Predicting Depression from Synthetic Mental Health Survey Data with Classical and Neural Models

Harjas Kaur 730559291
Sanjana Nalla 730573834
Nidhi Padala 730574464
Tejaswi Paladugu 730574505

Abstract

Depression is a prevalent mental health condition, and large-scale survey data provide an opportunity to study which demographic, academic, and lifestyle factors are associated with elevated risk. This work investigates depression prediction from a synthetic mental health survey using standard classification models on tabular data with demographic, academic, lifestyle, and self-reported mental health features, together with a binary Depression label for each respondent. We compare four approaches: Gaussian Naive Bayes, logistic regression with one-hot encoded categorical variables, a principal component analysis (PCA) based classifier trained on a two-dimensional projection of the numerical features, and a feedforward neural network. All models are evaluated using stratified 5-fold cross-validation and an 80/20 stratified validation split, with accuracy and ROC AUC as the primary metrics. Gaussian Naive Bayes provides the weakest baseline, whereas logistic regression and the PCA-based classifier achieve very high accuracy and near-perfect ROC AUC. The feedforward neural network attains slightly lower accuracy but produces informative probability estimates, indicating that relatively simple linear structure is sufficient to capture most of the predictive signal in this synthetic dataset.

1 Introduction

Depression is a widespread mental health condition that can disrupt daily routines, compromise academic performance, and impact long-term well-being, particularly for younger adults and students [1]. Screening for depression usually relies on self-report questionnaires and clinical interviews. These methods are beneficial, but they are also time-intensive, difficult to scale, and can still misclassify people who are experiencing significant or abnormal symptoms [2]. However, the growing availability of large survey datasets and standard machine learning libraries makes it possible to examine, in a more systematic way, how demographic, academic, and lifestyle factors are associated with depression.

In this project, we use a dataset from the Kaggle Playground Series (Season 4, Episode 11: "Exploring Mental Health Data"). Each row represents a synthetic mental health survey response generated by a deep learning model trained on an underlying depression survey. For each respondent, the dataset reports demographic characteristics, measures of academic and work pressure, sleep and diet patterns, indicators of financial stress, and several mental health-related items, along with a binary label indicating whether the person is categorized as depressed. This setup allows for a supervised learning problem: given the observed features, we aim to predict the depression label on a held-out test set, using accuracy as the primary evaluation metric.

We treat this dataset as a way to explore and compare several modeling approaches. Specifically, we consider four methods: a Gaussian Naive Bayes model as a simple probabilistic baseline; a logistic regression model with one-hot encoded categorical variables as a more expressive yet still interpretable linear classifier; a two-dimensional principal component analysis (PCA) projection to visualize structure in the feature space; and a feedforward neural network to capture potential

non-linear interactions between variables. Our goal is to examine how these models perform on the same prediction task and how well they separate depressed from non-depressed respondents.

2 Related Work

Several recent studies have used machine learning models to predict depression from large health datasets. Li et al. [3] built logistic regression, LASSO, and random forest models using National Health and Nutrition Examination Survey (NHANES) data for adults with obstructive sleep apnea, and showed that these models can achieve strong discrimination while also highlighting clinical risk factors associated with depression. Their work illustrates how standard tabular features on demographics, health status, and lifestyle can support accurate prediction of depression risk.

Separately, Gonzales et al. [4] review how synthetic data are being used in health care for tasks such as simulation studies, algorithm development, and public data release. They argue that synthetic data can make it easier to share health-related datasets while still protecting privacy, but also note that synthetic datasets may not perfectly replicate the statistical properties of the original data. Our project is aligned with this perspective: we treat the synthetic mental health survey as a safe environment for comparing modeling approaches, rather than as a source for building a deployable clinical tool.

3 Methods

We use the synthetic depression dataset described in the introduction. The main file contains an ID, a `Name` column, a set of feature columns, and a binary target `Depression`. For all models, we drop the ID and name fields and treat `Depression` as the label $y \in \{0, 1\}$; the remaining columns form the feature matrix X.

We group features into numerical variables (e.g., age, academic and work pressure, CGPA, study and job satisfaction, work/study hours, financial stress) and categorical variables (e.g., gender, student vs. working professional, city, degree, profession, sleep duration, dietary habits, family history of mental illness, and suicidal-thoughts responses).

For the scikit-learn models (Naive Bayes, logistic regression, and the PCA-based classifier), we use a shared preprocessing pipeline: numerical features are imputed with the median and standardized; categorical features are imputed with the most frequent category. For Gaussian Naive Bayes, categorical variables are then mapped to integer codes via an ordinal encoder. For logistic regression and the PCA-based classifier, categorical variables are instead one-hot encoded so that each category becomes a binary indicator.

For the neural network, we use a slightly more customized preprocessor. Numerical features are standardized as before. Categorical features are split into two groups: some (such as gender, student vs. working professional, city, sleep duration, and family history of mental illness) are one-hot encoded; others (such as degree, profession, dietary habits, and suicidal-thoughts responses) are converted to integer labels. Missing categorical values are filled with a placeholder token before encoding, and missing numerical values are filled with a constant before scaling. The result is a fully numeric design matrix that we convert to tensors for training.

We consider four models that cover different parts of the course: Gaussian Naive Bayes, logistic regression, principal component analysis (PCA) plus a simple classifier, and a feedforward neural network.

3.1 Gaussian Naive Bayes

Gaussian Naive Bayes is used as a simple baseline. Given a feature vector $x = (x_1, \ldots, x_d)$ and label $y \in \{0, 1\}$, the model assumes conditional independence of features given the label and factorizes

$$P(y \mid x) \propto P(y) \prod_{j=1}^{d} P(x_j \mid y). \tag{1}$$

Each likelihood term is modeled as a Gaussian with class-specific mean and variance estimated from the training data. After preprocessing, we fit a Gaussian Naive Bayes classifier and treat its performance as a lower bound that more flexible models should improve on.

3.2 Logistic Regression with one-hot encoding

Our main linear model is logistic regression on the one-hot encoded representation. After preprocessing, each respondent is represented by a vector $z \in \mathbb{R}^p$. Logistic regression models the probability of depression as

$$P_\theta(y = 1 \mid z) = \sigma(w^\top z + b), \quad \sigma(t) = \frac{1}{1 + e^{-t}}, \tag{2}$$

where w and b are learned parameters. We fit the model by minimizing the regularized negative log-likelihood with an L_2 penalty on w. This model gives a strong, relatively interpretable baseline: the coefficients indicate how different features and categories are associated with the depression label.

3.3 Principal Component Analysis (PCA)

To explore structure in the feature space, we apply unsupervised principal component analysis (PCA) to the standardized numerical features. Let $X \in \mathbb{R}^{n \times d}$ denote the numeric data. PCA finds orthogonal directions that maximize projected variance; we keep the first two principal components and compute scores $Z = XW \in \mathbb{R}^{n \times 2}$, where W contains the top two eigenvectors. We mainly use this 2D representation for visualization, plotting respondents in the $(\text{PC1}, \text{PC2})$ plane and coloring points by depression status to see how much class separation appears in a low-dimensional linear embedding. As a simple extension, we also fit logistic regression on these two components to see how much predictive signal survives under this aggressive dimensionality reduction.

3.4 Neural Network Classifier

Our most flexible model is a feedforward neural network implemented in PyTorch. After preprocessing, each input is a vector $x \in \mathbb{R}^p$. The network has three hidden layers with ReLU activations, batch normalization, and dropout, followed by a single output neuron. Concretely, we use layer sizes $p \to 128 \to 64 \to 32 \to 1$; the final scalar logit $\ell(x)$ is turned into a predicted probability $\hat{y} = \sigma(\ell(x))$. We train the network with mini-batch gradient descent using the Adam optimizer (learning rate on the order of 10^{-3} with small weight decay) and the binary cross-entropy loss with logits. We train for a fixed number of epochs with a moderate batch size and monitor the training loss to make sure learning is stable.

3.5 Evaluation

For the classical models (Naive Bayes, logistic regression, and the PCA-based version), we use stratified k-fold cross-validation with $k = 5$, which preserves the class balance in each fold. We report the mean and standard deviation of accuracy and ROC AUC across folds. We also create an 80/20 stratified train–validation split and fit each classical model on the training portion, then compute validation accuracy, ROC AUC, and a standard classification report on the held-out portion.

For the neural network, we train on the full preprocessed training set and report training accuracy and ROC AUC on that set, and we examine the distribution of predicted probabilities on the unlabeled test data (for example, by grouping outputs into low, medium, and high-confidence ranges).

4 Results

All four models were tested on the same prediction task, and their performance shows the differences in how simple or complex each model is. As expected, Gaussian Naive Bayes served as a lower baseline, performing noticeably worse than the other models. Logistic regression improved on this by taking advantage of the more detailed one-hot encoded features, which allowed it to capture relationships between demographic, academic, and lifestyle variables.

The PCA-based classifier performed surprisingly well given that it relied on just two principal components. With only this 2D projection, the model reached a cross-validation accuracy of 0.9901 and a CV ROC AUC of 0.9993. On the held-out validation set, it achieved an accuracy of 0.9903 and a ROC AUC of 0.9992. The classification report shows that the model handled both classes consistently: the non-depressed class had a recall of 1.00, and the depressed class achieved a recall of 0.97, leading to an overall F1-score of 0.99.

The feedforward neural network also learned meaningful structure in the data. Training loss steadily decreased from around 0.19 to about 0.15 over 100 epochs, indicating stable optimization. On the training set, the network reached an accuracy of 0.9399 and a ROC AUC of 0.9763. The distribution of predicted probabilities showed clear separation between classes, with most outputs pushed close to either 0 or 1. Visual inspection of the first 100 predictions showed that the model generally matched the true labels.

When the neural network was applied to the full test set of 93,800 examples, the mean predicted depression probability was 0.2220. Out of all cases, 19,212 (20.48%) were labeled as positive. A confidence breakdown showed that 69,898 predictions fell into the low-confidence range, 8,536 into medium confidence, and 15,366 into the high-confidence range. High-confidence cases tended to include many of the model's strongest positive predictions, while low-confidence outputs were mostly associated with non-depressed cases.

Overall, the results show that both the linear methods and the neural network were able to extract substantial signals from the dataset. The PCA-based classifier, despite using only two components, performed unexpectedly well, while the neural network gave more detailed probability scores and helped show how confident the model was in its predictions.

5 Conclusion

In this project, we compared several machine learning models to see how well they could predict depression using a large survey dataset. Each model brought something different to the table. Gaussian Naive Bayes gave us a simple baseline, while logistic regression performed much better thanks to the more detailed one-hot encoded features. The PCA-based classifier stood out because it worked extremely well even after reducing the data to only two components. The neural network did not outperform the linear models in accuracy, but it did give more detailed probability scores and made it easier to see how confident the model was in each prediction.

Overall, the results show that depression risk in this dataset can be predicted very accurately, even with fairly simple models. The strong performance of logistic regression and the PCA-based classifier suggests that a lot of the useful information in the dataset follows patterns that are close to linear. The neural network, while not the top performer, still provided helpful insight into how certain the model was about each prediction.

There are some limitations to keep in mind. The dataset is synthetic, so it does not perfectly reflect real-world mental health data. Because of that, these results might not transfer directly to actual clinical settings or real survey responses. In the future, it would be useful to test these models on real depression screening data, try more complex architectures, or look more closely at which features matter most when predicting depression.

In the end, the study shows that machine learning can pick up strong patterns in mental health survey data and that different modeling choices can impact both accuracy and how easy the results are to interpret.

References

[1] Fernandes, M. D. S. V., Mendonça, C. R., da Silva, T. M. V., Noll, P. R. E. S., de Abreu, L. C., & Noll, M. (2023). Relationship between depression and quality of life among students: a systematic review and meta-analysis. *Scientific Reports*, 13(1), 6715. `https://doi.org/10.1038/s41598-023-33584-3`

[2] Zimmerman, M. (2024). The value and limitations of self-administered questionnaires in clinical practice and epidemiological studies. *World Psychiatry*, 23(2), 210–212. `https://doi.org/10.1002/wps.21191`

[3] Li, E., Ai, F., & Liang, C. (2024). A machine learning model to predict the risk of depression in US adults with obstructive sleep apnea hypopnea syndrome: A cross-sectional study. *Frontiers in Public Health, 11*, 1348803. https://doi.org/10.3389/fpubh.2023.1348803

[4] Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digital Health, 2*(1), e0000082. https://doi.org/10.1371/journal.pdig.0000082