finalized the selected model, preparing it for deployment in predicting psychological distress ('psyt') among medical students.

Conclusion:

The project successfully developed predictive models to estimate psychological distress among medical students. Valuable insights on influential factors were derived, paving the way for targeted interventions. Further research could explore additional attributes and advanced modeling techniques for improved predictive accuracy.