Auditing Depression Prediction Automated Decision System

Final Project Report Submitted
in Partial Fulfillment of the Requirements
for the course

DS-GA 1017 Responsible Data Science

by

Judy Yang
hy1331@nyu.edu

Era Sarda
es6790@nyu.edu

under the guidance of

Instructors
Prof. Emily Black and Lucas Rosenblatt

to the

CENTER FOR DATA SCIENCE
NEW YORK UNIVERSITY
NEW YORK - 10012, NY

Abstract

This report presents a fairness audit of a high-performing ensemble machine learning model developed for depression prediction, originally designed for a Kaggle competition using synthetic mental health data. While the model achieves high predictive accuracy by combining AdaBoost, CatBoost, XGBoost, Logistic Regression, and LGBMClassifier, our audit reveals critical disparities in subgroup performance. Through detailed analysis of feature correlations, missing data patterns, and regional representation, we identify structural and cultural biases in the dataset, particularly affecting employment status and age subgroups. Stakeholder perspectives (patients, families, professionals, and institutions) are also mapped to understand real-world implications. Fairness evaluations using Fairlearn's MetricFrame and ThresholdOptimizer expose significant performance gaps: students experience high false positive rates, while working professionals face high false negative rates. Post-processing mitigation with equalized odds constraints reduces these disparities, though at a cost to overall accuracy. LIME is further applied to interpret individual predictions, uncovering group-specific reasoning patterns. This audit highlights the need for fairness-aware evaluation, culturally grounded feature design, and subgroup-specific calibration when deploying automated decision systems in high-stakes domains like mental health.

Contents

Background

Depression and other mental illnesses have long been known to exhibit significant gender disparities in their reported prevalence (Astbury, 1999). In the case of depression — including both depressive symptoms and Major Depressive Disorder (MDD) — research in the United States has shown that "females experience 1.5- to 3-fold higher rates than males beginning in early adolescence" (American Psychiatric Association, 2013, p. 165). However, studies have also found that this gender gap is considerably smaller in many developing countries (Sartorius et al., 1983). This discrepancy raises the question of whether systematic biases—such as those introduced through data collection methods (e.g., census formats) or through Western gender norms and stereotypes—may be contributing to the observed differences in depression prevalence.

Despite this possibility, the gender disparity in depression has often been accepted as fact within the field of clinical psychology and embedded into the "gold standard" of psychiatric diagnosis, the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) (Hunt et al., 2018). Given these potential underlying biases in how mental health data is framed and interpreted, it is reasonable to expect that machine learning (ML) models trained to predict depression may also exhibit gender-based biases.

In this project, we propose to audit a high-performing ML model developed by Mahdi Ravaghi (Ravaghi, 2024), which won first place in Season 4, Episode 11 of Kaggle's Playground Series Competition, titled *Exploring Mental Health Data* (Reade & Park, 2024). The competition tasked participants with predicting mental health outcomes using a structured dataset, and submissions were evaluated solely based on accuracy. Ravaghi's ensemble-based model achieved the highest accuracy among more than 2,500 submissions.

To evaluate both the performance and fairness of this model, we conducted a series of post hoc analyses including accuracy benchmarking, group-specific fairness metrics, and model interpretability techniques. We used Fairlearn's `MetricFrame` to assess disparities in predictive outcomes across sensitive groups, applied post-processing mitigation through threshold optimization to address identified inequities, and used LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) to interpret how feature contributions vary at the individual level. While gender was our initial focus as the sensitive attribute, exploratory analysis led us to shift attention to employment status — specifically, whether individuals were working professionals or students — as this distinction revealed more pronounced disparities in model performance. Together, these methods aim to reveal not only how well the model performs, but also whom it serves—and whom it may potentially overlook or misclassify.

About the Dataset

The dataset used in the Kaggle competition (Rao et al., 2024) includes a range of predictors relevant to mental health. For the purposes of our fairness audit, we consider `Gender` as the primary sensitive attribute. In addition to gender, features such as profession, degree, and dietary habits may act as proxy variables, potentially capturing implicit gender-related signals.

Beyond individual attributes, the model's performance may also be influenced by geographic and cultural biases embedded in the dataset. The data originates from the Indian subcontinent, but distribution plots [1, 2] suggest it is disproportionately sampled from major urban centers in North and Central-West India. Areas such as smaller cities, rural regions, and the South, Central-East, and North-East are notably underrepresented. This sampling imbalance raises concerns about whether the dataset reflects the full regional diversity of India, and whether the trained model can generalize well across culturally and geographically distinct populations.

Importantly, the dataset is **synthetically generated**. As a result, biases in regional distribution may stem from (i) the original data used for synthesis, (ii) the design of the generation algorithm, or (iii) deliberate sampling decisions. Yet, there appears to be no clear rationale for the selection of included cities. The sampled regions do not consistently correspond to city size, population density, state capitals, metropolitan zones, or any coherent regional grouping. This absence of justification limits confidence in the model's representativeness and generalizability.

Furthermore, the dataset omits important cultural and familial dimensions that are particularly salient in the Indian context. In many Asian societies, family structure, interdependence, and relationship quality play central roles in shaping psychological well-being. For instance, Lu and colleagues (Lu et al., 2010) demonstrate that even the conceptualization of depression differs significantly between Chinese and American populations, reinforcing the importance of culture in mental health modeling.

To that end, variables capturing intergenerational living, number of dependents, or relationship satisfaction would have been valuable additions to the dataset. Their inclusion could offer more culturally grounded insights into depression risk, especially for a demographically diverse population like India's. Without these features, the model may overlook key factors that contribute to mental health outcomes in the region.

See Appendix A for further details on data distribution and regional representation.

Implementation and Validation

This section presents our understanding of the implementation of the depression prediction model developed for the Kaggle Playground Series S4E11 competition. The code was written by Mahdi Ravaghi and publicly released as the winning solution. Our goal here is to demonstrate a high-level understanding of how the system operates, how the data was processed, and how the model was validated for its stated objective of maximizing predictive accuracy.

3.1 Data Cleaning and Pre-processing

The original pipeline involved substantial pre-processing to prepare the dataset for machine learning. Categorical features were cast as `category` dtype, with category alignment enforced across the training and test datasets. For certain models (e.g., AdaBoost and XGBoost), ordinal encoding was applied to categorical variables. Continuous features with missing values were imputed using mean imputation. The final feature matrix was standardized for models that were sensitive to feature scaling (e.g., Logistic Regression and Ridge Regression). Importantly, gender and other categorical fields were preserved as-is to allow downstream auditing of subgroup behavior. The data itself is synthetic, generated from an undisclosed private dataset, and does not include real personal identifiers.

3.2 High-level Implementation of the System

The system follows a stacked ensemble architecture. The first layer (L1) consists of five base models: AdaBoost, CatBoost, XGBoost, LightGBM, and Logistic Regression. Each of these models was trained independently and generated probability outputs on the validation set. These predictions were passed to a second-level ensemble (L2), where two models—Logistic Regression and Ridge Regression—were trained using the logit-transformed probabilities of the base models as inputs. These L2 models effectively learned how to weight the base model predictions. Finally, an L3 ensemble combined the outputs of the L2 models using a weighted average, with the optimal weights learned via hyperparameter tuning (using Optuna). This layered design was intended to maximize accuracy while leveraging the strengths of diverse classifiers.

The original implementation used stratified k-fold cross-validation to evaluate model performance and select hyperparameters. This approach is standard in ML competitions and helps ensure robust performance estimates by averaging accuracy across multiple validation splits.

3.3 Validation and Alignment with Stated Goals

The ADS was validated using stratified cross-validation, and the stated goal—maximizing accuracy—was achieved, with the final model reaching approximately 94–96% average accuracy across folds. However, since the competition's test set did not contain ground-truth labels, external validation and fairness analysis were not originally conducted. To audit the model for fairness, we created a fixed train-validation split from the training data. This enabled the use of fairness metrics such as selection rate, true/false positive rates, and enabled consistent subgroup analysis via Fairlearn's `MetricFrame`.

We also discovered that cross-validation introduces variability in subgroup representation across folds, complicating fairness analysis. A fixed validation set allowed for reproducible and interpretable group-level metrics. Figure 3.1 shows model performance under cross-validation, and Figure 3.2 shows performance under our fixed split.

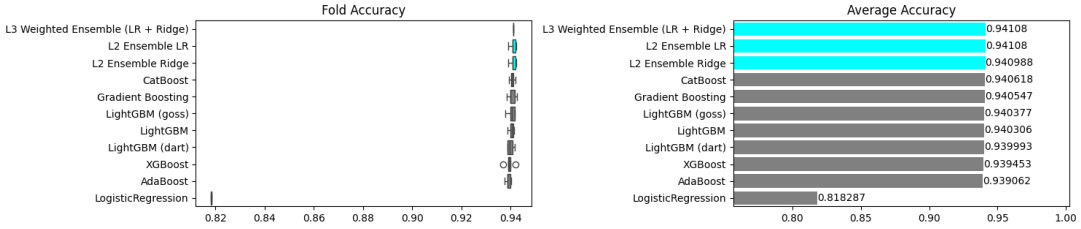


Fig. 3.1: Cross-validated accuracy comparison of different models under the original ensemble pipeline. The L3 weighted ensemble achieves the highest mean accuracy.

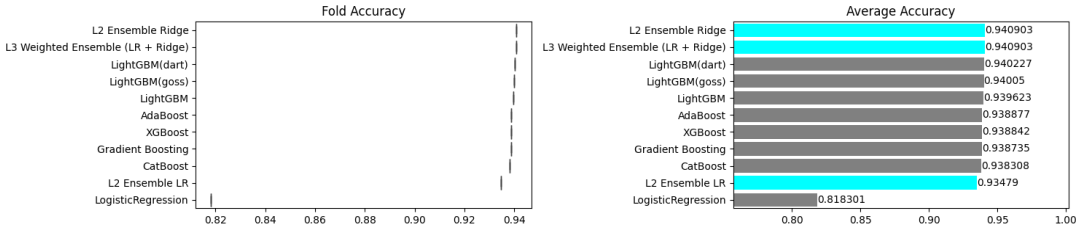


Fig. 3.2: Accuracy comparison of models using a fixed train-validation split. Overall rankings remain consistent, but this approach enables fair and consistent subgroup evaluation.

That said, our fixed-split strategy may have introduced information leakage or artifacts, as the data is synthetic and its generative process is not fully disclosed. Future work could involve generating privacy-preserving synthetic datasets using differentially private mechanisms or correlation-aware generative models.
In summary, the implementation was accurate and sophisticated in terms of ensemble architecture and optimization, but fairness and robustness to distributional shift were not considered in the original validation pipeline. Our re-validation using a fixed split enabled a more thorough fairness audit aligned with the needs of a real-world ADS.

Audit Outcomes

4.1 Experiments

Initial experiments showed that dropping the `Name` and `City` features from the dataset had no significant impact on the model's overall accuracy as well as other metrics like FNR/FPR and fairness metrics as can be seen in Figure 4.1. This suggests that these features did not carry meaningful predictive information and that the model is not overly reliant on potentially privacy-sensitive attributes. Other relevant results w.r.t. overall and genderwise FNR/FPR/Selection Rate Difference are included in Appendix A.2.1 (Figures A.13, A.14).

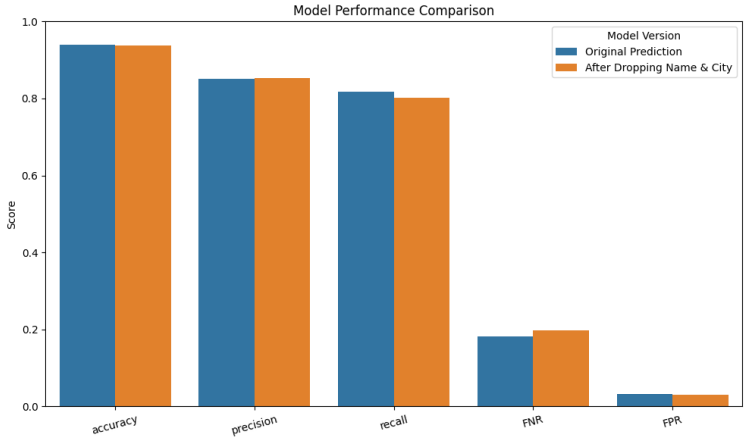


Fig. 4.1: Results before and after dropping sensitive features

An interesting observation emerged from examining how the two sub-models : Gradient Boosting (GB) and AdaBoost (ADB) handled missing data. These models imputed missing values in numerical columns such as `Academic Pressure`, `Work Pressure`, `CGPA`, `Study Satisfaction`, and `Job Satisfaction` using mean imputation. While this method is common, it can be problematic in our context. For example, imputing the average academic pressure of students for a working professional lacks contextual validity, and vice versa. Yet, when the ensemble model was retrained and evaluated separately for students and working professionals (after dropping the irrelevant features for each group), the resulting performance metrics were close to the original scores (Figure 4.2). This implies a degree of robustness in the overall ensemble model architecture, likely due to the averaging effects of combining multiple learners (Black et al., 2021). However, another observation to note here was that the outputs of weighted probabilities of the ensemble model go out of the range of [0,1] for various datapoints of working professional group specifically. The unbounded outputs are due to the application of Ridge and Logistic regression on base models' prob-

abilites, suggesting the need for post-hoc calibration (See figure A.15 in Appendix A.3). Although such changes are beyond the scope of this audit, we recommend applying sigmoid transformation or calibration techniques like Platt scaling to ensure the outputs conform to probabilistic interpretations.

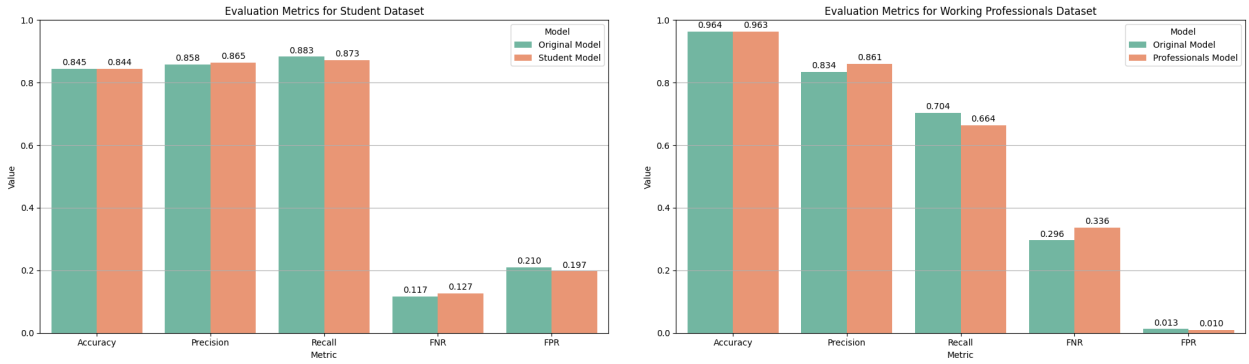


Fig. 4.2: Metrics Comparison before and after dropping non-relevant features for Students and Working Professionals

4.2 Model Fairness

To evaluate the fairness of our model, we passed its predictions into `Fairlearn`'s `MetricFrame`, using `Gender` as the sensitive feature. This allowed us to measure not only overall accuracy but also group-specific metrics such as selection rate, false positive rate (FPR), and true positive rate (TPR). These metrics were chosen because they reflect how differently the model behaves across demographic groups in ways that affect real-world outcomes. For example, TPR captures whether the model consistently identifies depression when it is truly present, while FPR reflects the cost of wrongly flagging individuals as depressed. Selection rate offers a lens into how often each group is predicted as positive, regardless of correctness—highlighting disparities in model attention or prioritization. These group-wise metrics provide a more granular understanding of equity than accuracy alone, especially in high-stakes contexts like mental health screening. Contrary to our initial expectations, the model exhibited minimal disparity across gender groups.

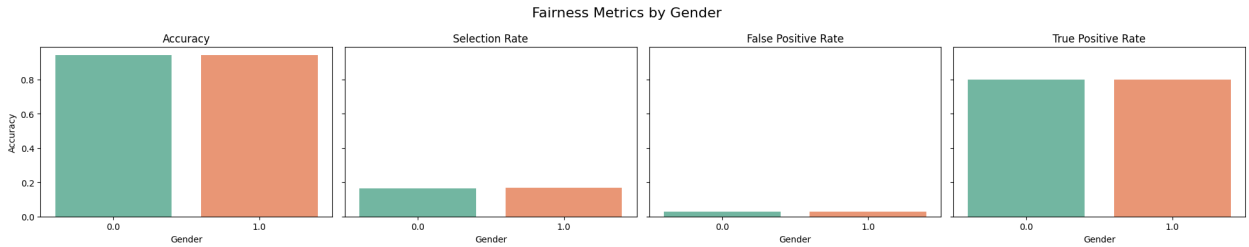


Fig. 4.3: Fairness metrics of the model, with Gender as the sensitive feature.

As shown in Figure 4.3, both male and female groups achieve comparable accuracy, selection rate, and error rates. These results suggest that the model is *not* exhibiting strong gender-based bias in its predictions, despite broader societal patterns that might suggest otherwise.

To further validate this finding, we examined fairness across subgroups defined by both `Gender` and employment status (student or working professional). Once again, we observed no meaningful disparities, as illustrated in Figure 4.4.

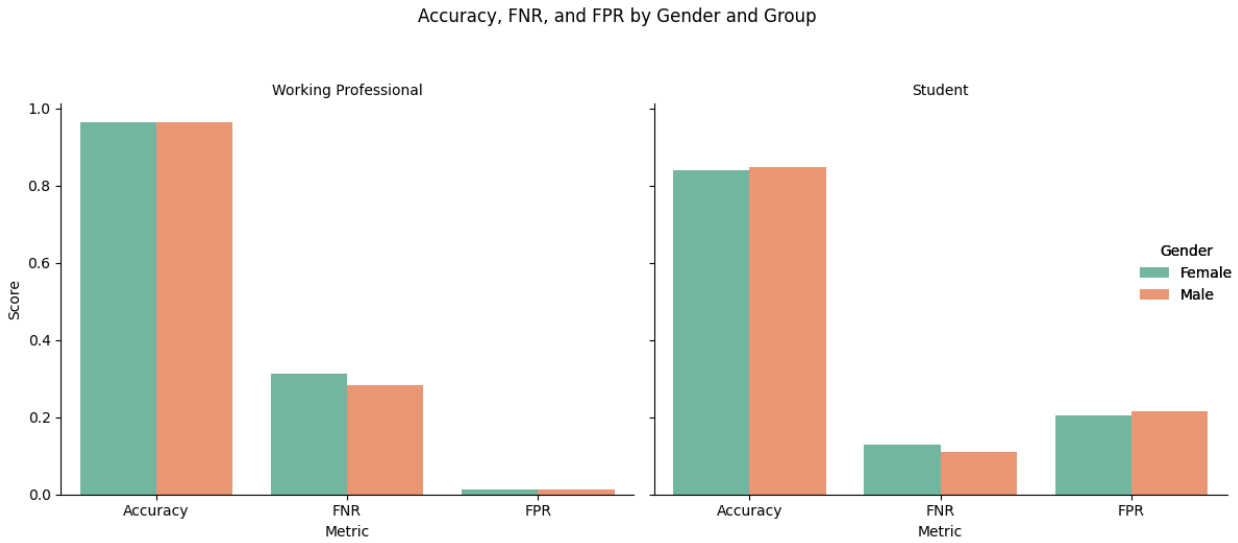


Fig. 4.4: Fairness metrics of the model across employment subgroups, with Gender as the sensitive feature.

Given the lack of observed gender disparity, we decided to explore alternative sensitive features. Our choice to use `Working Professional or Student` status as the next axis of analysis was guided by our initial exploratory data analysis. There, we uncovered a strong relationship between age and depression, with younger individuals — particularly students — exhibiting significantly higher depression rates compared to older working professionals. This demographic divide suggests a plausible source of structural disparity that warrants deeper fairness evaluation. By shifting the sensitive feature to employment status, we aim to capture this potentially meaningful divide and investigate whether the model is disproportionately disadvantaging one group over the other.

We again used `Fairlearn`'s `MetricFrame` to compute group-wise metrics, this time stratified by employment status. As shown in Figure 4.5, the results suggest a measurable disparity. While working professionals (group label 1) achieve higher overall accuracy, their selection rate is substantially lower. In contrast, students (group label 0) are selected by the model at a much higher rate but also experience a higher false positive rate. Notably, their true positive rate is also significantly higher, which could suggest overprediction of the positive class for this group.

These results suggest that, unlike with gender, the model's predictions vary meaningfully

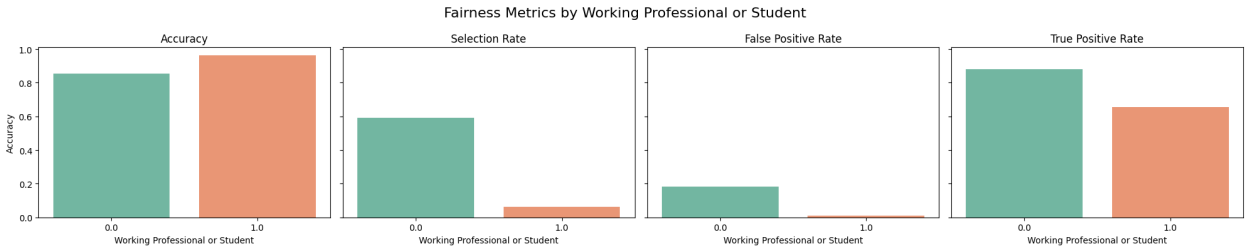


Fig. 4.5: Fairness metrics of the model, with Working Professional or Student as the sensitive feature.

across employment status. This raises potential fairness concerns, particularly in how the model may be over-identifying depression among students while under-identifying it among working professionals. Given the real-world implications of such predictions in mental health interventions, these disparities warrant further investigation and potential mitigation.

To further explore the disparities observed between students and working professionals, we computed and visualized separate confusion matrices for each group. These plots, shown in Figure 4.6, provide a more detailed look at how the model behaves in terms of false positives, false negatives, and overall predictive performance for each subgroup.

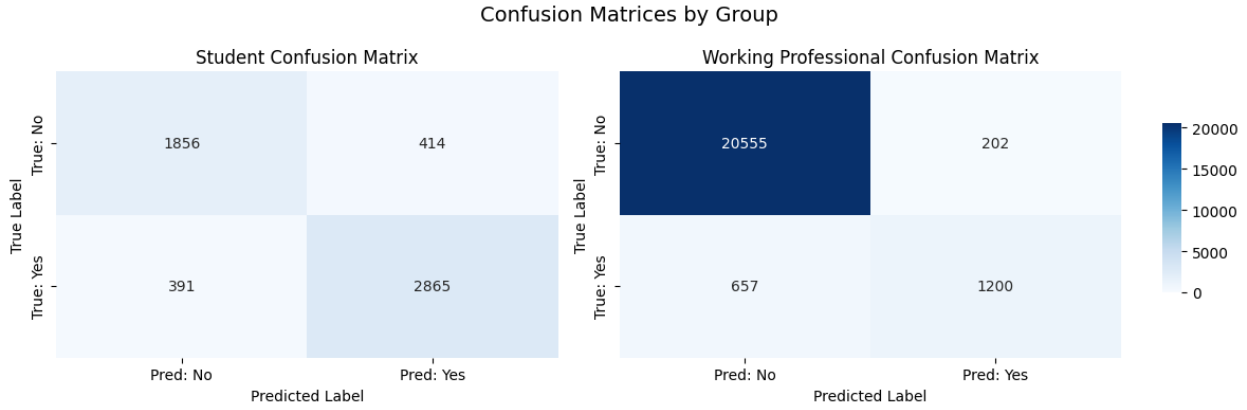


Fig. 4.6: Group-specific confusion matrices with Working Professional or Student as the sensitive feature.

Specifically, working professionals not only had a notably higher accuracy ($\sim$ 96%) than students ($\sim$ 84%), but also higher FNR ($\sim$ 30%), indicating that depressive cases within this group were more likely to go undetected. Additionally, students experienced higher FPR ($\sim$ 21%), meaning non-depressive individuals within this subgroup were more likely to be misclassified as depressive.

In the student group (left panel), we observe a relatively balanced distribution of false positives and false negatives. While the model correctly identifies a large number of depressed students (2,865 true positives), it also generates a considerable number of false positives (414) and false negatives (391). This indicates that the model is relatively sensitive to depression

in students, but it also errs in both directions.

By contrast, the confusion matrix for working professionals (right panel) reveals a different pattern. Here, the model shows a strong bias toward predicting the negative class: it identifies a large number of true negatives (20,555) and makes very few false positives (202). However, this comes at the cost of a much higher number of false negatives (657) and fewer true positives (1,200) compared to the student group.

Taken together, these results suggest that the model is more conservative in predicting depression among working professionals. It may be under-identifying true cases in this group, potentially due to differences in underlying feature distributions or label prevalence between students and working professionals. This finding complements the fairness metrics shown earlier and highlights a potential fairness concern: while the model performs well on average, it may systematically fail to detect depression in one group more than the other. This warrants further investigation, especially in high-stakes applications where underdiagnosis can have serious consequences.

To quantify the observed disparities between students and working professionals, we computed group-level disparity scores using a series of metrics provided by the `fairlearn.metrics` module. These metrics capture the absolute difference in model behavior across the two groups, using `Working Professional or Student` as the sensitive feature. The results are presented in Table 4.1.

Table 4.1: Group-wise performance metrics for the L3 ensemble model using `Working Professional or Student` as the sensitive feature.

	Accuracy	Selection Rate	True Positive Rate	False Positive Rate
Student (0)	0.854300	0.603900	0.888800	0.195200
Working Professional (1)	0.962000	0.062800	0.651100	0.010200

As shown in the table, the most pronounced disparity appears in the selection rate, with a difference of 0.5314 between groups. This indicates that one group is far more likely to be predicted as positive than the other. Substantial differences are also observed in the true positive rate (0.2337) and false positive rate (0.1726), suggesting that the model's sensitivity and error patterns vary meaningfully across groups. Even the accuracy difference, which is typically more stable across subpopulations, exceeds 10%, underscoring that the model's predictive performance is not uniform between students and working professionals.

Together, these metrics provide strong evidence of group-level disparities in how the model assigns predictions. The magnitude of these differences suggests that additional investigation and potentially mitigation are warranted, especially in applications where equity in prediction outcomes is critical.

Again to emphasize, when model was retrained and validated separately for professionals and students (with dropping non-relevant features), these high values of FNR and FPR respec-

tively were maintained (4.2). This suggests that the *data* and not just the model architecture drives this disparity. This may be due to the reasons like: (i) Student responses may be more inconsistent or self-reported (and for professionals they might come from clinical diagnoses), leading to noisy ground truth, making the model learn unreliable patterns, causing unstable FPR/FNR tradeoff. Figure 4.7 indicates that since ground truth depression for students is much higher than `Not Depressed`, and even 7-fold higher than Depressed professionals, the model might have overfit, and led to poor generalization.

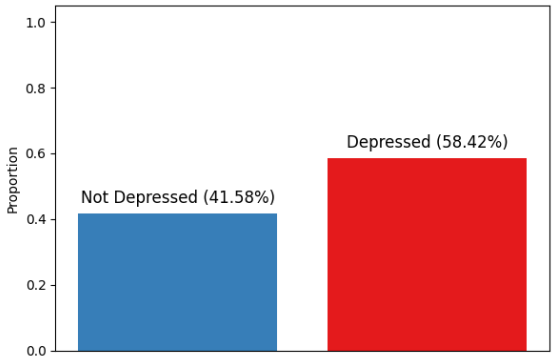


Fig. 4.7: True Counts in Training Data (Students)

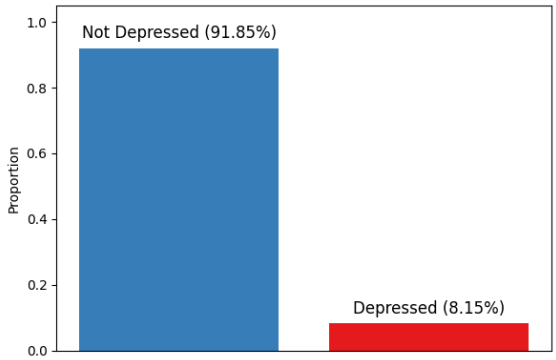


Fig. 4.8: True Counts in Training Data (Working Professionals)

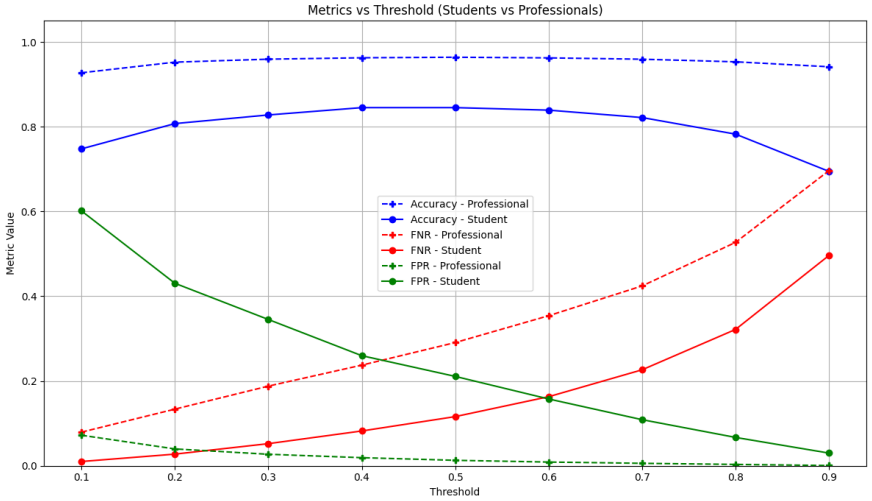


Fig. 4.9: How metrics change with Classification Threshold

(ii) Additionally, the decision threshold (probability less than 0.5 is classified 0, and more than 0.5 is 1) is neither optimized nor group-specific. The same threshold might cause under-detection in professionals (hence high false negative rate) and over-detection in students (high false positive rate). This suggests that overall and group-specific calibration might be necessary. Figure 4.9 clearly indicates the significant gap in accuracy, FNR, and FPR values between students (bold lines) and working professionals (dotted lines). Though the

threshold 0.5 optimizes the gap in accuracy, it doesn't balance the fairness needs of the model. However, an important observation is that even when threshold is modified, there is not much change in accuracy for both the groups, but there can be significant changes in fairness performances, if threshold is optimized. For example, as the threshold increases, false positive rate difference decreases with coming close to 0 at threshold 0.9.

In acccordance with above observations and to address the fairness disparities observed across employment groups, we applied `Fairlearn`'s `ThresholdOptimizer`, a post-processing mitigation technique designed to enforce fairness constraints without retraining the underlying model. In our case, we selected the `equalized_odds` constraint, which seeks to equalize both true positive and false positive rates across groups. This choice was motivated by our prior audit, which revealed substantial disparities in sensitivity and error rates between students and working professionals.

Table 4.2 reports the group-wise performance metrics after threshold optimization. While overall accuracy decreases for the student group, the true positive rates for both groups become closely aligned (0.5992 for students vs. 0.5918 for working professionals), and false positive rates converge as well (0.0370 vs. 0.0385). These results suggest that threshold optimization effectively reduces disparity in model behavior across groups, trading off a modest reduction in predictive performance for improved fairness in classification outcomes.

Table 4.2: Post-mitigation fairness metrics for the L3 ensemble using `Working Professional or Student` as the sensitive feature, optimized for equalized odds.

	Accuracy	Selection Rate	True Positive Rate	False Positive Rate
Student (0)	0.748600	0.368300	0.599200	0.037000
Working Professional (1)	0.931100	0.084000	0.591800	0.038500

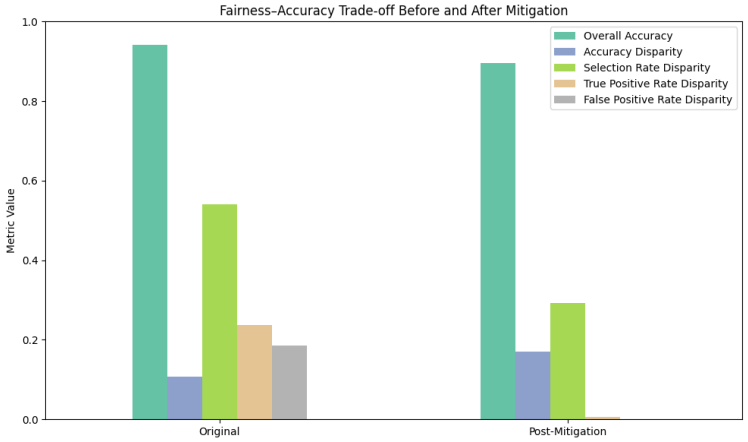


Fig. 4.10: Fairness–accuracy trade-off before and after applying threshold optimization for equalized odds.

To visualize this trade-off, we compare the original ensemble model and the mitigated version across key fairness and accuracy metrics. As shown in Figure 4.10, threshold optimization leads to a small reduction in overall accuracy (from 94% to 90%) but results in substantial improvements in fairness: disparities in true and false positive rates nearly vanish, and selection rate disparity decreases by over 40 percentage points. These shifts illustrate the practical cost of enforcing group fairness and highlight the importance of balancing equity and performance in sensitive applications like mental health screening.

4.3 Prediction Interpretation

To better understand how our ensemble model makes predictions—and whether those decisions align with human expectations—we applied LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016). LIME explains individual predictions by locally approximating the model with an interpretable surrogate, such as a linear model. However, our final L3 ensemble combines two second-level learners: `l2-ensemble-lr` (logistic regression) and `l2-ensemble-ridge` (ridge regression), using a weighted average of their predicted probabilities. While logistic regression exposes the `predict_proba` interface needed for LIME, ridge regression does not natively provide probabilistic outputs. Consequently, for interpretability purposes, we restricted LIME analysis to the logistic regression component of the ensemble.

This choice introduces an important caveat: as shown in Figure 4.11, the final ensemble assigns only **1.4%** of the total weight to the logistic regression model, while the ridge regression component dominates with 98.6%. Therefore, while LIME offers useful intuition into local feature contributions under a linear model, the resulting explanations must be interpreted with caution. They reflect the behavior of a minor contributor to the ensemble, and thus do not fully represent the decision logic of the final model.

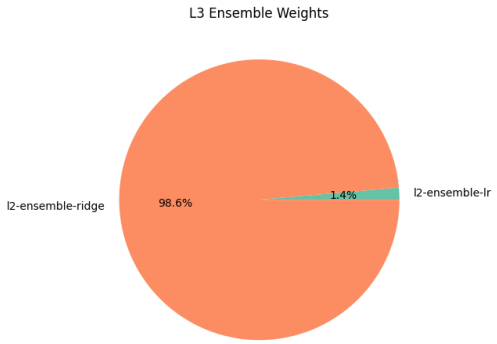


Fig. 4.11: L3 ensemble weights assigned to the second-level models. The final prediction is heavily dominated by `l2-ensemble-ridge`.

Despite this limitation, LIME still provides a transparent lens into how one interpretable component processes the data, and allows us to inspect patterns of misclassification and

borderline decisions in a structured way.

We used LIME to analyze three representative validation cases: a misclassified working professional, a misclassified student, and an uncertain but correct prediction. These examples were selected to probe potential group-specific failure modes and examine the model's behavior near the decision boundary.

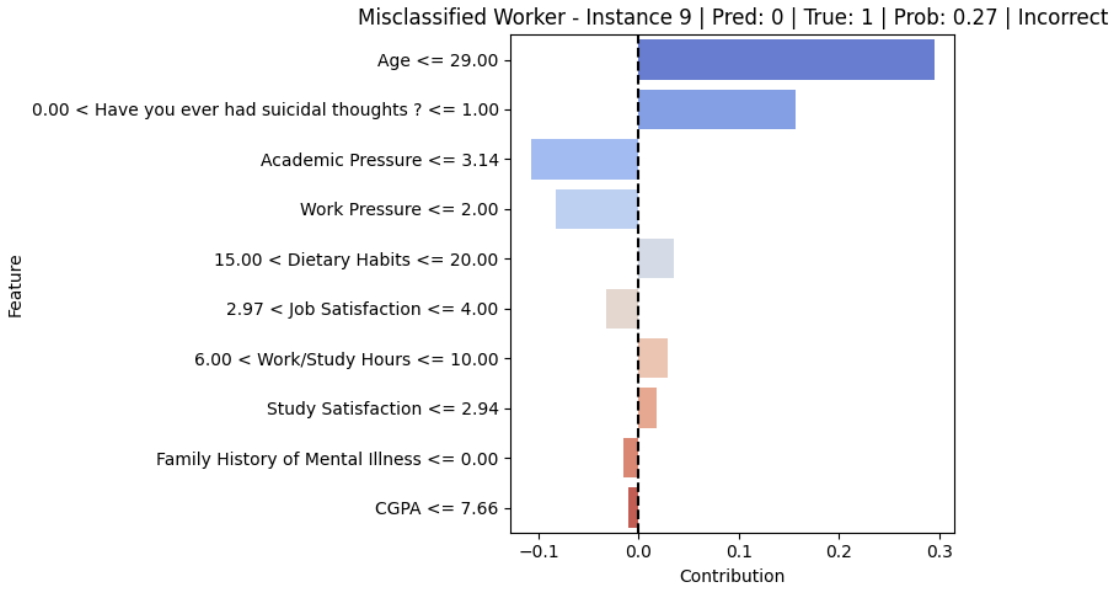


Fig. 4.12: LIME explanation for a misclassified working professional (true: depressed, predicted: not depressed). Blue bars indicate features contributing to the negative class ("No"), red bars indicate features contributing to the positive class ("Yes").

In the first example 4.12, a working professional was incorrectly classified as not depressed despite self-reporting suicidal ideation and experiencing work pressure. The predicted probability was just 27%. LIME revealed that older age and low work pressure were dominant negative contributors, outweighing more concerning signals. This case illustrates how the model may be biased toward under-identifying depression among working professionals when protective features are present, even alongside serious risk indicators.

In the second case 4.13, a student was incorrectly classified as depressed, with a predicted probability of 64%. While this individual showed low financial stress and no family history of mental illness, the model was strongly influenced by age, suicidal ideation, and academic pressure, factors it learned to associate with depression in younger populations. Although understandable, this case demonstrates how the model may over-identify depression among students, consistent with the earlier finding of selection rate disparity by employment group.

Lastly 4.14, we examined a borderline case with a correct prediction. Here, the model output a probability of 50% for depression, ultimately predicting the correct label. No single feature dominated the explanation; instead, moderate academic pressure and financial stress were

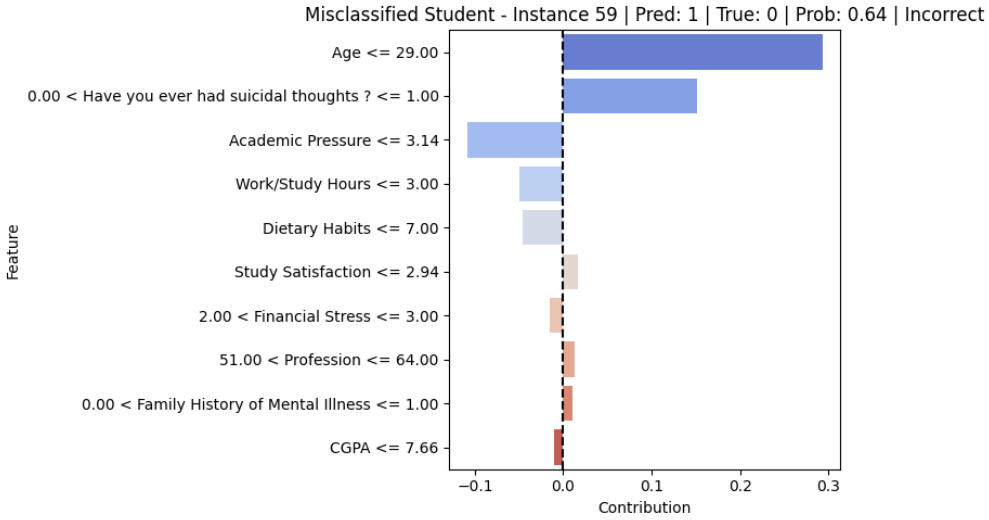


Fig. 4.13: LIME explanation for a misclassified student (true: not depressed, predicted: depressed).

offset by the absence of suicidal thoughts. This example highlights the model's capacity to handle ambiguity by balancing weak signals, an encouraging sign, though such behavior may also be fragile.

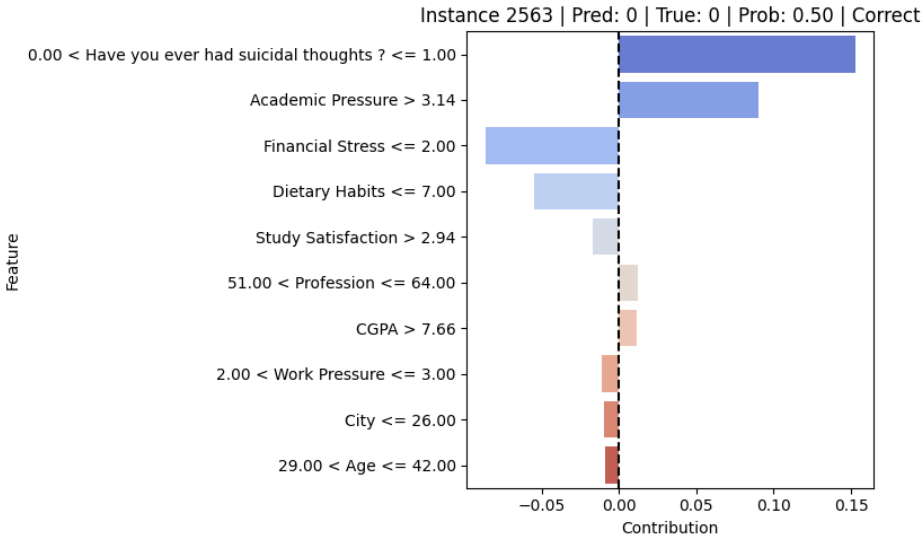


Fig. 4.14: LIME explanation for a correct but uncertain prediction (true: not depressed, predicted: not depressed, probability: 50%).

These local explanations lend transparency to model behavior, reinforce the fairness concerns raised earlier, and offer valuable guidance for future model debugging and refinement.

Conclusion

While the model audited in this project demonstrates strong predictive accuracy, we find that the underlying dataset poses significant challenges for responsible deployment in an automated decision system (ADS). The data is synthetically generated with unclear provenance and shows substantial geographic and demographic biases, particularly overrepresentation of urban centers in North and Central-West India and underrepresentation of rural, southern, and northeastern regions. Moreover, culturally relevant variables such as family structure, interpersonal relationships, and stigma-related experiences, which are central to understanding depression in the Indian context, are missing. As a result, while the dataset is useful for experimentation and benchmarking, it is not appropriate for real-world, fairness-sensitive applications without considerable expansion and refinement.

The implementation itself is robust in terms of accuracy, with the final ensemble achieving over 94% on held-out validation data. However, fairness evaluations tell a different story. While gender-based disparities were minimal, significant disparities emerged along employment status: students were over-identified as depressed (high false positive rate), while working professionals were often missed (high false negative rate). These findings were confirmed through `MetricFrame` audits, confusion matrices, and local explanations using `LIME`. To address these disparities, we applied `Fairlearn`'s `ThresholdOptimizer` with an equalized odds constraint. This post-processing method successfully reduced group-level disparities in true and false positive rates, at the cost of a modest drop in accuracy ($\sim 4\%$). Given the sensitive nature of mental health prediction, we believe this is an acceptable trade-off. Stakeholders such as clinicians, patients, and policy makers would likely find fairness metrics, particularly false negative and false positive rates, more meaningful than accuracy alone, as these capture the real-world harms of misclassification.

Despite improvements, we would not recommend deploying this model in the public sector or industry in its current form. The lack of transparency in data sourcing, insufficient cultural representation, and observed subgroup disparities make it unsuitable for high-stakes decision-making without further development. Future iterations should incorporate richer, more representative data; include culturally and socially relevant features; apply fairness interventions during model training rather than solely post hoc; and calibrate prediction outputs to support informed threshold selection. Only with these enhancements can such a model be responsibly used to support equitable and trustworthy mental health screening at scale.

Acknowledgements

We would like to acknowledge our teaching assistant, Manasvin Anand, for his valuable guidance, prompt feedback, and continuous support throughout this audit project. His insights were instrumental in helping us navigate both technical challenges and conceptual clarity during our work.

Bibliography

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders, fifth edition (DSM-5)*. https://doi.org/10.1176/appi.books.9780890425596

Astbury, J. (1999). *Gender and mental health*. Citeseer.

Audit experiments. (2025). https://colab.research.google.com/drive/1kkDwgBXMm_YddiM6oinRitEad7rLdPg-?usp=sharing

Black, E., Leino, K., & Fredrikson, M. (2021). Selective ensembles for consistent predictions. *arXiv preprint arXiv:2111.08230*. https://doi.org/https://doi.org/10.48550/arXiv.2111.08230

Data distribution plots (regional). (2025). https://drive.google.com/drive/folders/1JqDUR4_2FxzjZQRurSWNlxaeilamtMi2?usp=sharing

Hunt, J. C., Chesney, S. A., Jorgensen, T. D., Schumann, N. R., & deRoon-Cassini, T. A. (2018). Exploring the Gold-Standard: Evidence for a Two-Factor Model of the Clinician Administered PTSD Scale for the DSM-5. *Psychological trauma : theory, research, practice and policy*, *10*(5), 551–558. https://doi.org/10.1037/tra0000310

Input-output on data. (2025). https://colab.research.google.com/drive/1QlzYtKmta2uCyVRcRLXgUYePT
usp=sharing

Lu, A., Bond, M. H., Friedman, M., & Chan, C. (2010). Understanding cultural influences on depression by analyzing a measure of its constituent symptoms. *International Journal of Psychological Studies*, *2*(1). https://doi.org/10.5539/ijps.v2n1p55

M., K., Balaji, E., & I., S. H. (2024). Work- life balance and factors impacting work performance among women in bengaluru's it sector. https://doi.org/10.9734/ajeba/2024/v24i111555

Mirowsky, J., & Ross, C. E. (1992). Age and Depression. *Journal of Health and Social Behavior*, *33*(3), 187. https://doi.org/10.2307/2137349

P, R. A., & Perwez, S. K. (2023). An empirical analysis of work-life balance on work from home during covid-19 pandemic: A comparative study on men and women. https://doi.org/10.2174/0118743501275173231023102400

Rao, R., Gujral, D., & Pathak, S. (2024). Depression Survey/Dataset for Analysis.

Ravaghi, M. (2024, December). S04E11 — Mental Health Prediction — Ensemble.

Reade, W., & Park, E. (2024). Exploring mental health data.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Sartorius, N., et al. (1983). *Depressive disorders in different cultures: Report on the WHO collaborative study on standardized assessment of depressive disorders*. World Health Organization.

Talamala, S. (2024). Locus of control and job satisfaction: A comparative study between male and female employees in it sector bangalore. https://doi.org/10.61707/94ysc281

Data Distribution

(Referring to the ("Input-Output on Data", 2025))

The data is synthetically generated by a deep learning model. The data is distributed similarly to the original data that the model was trained on (Ravaghi, 2024).

Col#	Column name	datatype	Feature values
0	id	int64	[0, 1, 2, 3, 4, 5, ...
1	Name	object	[Aaradhya, Vivan, Yuvraj, ...
2	Gender	object	[Female, Male]
3	Age	float64	[49.0, 26.0, 33.0, ...
4	City	object	[Ludhiana, Varanasi, Visakhapatnam, ...
5	Working Professional or Student	object	[Working Professional, Student]
6	Profession	object	[Chef, Teacher, nan, Business Analyst, ...
7	Academic Pressure	float64	[5.0, 4.0, 3.0, 2.0, 1.0]
8	Work Pressure	float64	[5.0, 4.0, 3.0, 2.0, 1.0]
9	CGPA	float64	[0-10]
10	Study Satisfaction	float64	[5.0, 4.0, 3.0, 2.0, 1.0]
11	Job Satisfaction	float64	[5.0, 4.0, 3.0, 2.0, 1.0]
12	Sleep Duration	object	[More than 8 hours, Less than 5 hours, 5-6 hours, ...
13	Dietary Habits	object	[Healthy, Unhealthy, Moderate, ...
14	Degree	object	[BHM, LLB, B.Pharm, BBA,...
15	Have you ever had suicidal thoughts ?	object	[No, Yes]
16	Work/Study Hours	float64	[1.0, 2.0, ..., 11.0]
17	Financial Stress	float64	[5.0, 4.0, 3.0, 2.0, 1.0]
18	Family History of Mental Illness	object	[No, Yes]
19	(Target Value) Depression	int64	[0, 1]

A.1 Null Values

The dataset contains significant missing values for several attributes, including 'Profession', 'Academic Pressure', 'Work Pressure', 'CGPA', 'Study Satisfaction', 'Job Satisfaction', 'Dietary Habits', 'Degree', and 'Financial Stress'.

The missing values can be understood as follows. Students often leave fields such as 'Profession', 'Work Pressure', and 'Job Satisfaction' blank because these attributes are not relevant to their situation, or, in some cases, individuals may not feel comfortable sharing details about their jobs. On the other hand, working professionals tend to omit values for 'Academic Pressure', 'CGPA', and 'Study Satisfaction', as these fields are similarly irrelevant to them.

Other attributes might simply be missing due to the synthetic data generation process.

Else, in the case of missing values for 'Degree' (two rows), both individuals were identified as working professionals who also chose not to specify their professions. This may suggest discomfort in disclosing qualifications.

The missing values for 'Dietary Habits' and 'Financial Stress' (four rows each), however, remain unexplained. These omissions do not appear to follow a clear pattern.

A.2 Data Analysis

As part of the exploratory data analysis (EDA), we plotted the distribution of all features, as well as all features by gender (sensitive attribute) and depression (target). Considering the large number of plots, we only included figures that are essential or surprising in relation to this project in this report. All figures can be found in the Google Colab Notebook.

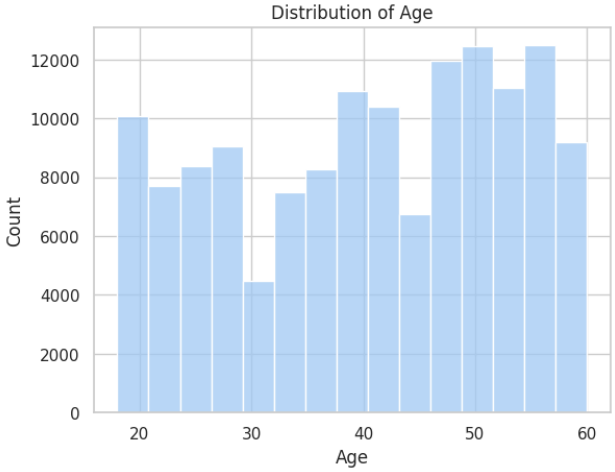


Fig. A.1: Distribution of age of the participants.

Figure A.1 is the general distribution of age of the sample. As you can see, the distribution is approximately uniform, with slightly more older than younger adults.

Figure A.2 is the distribution of gender of the sample. As you can see, there are more males than females in this sample.

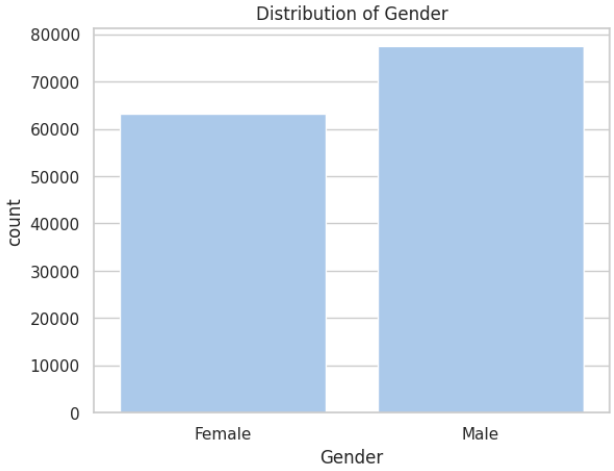


Fig. A.2: Distribution of gender of the participants.

Similarly, the distribution of depression is presented as figure A.3. Not surprisingly, the rate of depression is relatively low, which is consistent with our expectation. However, the prevalence of depressed individuals in this sample is 18.2%, which is way higher than the population average. Considering the data is synthetically generated with the purpose of depression detection, the depression prevalence can be explained.

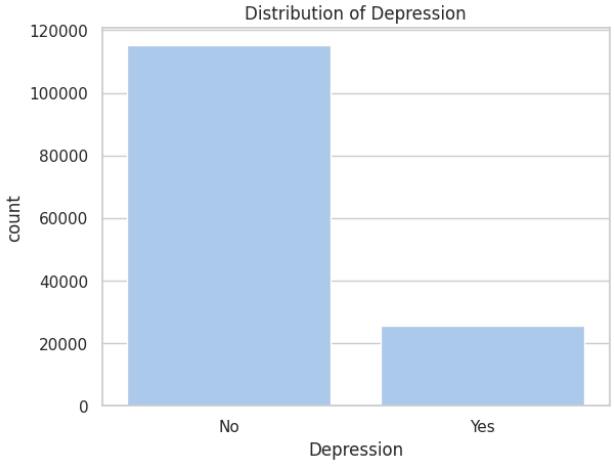


Fig. A.3: Distribution of depression.

To get a basic sense of how depression distributes relative to our sensitive attribute, gender, we plotted depression by gender. The resulting figure is figure A.4. Contrary to our expectation and general consensus (American Psychiatric Association, 2013), there are slightly more depressed males than there are females. One potential explanation is that the data is simulating features of Indian population, which could have a different gender prevalence when it comes to depression than the one based on almost completely U.S. and Western

context.

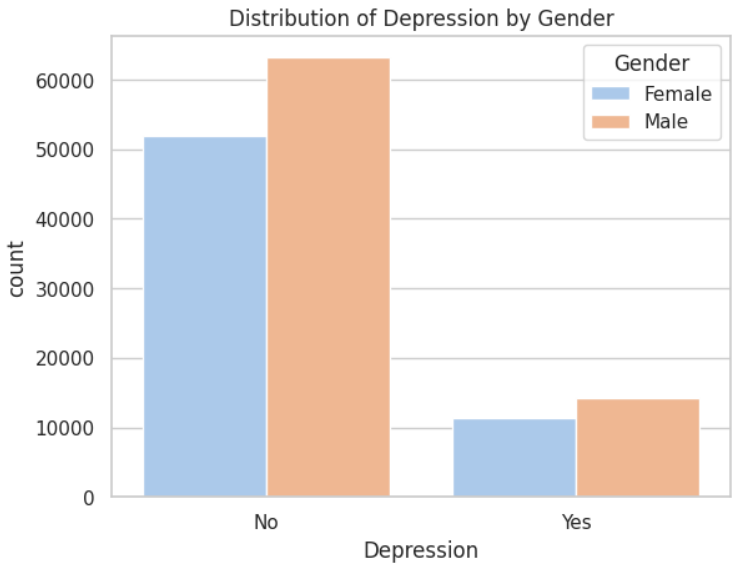


Fig. A.4: Depression by Gender.

Another interesting aspect of this dataset we discovered through EDA is the distribution of age in terms of depression. See figure A.5. The mean age of depressed individuals is substantially lower than that of the not depressed ones, at around 23 years of age, while the mean age of not depressed participants is around 35. The age distribution aligns with published studies on the age of onset of depression (Mirowsky & Ross, 1992), where there is a large spike in depression rate for people between the ages of 17 and 25, and again for people that are over 80 years of age. Though there are more working professionals than students in this dataset (results not shown), there is a much higher rate of depression in young adults. This trend indicates a possibility of treating age as a sensitive attribute, as it might lead to imbalanced model prediction based on age alone. To explore this possibility, we calculated the point-biserial correlation of age and depression, which we will present and report in the later sections.

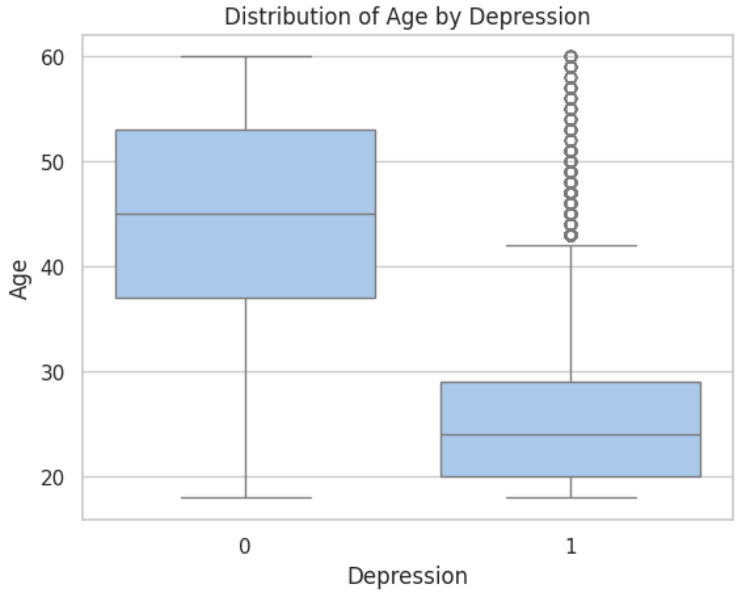


Fig. A.5: Distribution of age by depression.

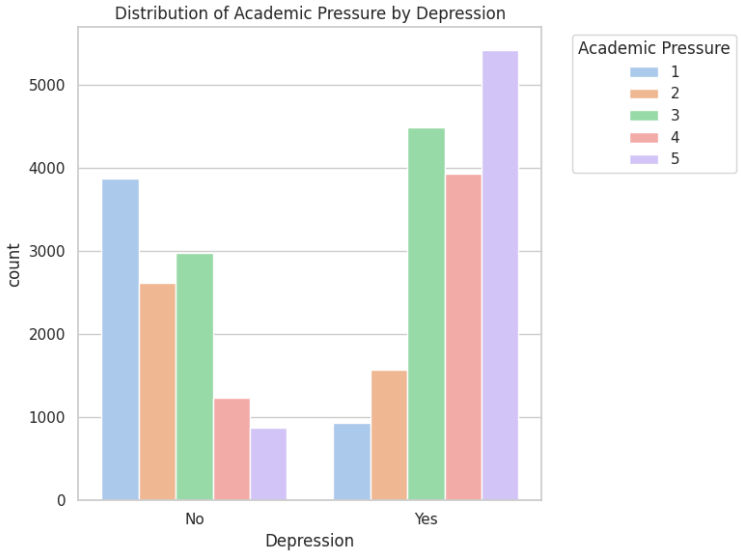


Fig. A.6: Distribution of academic pressure and depression.

Figure A.6 illustrates the relationship between academic pressure and depression. The distribution for non-depressed individual is clearly negatively skewed, while the distribution for depressed individual is the direct opposite. This result makes intuitive sense, as one would expect the perceived stress would have a relationship with depression. Similar trends are observed in the relationship between work pressure and depression, and opposite trends are observed in the relationship between study satisfaction and job satisfaction and depression, in the sense that depressed individuals reported less satisfaction in their study or job than

non-depressed individuals.

A.2.1 Feature Correlation

As illustrated in Figure A.5, the sigmoid curve depicted in Figure A.7 clearly demonstrates that the probability of depression is higher among young adults. This can be attributed to the fact that students exhibit significantly higher rates of depression compared to working professionals (Figure A.8). Furthermore, the correlation matrix shown in Figure A.9 emphasizes that attributes such as unhealthy dietary habits, inadequate sleep duration, and the occurrence of suicidal thoughts are positively correlated with each other and are notably associated with students.

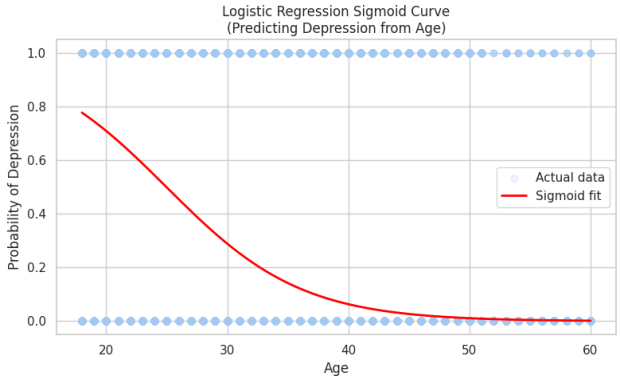


Fig. A.7: Predicting Depression from Age

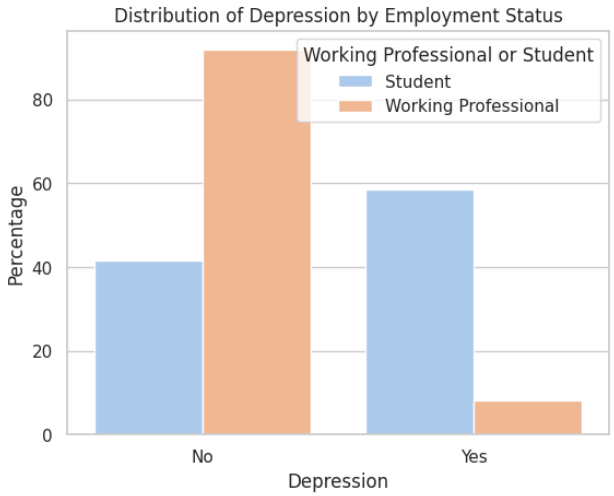
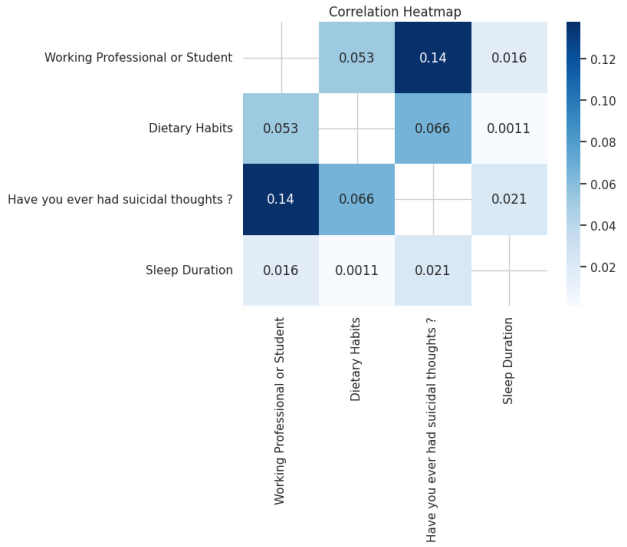


Fig. A.8: Distribution of Employment Status vs Depression



Student:1, Working Professional:0
Unhealthy: 2, Moderate: 1, Healthy:0
Sleep : < 5 hours:3, 5-6 hours:2, > 8 hours: 1, 7-8 hours: 0
Suicidal Thoughts: Yes: 1, No: 0

Fig. A.9: Correlation of Employment Status, Dietary Habits, Sleep Duration and Suicidal Thoughts.

Our analysis as shown in Figure A.10 also reveals a positive correlation between gender (-1 for males, 1 for females) and work pressure, consistent with prior studies indicating women face higher work pressure due to overlapping professional and caregiving roles, especially during remote work setups (P & Perwez, 2023).

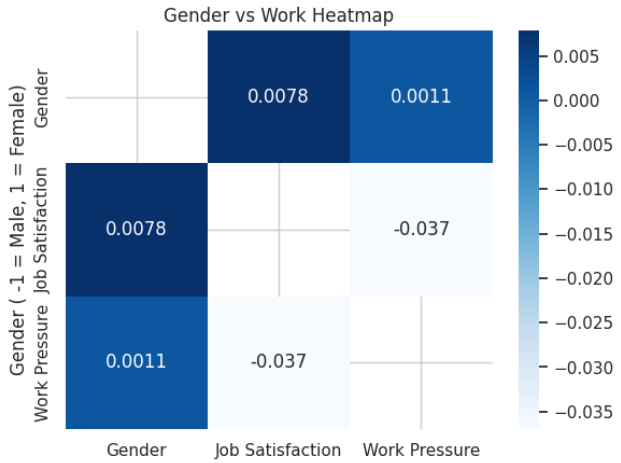


Fig. A.10: Distribution of Gender vs Work Pressure and Job Satisfaction

Interestingly, the data also indicates a positive correlation between gender and job satisfaction. This observation appears to contradict prior studies that associate lower job satisfaction among women with discriminatory organizational practices and societal expectations (M. et al., 2024). One potential explanation for this contradiction might lie in the study's

context—data was collected from selected urban cities in India, where unique factors could influence these findings. For instance, exceptions in prior research suggest that variables such as locus of control can sometimes neutralize gender disparities in job satisfaction (Talamala, 2024).

Another aspect of the data distribution is shown in Figure A.11 and A.12 , which suggests that dataset primarily includes data from major cities in larger states of India, predominantly from the Central North and Central-West regions. This lack of data from small-to-medium cities, rural areas, and regions like the South, Central-East, and North-East India might be an unrepresentative sample.

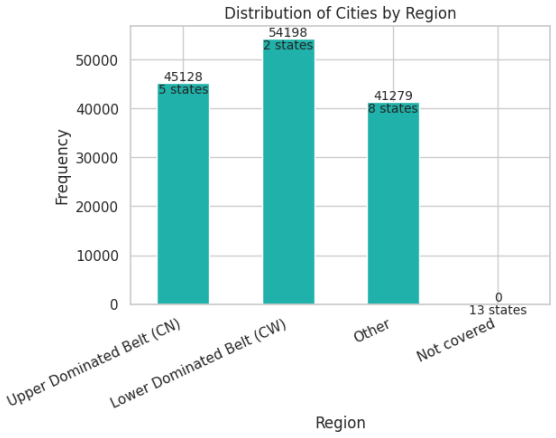


Fig. A.11: Data Distribution Region-wise

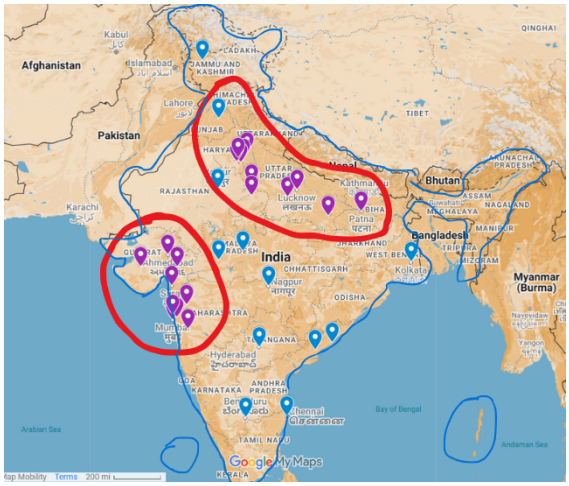


Fig. A.12: Map Annotation of Training Data, depicting biased selection

A.3 Post Training Analysis

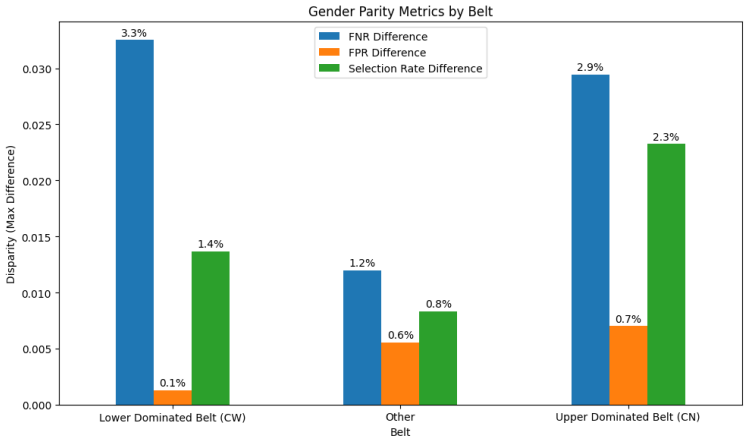


Fig. A.14: Fairness w.r.t. Gender for Region After Model Training

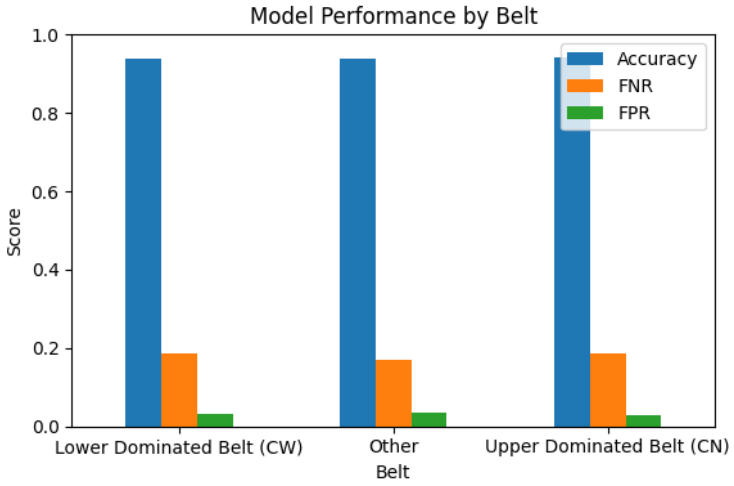


Fig. A.13: Overall Performance w.r.t. Region After Model Training

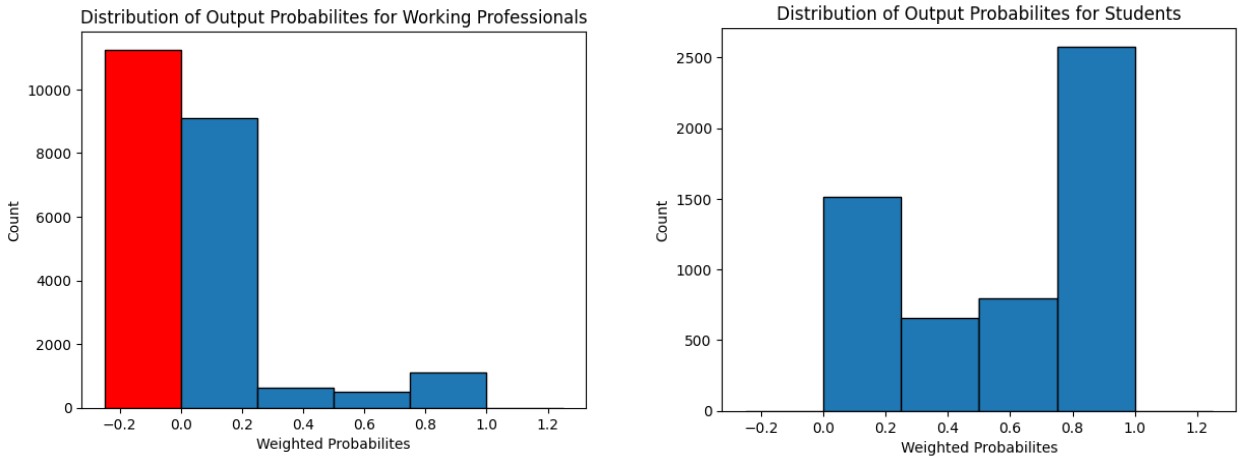


Fig. A.15: The output "probabilites" for Working Professionals are outside the range of [0,1], with a significant count in less than 0, which cannot be ignored or considered outliers. This imply that there is an in-processing need for scaling such distribution, as they don't clearly justify the predicted outcome. This also strengthens the argument for optimizing threshold for classification.