UC San Diego

Mental Health Analysis and Prediction

Group 20: Tianruo Sang, Wenhao Zhao, Xiyu Wang

March 12, 2025

Contents

Background

Basic Information:

- **Mental health issues** like depression and anxiety are rising due to **work, academic, and financial stress**.
- Many avoid seeking help due to **stigma and lack of resources**.
- **Traditional research** is subjective—data-driven analysis uncovers **key risk factors** and predicts **mental health outcomes**.
- Machine learning helps in **early detection**, guiding **better policies and support systems**.
- To achieve this, we analyze a large-scale mental health dataset.

Background

Dataset Overview:

- Generated from an anonymous survey (Jan–June 2023) studying depression risk factors in adults.
- **140,700 training & 93,800 test samples** covering students and professionals.
- **Key Features:**

  Demographics: Age, gender, city.

  Work & Study: Profession, job/study satisfaction, work/study hours.

  Mental Health: Suicidal thoughts, family history, depression.

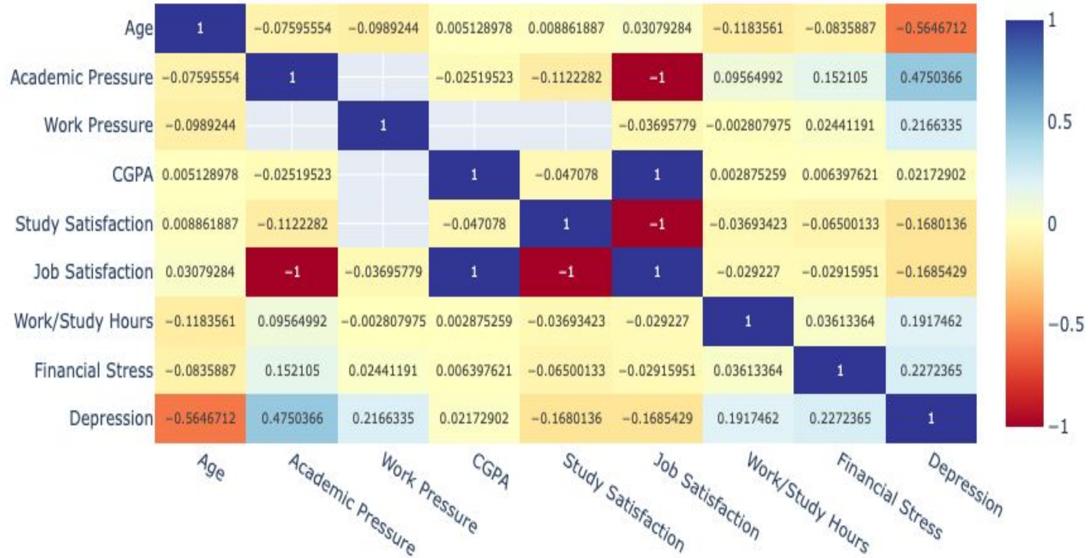
  Lifestyle & Stress: Sleep, diet, financial, academic & work pressure.

- Includes **numerical & categorical data**, requiring preprocessing for analysis and various cleaning methods for different models.

UC San Diego

Data Analysis

Heatmap of correlation matrix

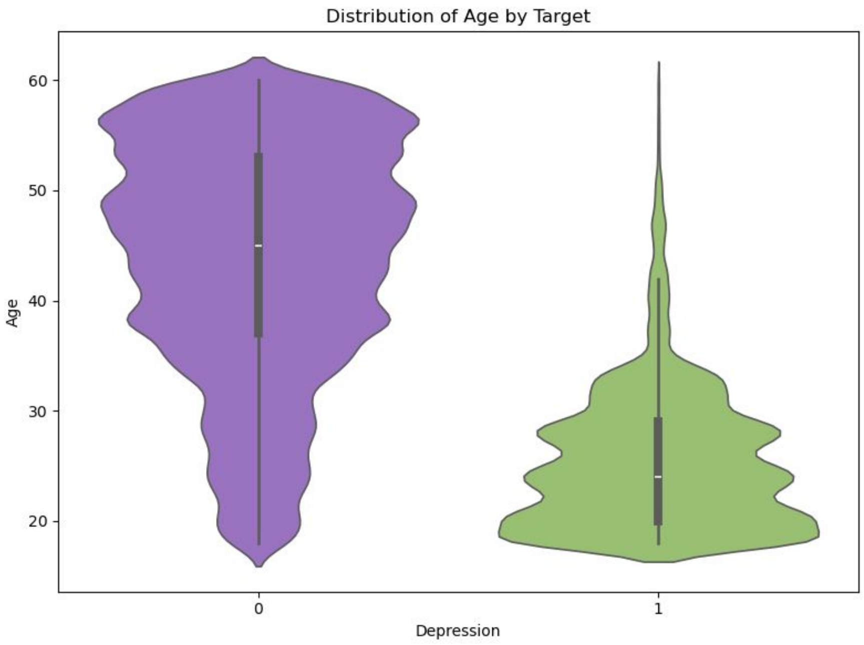


+1 : a perfect positive linear relationship
0 : no linear relationship
–1 : a perfect negative linear relationship

<u>Notable Relationship</u>:

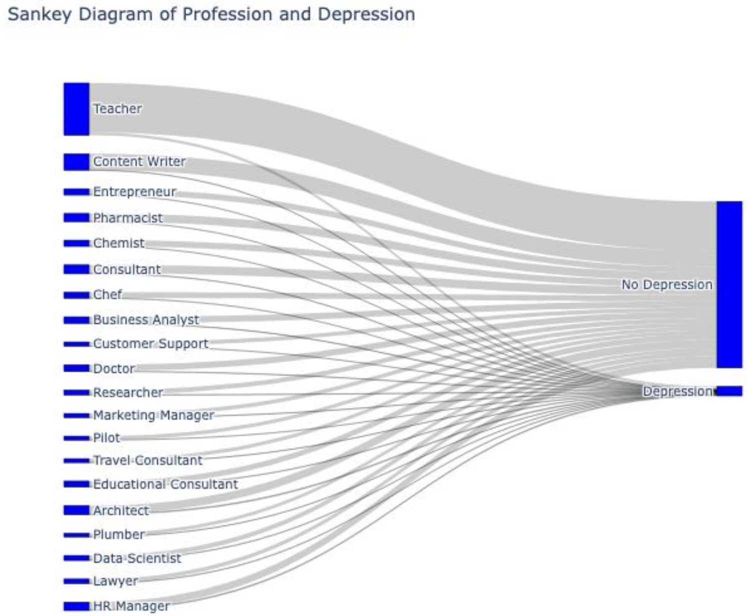
- **Age** (around –0.56, ***strongest***), meaning older individuals in this dataset tend to report lower levels of depression.
- **Academic Pressure** (≈ +0.48), indicating that those with higher academic pressure are more likely to be depressed.
- **Financial Stress** & **Work Pressure** (both mildly positive, around +0.22), so higher financial or work-related stress is associated with higher depression levels.

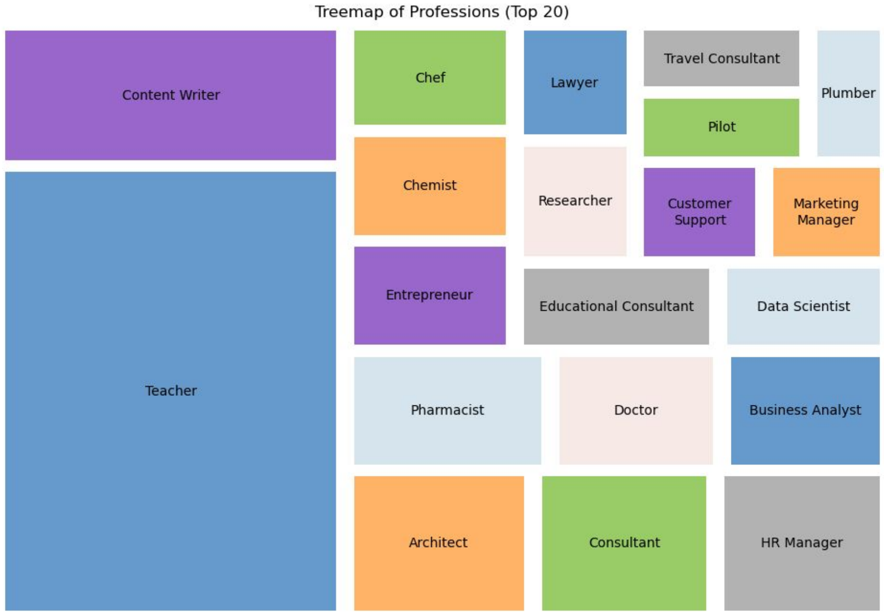
Data Analysis



- People with ***no*** depression tend to be **older**. Their age distribution stretches from roughly the late teens/early 20s up to the 60s, with a median somewhere around the early 40s.
- Those ***with*** depression form a somewhat **younger** (and slightly narrower) distribution, with a median closer to the early 30s.

Data Analysis





Teacher appears to be the single largest profession block who is more likely to be depressed , followed by **Content Writer**.

Model Prediction

UC San Diego

Model 1 - Random Forest:

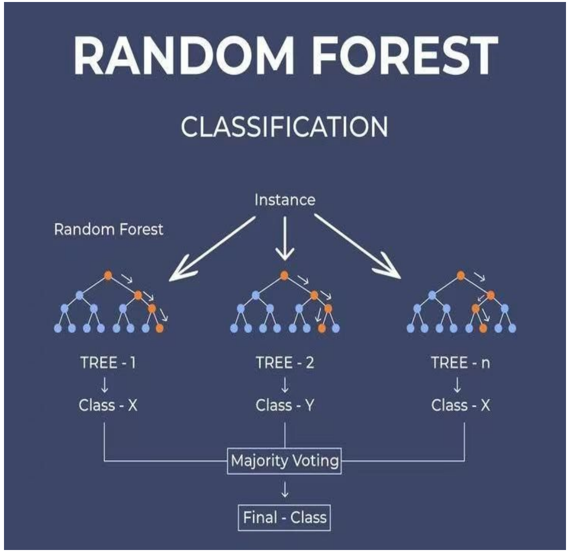
- An ensemble learning method that enhances accuracy and reduces overfitting.
- Handles both categorical and numerical data efficiently

1. Data Cleaning & Preparation

- **Fill the Missing Values:** Median imputation (numerical)& mode imputation (categorical).
- **Duplicates & Outliers:** Removed redundant entries; treated extreme values.
- **Encoding:** One-Hot Encoding (nominal), Label Encoding (ordinal).
- **Train-Test Split:** 80% training, 20% validation (Stratified Sampling).

2. Random Forest Model & Hyperparameter Tuning

- **Baseline Model:** Initial Random Forest with default parameters.
- **Grid Search Optimization:** Tuned key hyperparameters, achieving:
  - **Best Parameters:** n_estimators = 300, random_state = 42
  - **Cross-Validation:** Ensured stability of optimized model.



Model Prediction

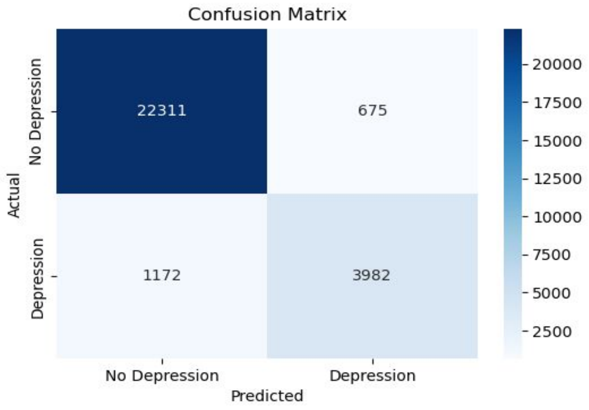
3. Results and Visualization

- **Model Evaluation Metrics**

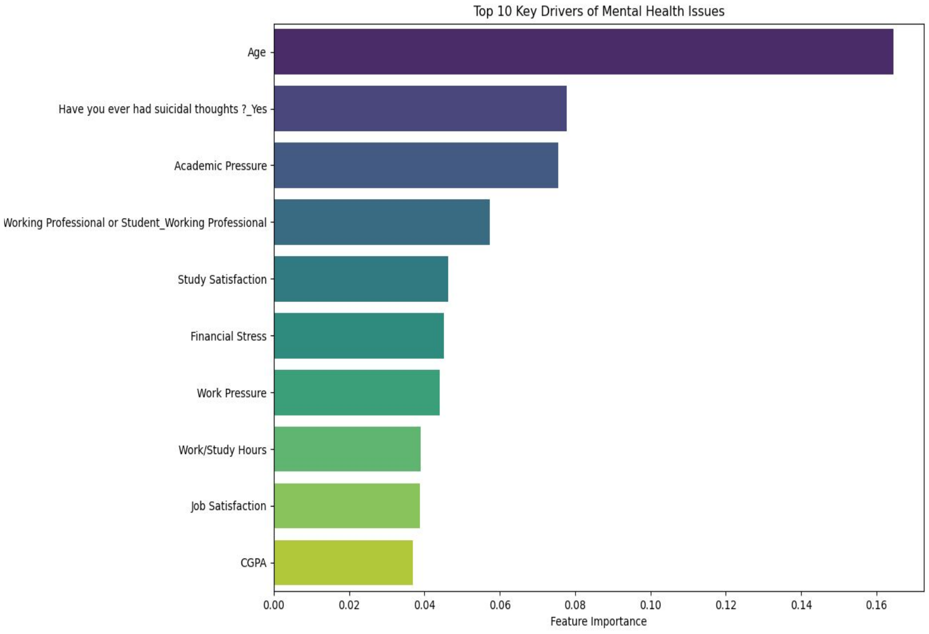
Accuracy: 93.4%; **Precision:** 0.93 (weighted)

Recall: 0.93 (weighted) ; **F1-score:** 0.93 (weighted)

- **Confusion Matrix Insights**



- **Feature Importance Plot**



Model Prediction

Model 2 - Voting Classifier:

1.Removing Unrelated Columns and Merging Similar Columns

- Columns such as `'id'`, `'Name'` are removed from both the training and test datasets as they are not relevant.
- The `'Work Pressure'` and `'Academic Pressure'` columns are combined into a single `'Pressure'` column .

2.Encoding Categorical Variables

- Applied target encoding to `'City'` and `'Profession'`.

3.Handling Missing Values & Scaling

- Median imputation, standardization, and conversion to `float32`.

4.Outlier Detection & Removal

- Used `IsolationForest` (4% contamination) to filter outliers from training data.

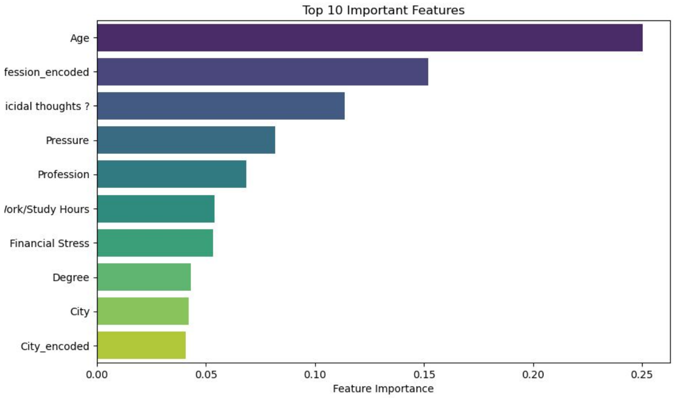
Model Prediction

Model Training:

[RandomForestClassifier: n_estimators=50]

[LogisticRegression: Solver='saga' with max_iter=1000]

[LinearSVC: A fast linear SVM implementation]



Combine Models using Voting Classifier:

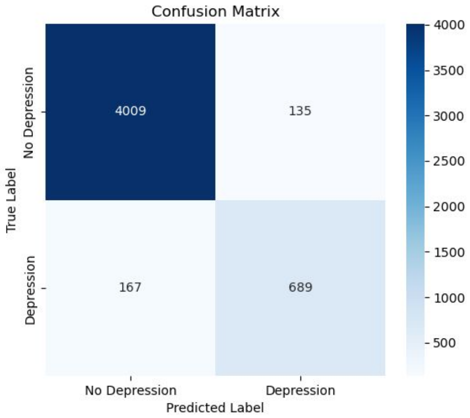
- `VotingClassifier` integrates the three classifiers with **hard voting**, meaning it predicts based on majority voting (without probability weighting).
- This improves robustness by leveraging the strengths of multiple models.

Results:

Training Accuracy: 0.9396

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.97	0.96	4144
1	0.84	0.80	0.82	856
accuracy			0.94	5000
macro avg	0.90	0.89	0.89	5000
weighted avg	0.94	0.94	0.94	5000



Model Prediction

UC San Diego

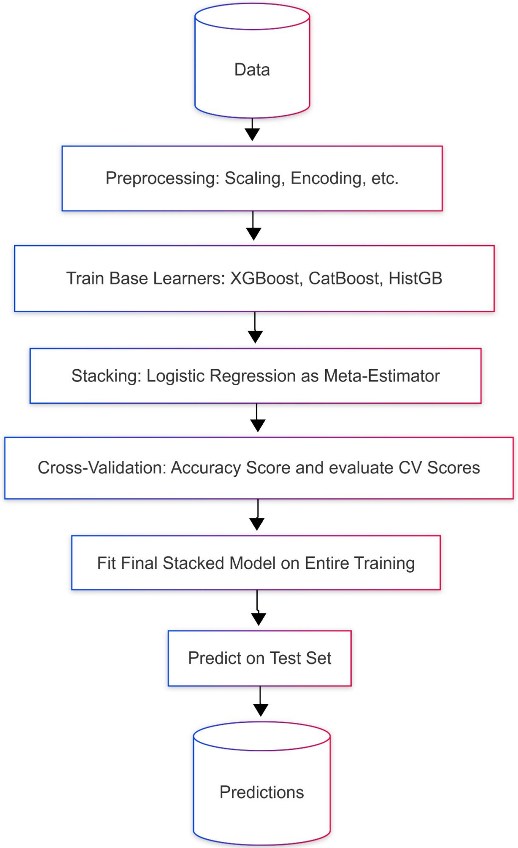
Model 3 - Stacked Ensemble Model

1. Parameter Tuning & Stacking

- **Base estimators :** CatBoost, XGBoost, and HistGradientBoosting.
- **Parameter tuning :** Hyperparameters have been tuned separately to achieve better predictive performance.
- **Stacking classifier :** Take the outputs base models as inputs to a Logistic Regression model.
- **Logistic regression :** Learn how best to combine the predictions of the three base models to produce a final prediction.

2. Cross-validation & Scoring

- **Evaluation :** Split the data into 5 folds, train on four folds, and validate on the remaining one.
- **Scoring** : Uses `accuracy_score` to measure how well the stacked ensemble classifies the target labels.



Model Prediction

3. Results and Visualization

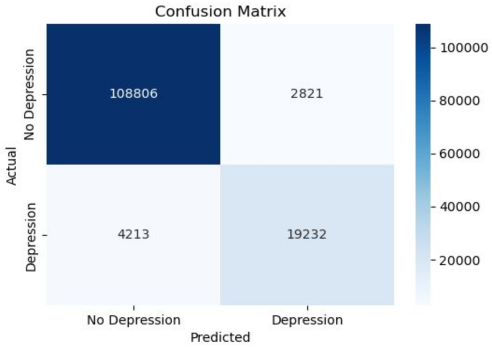
- **Model Evaluation Metrics**

  Cross-Validation Scores: [0.94 0.94 0.94 0.94 0.94]

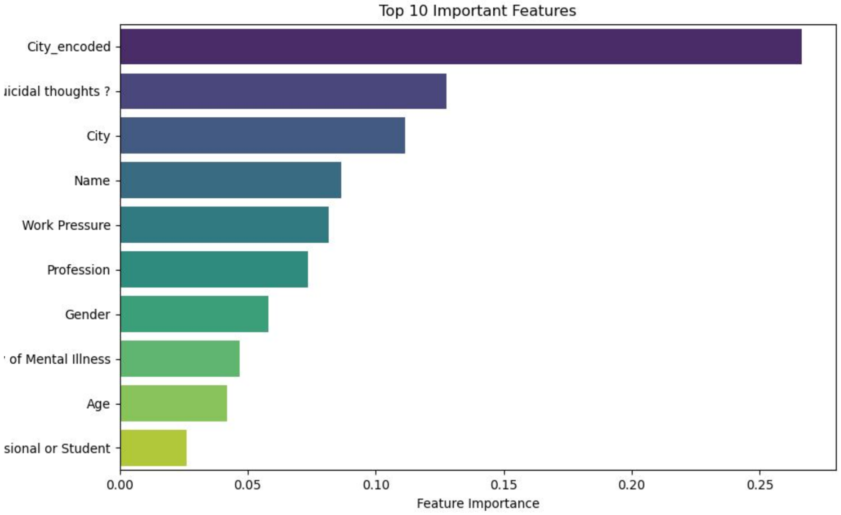
  Mean CV Accuracy: 0.9437

  Standard Deviation of CV Accuracy: 0.0013

- **Confusion Matrix Insights**



- **Feature Importance Plot**



Summary

In this project, we investigate a large-scale mental health dataset, do data analysis work on it to get an overview on the relationship between various features and depression, and then build and evaluate three classifier models to predict depression. We also identify key features that are most likely to lead to depression from these models. We aim for our results to be helpful in clinical diagnosis and analysis, ultimately reducing the risk of depression in individuals.

Thank you!