Analysis and Predictive Modeling of Depression Risk

Matthieu DONATI ARRANZ

April 9, 2025

Case Study 3

Toulouse School of Economics
Magistère Economiste-Statisticien 3A

Abstract

This report presents an analysis of a dataset aimed at understanding the factors contributing to depression risk among adults. The goal is to apply various predictive modeling techniques to assess depression risk based on demographic and lifestyle factors, with a focus on selecting the best-performing model. The report evaluates the performance of models such as Logistic Regression, Random Forest, comparing their effectiveness through cross-validation and various performance metrics. The best model is chosen based on these results, and conclusions are drawn regarding its applicability in predicting depression risk.

Contents

1 Previous Studies

In this section, we review existing literature on depression risk factors and predictive models applied to mental health data. We explore studies that have highlighted demographic, lifestyle, and psychological factors as contributors to depression, as well as models that have been used to predict mental health outcomes.

1.1 Previous Studies on Depression and Mental Health

Ali and Khoja [2019] demonstrated the growing evidence of air pollution's impact on depression. Air pollution has serious neurocognitive effects and is linked to mental disorders such as depression. Particulate matter induces brain inflammation and oxidative stress, contributing to depression.

Olfati et al. [2024] fond that with only the variable "Sleep" they obtained $R^2 = 0.18$ and adding Anxiety Level a $R^2 = 0.45$.

2 The Data at Hand

2.1 Data Description

The dataset used in this analysis was collected as part of a survey conducted between January and June 2023. The survey targeted individuals across various cities and aimed to understand the factors contributing to depression risk. The dataset includes both demographic and lifestyle features, as well as a binary target variable indicating whether a person is at risk of depression. The data was found on Kaggle (Rao [2023])

2.2 Variable Description

The dataset contains the following features:

- **Name:** Categorical variable representing the participant's name.
- **Gender:** Categorical variable (Male/Female).
- **Age:** Continuous variable representing the participant's age.
- **City:** Categorical variable indicating the respondent's city of residence.
- **Working Professional or Student:** Categorical variable indicating whether the participant is a working professional or a student.
- **Profession:** Categorical variable specifying the participant's profession (if applicable).
- **Academic Pressure:** Ordinal variable indicating the level of academic pressure experienced (1 to 5).
- **Work Pressure:** Ordinal variable indicating the level of work pressure experienced (1 to 5).
- **CGPA:** Continuous variable representing the participant's cumulative grade point average.
- **Study Satisfaction:** Ordinal variable indicating satisfaction with studies (1 to 5).
- **Job Satisfaction:** Ordinal variable indicating satisfaction with their job (1 to 5).

- **Sleep Duration:** Categorical variable representing intervals of Sleep Duration
- **Dietary Habits:** Categorical variable indicating the participant's dietary preferences (Healthy, Unhealthy, Moderate).
- **Degree:** Categorical variable specifying the participant's highest degree achieved.
- **Have you ever had suicidal thoughts?:** Binary variable indicating whether the participant has ever had suicidal thoughts (Yes/No).
- **Work/Study Hours:** Continuous variable representing the number of hours worked or studied per week.
- **Financial Stress:** Ordinal variable indicating the level of financial stress experienced (1 to 5).
- **Family History of Mental Illness:** Binary variable indicating whether there is a family history of mental illness (Yes/No).
- **Depression (target):** Binary variable indicating depression risk (Yes/No).

3 Data Processing

In this section, the data was processed through a series of transformations to ensure consistency and completeness, as well as to derive new features for analysis. The following steps were undertaken:

3.1 Handling Missing Values

- **Pressure Columns**: The 'Academic Pressure' and 'Work Pressure' columns were filled with zeros for any missing values. These were then summed to create a new 'Pressure' column representing the combined academic and work pressure.
- **Activity Satisfaction**: The missing values in the 'Study Satisfaction' and 'Job Satisfaction' columns were replaced with zeros, and the sum of these columns was stored in the new 'Activity Satisfaction' column.
- **CGPA**: Missing values in the 'CGPA' column were replaced with the mean CGPA of the dataset.

3.2 Categorical Data Processing

- **Profession**: Entries in the 'Working Professional or Student' column were used to assign a value of 'Student' to individuals identified as students. Additionally, any typos in the 'Profession' column were corrected (e.g., 'Finanancial Analyst' was corrected to 'Financial Analyst').

- **Mental Health Indicators**: The 'Depression' and 'Have you ever had suicidal thoughts?' columns were transformed into binary variables with '1' indicating a "Yes" response and '0' for "No."

- **Dietary Habits**: The 'Healthy Diet' and 'Unhealthy Diet' columns were created based on the dietary habits of the respondents, with '1' indicating a positive match and '0' for a negative match.

3.3 Feature Engineering

- **Degree Classification**: The 'Degree' column was mapped to a higher-level category using predefined classifications (e.g., "Undergraduate," "Postgraduate," "Doctorate," etc.). This was achieved by creating a dictionary that mapped each degree to its category, which was then applied to the 'Degree' column to create a new 'Degree Cat' column.

- **State Mapping**: Cities were mapped to their respective states using a dictionary, and a new 'State' column was created accordingly.

- **AQI (Air Quality Index)**: The Air Quality Index (AQI) data for each city was scraped from an external website: wunderground and mapped to the 'City' column.

3.4 Data Transformation

- **Sleep Duration**: The 'Sleep Duration' column, which had categorical data such as '7-8 hours,' was mapped to numerical values using a dictionary that represented average hours for each category.

- **Job Classification**: The 'Profession' column was mapped to broader job categories (e.g., "Finance & Accounting," "Education," "Engineering," etc.) based on predefined job classifications. A new column 'Job Cat' was created for this classification.

3.5 Name Processing

- **Name Normalisation**: In the 'Name' column, names that appeared less than 15 times were grouped into a category labeled "Other."

3.6 Dropping Irrelevant Columns

Several columns that were deemed unnecessary for further analysis (such as 'Academic Pressure,' 'Work Pressure,' 'Study Satisfaction,' 'Job Satisfaction,' 'Working Professional or Student,' 'Dietary Habits,' 'Gender,' 'City,' 'Profession,' and 'Degree') were dropped to reduce dimensionality.

4 Simple Statistics from the Data

In this section, we explore the basic statistics and distributions of the dataset to understand the underlying patterns.

The dataset contains several variables that represent different aspects of the respondents' characteristics, including demographic details, lifestyle factors, and measures of mental well-being. Below is a summary of the key descriptive statistics for these variables, which include both quantitative and qualitative aspects.

4.1 Quantitative Data

Let's first starThe table presents the count, mean, standard deviation, and percentiles (25%, 50%, and 75%) for each variable, along with the minimum and maximum values observed in the dataset.

Variable	**Count**	**Mean**	**Std**	**Min**	**25%**	**50% (Median)**	**75%**	**Max**
Age	2,556	39.04	12.26	18	28	39	50	60
CGPA	2,556	7.57	0.65	5.03	7.57	7.57	7.57	10.00
Sleep Duration (hours)	2,556	6.73	2.24	4.00	4.00	7.50	7.50	10.00
Work/Study Hours (hours)	2,556	6.02	3.77	0.00	3.00	6.00	9.00	12.00
Financial Stress	2,556	2.97	1.42	1.00	2.00	3.00	4.00	5.00
Pressure	2,556	3.02	1.41	1.00	2.00	3.00	4.00	5.00
Activity Satisfaction	2,556	3.03	1.41	1.00	2.00	3.00	4.00	5.00
AQI 2024	2,556	186.2	121.4	61.0	97.0	115.0	327.0	403.0

Table 1: Summary statistics for quantitative variables

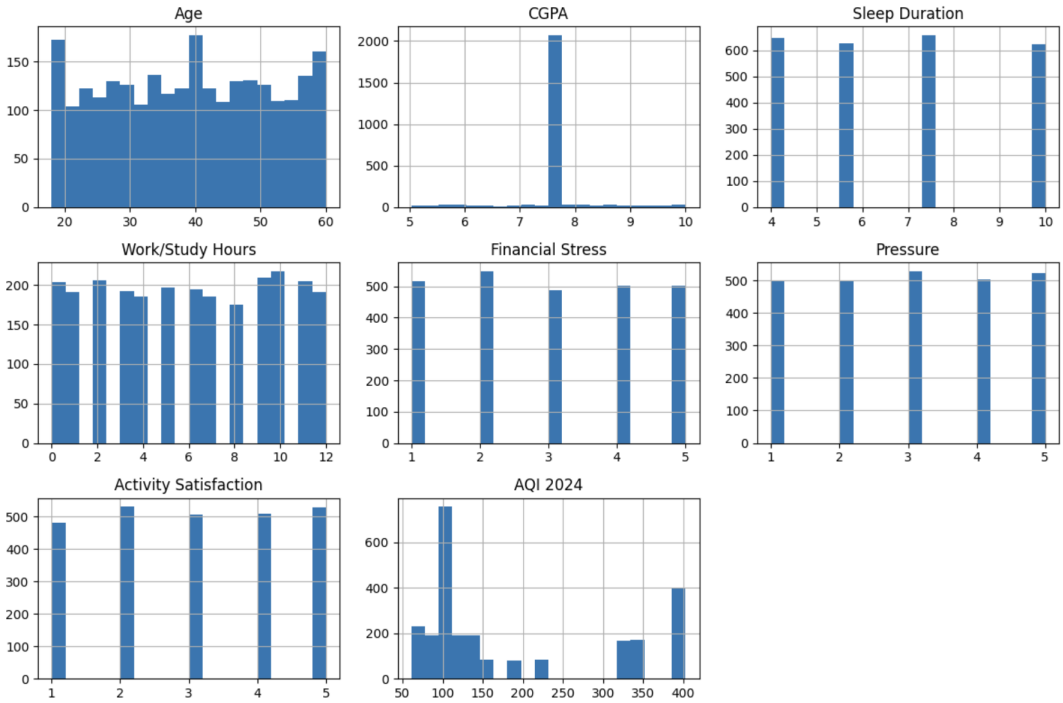

Figure 1: Histograms Quantitative Variables

All the categories are rather balanced apart from CFPA. But this is probably because all the non-students were assigned the mean grade, which represents more than 80% of the sample. Quality of Air Index is also crushed to the left because of the quality of the air in Delhi that is awful (QAI supérieur to 400).

4.2 Dummy Variables

The table below presents the proportions of the dummy variables in the dataset. Each variable has been encoded as a binary indicator (1 for true and 0 for false), and the proportions represent the percentage of respondents who fall into each category.

Variable	Proportion (%)
Student	19.64
Have you ever had suicidal thoughts?	48.87
Family History of Mental Illness	48.71
Male	52.15
Healthy Diet	32.94
Unhealthy Diet	34.51
Depression	17.80

Table 2: Proportions of Each Dummy Variable

4.3 Categorical Variables

In this subsection, we present the counts for key categorical variables in the dataset, specifically for the degree categories, states, and job categories. These counts give an overview of the distribution of respondents based on their educational background, location, and job classification.

Degree Category	Count
Undergraduate	1109
Postgraduate	942
School	275
Med School	149
Doctorate	81

Table 3: Degree Category Distribution

Many Undergraduate and PostGraduate. School, Medschool and especially Doctorate have less occurences.

State	Count
Maharashtra	615
Uttar Pradesh	496
Gujarat	356
Madhya Pradesh	163
Jammu and Kashmir	102
Bihar	89
Tamil Nadu	88
Punjab	88
West Bengal	87
Telangana	85
Andhra Pradesh	84
Rajasthan	80
Karnataka	76
Delhi	74
Haryana	73

Table 4: State Distribution

Job Category	Count
Student	502
Business School	434
Education	411
Creative Arts	274
Engineering	237
Unknown	171
Healthcare	141
Finance & Accounting	102
Law	90
Travel	85
Manual	68
Customer Service	41

Table 5: Job Category Distribution

We can observe that some regions are much more represented than others, such as Maharashtra (615), Uttar Pradesh (496), Gujarat (356).

We can also observe that the Data Set contains many Student (502).

4.4 Relation between Quantitative Explanatory Variables

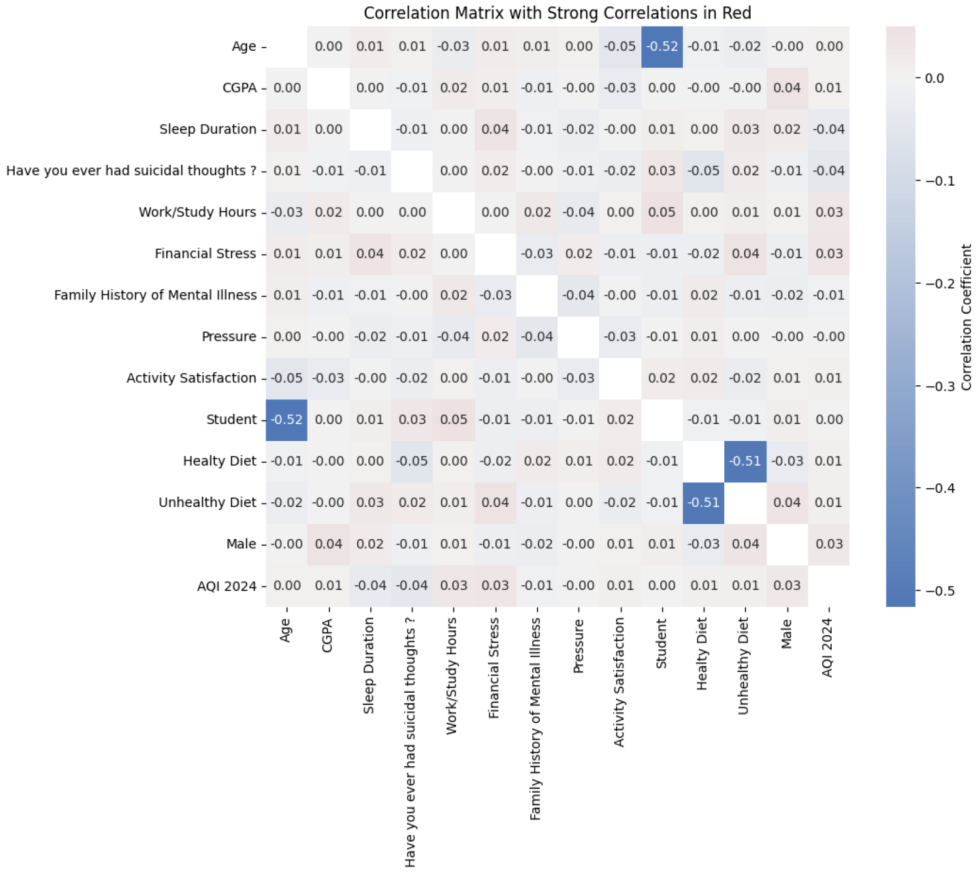

Figure 2: Correlation Matrix: Quantitative Explanatory Variables

Apart from Student with Age that are correlated negatively (not surprising) and Healthy Diet with Unhealthy Diet (not very shocking either as their are incompatible) all the other variable seems rather uncorrelated to each other which is nice.

4.5 Links betweeen Variable of Interest "Depression" and Explanatory Variables

The Student's t-test is used to compare the means of two independent samples to determine if a statistically significant difference exists between the two groups. The test is based on the following hypotheses:

- Null hypothesis (H_0): The means of the two groups are equal.
- Alternative hypothesis (H_1): The means of the two groups are different.

The t-statistic is calculated as follows:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\overline{X_1}$ and $\overline{X_2}$ are the means of the two groups,
- s_1^2 and s_2^2 are the variances of the two groups,
- n_1 and n_2 are the sample sizes of the two groups.

The larger the absolute value of the t-statistic, the more likely it is that the means of the two groups are significantly different.

The p-value measures the probability of observing a difference as extreme as the one calculated, assuming the null hypothesis is true. If the p-value is less than a significance threshold (usually 0.05), the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected.

$$\text{p-value} = P(\text{t-value} \geq \text{observed value} | H_0 \text{ true})$$

If the p-value is less than 0.05, we consider the difference between the groups to be statistically significant. Otherwise, there is not enough evidence to reject the null hypothesis.

Variable	t-statistic	p-value	Difference
Age	28.04	4.799×10^{-151}	True
CGPA	-1.31	0.191	False
Sleep Duration	3.09	2.046×10^{-3}	True
Have you ever had suicidal thoughts ?	-14.84	8.000×10^{-48}	True
Work/Study Hours	-7.80	9.209×10^{-15}	True
Financial Stress	-8.49	3.459×10^{-17}	True
Family History of Mental Illness	-0.97	0.332	False
Pressure	-12.63	1.599×10^{-35}	True
Activity Satisfaction	8.80	2.560×10^{-18}	True
Student	-23.30	4.726×10^{-109}	True
Healthy Diet	4.40	1.111×10^{-5}	True
Unhealthy Diet	-5.14	3.022×10^{-7}	True
Male	-0.38	0.701	False
AQI 2024	1.29	0.197	False

Table 6: Results of t-tests for Quantitative Variables with Depression

If we stopped here, we would conclude that Grades, "Family History of Mental Illness", Gender,and Air Quality have not influence on Depression.

The Chi-Square test is used to determine whether there is a significant association between two categorical variables. It is based on the following hypotheses:

- Null hypothesis (H_0): The two categorical variables are independent.
- Alternative hypothesis (H_1): The two categorical variables are dependent.

The Chi-Square statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

- O_i is the observed frequency for category i,
- E_i is the expected frequency for category i.

The larger the value of the Chi-Square statistic, the more likely it is that the two variables are dependent.

The p-value indicates the probability of obtaining a Chi-Square statistic as extreme as the one calculated, assuming the null hypothesis is true. If the p-value

is less than a significance threshold (usually 0.05), the null hypothesis is rejected, indicating that the variables are dependent.

$$\text{p-value} = P(\chi^2 \geq \text{observed value} | H_0 \text{ true})$$

If the p-value is less than 0.05, we reject the null hypothesis, suggesting that the two categorical variables are dependent. If the p-value is greater than 0.05, the null hypothesis cannot be rejected, suggesting that the variables are independent.

Variable	Chi-square	p-value	Independence
Name	26.57	0.935	True
Degree Cat	169.3	0.000	False
State	23.81	0.048	False
Job Cat	522.3	0.000	False

Table 7: Results of Chi-Square Tests for Qualitative Variables with Depression

Degree_Cat, State and Job_Cat seems to be connected to the "Depression" especially "Degree_Cat" and "Job_Cat" that have very small $p - values$.

The following table presents the proportions of individuals with and without depression for each degree category, as well as the total sample size (n).

Degree Category	**No Depression**	**Depression**	**n**
Doctorate	0.8765	0.1235	81
Med School	0.8658	0.1342	149
Postgraduate	0.8715	0.1285	942
School	0.5418	0.4582	275
Undergraduate	0.8395	0.1605	1109

Table 8: Proportions of Depression by Degree Category

Remark: The higher the eduction, the happier. The proportion of Depression among people at School is enormous.

The following table presents the proportions of individuals with and without depression for each state, along with the total sample size (n).

State	No Depression	Depression	n
Andhra Pradesh	0.8333	0.1667	84
Bihar	0.8202	0.1798	89
Delhi	0.8784	0.1216	74
Gujarat	0.8118	0.1882	356
Haryana	0.8630	0.1370	73
Jammu and Kashmir	0.7647	0.2353	102
Karnataka	0.8684	0.1316	76
Madhya Pradesh	0.8037	0.1963	163
Maharashtra	0.8179	0.1821	615
Punjab	0.8295	0.1705	88
Rajasthan	0.8750	0.1250	80
Tamil Nadu	0.8295	0.1705	88
Telangana	0.6706	0.3294	85
Uttar Pradesh	0.8468	0.1532	496
West Bengal	0.8046	0.1954	87

Table 9: Proportions of Depression by State

Remark: Proportion in "Telangana", "Jannu and Kashmir" are much higher than the national average. People from Delhi seems happier

The following table presents the proportions of individuals with and without depression for each job category, along with the total sample size (n).

Job Category	No Depression	Depression	n
Business School	0.9240	0.0760	434
Creative Arts	0.9124	0.0876	274
Customer Service	0.9512	0.0488	41
Education	0.9173	0.0827	411
Engineering	0.9198	0.0802	237
Finance & Accounting	0.9118	0.0882	102
Healthcare	0.9645	0.0355	141
Law	0.8778	0.1222	90
Manual	0.9265	0.0735	68
Student	0.4980	0.5020	502
Travel	0.9529	0.0471	85
Unknown	0.6667	0.3333	171

Table 10: Proportions of Depression by Job Category

Student have a much rate of Depression than other people in general (almost

50 %!). People that have not specified their occupation also have a higher rate of depression (33.33%).

We will not use "Name" for the Predictions.

5 Tested Predictive Models

In this section, we discuss the different predictive models tested on the dataset.

5.1 Logistic Regression

5.1.1 Theory

Logistic regression is a common method used for binary classification problems, where the goal is to predict the probability of an outcome. It uses the logistic function (also called the sigmoid function), which maps any real number into a value between 0 and 1.

The logistic regression model can be written as:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n))}$$

Where:

- $P(y = 1 \mid \mathbf{x})$ is the probability of the positive class (e.g., depression),
- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the feature vector,
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients to be learned by the model.

The model estimates the probability of the outcome, and a threshold (usually 0.5) is used to classify the result (0 or 1).

5.1.2 Ridge, Lasso, and Elastic-Net Regularisation applied to Logistic Regression

Regularisation techniques are used to prevent overfitting by adding a penalty to the model, which helps keep the model simple and generalisable.

- **Ridge Regularisation (L2):** In ridge regression, the penalty term is proportional to the square of the coefficients. This is added to the loss function:

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} [y_i \log P(y_i \mid \mathbf{x}_i) + (1 - y_i) \log(1 - P(y_i \mid \mathbf{x}_i))] + \lambda \sum_{j=1}^{p} \beta_j^2$$

  Where:

  - $\mathcal{L}(\beta)$ is the loss function,

- λ is the regularisation parameter that controls how much the coefficients are penalised,
- β_j^2 is the squared value of each coefficient.

Ridge regularisation helps reduce the influence of less important features by making their coefficients smaller.

- **Lasso Regularisation (L1):** Lasso regression uses the sum of the absolute values of the coefficients as the penalty term:

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} [y_i \log P(y_i \mid \mathbf{x}_i) + (1 - y_i) \log(1 - P(y_i \mid \mathbf{x}_i))] + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso can force some coefficients to be exactly zero, effectively selecting the most important features for the model.

- **Elastic-Net Regularisation:** Elastic-net combines both Lasso and Ridge regularisation. The penalty term is a mix of the L1 and L2 penalties:

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} [y_i \log P(y_i \mid \mathbf{x}_i) + (1 - y_i) \log(1 - P(y_i \mid \mathbf{x}_i))] + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

Where:

- λ_1 and λ_2 control the strength of the L1 and L2 penalties, respectively.
- Elastic-net is useful when there are many correlated features.

5.2 Random Forest

5.2.1 Theory

Random Forest is a method that builds several decision trees and combines them to make more accurate predictions. Each decision tree is built using a random sample of the data, and it randomly selects features when making splits. This randomness helps prevent overfitting and increases the robustness of the model.

The final prediction from a Random Forest model is the average (or most common class for classification tasks) of the predictions made by each tree:

$$\hat{y}_{\text{RF}} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t$$

Where:

- $\hat{y}_t$ is the prediction from the t-th decision tree,
- T is the total number of trees in the forest.

Random Forest is often effective for data with many features and complex interactions.

5.3 Model Evaluation: 10-Fold Cross-Validation, GridSearch, and Accuracy

To determine the best-performing model, we employed 10-fold cross-validation, which is a robust method for assessing model performance. In 10-fold cross-validation, the dataset is randomly split into 10 equal-sized subsets (folds). The model is trained on 9 of these folds and tested on the remaining fold. This process is repeated 10 times, with each fold being used once as the test set. The results from each fold are averaged to provide a more reliable estimate of model performance, reducing the risk of overfitting.

To optimise the models further, we used GridSearch, a technique that systematically searches for the best hyperparameters by evaluating multiple combinations. GridSearch exhaustively tests a predefined set of parameters (such as regularisation strength, tree depth, etc.) and selects the set that produces the best performance according to a chosen evaluation metric.

Finally, accuracy is used as a primary metric to evaluate model performance. Accuracy measures the proportion of correct predictions (both true positives and true negatives) out of all predictions:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

By combining 10-fold cross-validation with GridSearch and accuracy, we can ensure that the chosen model not only performs well but also generalises effectively to unseen data.

6 Results

Penalty	Ridge	Lasso	Elastic Net
9.840e-01	9.836e-01	9.851e-01	9.851e-01

Table 11: Logistic Regression Performance with Different Regularisation Techniques

6.1 Results

6.1.1 Log Regression

The table above shows the accuracy of logistic regression models with various regularisation techniques. The models were trained using the 'saga' solver, a random seed of 10, and a grid search over the following hyperparameters:

- **C**: [0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000] (C = $1/\lambda$)
  For Ridge, Lasso, Elastic - Net
- **L1 ratio**: [0.0, 0.25, 0.5, 0.75, 1.0] *for Elastic Net*

The results indicate that all regularisation techniques (None, Ridge, Lasso, and Elastic Net) performed similarly with slightly higher accuracy for the Elastic Net and Lasso regularisation. Lasso and Elastic Net display the exact same result because L1 ratio = 1 as best tested L1 parameter. When an elastic Net has a L1 ratio = 1 it becomes also a Lasso regularisation.

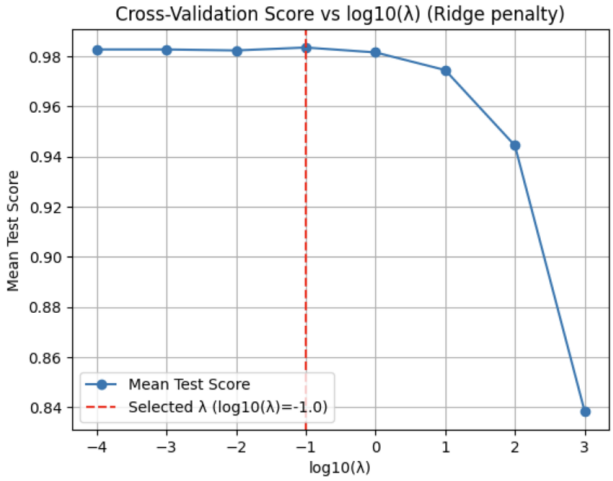

Figure 3: Log Regression with Ridge Penalty CV on λ

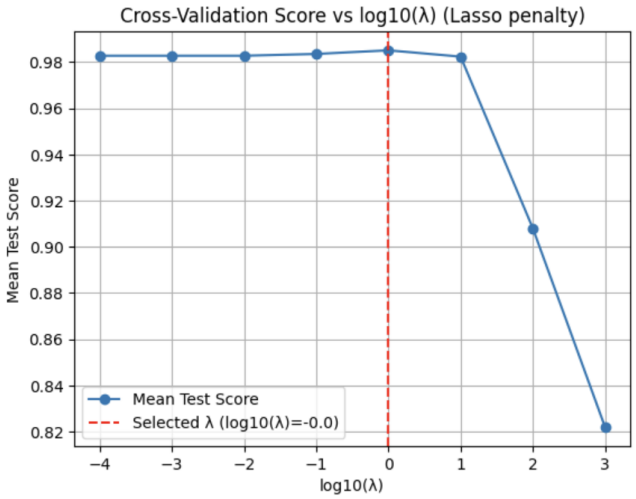

Figure 4: Log Regression with LASSO Penalty CV on λ

Remark: Best Tested Params: LASSO with $\lambda = 0$

6.1.2 Random Forest

Hyperparameters for Random Forest Classifier

In this subsection, we explain the main hyperparameters used in tuning a Random Forest model.

The **n_estimators** hyperparameter specifies the number of trees in the forest. A higher number of trees generally improves the model's performance by reducing variance, but it also increases computational cost. Typical values to consider are 10, 50, 100, and 200.

The **max_depth** hyperparameter controls the maximum depth of each individual decision tree in the forest. Setting this parameter limits how deep the tree can grow, preventing overfitting by forcing the model to make decisions with fewer levels of complexity. If **max_depth** is set to **None**, the trees will expand until they contain less than **min_samples_split** samples at a leaf node. Possible values are **None**, 10, 20, and 30, with higher values allowing trees to grow deeper.

The **min_samples_split** parameter defines the minimum number of samples required to split an internal node. A lower value allows the tree to grow deeper, leading to higher variance and potential overfitting. A higher value forces the model to make splits based on larger groups of data, reducing the tree depth and helping avoid overfitting. Typical values are 2, 5, and 10.

These hyperparameters can be fine-tuned through techniques like grid search to find the combination that provides the best performance for a given dataset.

The seed was set at 10.

Tested hyperparameters:

- n_estimators: [10, 50, 100, 200]
- max_depth: [None, 10, 20, 30]
- min_samples_split': [2, 5, 10]

Best Result: 0.9362

Obtained with the following tuning:

max_depth= None, min_samples_split= 2, n_estimators= 200

7 Best Model Observations

Notebook: Donati [2024] $R^2 = 0.92$, The best model we observe was the Log Regression Lasso with $\lambda = 1$: Here are the variable and their respective coefficients:

$$
\begin{aligned}
\hat{z} = & -1.537 \times 10^{1} \\
& -1.014 \times 10^{1} \cdot \text{Age} \\
& +2.631 \times 10^{-2} \cdot \text{CGPA} \\
& -1.769 \cdot \text{Sleep Duration} \\
& +6.346 \cdot \text{Have you ever had suicidal thoughts ?} \\
& +2.948 \cdot \text{Work/Study Hours} \\
& +3.642 \cdot \text{Financial Stress} \\
& +1.570 \cdot \text{Family History of Mental Illness} \\
& +5.343 \cdot \text{Pressure} \\
& -4.376 \cdot \text{Activity Satisfaction} \\
& +1.111 \cdot \text{Student} \\
& -1.021 \cdot \text{Healthy Diet} \\
& +1.258 \cdot \text{Unhealthy Diet} \\
& +4.860 \times 10^{-2} \cdot \text{Male} \\
& -1.816 \times 10^{-1} \cdot \text{Degree Cat\_Med School} \\
& -2.388 \times 10^{-1} \cdot \text{Degree Cat\_Postgraduate} \\
& +8.283 \times 10^{-2} \cdot \text{Degree Cat\_School} \\
& +2.693 \times 10^{-2} \cdot \text{State\_Bihar} \\
& -2.307 \times 10^{-1} \cdot \text{State\_Delhi} \\
& +4.721 \times 10^{-3} \cdot \text{State\_Haryana} \\
& +6.445 \times 10^{-2} \cdot \text{State\_Madhya Pradesh} \\
& -8.036 \times 10^{-2} \cdot \text{State\_Maharashtra} \\
& +7.580 \times 10^{-2} \cdot \text{State\_Punjab} \\
& +3.805 \times 10^{-2} \cdot \text{State\_Rajasthan} \\
& -1.202 \times 10^{-1} \cdot \text{State\_Tamil Nadu} \\
& -6.994 \times 10^{-3} \cdot \text{State\_Telangana} \\
& -9.359 \times 10^{-2} \cdot \text{State\_Uttar Pradesh} \\
& +1.053 \times 10^{-1} \cdot \text{State\_West Bengal} \\
& +5.633 \times 10^{-2} \cdot \text{Job Cat\_Creative Arts} \\
& +1.950 \times 10^{-1} \cdot \text{Job Cat\_Education} \\
& +1.423 \times 10^{-1} \cdot \text{Job Cat\_Engineering} \\
& -3.558 \times 10^{-2} \cdot \text{Job Cat\_Healthcare} \\
& +1.796 \times 10^{-2} \cdot \text{Job Cat\_Law} \\
& +1.451 \times 10^{-1} \cdot \text{Job Cat\_Manual} \\
& +1.111 \cdot \text{Job Cat\_Student}
\end{aligned}
$$

$\hat{y} = \frac{1}{1+\exp -\hat{z}}$, when $z \to \inf$ then $\hat{y} \to 1$ and when $z \to -\inf$ then $\hat{y} \to 0$

- It is observed that the older, the better.
- Healthy food is beneficial not only for the body but also for the spirit, as stated in *Mens sana in corpore sano*.

- Students tend to be more depressive than the general population.
- Pressure plays a very important role with a coefficient of 5.343.
- Activity Satisfaction is crucial in combating depression, with a coefficient of -4.376.
- Financial Stress can also have a significant impact.
- One of the most important variables is: "Have you ever had suicidal thoughts?", with a coefficient of 6.346.
- As noted in Olfati et al. [2024], lack of sleep can contribute to depression, as can long working hours.
- Unlike Ali and Khoja [2019], we do not observe an influence of air quality on depression.
- Despite having many explanatory variables, several have been removed due to small coefficients, which would have remained with Ridge regularization.

Conclusion

The goal of this study was to predict the risk of depression based on various explanatory variables. In this study, we first processed the data, followed by performing simple univariate and bivariate statistical analyses. We then tested several models, including Logistic Regression, Random Forest, and various regularisation techniques such as Lasso, Ridge, and Elastic-Net. Our results indicated that Lasso performed the best overall, closely followed by Logistic Regression without penalty, Ridge, and Elastic-Net. However, Log Regression models clearly outperformed Random Forest.

Among the selected models, we found that the most influential variables included being a student, pressure, activity satisfaction, financial stress, and suicidal thoughts, with the latter being the most significant predictor of depression. This model provides valuable insights into the factors contributing to depression and highlights the importance of addressing mental health in different populations.

8 Bibliography

References

Noor A. Ali and Abdullah Khoja. Growing evidence for the impact of air pollution on depression. *Ochsner Journal*, 19(1):4, Spring 2019. doi: 10.31486/toj.19.0011.

Matthieu Donati. Ed3 notebook, 2024. URL `https://colab.research.google.com/drive/1fiA8h_KyNRoyiGM5BKEVV3oJgrlK96nb`. Accessed: 2024-11-26.

Mahnaz Olfati, Fateme Samea, Shahrooz Faghihroohi, Somayeh Maleki Balajoo, Vincent Küppers, Sarah Genon, Kaustubh Patil, Simon B. Eickhoff, and Masoud Tahmasian. Prediction of depressive symptoms severity based on sleep quality, anxiety, and gray matter volume: a generalizable machine learning approach across three datasets. *eBioMedicine*, 108:105313, 2024. ISSN 2352-3964. doi: https://doi.org/10.1016/j.ebiom.2024.105313. URL `https://www.sciencedirect.com/science/article/pii/S2352396424003499`.

Radhika Rao. Depression survey dataset for analysis, 2023. URL `https://www.kaggle.com/datasets/sumansharmadataworld/depression-surveydataset-for-analysis/data`. Accessed: 2024-11-26.