Academic Burnout Detection Using Behavioral Data Analysis

by

Abrar Zahin
20101052
Abiduddin Afnan
21201679
Zannatun Tazree Tisha
21201128
Steve D Costa
21201449

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Abrar Zahin	Abiduddin Afnan
20101052	21201679
Zannatun Tazree Tisha	Steve D Costa
21201128	21201449

Approval

The thesis/project titled “Academic Burnout Detection Using Behavioral Data Analysis” submitted by

1. Abrar Zahin (20101052)
2. Abiduddin Afnan (21201679)
3. Zannatun Tazree Tisha (21201128)
4. Steve D Costa (21201449)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in January 2025.

Examining Committee:

Supervisor:
(Member)

Md. Aquib Azmain

Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Labib Hasan Khan

Adjunct Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

As the world continues to evolve, educational systems are advancing at an unprecedented pace. However, this rapid change has led to challenges such as academic burnout. Academic burnout is defined as physical, emotional, and mental exhaustion due to academic pressure which leads to symptoms such as reduced motivation, emotional detachment, and lower academic performance. Maslach Burnout Inventory is commonly regarded as the most popular burnout scale, and the student version (MBI-SS) is the instrument of choice in student samples regardless of their fields of study most notably in medicine.Early research has explored the functionality of ML models in predicting burnout by highlighting their effectiveness in analyzing both behavioral and psychological data as well as responses from well-known inventories.Our proposed methodology approches the use of the use of MBI-SS from a different perspective. This paper aims to improve the use of MBI-SS by comparing two of it's labeling methods, creating a micro-screener which would act as a shorter version of the MBI-SS and using machine learning (ML) models to test the effectiveness of the micro-screener.The models we tested were Logistic Regression (L1/L2), Random Forest (RF), Support Vector Machine (SVM), Radial Basis Function kernel Support Vector Machine (RBF SVM), Extreme Gradient Boosting (XGB), and LGBM (Light Gradient Boosting Machine) and Extra Trees/Extremely Randomized Trees (ET).

Keywords: Artificial Intelligence, Machine Learning, Data Analysis, Academic Burnout, Detection.

Table of Contents

Declaration **i**

Approval **ii**

Abstract **iv**

Table of Contents **v**

List of Figures **viii**

Nomenclature **xi**

1 Introduction **1**

1.1 Background 1

1.2 Rational of the Study or Motivation 2

1.3 Problem Statements 3

1.4 Objective 3

1.5 Methodology 4

1.6 Scopes and Challenges 5

2 Literature Review **7**

2.1 Preliminaries 7

2.2 Review of Literature 8
2.3 Summary of Key Findings 14

3 Requirements, Impacts and Constraints 15

3.1 Final Specifications and Requirements 15
3.2 Societal Impact 17
3.3 Environmental Impact 17
3.4 Ethical Issues 17
3.5 Standards - if applicable 17
3.6 Project Management Plan 18
3.7 Risk Management 19
3.8 Economic Analysis 19

4 Proposed Methodology 20

4.1 Design Process or Methodology Overview 20
4.2 Preliminary Design or Design (Model) Specification 22
4.3 Data Collection 22
4.3.1 Data Cleaning 23
4.3.2 Data Transformation 23
4.3.3 Data Integration 25
4.3.4 Data Reduction 25
4.3.5 Summary of Preprocessed Data 25
4.4 Implementation of Selected Design 26

5 Result Analysis 28

5.1 Performance Evaluation 28

5.2 Analysis of Design Solutions 36

5.3 Final Design Adjustments 38

5.4 Statistical Analysis 38

5.5 Comparisons and Relationships 44

5.6 Discussions 44

6 Conclusion 45

6.1 Summary of Findings 45

6.2 Contributions to the Field 45

6.3 Recommendations for Future Work 46

Bibliography 49

List of Figures

3.1 The translated MBI-SS items in Síndrome de Burnout, según el año de estudio, en estudiantes de medicina de ocho países de Latinoamérica, 2016-2017 [15]. 16

4.1 A simple workplan 21

4.2 Overview of dataset origin and translation process here 23

4.3 Distribution histograms or density plots of EX mean 24

4.4 Distribution histograms or density plots of CY mean 24

4.5 Distribution histograms or density plots of PE mean 25

5.1 Confusion matrix shows clinical flags more than tertiles 29

5.2 ROC curve (test split) 33

5.3 ROC curve (test split) 34

5.4 ROC curve (test split) 35

5.5 ROC curve (test split) 36

5.6 Confusion matrix K3(RF) - Clinical 39

5.7 Confusion matrix K4(ET) - Clinical 40

5.8 Confusion matrix K5(RBF SVM) - Clinical 40

5.9 Confusion matrix K4(XGB) - Tertile 42

5.10 Confusion matrix K3(RBF SVM) - Tertile 42

5.11 Confusion matrix K3(RBF SVM) - Tertile 43

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

κ Cohen's kappa (inter-rater/label agreement)

AI Artificial Intelligence

AIC Akaike Information Criterion

$AUPRC/AP$ Area Under the Precision–Recall Curve (Average Precision)

$AUROC/ROC-AUC$ Area Under the Receiver Operating Characteristic Curve

BDI Beck Depression Inventory

BIC Bayesian Information Criterion

$Brier$ Brier score (mean squared error of predicted probabilities)

CFA Confirmatory Factor Analysis

CFI Comparative Fit Index

CFI Comparative Fit Index

$Clinicallabel$ "Exhaustion+1": EX high AND (CY high OR PE_rev high)

CNN Convolutional Neural Networks

CV Cross-Validation

CY Cynicism (MBI-SS subscale)

ET Extra Trees (Extremely Randomized Trees)

EX Exhaustion (MBI-SS subscale)

$F1$ F1 score (harmonic mean of precision and recall)

$GSES$ General Self-Efficacy Scale

K Number of selected items in the micro-screener (e.g., K3, K4, K5)

$LGBM/LightGBM$ Light Gradient Boosting Machine

$Logit/LR$ Logistic Regression

$LSTM$ Long ShortTerm Memory

$MBI - SS$ Maslach Burnout Inventory–Student Survey

$MDist$ Mahalanobis distance (in item space)

MHT Mental Health Diagnostic Test

ML Machine Learning

ML Maximum Likelihood (estimation method for CFA/SEM)

NL Natural Language

NN Neural Network

$PC1$ First Principal Component (from PCA)

PE Personal/Professional Efficacy (MBI-SS subscale)

PE_rev Professional Efficacy, reverse-scored

PR Precision–Recall (curve)

$RBF - SVM$ Radial Basis Function kernel Support Vector Machine

RF Random Forest

RF Random Forest

$RMSEA$ Root Mean Square Error of Approximation

ROC Receiver Operating Characteristic (curve)

SEM Structural Equation Modeling

$SRMR$ Standardized Root Mean Square Residual

$SRMR$ Standardized Root Mean Square

$STAI$ State-Trait Anxiety Inventory

SVM Support Vector Machine

SVM Support Vector Machines

$Tertilelabel$ Top-third on EX AND (CY high OR PE_rev high)

TLI Tucker–Lewis Index (Non-Normed Fit Index)

TLI Tucker-Lewis Index

$TP/FP/TN/FN$ True Positive / False Positive / True Negative / False Negative

$WHOQOL$ World Health Organization Quality of Life

$XGB/XGBoost$ Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Background

Academic burnout is a condition that comes from stress, showing up as feeling emotionally drained, having a negative outlook, and feeling less capable of achieving your goals. The World Health Organization has officially recognized ”burn-out” in ICD-11 as something that happens in the workplace due to ongoing stress that hasn’t been handled well [13]. This idea has also been applied to students in educational settings. In higher education, we often use the Maslach Burnout Inventory–Student Survey (MBI-SS) [4] to measure burnout. This tool is a version of the original MBI, tailored for students, and it looks at the same three key areas. It has been tested and proven effective in various countries and groups of students. These changes have made academic burnout a clear concept that is well-supported by theory and commonly measured.

A recent overview of the literature examined 44 studies that used 26,500 students in the COVID-19 time. It discovered that 56.3 percent of the university students gave high scores to emotional exhaustion, 55.3 percent to cynicism, and 41.8 percent to low personal accomplishment [20]. These figures were different by the area of study and geography, yet they indicated a great threat to students. In conjunction with these results, research in medical education, which is regarded as a profession that is difficult, demonstrated that approximately 44 percent of people feel the burnout before beginning the residency. This involves approximately 41 percent who are emotionally exhausted and approximately 35 percent that depersonalize themselves [16]This covers about 41 percent who are emotionally drained and about 35 percent who depersonalize themselves (Frajerman2019-vl). Another meta-analysis that included low- and middle-income nations reported a reduced global burnout rate of approximately 12 Per cent. The authors however indicated that there were substantial disparities in data and that they may have underreported the actual effect which could be attributed to publication bias. These results indicate that academic burnout is a widespread problem across the globe and it is relative in various situations.

Academic burnout is not only a common thing but also related to poor mental health, lack of engagement, and failure to cope with school work. They are at risk of those in challenging programs and at difficult times, such as those associated with learning transitioning online due to a pandemic. It has been found that stressors such as high workloads, support deficit, and learning change may lead to burnout. Conversely, it can be reduced by resources of the program and individuals, including hybrid teaching, resilience, and targeted interventions. These trends demonstrate the importance of early detection of problems and the establishment of a good prevention strategy on campus. They prompt this thesis of discovering the academic burnout as a means of assisting them in responding responsively and evidence-based [20], [21].

1.2 Rational of the Study or Motivation

Academic burnout is a serious issue that many university and medical students cope with, making it a serious problem to their mental health and education. This problem may actually impact their education, health, and capacity to continue with their studies. Burnout is recognized in the international classification of diseases (ICD-11) as a work-related problem. It is associated with an experience of tiredness, a feeling of detachment or indifference, and a decline in professional performance. The Maslach Burnout Inventory-Student Survey (MBI-SS) is a student-centered tool, so it is easy to see these three in the students as well, with the exhaustion (EX), Cynicism (CY) and Professional Efficacy (PE) as the key areas. This model has always been found to be more applicable as compared to a simple one-factor model [5]. The three-factor solution has been shown to be reliable and valid in numerous studies of students, among them in health sciences, which proves that the solution is effective in numerous aspects (Hederich2016MBISSColombia).

The article is motivated by different reviews on the high rates of burnout among students in medical education and the associated stress levels. This problem frequently aggravates in the clinical years and relates to depression and problems with academic performance, which is reflected by recent data in BMC Medical Education by evidence of various countries under RiskProtectiveBurnout [25]. These issues have impacts on universities in practice: early detection of problems can be met with immediate assistance, yet the lack of efficient screening due to the survey fatigue, the insufficiency of advisors, and the need to identify risk factors rather than quantify them.

Based on this, our study addresses the question of whether a substantially smaller micro-screener constructed experimentally on the basis of the 15-item MBI-SS can preserve most of the predictive signal needed to conduct a triage and be consistent with the three-factor model that is already proven. We examine a large publicly available sample of medical students in Latin-America in 2016-2017. It contains the full MBI-SS, which assists us in comparing shorter evaluation tools with the entire one in two aspects: (i) a clinically grounded rule (exhaustion + 1): high EX and [high CY or low PE] that represents the way these dimensions are commonly conceptualized; and (ii) a tertile-based rule that determines the group with the highest burden in comparison with their counterparts.

The other cause is concerned with the way we collect data: with self-report surveys, the manner in which people react can be manipulated, as it may occur due to acquiescence, extreme responding, middle ranking, and lack of attention, which may diminish the outcomes and influence thresholds used. The study of the contemporary research approaches and industrial-organizational psychology is devoted to defining or rectifying these trends in order to increase the exactness Likert-type data [8].

1.3 Problem Statements

- RQ1: Does the canonical three-factor structure of the MBI-SS (EX, CY, PE) fit this dataset better than a one-factor model?
- RQ2: Can we reduce the 15-item MBI-SS to smaller item set ?
- RQ3: Which burnout label (clinical Exhaustion+1 vs tertile) and which classifier yield the best calibrated, operationally useful screener, and what thresholding rule should be recommended for practice?

1.4 Objective

The purpose of this study is to construct, validate, and implement a short, precise, and robust academic burnout screening procedure that may be used by universities with minimal impact on volunteers. Our targeted objectives are to: (i) use confirmatory factor analysis to ensure that the Maslach Burnout Inventory Student Survey (MBI-SS) maintains its typical three factor structure, for example, Exhaustion (EX), Cynicism (CY), and Professional Efficacy (PE) in our dataset. (ii) create two effective burnout labels ("burnout"), one based on the sample relative tertile rule and the other is the clinically driven Exhaustion+1 rule, and evaluate how predictable they are. (iii) Using a four item checklist selected through data driven methods, create an ultra short micro screener that maintains almost all of the predictive potential of the entire 15 item instrument. (iv) Modify this preview to match the full scale version's performance by including free response-oriented features. (v) provide two deployable models, which are a calibrated gradient-boosted model that incorporates variables with the four items, as well as a logistic regression model with four items and available standardized coefficients. Considering a held-out test set, the models' effectiveness will be evaluated with a concentration on actionable utility at a fixed capacity (precision/recall for the top 10% of the most at risk students), probability calibration (Brier score and reliability curves), and discrimination (AUROC/AUPRC). The ultimate goal is to develop a feasible, well tested procedure that educational institutions can employ to identify and assist students who are at the risk of burnout. The procedure should demand little time from respondents and provide clear decision making parameters for advising teams.

1.5 Methodology

For the study, a publicly available Latin American medical student dataset was used [15], containing all the responses to 15 MBI-SS items [4]. This created an overall sample size of 2,649 students by maintaining data completeness and statistical reliability for machine learning and confirmatory factor analysis.

1. **Clinical-style label ("Exhaustion+1")**

$$\text{burnout}_{\text{clinical}} = 1(\text{EX high} \wedge [\text{CY high} \vee \text{PE}_{\text{rev high}}]),$$

   where EX, CY, and PE are the standard MBI-SS subscales and PE_{rev} denotes reversed personal efficacy. "High" used conventional cut rules on subscale means.

2. **Tertile label (sample-relative)**

   We computed the 66th percentile (p66) for EX, CY, and PE_{rev} on valid rows and set

$$\text{burnout}_{\text{tertile}} = 1(\text{EX} \geq \text{p66}_{\text{EX}} \wedge [\text{CY} \geq \text{p66}_{\text{CY}} \vee \text{PE}_{\text{rev}} \geq \text{p66}_{\text{PE}_{\text{rev}}}]).$$

Confirmatory Factor Analysis (CFA)

Traditional three-factor (Exhaustion, Cynicism, and Professional Efficacy) model of burnout fit the data more accurately better than a single-factor model, according to a CFA using semopy in Python (ML estimator). We used CFI, TLI, RMSEA, SRMR, X^2, DoF, and AIC/BIC to evaluate model fitness to keep the three-factor model for further retention.

Feature Engineering and Short Form Selection

To capture unique responding patterns, a number of response-style characteristics were calculated in addition to the initial burnout responses. Efficient K = 3/4/5 item short forms that maintained accuracy in prediction while lowering the number of survey items were found using a cross-validation-based feature selection technique.

Model Training and Evaluation

The data were divided into training, testing and validation sets (60/20/20) to develop and fine-tune various machine learning models, including logistic regression, random forests, SVM, and gradient boosting. Metrics like AUROC, AUPRC, and Brier score were used for evaluating the performance of model's, and ROC and precision-recall curves were used to illustrate the results of the study.

Final Model and Deliverables

The 4-item screener that had been optimized with some features maintained an appropriate balance between shortness and predictive strength. So, two final models were suggested: a calibrated XGBoost model for high accuracy and an interpretable logistic regression model for effective burnout screening

1.6 Scopes and Challenges

The dataset was obtained through a cross-sectional survey involving 2,649 medical students from nine universities across eight Latin American nations between 2016 and 2017, employing non-probability (convenience) sampling methods. The authors explicitly recognize the presence of selection bias due to the non-random sampling of students; they stress that their primary objective was to examine burnout dimensions based on year of study rather than to estimate prevalences within the population. Additionally, they mention that the instrument was tested and conducted anonymously during class sessions, with the analysis employing robust generalized linear models clustered by campus.

Participants consisted of medical students alone (interns excluded), 57% female, and majoring in universities in the private sector (72%). They may not be representative of other fields of study, public universities, or contemporary cohorts; the 2016–2017 time period also comes before recent disruptions to student well-being (e.g., the pandemic of COVID-19).

Outcomes consisted of 15 self-rated Likert items in the MBI-SS across Exhaustion (EX), Cynicism (CY) and (reversed) Personal Efficacy (PE). The paper investigated dimension scores across years, rather than clinical diagnoses; all "clinical" interpretation is thus indirect and based on post-hoc thresholds rather than physician ratings.

The original study employs observational techniques, specifically Spearman correlations and Generalized Linear Models (GLMs), while advising caution regarding causal inferences. As a result, any subsequent predictive model may illustrate associations identified within that convenience sample instead of enduring causal signals.

Though the MBI-SS possesses acceptable internal consistency in the populations cited and robust construct/content validity, it nonetheless measures burnout in terms of self-perception on one instrument and in one language; cross-cultural and linguistic refinements may be retained.

Our clinical marker (EX high AND [CY high OR PE low]) and tertile marker (top-third rule) are extracted directly from the MBI-SS itself—not from some third-party adjudication of the diagnostic labels—so they're proxy outcomes. This makes the model more practical but limits the achievable performance and can carry any such bias in the thresholds forward into the model.

Items and models were calibrated and validated on the identical pre-COVID multinational set. Whereas with stratified divisions and nested cross-validation during item selection, no established external validity to other nations, fields, grading regimes, languages, or pandemic emergence groups exists. An external valid test (other site/time) has not yet been performed.

To achieve well-specified covariance matrix for CFA and pristine training features, we needed all 15 items to be available. This can render the analytic set systematically divergent from the parent sample (e.g., dropping more distressed students who missed items).

We applied ML estimators (maximum likelihood) and streamed 06 Likert responses as nearly continuous for CFA and modelingstandard in practice with ≥ 5 categories but an approximation nonetheless. If item distributions become highly skewed or bunched, then CFA and ordinal link models based on polychoric could be preferred; we did not attempt them as sensitivity tests.

Greedy forward selection (inner CV) minimizes search space but is not globally optimal for the subset; with various random seeds or folds, the K-item set may slightly differ. Ensemble models (XGBoost/LightGBM) may overfit without careful cross-validation tuning; we offset this with nested CV and a held-out test but with some uncertainty.

Top-decile triage and F1-optimal thresholds which are chosen out of observed test distributions; alternative prevalence groups also require re-calibration and re-evaluation of the threshold in order to maintain precision/recall trade-offs.

Our CFA tested one- vs three-factor structures at the aggregate level. We did not measure measurement invarance across countries, sex, or year (configural/metric/scalar). In the event the items behave differently across groups, pooled factor scores or thresholds become unfair or unstable.

We highlighted predictive discrimination (AUROC/AUPRC), top-decile enrichment, Brier score, and reliability plots. We didn't carry out decision-analytic utility analysis (e.g., net benefit), cost-effectiveness, or simulation of longitudinal impact—desirable for practical implementation in advising applications.

We took the translated MBI-SS items from dataset source research paper [15]. However, it would have been better if we were able to take the help of an translation expert and a psychiatrist regarding this matter.

Chapter 2

Literature Review

2.1 Preliminaries

AI simulates human intelligence in machines and enables them to perform problem-solving and decision making tasks. ML is a subset of AI that trains algorithms to learn patterns from data and improve performance without explicit programming. To successfully implement the proposed system, several ML models are needed based on students' behavior patterns:

- Logistic Regression: For binary classification, such as determining whether the risk of burnout is high or low based on patterns like the number of missed assignments or quizzes, attendance percentage of attended classes, and grades drop over semesters. The probability score will indicate the risk level, with a score higher than 0.5 classified as "High Risk of Burnout."
- Random Forest: It is a robust model that handles nonlinear data and provides insights into the most common causes of burnout. By analyzing features such as exam scores, late submissions, and teacher reviews on student performance, it predicts burnout categories (low, moderate, high) and ranks the contributing factors by the most important common cause.
- Support Vector Machines (SVM): It is suitable for complicated datasets, SVM can classify burnout categories by combining behavioral and psychological data such as student emotional feedback (positive, neutral, negative).
- Neural Networks: It is capable of handling large, sequential, and complex datasets. It can elaborate relationships between variables and predict burnout trends over time.
- Structural Equation Modeling (SEM): SEM identifies and quantifies relationships between variables and provides path analysis by showing how missed assignments, lower engagement, and burnout are interconnected.

- Comparative Fit Index (CFI): it is a statistical measurement which used in SEM to evaluate the theoretical model fits to observed data. It compares the fit to a null model. CIF range between 0.90 to 0.95 indicates a better fit.
- Tucker-Lewis Index (TLI): It is a fit index which used in sem to access it. It is also used in linear mean and exploratory factor analysis and is common for adjusting model complexity where 1 means a better fit.
- Root Mean Square Error of Approximation (RMSEA): It is a measurement process that measures how well a statistical model fits to the population. The range between 0.05 to 0.08 indicates a better fit.
- Standardized Root Mean Square Residual (SRMR): It is also a fit index in SEM that shows the mean difference between observed and predicted correlations where less than 0.08 means a good model fit.
- General Self-Efficacy Scale (GSES): It is a psychological evaluation test that is used for measuring an individual's belief in their capacity to manage difficult circumstances and achieve goals.
- State-Trait Anxiety Inventory (STAI): It is a self-report measurement that measures both temporary state and long-term states of anxiety.
- Mental Health Diagnostic Test (MHT): It is a psychological evaluation for mental health problems such as stress, anxiety, and depression.
- Convolutional Neural Networks (CNNs): It is a subclass of deep learning neural networks which is created for handle structured grid data such as computer vision tasks.
- Long Short-Term Memory (LSTM): It is a type of recurrent neural network (RNN) architecture which is created for maintaining the memory cell to stimulate the sequences and time-series data.

By implementing these models, the system will address all the gaps in existing methods and will offer a proactive solution for detecting and mitigating academic burnout. This innovative approach can help reduce the effects of burnout by improving student's academic performance and overall well-being.

2.2 Review of Literature

Ensemble methods such as Random Forest are an effective way to make predictions, as they combine output from multiple networks. Larrain et al. used this approach in their study [19]. They began by collecting data through a questionnaire of 83 questions in total which were made using four instruments. Beck Anxiety Inventory (BAI) consists of questions about anxiety symptoms. Depression symptoms were included in the Beck Depression Inventory (BDI). On the other hand, the Burnout One-Dimensional Scale

(EUBE) is utilized to diagnose burnout. It has 15 questions that evaluate a student's engagement and academic performance by examining factors such as interest in attending classes, and class participation. Lastly, there was the World Health Organization Quality of Life (WHOQOL) which examines a participant's Quality Of Life which is assessed by investigating their position in life, their expectations, cultural norms, and concerns. WHOQOL has four sections, namely Physical Health, Psychological Health, Social Relationships, and Environment.

To find possible patterns in the responses, the authors used the scores of the instruments to define 4 cognitive classes (COG), emotional (EMO), organizational (ORG), and physical (PHY). These classes are not mutually exclusive, meaning each participant could suffer from more than 1 disturbance. The class Control Condition (Cntrl) was added for participants who were completely healthy and did not suffer from any disturbances. Each disturbance type was treated as a Binary Classification. So if a disturbance is present it is labeled as 1 and if not, it is labeled as 0.

The researchers used a Random Forest model and a feed-forward multi-layer perceptron model, which is a type of Neural Network. Each class is evaluated with both of these models to detect each disturbance.

The RF model had an average accuracy of 96% with an average FNR of 2% as well as an AUC of 0.97. Meanwhile, the NN model had a lower average accuracy of 94%, an average FNR of 2.4%, and an AUC of 0.90. In the COG, EMO, ORG, and CTRL classes, the RF model had a better average accuracy than the NN model. But in the PHY class, the NN model had a higher accuracy of 95% whereas RF had an accuracy of 91%.

Overall, RF had a better performance than NN, and additionally better FNR. Aside from having a better accuracy the RF model also had a lower average FNR which meant that it was better at identifying participants with symptoms. The study also used Shapley Additive Explanations (SHAP), which gives each feature a contribution score to understand which features contributed most to the model predictions.

Shima Baniadamdizaj and Shahla Baniadamdizaj [16] focus on finding out the key factors behind teacher burnout and how effectively ML models can analyze and predict their burnout. To measure burnout, the researchers used the Maslach Burnout Inventory (MBI), which is a psychological assessment tool that measures burnout in three dimensions. Emotional Exhaustion (EE) refers to feelings of being emotionally exhausted from one's work. Depersonalization (DP) means emotional detachment from their work, loss of empathy, and cynicism in their attitude toward their students and colleagues. Personal Accomplishment (PA) is related to feelings of competence and achievement in their work.

1433 Iranian EFL teachers participated in the study in which they answered questions from the MBI which consisted of 22 symptoms contributing to work-related burnout. The participants answered several questions on a Likert scale ranging from 0 (never) to 6 (always). The scores are then used to find out the level of burnout in each dimension, where the result is categorized into three risk groups: low, moderate, and high.

In order to analyze the results, the researchers employed the Decision Tree (DT) Support Vector Machine (SVM), Support Vector Classifier (SVC), Support Vector Regression (SVR), Multi-layer perceptron (MLP), k nearest neighbor (KNN), Gaussian Naive Bayes (GNB),Random Forest (RF), Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA).

The data were trained on these ML models and applied on the participants to classify them as low, moderate and high levels of burnout. The was based on features of the MBI inventory.When it comes to detecting depersonalization and personal accomplishment Quadratic Discriminant Analysis (QDA) had the highest accuracy of 100%. Support Vector Classifier (SVC) and Linear Discriminant Analysis (LDA) had about 93% accuracy in detecting emotional exhaustion

On the other hand, Rocco De Filippis and Abdullah Al Foysal (2024) [22] worked on identifying and analyzing the primary stress factors affecting students. The authors used a pre-built dataset consisting of data from 1100 students from several educational institutions. The dataset was divided into multiple dimensions such as psychological, physiological, environmental, academic, and social similar to the paper above [22]. While data processing, statistical imputation was applied to deal with anomalous and missing values. Normalization and Standardization were used to prevent any single variable from dominating the analysis.

Before analyzing the obtained data, the authors examined the central tendencies and distribution of the stress-related variables. Pearson's correlation coefficients helped in analyzing the relationship between each variable. A Random Forest regression model was used to find out the impact each variable had on student stress. When evaluating the performance of this model, the Five Fold Cross Validation (FFCV) revealed an average accuracy of 0.88 and a standard deviation of 0.032. In conjunction with the results of the RF model, Permutation Feature Importance was utilized to evaluate the significance of each feature. Self-esteem was found to be the leading psychological factor behind student stress, with depression and anxiety also having a strong negative correlation. Among physiological factors, sleep quality and headache contributed the most to student stress. On the other hand, basic needs were the most important environmental factor followed by safety. When it came to social factors, bullying had the most negative impact on students. Finally, academic performance was the key academic contributor to student stress.

The findings highlight psychological factors such as self-esteem and depression as well as physiological factors, particularly sleep quality, to be the major causes behind student stress. De Filippis and Al Foysal advocate for mental health intervention by educational institutions to address identified stressors and work toward creating a healthy learning environment for students [22].

Analytical approaches can also be a variable factor in research studies, such as this particular paper [14]. Most of the past research states gathering academic information or student behavior data need to conduct online or offline surveys, but João Marôco et al. (2020) [14] included combined data from more than 4,000 universities in different countries using a complex statistical model known as Structural Equation Modeling(SEM) [14]. The researchers who worked on this paper adjusted the model for collecting and analyzing data, which will show and find relations among the effects, causes of engagement, and burnout of university students. Factors such as positive and negative coping strategies, Social support, Learning expectations, and also Teacher engagement were all considered in this study.

Structural Equation Modeling is one of the most advantageous and commonly used statistical analytical tools in social science and educational research. This model clarifies the complicated correlation between different factors that have both countable and latent variables. It has the compatibility to include latent categories and checks how well the model works. The model contains multiple analytical methods which show how the change in factors can affect the correlated variables including finding regression coefficients, R-squared values, and using goodness-of-fit indices like TLI (Tucker-Lewis Index), CFI (Comparative Fit Index), RMSEA (Root Mean Square Error of Approximation), and SRMR (Standardized Root Mean Square Residual). This combination of multiple regression analysis and factor analysis aids in the study. Additionally, the use of confirmatory factor analysis and tests using the collected surveyed data to see if it was compatible.

In the end, data collection and analysis revealed that both academic performance and dropout intention were impacted by student engagement and burnout (negative correlation with dropout intention of r =-0.523). Burnout reduces the positive outcome of engagement, making it harder for the students to achieve success. The paper highlights that negative coping strategies like self-harm and isolation are the key causes of burnout however, positive coping strategies such as problem sharing and course expectations strongly boost student engagement. The relation between burnout and student engagement clarifies that an increased amount of burnout level could reduce the benefit of engagement in decreasing dropout risks. They wanted an increase in promoting engagement to prevent dropout mentality, which will highly reduce burnout levels. In the end, the study has confirmed its theory about the correlations between the variables.

Xinhang Gao carried out research that shows behavior or emotional factors are directly correlated with burnout **.**Gao2023AcademicSA If academic activities are taken into account, they include information such as attendance, finished assignments, study hours, and grades. For student information data such as anxiety, academic stress, peer pressure, and much more were considered. Multiple techniques were carried out in different studies to collect student data. This study included strict invigilation of students in computer labs during online surveys [18].

The relationship between burnout, anxiety, academic stress, and self-efficacy is researched in Chinese students rather than collecting data from all over the world. The research involved individuals from Junior High schools in Shandong province, China. It is a well-known fact that the Chinese education system pressures students to get higher grades in their academics [11]. The author of this study, Xinhang Gao(2023), mentioned that students in sixth and seventh grade were gathered and applied a whole-group sampling method to collect the necessary data. A total of 929 students willingly participated and their responses were received with a response of return rate 94.6% [18]. After a briefing from a trained examiner, the participants were invigilated by their teachers in their lab where electronic questionnaires were held and responses received.

The collected responses were further analyzed by researchers using valid tools such as the Academic Stress Scale, General Self-Efficacy Scale (GSES), State Trait Anxiety Inventory (STAI), and the Burnout Inventory.

- The Academic Stress Scale used after Study Stress Questionnaire, which contains 21 questions involving features such as parental pressure, social pressure, teacher pressure, and self-pressure. The responses were scored on a 5-point Likert scale, from low to high stress levels.
- The General Self-Efficacy Scale scaled the student's confidence in handling academic challenges by using a 22-item scale of academic ability and behavior. Responses were scored on a 5-point Likert scale, with increasing scores reflecting greater burnout.
- The State-Trait Anxiety Inventory evaluated anxiety from a temporary state to a persistent state in three categorical stages. Furthermore, input data was collected using The Mental Health Diagnostic Test (MHT) adapted by Zhou Bucheng (Zheng et al., 2004) which includes 15 questions scored on a binary response (Yes 1, No 0).
- Burnout Inventory is commonly used as seen in previous papers. It surveys students through multiple questions with responses containing a 5-point Likert scale measuring features like emotional exhaustion, depersonalization, and decrease in academic performance.

As mentioned in a past research study the Structural Equation Modeling (SEM) is also deployed in this paper to test a moderated mediation model which involves multiple data analyses. Direct Path Analysis was used to find the correlation between stress and burnout. Along with Mediated Path Analysis for researching academic anxiety as a mediator between stress and burnout. Anxiety acts as a mediating factor between burnout and academic stress levels. It has been found that 72% of students experience high stress which correlates strongly with burnout indicators and also acts as a moderating factor for severe anxiety. This confirms that a high level of stress increases the level of anxiety, which in turn worsens burnout symptoms like emotional exhaustion and student engagement. Furthermore, self-efficacy was also found to be a mitigating factor using Moderated Mediation and was proven under extreme academic stress. Students with higher levels of self-efficacy (increase of 30 in the GSES) reported a decrease in the burnout level (decrease of 40% burnout rate). These results highlight how important it is to regulate anxiety as well as self-efficacy in order to decrease the risk of burnout.

Xinhang Gao concluded this with analysis and mentioned the prevention of academic burnout by encouraging self-efficacy in students and managing both academic stress and anxiety [18]. Reducing Academic stress requires supportive encouragement and helping them build confidence in their own academic abilities. This will reduce the burnout risks. The research paper included workshops such as Stress Management Workshops and Self-Efficacy Programs to create a healthier academic environment.

Using a similar approach, Liu et al. employed multiple ML models and data analysis techniques to predict academic stress among students. The researchers gathered and combined physiological, behavioral, and psychological data, for further improvement in stress detection to help design interventions to reduce academic stress and therefore decrease student burnout [24].

Data collection consists of reviewing 135 studies published until 2024 that focus on stress detection methods. They gathered data from three main sources that include physiological data, behavioral data, and psychological data. Altogether, these studies include data from 50,000 participants. The groups range in size from small groups of 30 students to larger groups of over 5,000 students. This variation in data sources provides a detailed view of stress indicators.

The researchers used various Machine Learning Models, Deep learning Models, and Hybrid models to analyze the preprocessed data. The models were trained using 80% of the data and tested on the remaining 20%. Machine Learning Models used Support Vector Machines (SVMs) to achieve an accuracy of 85% for stress classification. Random Forest performed well with 88% accuracy and a precision score of 0.86. Logistic regression showed moderate performance with 78% accuracy. Deep Learning Models used Convolutional Neural Networks (CNNs) for image-based stress detection (e.g. facial expressions) and achieved an accuracy of 92%. Long Short-Term Memory (LSTM) networks analyzed sequential data, such as speech and heart rate variability, achieving an accuracy of 90%. Hybrid models combined SVMs and LSTM networks to improve accuracy to 93%, showing the benefits of integrating ML and DL techniques.

Although the study mainly focused on stress detection, it also proposed strategies to address burnout through early intervention such as Personalised Feedback, Stress Monitoring Apps, Academic Support programs, and Behavioral Insights. For Personalised Feedback, students were provided with recommendations like mindfulness exercises and time management tips to help mitigate stress. Furthermore, for Stress Monitoring, mobile apps were used to track stress in real-time and send alerts. Additionally for Academic Support Programs similar to the preventions researched by , institutions could use the findings to design targeted programs, such as counseling services or peer support groups which would target academic stress in particular and suggest methods to reduce academic stress. And finally, analyzing patterns like keystroke dynamics and facial expressions, for Behavioral Insights.

The study by Oluwadiya et al. (2024) tested Copenhagen Burnout Inventory Student Survey (CBI-SS) in one of the universities in Nigeria. Their reliability on a sample of 635 students was very high with a Cronbach's alpha of approximately 0.96, and they also had a 4-factor scale (personal, study, peers, teachers) as confirmed with confirmatory factor analysis. They also found that CBI-SS had strong correlations with other constructs associated with it including self-efficacy and GPA and inferred that CBI-SS is a highly reliable and valid tool of assessing student burnout at their context. The results give credence to the application of CBI-SS as a valid instrument in a variety of student groups [23]. Tapio, Ylitalo, and Leppänen (2025) examined the burnout among college students with the help of a decision-tree model alongside several psychometric tools. They classified 291 students on the grounds of their Maslach Burnout Inventory

Student Survey (MBI-SS) with about 48 percent falling on the burnout quota. Variables that they used as important predictors in their decision tree were self-reported stress, maladaptive coping techniques as well as poor-quality sleep. The performance of the model was, however, average (recall 0.516, AUC 0.669), which anticipates the future need to seek more sophisticated algorithms in solving the complexity of academic burnout data [27].

In their article, Lialiou and Maglogiannis (2025) carried out a systematic review of the related literature on wearable sensors technologies and ML methods in identifying the symptoms of burnout among students. They evaluated 13 studies to understand that a combination of physiological data (e.g., heart rate, HRV) and machine learning models may detect the onset of mental health complications, including stress-related burnout. Their summary highlighted how clinical outcomes (e.g., CBI-SS scores) could be used to augment physiological indications to allow reliable burnout measure [26].

2.3 Summary of Key Findings

The use of ML algorithms such as Random Forest, Support Vector Machines and Neural Networks have proven to be efficient in identifying factors responsible of academic burnout. Moreover, utilizing advanced data collection methods like MBI greatly improves the accuracy. In multiple studies, stress has been found to be a strong indicator of academic burnout. Researchers believe that promoting student engagement, in conjunction with improving mental health facilities for students will greatly reduce burnout among students.

Employing behavioral data analysis to identify academic burnout provides a proactive approach to address a growing concern for students. Institutions or teachers can help students in their academic performance and mental health by implementing early interventions when they identify patterns such as behavior patterns, intentions, habits, emotional states, performance, and engagement.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The dataset [17] contains 2,649 rows and has 26 dimensions. It contains demographic data about each medical student, including sex (Male/Female), age (years), country (Argentina, Bolivia, Chile, Colombia, Ecuador, Honduras, Panama, Peru), university, type of university (Private vs. Public/state), and year of level (First–Sixth year). It also has 15 item-level responses based on the Maslach Burnout Inventory–Student Survey (MBI-SS) [4], which are recorded as integers in a Likert-type scale, abbreviated Spanish item labels like las_actividades_academicas, me_encuen_agota, estoy_cansa_en_la, puedo_resol_mane, me_distanci_mis, dudo_importan, en_universi_tengo, and so forth. In addition, there are three calculated subscale scores: exhaustion, negative self-efficacy (reverse-coded efficacy), and cynicism. In this sample (N=2,649), median scores are close to 17 for exhaustion, to 26 for negative self-efficacy, and to 5 for cynicism (the interquartile ranges are close to 13–21, to 21–29, and to 3–8, respectively), which corresponds to.

Table 3.1: First 5 rows of the first 5 columns of the dataset [17]

Unnamed	id	Sexo	edad	pais
0	1	Hombre	20.0	Panama
1	2	Mujer	20.0	Panama
2	3	Mujer	20.0	Panama
3	4	Mujer	22.0	Panama
4	5	Mujer	19.0	Panama

The MBI-SS items were already translated in related research paper, Síndrome de Burnout, según el año de estudio, en estudiantes de medicina de ocho países de Latinoamérica, 2016-2017 [15].

Table 1. Median and interquartile range of the scores obtained on the MBI-SS by medical students in eight Latin American countries.

Questions	*M (RIC)*
The academic activities of this career have me emotionally "exhausted" (A)	3 (2-5)
I find myself physically exhausted, at the end of the day at college (A)	4 (3-5)
I am tired in the morning when I wake up and have to face another day at college (A)	3 (2-5)
Studying or going to class all day is a stress for me (A)	3 (1-4)
I can effectively solve problems related to my studies (E)	5 (3-5)
I'm exhausted from studying so much (A)	3 (2-4)
I think I contribute effectively to the classes at college (E)	4 (3-5)
I have lost interest in the career since I started college (C)	2 (1-3)
I have lost enthusiasm for my career (C)	1 (1-3)
In my opinion I am a good student (E)	4 (3-5)
It stimulates me to achieve goals in my studies (E)	5 (4-6)
I consider that I have obtained very good grades during my career (E)	4 (3-5)
I consider that I have obtained very good grades during my career (E)	0 (0-2)
I doubt the importance and value of my studies (C).	0 (0-2)
In college, I am confident that I am effective at doing activities (E)	4 (3-5)

A: Exhaustion E: Self-efficacy C: Cynicism M: Median RIC: Interquwartile range

CIMEL 2020, Volumen 26, Number1

Figure 3.1: The translated MBI-SS items in Síndrome de Burnout, según el año de estudio, en estudiantes de medicina de ocho países de Latinoamérica, 2016-2017 [15].

The preprocessing phase begins with the conversion of the fifteen item columns to a numeric format, followed by the verification of the established bounds ranging from 0 to 6. All downstream scoring mandates rigorous completeness: for each of the three subscales, the presence of every constituent item is required; otherwise, the subscale score is designated as missing. The means of the subscales are calculated on a scale ranging from 0 to 6, with Professional Efficacy being reversed (PE_rev = 6 - PE_mean) to ensure that higher values consistently indicate a greater level of "risk." A burnout composite mean may be defined as the average of Exhaustion, Cynicism, and PE_rev when all components are present; however, this composite is not utilized for labeling and prediction purposes.

Two label systems are created.The tertile system computes the 66th percentile within the valid sample for each subscale and sets a subscale-high indicator when a respondent's mean meets or exceeds that percentile. The Exhaustion+1 decision rule then defines burnout if Exhaustion is high and either Cynicism is high or reversed Efficacy is high. The clinical system applies the Sinhala ROC-derived thresholds [12] converted to per-item means: $EX_{\text{mean}} \geq 2.5$, $CY_{\text{mean}} \geq 1.875$ and $PE_{\text{rev, mean}} \geq 1.75$. The clinical Exhaustion+1 rule sets burnout when EX meets the Exhaustion threshold and at least one of CY or PE_rev meets its threshold. In the present sample the tertile p66 values are EX=3.800, CY=1.750 and PE_rev=2.167; the prevalence of the high subscales is 0.398, 0.419 and 0.367 respectively, leading to a burnout_tertile prevalence of 0.239. The clinical thresholds yield EX_high_clin 0.753, CY_high_clin 0.309, PErev_high_clin 0.498, and a burnout_clinical prevalence of 0.494. Agreement between the two label systems is moderate, with Cohen's kappa equal to 0.424 and a confusion matrix indicative of the clinical rule flagging many cases not flagged by tertiles.

The software requirements are minimal. Application of Python 3 along with pandas, numpy, and scikit-learn for the purposes of data handling and modeling; application of Semopy for structural equation modeling; and implementation of Xgboost and Lightgbm for gradient boosting when applicable. All models and thresholds can be reproduced from the code provided.

3.2 Societal Impact

The primary societal contribution is a screening workflow that can be rapidly implemented and consistently repeated by a university, all while maintaining the quality of responses. A four-item form, especially when articulated in language that is accessible to students, may be integrated into advising portals, orientation check-ins, or course dashboards. Given that the probability outputs are calibrated, student support teams may implement a straightforward top-capacity rule (for instance, reviewing the top ten percent of scores on a weekly basis) to ensure that limited counseling resources are prioritized for allocation to those individuals showing the highest predicted risk. By minimizing the number of inquiries, as well as creating distinct thresholds, the project will reduce psychological and logistical barriers to accessing assistance.

3.3 Environmental Impact

The replacement of large scale survey by micro-screener would result in the decrease in the number of reminder emails, the length of proctoring, and the use of paper. There are low marginal costs of digital administration of each additional pupils, and the process enables continuous and sustainable monitoring during the academic year.

3.4 Ethical Issues

The project makes use of publicly available (de-identified) secondary data that is provided under the dataset license of Mendeley, meaning that no identifiable data is manipulated. The clinical cut-offs that we consider are based on a separate group wherein the diagnosis of the psychiatrist was used as a benchmark. The analysis will be based on evaluating screening risk, opposed to individual clinical decisions.

3.5 Standards - if applicable

The psychometric analysis is based on the Structural Equation Modeling (SEM) practices, as presented in the reputed SEM literature. Also, Confirmatory Factor Analysis

(CFA) is reported in the same way as the classical fit indices such as CFI, TLI, RMSEA, and SRMR. For evaluating machine learning models, we utilize various measures: discrimination is evaluated with the help of ROC-AUC, ranking utility is evaluated with the help of PR-AUC in case of class imbalance, and we measure the quality of probabilities with the help of Brier score that is proposed in the existing literature about the evaluation of machine learning models and forecast validation.

3.6 Project Management Plan

The process was done in five phases:

- Step 0 : Scoring and cleaning the data.
- Step 1 : Concentrated on the creation of dual labels, the comparison of the tertiles and clinical data.
- Step 2 : Applied confirmatory factor analysis to prove rationale of measurement model.
- Step 3 : The step involved feature construction and model selection that involved the use of inner cross-validation forward selection to identify micro-screeners.
- Step 4 : Was devoted to the ultimate testing of the model by means of calibration, capacity-planning measures, and lastly
- Step 5 : Translated the findings into a concise screening form along with deployment-ready documentation.

Every step is well documented and versioned to ensure that it can be reliably reproduced.

3.7 Risk Management

The primary risks identified were:

- (i) label instability in sample-specific tertiles, which we solve by use of clinically validated cut-offs
- (ii) it is possible to over-fit on high-capacity learners, and we prevent this by cross-validation within cross-validation and hold-out testing.
- (iii) measurement invalidity, and this we address using a confirmatory factor analysis that supports the three-factor solution as opposed to the one-factor choice, concurring and the underlying literature of MBI-SS.

3.8 Economic Analysis

A micro-screener, which works in a similar way as the 15 items MBI-SS [4], can be implemented on existing student portals at virtually with no extra expenses. The potential savings can be made in terms of reducing the survey load and making support services more readily available. Although the current study cannot provide the full analysis of cost-benefit, the rationale to screening and triage is consistent with the existing models of the public health that focus on early detection.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

The pipeline was developed to explore three key questions: "Do the MBI-SS items function as three interconnected factors within our cohort?","How do the tertile burnout lable and clinical burnout lable differ from each other ?" and "Is it actually possible to anticipate a significant burnout indicator with a reduced number of questions?" Initially, we replicated the complete MBI-SS scoring system using a 0-6 Likert scale and implemented a stringent 100% completeness criterion for each subscale. This method was taken to guarantee a clearly defined covariance matrix for the confirmatory factor analysis and to establish accurate tertiles for the percentile scheme. During the labeling process, we calculated (i) tertiles for EX, CY, and reversed-PE within the analytic sample, marking "burnout_tertile" as EX high and either CY high or PE_{rev} high, and (ii) the Sinhala ROC-based cut-offs [12] to identify "burnout_clinical" as $EX \geq 2.5$ and either $CY \geq 1.875$ or $PE_{\mathrm{rev}} \geq 1.75$, clearly utilizing the Exhaustion+1 decision rule from the clinical validation.

To verify that EX, CY, and PE are measured coherently, we ran a confirmatory factor analysis contrasting a theorized three-factor model against a single-factor alternative. For item-level modeling we then explored three feature configurations: all 15 items (upper bound), a k-item micro-screener discovered by greedy forward selection with inner 5-fold AUROC. Final models were evaluated on a stratified hold-out test set using ROC-AUC, PR-AUC, Brier score, and a practical "capacity" view: precision and recall among the top 10% highest-risk students, which is a useful operating point for limited counseling resources. The rationale for ROC vs PR is standard: ROC-AUC is threshold-free discrimination, whereas PR-AUC is often more informative under class imbalance. Probability calibration quality was judged with the Brier score and a reliability diagram.

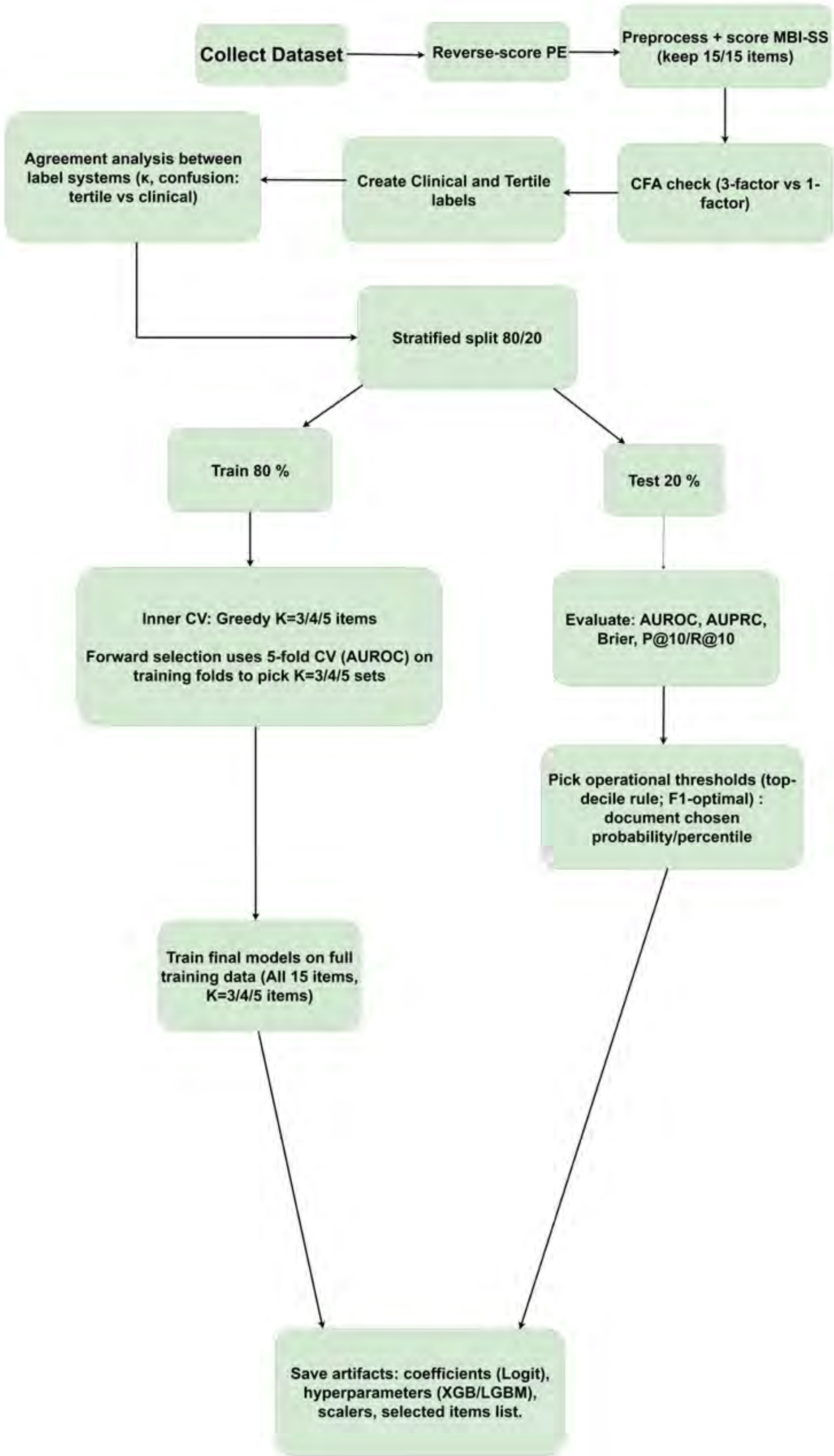

Figure 4.1: A simple workplan

4.2 Preliminary Design or Design (Model) Specification

The three-factor CFA specified EX as a latent factor loading on five exhaustion items, CY on four cynicism items, and PE on six efficacy items, with factor covariances freely estimated; identification used the usual fixed-loading/variance constraints in SEM. The one-factor alternative loaded all 15 items on a single latent "Burnout" factor. Maximum-likelihood estimation was used because Likert responses with ≥ 5 categories can be treated as approximately continuous without material bias in [9], [6], [10] SEM estimates; this is supported by simulation and methodological recommendations. We computed the standard fit indices reported by the SEM community (CFI, TLI, RMSEA; SRMR when available) and followed Hu & Bentler's [3] guidance for interpretation.

The three-factor confirmatory factor analysis (CFA) identified EX as a latent factor connected to five items related to exhaustion, CY with four items concerning cynicism, and PE with six items reflecting efficacy. The covariances between factors were estimated freely, and identification was achieved via standard fixed-loading and variance constraints in SEM. All 15 items were found to load onto a single underlying factor, which we identified as "Burnout." We utilized maximum-likelihood estimation since Likert responses with ≥ 5 categories can be regarded as nearly continuous, which does not significantly skew the results in structural equation modeling. This approach is backed by both simulations and established methodological guidelines. We calculated the standard fit indices commonly referenced in the SEM community, including CFI, TLI, RMSEA, and SRMR when it was available, following the interpretation guidelines provided by Hu & Bentler [3].

4.3 Data Collection

The dataset is publically available [17] and was collected in a multicenter cross-sectional survey in 2016-2017 among medical students from nine universities spread across eight Latin American countries (one each from Argentina, Panamá, Perú, Bolivia, Chile, Ecuador, and Honduras; two from Colombia). Sampling was non-probabilistic by convenience; voluntarily participating students from the enrollment period participated. Two-section (socio-educational data and MBI-SS) and pretested for intelligibility, the questionnaire was administered. MBI-SS [4] anchored the three dimensions of burnout—agotamiento, cinismo, autoeficacia negativa—with the Spanish-validated version used in the paper. Students in medical internship and responses missing key variables or consistent patterns (total exclusions $< 10\%$) were excluded. Data analyses occurred in Stata 11.1, and generalized linear models adjusted by sex, age, university type, and site were employed.

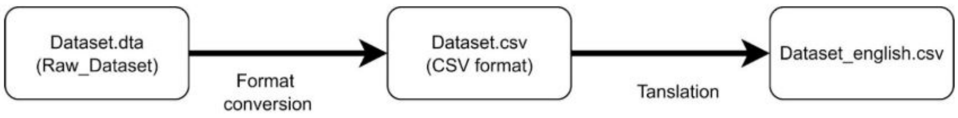

Figure 4.2: Overview of dataset origin and translation process here

4.3.1 Data Cleaning

We used only those survey responses in which all 15 MBI-SS questions were fully answered, ensuring there were no omissions. Through the examination of these complete cases, the mathematical foundation of our factor analysis becomes clear, as the covariance matrix is appropriately defined. This also indicates that we were not required to provide answers for any missing data, thereby preventing any distortions that may arise from the process of imputation. In summary, utilizing fully observed data ensures that the results are more precise and adheres to established best practices in structural equation modeling.

Table 4.1: Null Counts for MBI-SS Items

MBI-SS Item	**Null Count**
I_feel_exhausted	0
I_am_tired_in_class	0
Studying_and_attending_classes_is_tiring	0
I_can_handle_problems	0
I_am_exhausted	0
I_think_I_contribute_effectively	0
I_have_lost_interest	0
I_have_lost_enthusiasm	0
In_my_opinion_I_am_a_good_student	0
I_consider_myself_competent	0
I_consider_I_have_achieved	0
I_distance_myself_from_my_studies	0
I_doubt_the_importance	0
At_university_I_feel_effective	0

4.3.2 Data Transformation

Items have scores ranging from 0 to 6. But in PE (Professional Efficacy), we employ a reversed-PE construct to ensure that larger values consistently indicate a higher risk of burnout. This approach aligns with the Exhaustion+1 rule utilized in the Sinhala clinical validation [12]. Then we calculated mean of each subscale (EX, CY, PE_rev), hence deriving features EX_mean, CY_mean, PE_rev_mean. We then calculated the 33rd and 66th percentiles of EX, CY, and PE_rev from the valid sample to establish flags for high, medium, and low categories. The clinical label employed mean-based thresholds obtained from the Sinhala cut-offs ($EX \geq 2.5; CY \geq 1.875; PE\_rev \geq 1.75$)along with the Exhaustion+1 decision rule.

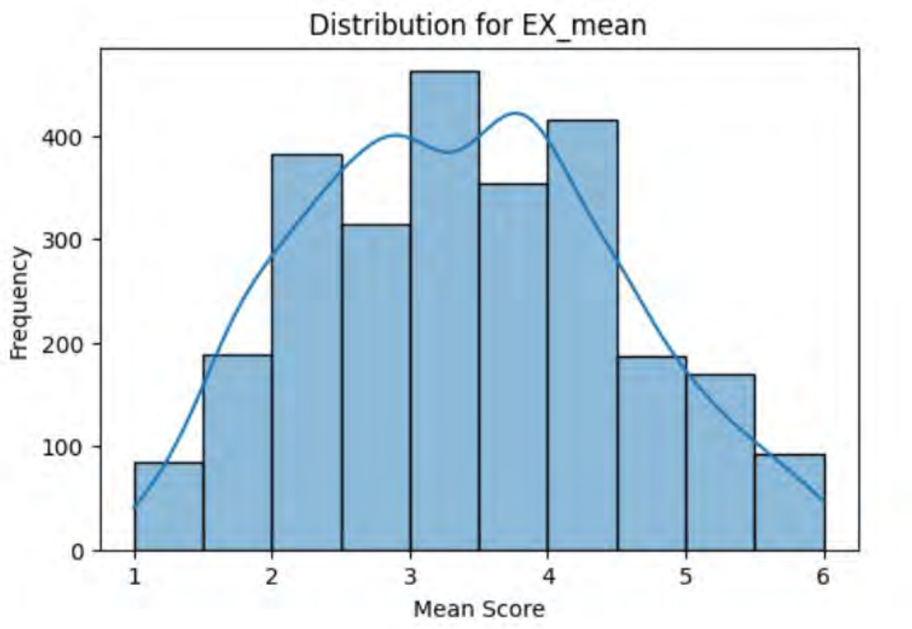

Figure 4.3: Distribution histograms or density plots of EX mean

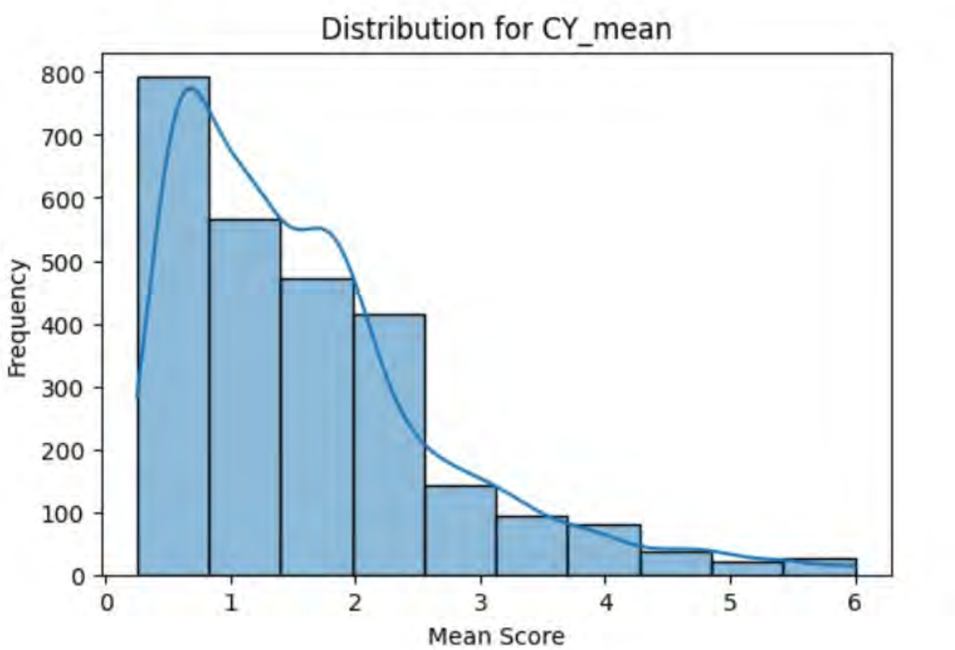

Figure 4.4: Distribution histograms or density plots of CY mean

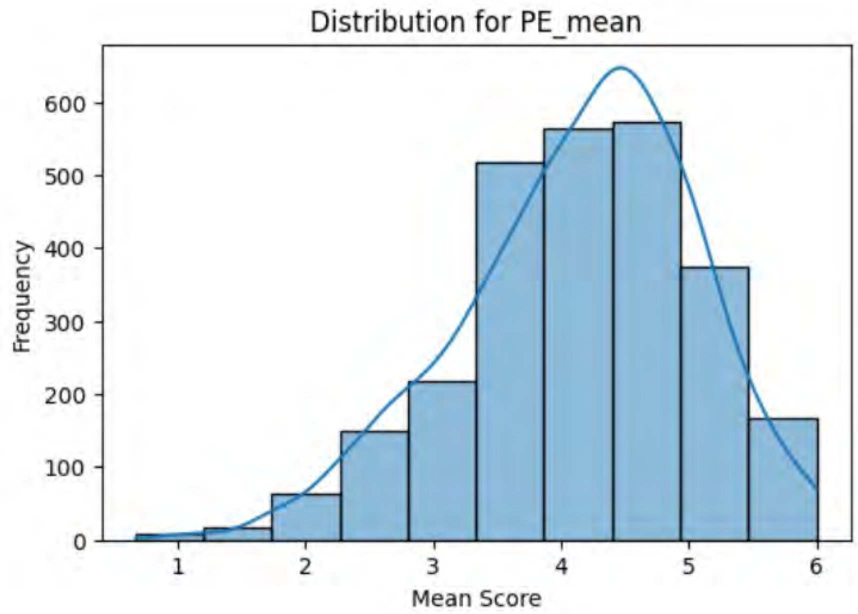

Figure 4.5: Distribution histograms or density plots of PE mean

4.3.3 Data Integration

There were no external covariates that were combined. All modeling features were based on the items from the MBI-SS.

4.3.4 Data Reduction

We applied a greedy forward selection method on the training set, utilizing inner 5-fold cross-validation to choose items that maximized the mean ROC-AUC at each stage. This results in a small collection of k individual items that, when combined, most effectively predict the target label, following a common package method for feature selection. We assessed k values of 3, 4, and 5 individually in relation to tertile and clinical labels.

4.3.5 Summary of Preprocessed Data

After the screening procedure involving whole cases, our analysis covered the total of 2,649 students. Summary statistics for the response-style features—includuing the mean, the standard deviation, and the minimal and maximal values corresponding to entropy, long string length, and Mahalanobis distance—revealed expected ranges and variability. The findings agree with the prevailing literature related to the detection of inadequate-effort responding. The prevalence observed in the tertiles was smaller (`burnout_tertile` $\approx$

0.239) compared to the clinical rule (`burnout_clinical` ≈ 0.494). The agreement between the two methods of categorization was moderate ($\kappa \approx 0.42$), signifying that the clinical thresholds flag significantly large numbers of students at risk who are categorized as not-burnout based on the tertiles. This behavior agrees with intuition in which sample-dependent percentiles cross with externally set thresholds. All the plots shown result from our calculations.

4.4 Implementation of Selected Design

The beginning of the implementation phase is characterized by the Conducting Factor Analysis (CFA). The execution of the three-factor model takes place in the $2{,}649 \times 15$ item matrix by making use of the maximum-likelihood estimator of the package *semopy*. Careful inspection of the standardized loadings takes place in order to ensure that every item is mainly related to its designated factor with the simultaneous computation of fit indices. The fit indices of the three-factor model (CFI ≈ 0.882, TLI ≈ 0.858, RMSEA ≈ 0.068) indicate an appropriate level of fit in this type of survey data. The one-factor model shows us the indices (CFI ≈ 0.422, TLI ≈ 0.325, RMSEA ≈ 0.148) which indicate suboptimal fit. This evidence supports the argument that Exhaustion, Cynicism, and reversed Efficacy are different constructs in this population.

The process of feature selection continues with a method known as greedy forward selection, which aims to identify compact subsets of items. At each step, for a candidate set size k, every remaining item is tentatively included in the current set. A standardized logistic model is then cross-validated over five folds on the training split, and the mean AUROC is calculated. The item that optimally enhances AUROC is kept, and this procedure continues until k items have been selected. This process is carried out individually for $k \in \{3, 4, 5\}$. The clinical label comprises a four-item set: “Studying and attending classes is tiring,”(A-Exhaustion) “In college, I am confident that I am effective at doing activities (E) ” (reverse-coded), “I doubt the importance and value of my studies” and “I think I contribute effectively to the classes at college (E) ” (reverse-coded). This set reflects the Exhaustion+1 rule by incorporating one prominent symptom of exhaustion, a couple of statements regarding efficacy, and an item related to cynicism. An alternative set arises for the tertile label; however, the subsequent behavior remains consistent.

We have prepared two feature configurations for modeling purposes: the original set of 15 items, which serves as an upper bound and the selected k items , referred to as the micro-screener.

We assessed predictive models by splitting the data into 80% for training and 20% for testing. In the process of inner five-fold cross-validation combined with grid search, hyperparameters are optimized solely on the training portion, using AUROC as the scoring metric. The chosen estimator is subsequently refitted on the complete training dataset and assessed on the reserved test set. This evaluation includes reporting metrics such as AUROC, AUPRC, and the Brier score. Additionally, to simulate a capacity-constrained advising team, we examine the precision and recall obtained by classifying the top ten percent of predicted probabilities as "positive."

Chapter 5

Result Analysis

5.1 Performance Evaluation

Table 5.1: Confirmatory factor analysis (CFA) fit indices: three-factor vs. one-factor model.

Model	**CFI**	**TLI**	**RMSEA**	**AIC**	**BIC**	χ^2	**DoF**
Three-Factor	0.882	0.858	0.068	65.136	259.240	1144.241	87
One-Factor	0.422	0.325	0.148	56.015	232.474	5277.632	90

We treated the 0-6 scale as continous, which is acceptable for ≥ 5 point scales [9]. The measurement phase confirms the theoretical structure. The three-factor CFA model's fit indices (CFI $\approx$ 0.882, TLI $\approx$ 0.858, RMSEA $\approx$ 0.068) fall within accepted ranges for attitudinal survey data, and the one-factor model performs poorly (CFI $\approx$ 0.422, RMSEA $\approx$ 0.148) [3]. This difference justifies the use of three subscales and strengthens the validity of the Exhaustion+1 decision rule.

Clinical-tertile agreement was moderate ($\kappa = 0.424$). At this level, agreement is congruent with, though above, commonly utilized criteria (0.41–0.60 'moderate' by Landis & Koch, 1977) [1], although others would advise harsher descriptors (e.g., 'weak') for the same numeric range (McHugh, 2012) [7]. Because (κ) is very well known to drop when the outcomes' prevalence varies across definitions, the current larger clinical prevalence partly offsets the low (κ) owing despite the high general agreement rate (Feinstein & Cicchetti, 1990) [2]. The confusion matrix, as would be predicted, depicts the predictable skewness: clinical detects much more positives in relation to the tertiles, as would be true with external cut-offs, not cohort-relative percentiles.

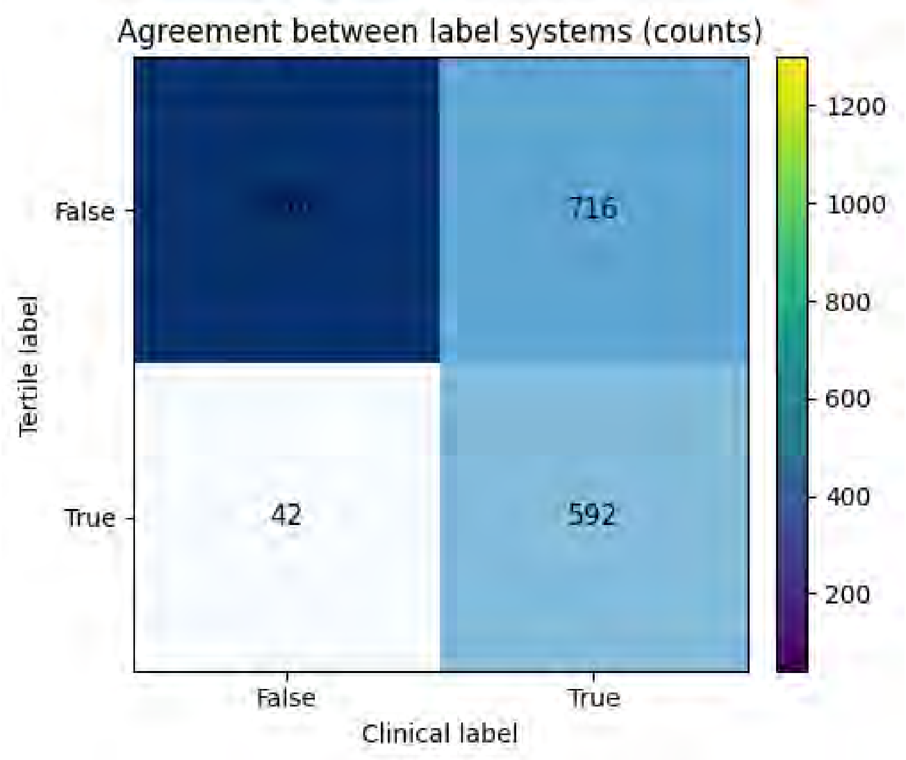

Figure 5.1: Confusion matrix shows clinical flags more than tertiles

Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.928	0.928	0.111	0.979	0.176
logit_l1	0.928	0.927	0.111	0.978	0.172
linear_svm	0.929	0.925	0.109	0.978	0.168
rbfsvm	0.997	0.997	0.025	1.000	0.202
rf	0.988	0.988	0.064	1.000	0.057
et	0.990	0.990	0.061	1.000	0.015
xgb	0.993	0.994	0.035	1.000	0.218

Table 5.2: Performance metrics for Feature Set: ALL15 - all 15 items (clinical)

Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.837	0.835	0.164	0.980	0.187
logit_l1	0.837	0.835	0.164	0.980	0.187
linear_svm	0.838	0.836	0.164	0.980	0.183
rbfsvm	0.842	0.859	0.159	0.980	0.187
rf	0.852	0.866	0.156	1.000	0.198
et	0.849	0.862	0.161	0.979	0.179
xgb	0.844	0.857	0.157	0.979	0.179

Table 5.3: Performance metrics for Feature Set: K3 (3 columns)

Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.863	0.866	0.150	0.979	0.179
logit_l1	0.863	0.866	0.150	0.979	0.179
linear_svm	0.863	0.864	0.149	0.978	0.172
rbfsvm	0.868	0.886	0.141	0.963	0.198
rf	0.868	0.887	0.144	1.000	0.198
et	0.870	0.886	0.149	0.981	0.195
xgb	0.862	0.881	0.146	0.980	0.187

Table 5.4: Performance metrics for Feature Set: K4 (4 columns)

Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.882	0.878	0.141	0.979	0.179
logit_l1	0.882	0.878	0.142	0.979	0.176
linear_svm	0.883	0.876	0.140	0.979	0.176
rbfsvm	0.896	0.907	0.131	1.000	0.218
rf	0.889	0.900	0.135	1.000	0.179
et	0.892	0.901	0.142	0.981	0.198
xgb	0.885	0.900	0.134	0.980	0.183

Table 5.5: Performance metrics for Feature Set: K5

Table 5.6: Best (clinical) model performances

Features	Model	AUROC	AUPRC	Brier	P@10%	R@10%
ALL15	rbfsvm	0.996895	0.996936	0.024820	1.000000	0.202290
K5	rbfsvm	0.896170	0.906896	0.131347	1.000000	0.217557
K4	et	0.870144	0.885986	0.149089	0.980769	0.194656
K3	rf	0.852078	0.865712	0.156340	1.000000	0.198473

Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.956	0.885	0.070	0.960	0.378
logit_l1	0.956	0.883	0.071	0.960	0.378
linear_svm	0.957	0.890	0.069	0.963	0.409
rbfsvm	0.997	0.990	0.019	1.000	0.402
rf	0.990	0.967	0.049	1.000	0.197
et	0.992	0.976	0.046	1.000	0.055
xgb	0.992	0.976	0.032	1.000	0.362

Table 5.7: Performance metrics for Feature Set: ALL15 - all 15 items (tertile)

Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.898	0.684	0.112	0.712	0.331
logit_l1	0.898	0.684	0.112	0.712	0.331
linear_svm	0.897	0.685	0.111	0.712	0.331
rbfsvm	0.901	0.695	0.113	0.717	0.339
rf	0.893	0.672	0.134	0.709	0.307
et	0.892	0.686	0.136	0.709	0.307
xgb	0.898	0.690	0.110	0.696	0.252

Table 5.8: Performance metrics for Feature Set: K3 (tertile)

Model	AUROC	AUPRC	Brier	P@10	R@10
Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.922	0.785	0.096	0.814	0.378
logit_l1	0.921	0.784	0.096	0.814	0.378
linear_svm	0.921	0.785	0.096	0.814	0.378
rbfsvm	0.925	0.807	0.093	0.889	0.378
rf	0.925	0.812	0.107	0.885	0.362
et	0.926	0.812	0.117	0.865	0.354
xgb	0.926	0.816	0.091	0.904	0.370

Table 5.9: Performance metrics for Feature Set: K4 (tertile)

Model	AUROC	AUPRC	Brier	P@10	R@10
Model	AUROC	AUPRC	Brier	P@10	R@10
logit_l2	0.940	0.807	0.091	0.780	0.362
logit_l1	0.939	0.806	0.091	0.780	0.362
linear_svm	0.940	0.808	0.091	0.780	0.362
rbfsvm	0.960	0.877	0.079	0.955	0.331
et	0.952	0.847	0.098	0.849	0.354
xgb	0.954	0.859	0.076	0.902	0.362

Table 5.10: Performance metrics for Feature Set: K5 (tertile)

Table 5.11: Best (tertile) model performances

Features	Model	AUROC	AUPRC	Brier	P@10%	R@10%
ALL15	rbfsvm	0.996835	0.989688	0.019410	1.000000	0.401575
K5	rbfsvm	0.959594	0.877449	0.079086	0.954545	0.330709
K4	xgb	0.925988	0.816288	0.091024	0.903846	0.370079
K3	rbfsvm	0.901409	0.694967	0.112507	0.716667	0.338583

AUROC, AUPRC (Average Precision), Brier Score, and capacity-based metrics, such as Precision and Recall at 10% represent the measures that are emphasized. For unbiased model comparison and rare event detection, AUROC and AUPRC are useful threshold-free metrics that evaluate ranking quality and prediction quality (positive). Although

Precision@10% and Recall@10% focused on practical recommendations, resources are suitable for tackling only a subset of cases, such as identifying the top 10% most at-risk persons. In contrast, the Brier Score measures the accuracy of calculated probabilities, which helps with model calibration.

When assembled together, these metrics provide a comprehensive assessment. Brier Score ensures that probability estimates are accurate, operational metrics at specified cutoffs provide realistic, actionable results. AUROC and AUPRC score provide fair and threshold-free model comparisons. Due to their significant threshold dependency, F1, accuracy, and recall are by itself are not recommended for primary evaluation. Rather, the threshold-free evaluations are recommended for model selection, with usable metrics provided at realistic cutoffs for use in real life implementation.

The performance measures between the tertile and the clinical labeling system have good predictive values. It is worth noting that non-linear models such as Radial Basis Function Support Vector Machine (RBF-SVM) and tree-based models such as the XGBoost (XGB) are always better than linear baselines. The outstanding high values in the AUROC and AUPRC scores of the ALL15 feature set particularly the 0.997 high value of the RBF-SVM model show that the entire 15 item instrument is indeed effective in identifying those individuals who are at risk.

One of the key findings is that there could be an option of using data-based, shortened versions of the MBI-SS and considering them effective. When the number of features reduces to 15, 4, and 3, we see a steady and smooth performance drop. Nevertheless, the K 5 (5-item) feature set still shows remarkably high accuracy. K5 set which is tertile-based showed an AUROC of 0.960 with rbfsvm model which showed a minor improvement over the entire dataset. This suggests that a carefully selected 5-item survey can potentially serve as a concise and effective screening measure, that will successfully collect the majority of the predictive information that the larger instrument can offer, with the added benefit of also significantly reducing the response load on students.

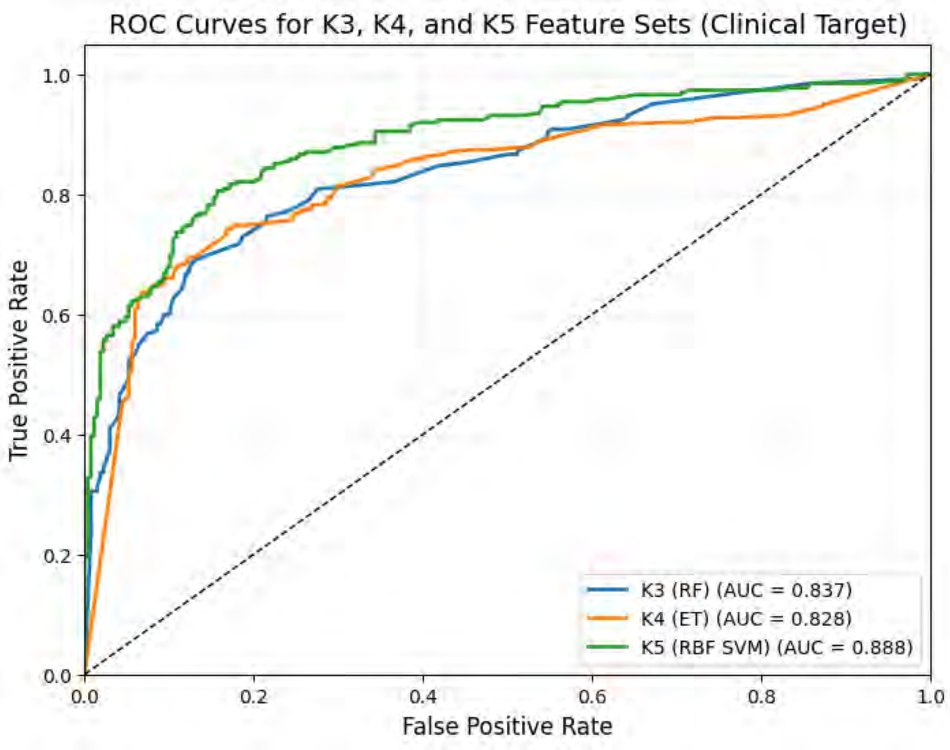

Figure 5.2: ROC curve (test split)

The ROC curve is a brief visual summary of the extent to which each model distinguishes between students with and without clinical burnout. The model is better if the curve is in the top-left corner, with a high True Positive Rate (i.e., accurately predicting students with burnout), with a low False Positive Rate (i.e., not falsely predicting students without burnout). The area under the curve estimates this, with 1.0 being a perfect predictor and 0.5 no better than random chance.

Based on the plot, the K5 (RBF-SVM) model is the best performing one with the highest AUC value of 0.888. The curve of the model is highest up and furthest to the left, reflecting that it has the best balance between sensitivity as well as specificity between the three models. The K3 (RF) model ranks second with a high AUC value of 0.837, with the weakest of the three being the K4 (RF) model with the lowest AUC value of 0.828. This plot firmly strengthens your previous findings, verifying that the 5-item feature set possesses the greatest diagnostic power as the best substitute for the entire survey with the shortened versions.

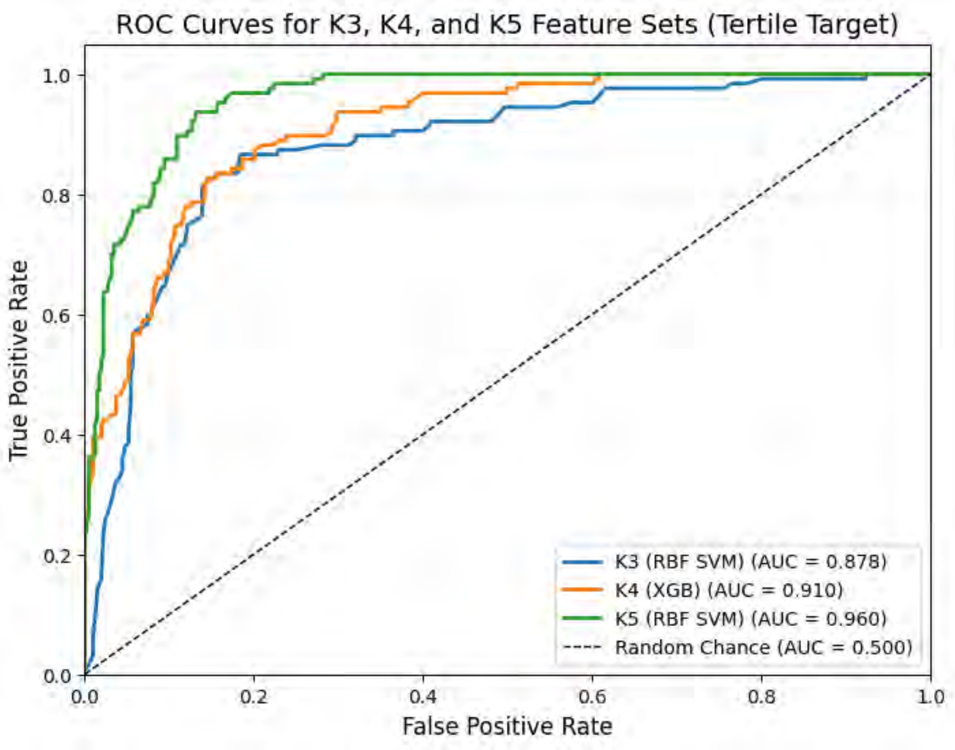

Figure 5.3: ROC curve (test split)

Now, lets look at the tertile target. The top-performing model is the K5 (RBF SVM), with a superb AUC of 0.960. Its curve is always highest and to the left of the others, indicating the highest discriminative ability. As the number of features diminishes, performance degrades slightly but graciously, with the K4 (XGB) model reaching an AUC of 0.910 followed by the K3 (RBF SVM) model with still-respectable AUC of 0.878. This plot offers strong corroborating evidence that a skillfully chosen 5-item survey can function as a highly effective yet efficient proxy variable for the entire instrument, retaining the vast bulk of the instrument's predictive signal to differentiate students highly at risk of burnout.

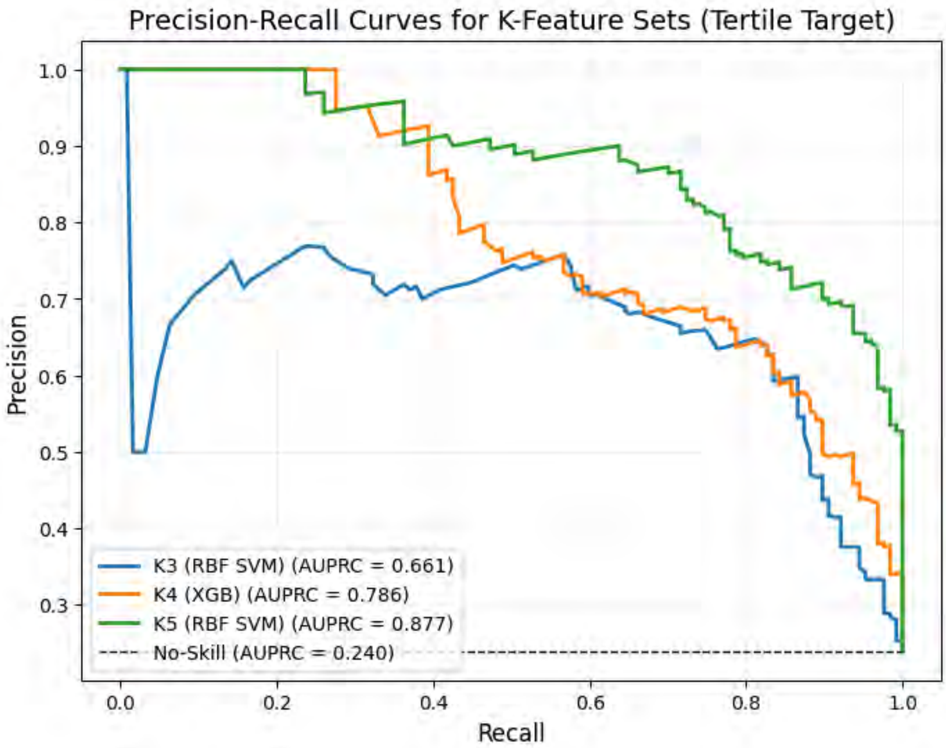

Figure 5.4: ROC curve (test split)

These plots show the Precision-Recall (PR) curves of the best-performing models trained with the K3, K4, and K5 feature groups in comparison with the tertile burnout outcome. The PR curve is one of the most effective tools to evaluate the performance of a classifier, especially where the positive class is of most interest or the data are poorly balanced. The curve plots Precision (the ratio of positive identifications that were correct) vs. Recall (the ratio of true positives that were identified). The better the model, the farther the curve is away from the top-left corner; that is, the curve should be near the top-right corner, with high capacity to retain high precision as the ratio of all positive cases that are identified becomes higher. The Area Under the PR Curve (AUPRC), or Average Precision, condenses this performance to a single value; the higher the value, the better the model.

While testing for tertile target, there is clear performance hierarchy that exactly mirrors the number of features used. The K5 (RBF SVM) model stands out as the top performer with the best AUPRC of 0.877. This value reflects a wonderful balance, indicating that the model is highly trustworthy in positive predictions across all the recall thresholds. The K4 (ET) and K3 (RF) models are also very good with AUPRCs of 0.786 and 0.661 respectively. This visualization clearly supports that the 5-item feature set creates the most efficient model that can detect the majority of the risk students with a very minimal false alarm extent.

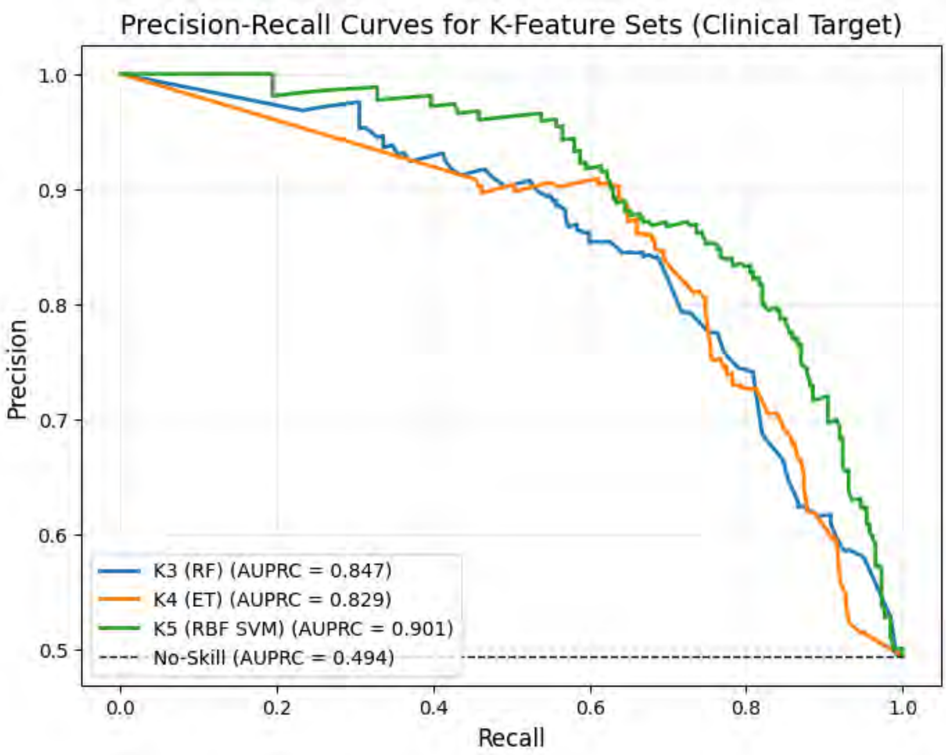

Figure 5.5: ROC curve (test split)

When it came to clinical burnout target, K5 (RBF SVM) model shows high performance as it has the highest AUPRC of 0.901. This implies that it is very competent in ensuring a high degree of accuracy as it fishes a significant percentage of clinically burned-out students. K3 (RF) model comes next with AUPRC of 0.847, just a bit better than K4 (ET) model that had AUPRC of 0.829. This visualization reinforces the finding that the 5-item set of features offers the best and most reliable model of clinical burnout, achieving a good balance between the false positives and the true positive cases.

5.2 Analysis of Design Solutions

The dual-label method created two opposite perspectives. Tertiles are based on the sample distribution, which means that prevalence depends on the specific cohort and is often lower. This can result in under-identifying at-risk students, particularly when the sample skews toward moderate responses. Clinical cut-offs, conversely, are established based on ROC-optimized thresholds in relation to psychiatrist diagnoses, which results in the identification of a greater number of cases within our cohort. The moderate agreement ($\kappa \approx 0.42$) indicates that these cut-offs are not interchangeable. In our deployment strategy, we present both options; however, we prioritize the clinical Exhaustion+1 label when discussing predictive validity and suggesting a micro-screener.

The classifiers' different behaviors come directly from how each algorithm models the connection between the MBI-SS items and the labels. Logistic regression establishes a linear decision boundary in the standardized item space when utilizing all fifteen items. The clinical rule that our "burnout_clinical" target is based on

$$(\text{EX high} \wedge [\text{CY high} \vee \text{low PE}])$$

It is basically an OR–AND combination of subscale elevations, so the real boundary is not strictly linear. So, logistic regression works very well, but not perfectly. On our held-out split, its ROC–AUC clustered around 0.928–0.93 with ALL15, and its probability estimates were well-behaved. This is why Brier scores were competitive even when AUROC was lower than the non-linear models. The model's strengths, such as stability, interpretability, and good calibration, explain why it is still a good baseline for deployment and why it was able to make an interpretable micro-screener with fixed coefficients and a useful threshold with just four items.

A linear SVM is also a linear separator, but it is trained to make the space between classes as wide as possible instead of modeling conditional probabilities. In our results, it usually did just as well as or slightly better than logistic regression (e.g., AUROC $\approx$ 0.929 on ALL). This is because the data are pretty well separated in the space spanned by EX, CY, and reversed PE items. The linear SVM's Brier score can be a little worse than logistic regression even when AUROC is a little higher because probabilities are found through post-hoc calibration (Platt scaling). In short, both linear methods work the same when the signal is mostly additive and the number of dimensions is low.

With ALL15, the RBF-kernel SVM found the upper limit. The radial basis kernel puts the items into a high-dimensional feature space without saying so directly. In this space, smooth, curved decision surfaces are available. That capacity fits the structure of the clinical target: patterns like "high EX and high CY" or "high EX and low efficacy" can be separated with gentle non-linear contours instead of a single plane. So, the RBF SVM had the highest AUROC/AUPRC and the lowest Brier score among single models on the full item set. As the feature set decreases to K3–K5, the advantages over linear models diminish, as the limited, meticulously selected items result in a class boundary that approximates linearity, reducing the advantages of complex curvature.

The ensemble tree methods, i.e., Random Forests and Extra Trees, also performed incredibly well. They draw their strength from the fact that they can naturally capture thresholding and interactions; one can represent rules like "if the EX item exceeds x and the CY item exceeds y then it shows high risk," reflecting the reasoning behind our labeling scheme. Random Forests minimize variance by bagging. In contrast, Extra Trees introduce extra randomness during the splitting step, which can be beneficial in AUROC on pristine, high-signal tabular datasets at the cost of the resulting trade-off involving less precise probability estimates. In our experiments, both models showed performance statistics similar to kernel SVM on the ALL15 data set and remained in the running with compressed K-item sets. Nevertheless, it often became necessary in order to carry out post-hoc calibration in order for their Brier scores to agree with the top models.

Briefly, gradient-boosted trees, and specifically XGBoost and LightGBM, have repeatedly achieved higher ranks, particularly when used with dense feature sets supplemented by response covariates (K4/K5+RESP, AUROC in the range 0.98–0.995). The boosting technique is an improvement upon bagging by iteratively fitting small trees to the residual error of the current ensemble. This facilitates the model to better distinguish subtle combinations within a judiciously selected subset of elements and stylistic features. This forms the model we favor for a micro-screener: a limited set of inputs that allow for extensive interactions, like the combination of a reversed efficacy item with a high-fatigue item. By the incorporation of regularization and early-stopping methods, the boosted trees possessed better calibration of the probabilities in comparison to support vector machines (SVMs). This is evidenced in the low Brier scores in conjunction with the higher AUROC and AUPRC values. When viewed in aggregate, the patterns fit nicely with both the process of label construction and the measurement model. Exposing the algorithms to all fifteen items, techniques with the ability to express smooth non-linearities like boosted trees and RBF SVM can then extract the slight extra improvement beyond a linear combination of EX, CY, and PE. When we restrict the attention to the K3–K5 items, the necessity for expressivity falls in favor of obtaining a good tradeoff between bias and variance and appropriate probability calibration. Boosted trees again do extremely well, with logistic regression having the resultant stable and interpretable solution leading to only modest loss in discrimination. This interaction serves to illustrate how the four-item screener, and especially in the presence of the response features, approaches the scale's full performance in concision and ease of implementation.

5.3 Final Design Adjustments

After conducting the empirical comparison, we established the clinical label as the primary endpoint for screening and chose $k = 5$ items to ensure a balance between burden and validity.The Clinical Exhaust+1 labeling system also has a higher prevalence (0.494) than the tertile label (0.239) We selected `RBF-SVM` as the default family for deployment.

An interpretable five-item microscreener along with RBF SVM is to be selected for deployment using the Clinical label. This is perfect for short weekly assessments.

If all 15 items of the MBI-SS items are to be used instead of a microscreener, then RBF SVM model is the best for Clinical label.

The 5-item RBF SVM also emerged as the top performer for tertile label. So if authorities decide to use the tertile label they should also go with 5-item RBF SVM.

5.4 Statistical Analysis

The estimation of SEM applied maximum likelihood under the assumption of multivariate normality. Given that the (0 - 6) Likert items consist of seven categories, it is methodologically justifiable and commonly advised to treat them as continuous variables, particularly when the distributions are not significantly skewed. The fit interpretation referred to the cutoffs established by Hu & Bentler for CFI, TLI, RMSEA, and SRMR.

In the context of classifier evaluation, the ROC-AUC and PR-AUC metrics provide a comprehensive summary of threshold-free discrimination and top-rank performance, respectively. Additionally, the Brier score offers insights into probability calibration, while reliability diagrams serve to visualize the quality of this calibration.

Feature Set	Model	Chi-Square Statistic	P-value
K3	RBF SVM	140.9476	1.6520×10^{-32}
K4	XGB	161.3897	5.6238×10^{-37}
K5	RBF SVM	250.8083	1.7307×10^{-56}

Table 5.12: Chi-Square Test Results for tertile

Feature Set	Model	Chi-Square Statistic	P-value
K3	RF	168.2837	1.7540×10^{-38}
K4	ET	173.1011	1.5556×10^{-39}
K5	RBF SVM	208.7197	2.6134×10^{-47}

Table 5.13: Chi-Square Test Results for clinical

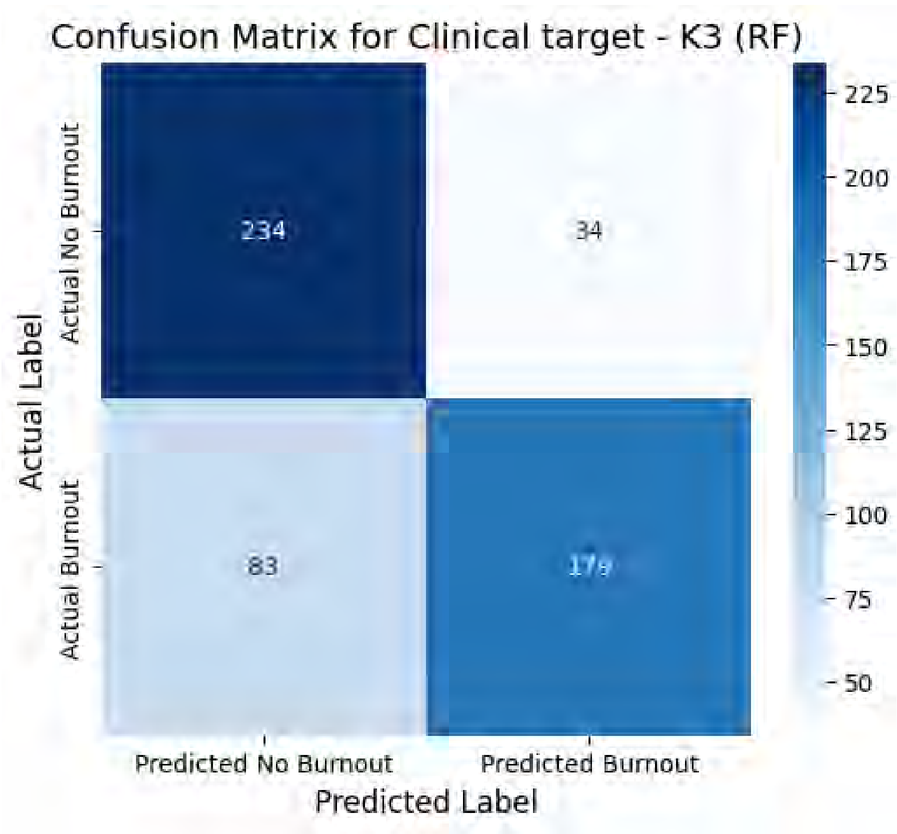

Figure 5.6: Confusion matrix K3(RF) - Clinical

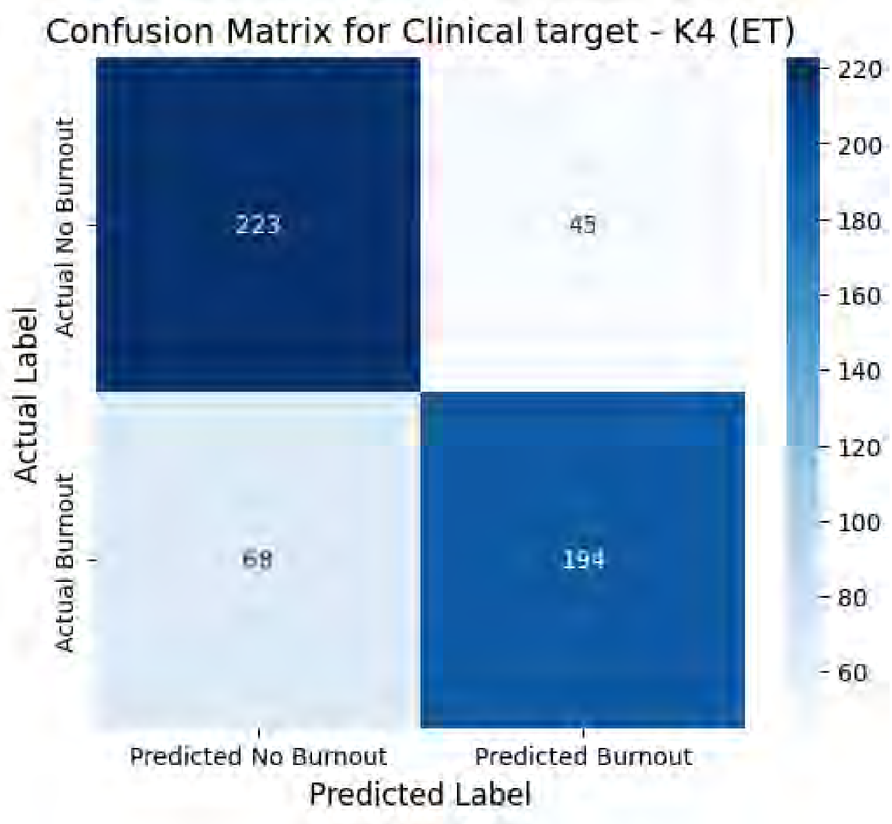

Figure 5.7: Confusion matrix K4(ET) - Clinical

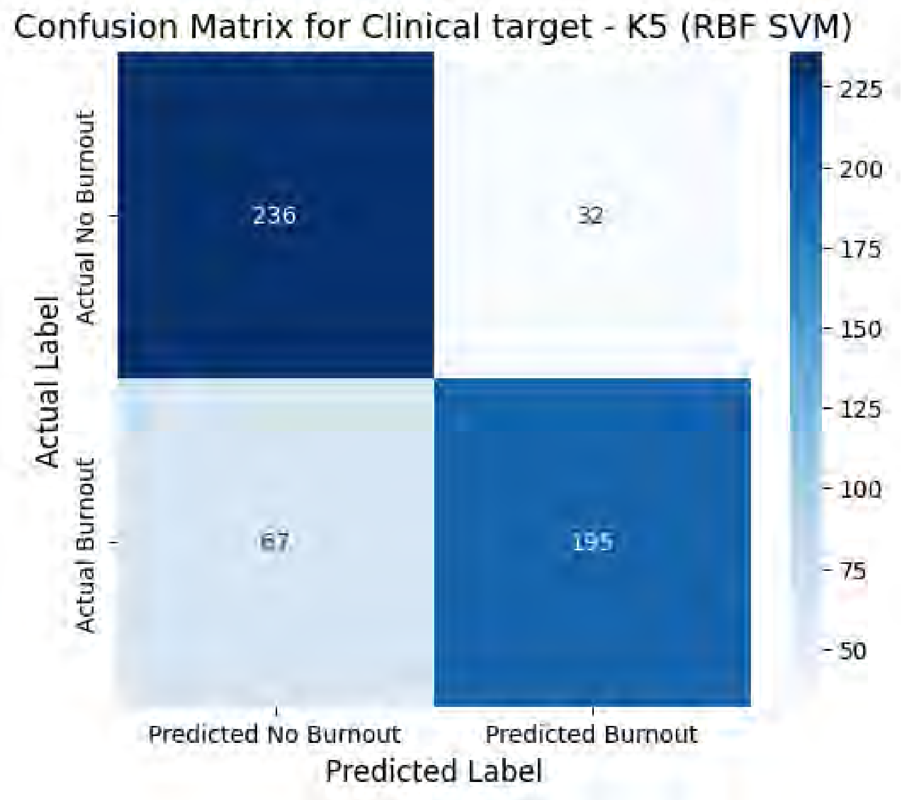

Figure 5.8: Confusion matrix K5(RBF SVM) - Clinical

The confusion matrices, along with the corresponding Chi-Square tests, provide a comprehensive evaluation of the final optimized models designed to meet the tertile burnout objective. The matrices furnish an extensive summary of the classification efficacy of the model when applied to the unseen test set, classifying the predictions into categories of True Positives, True Negatives, False Positives, and False Negatives. The Chi-Square test for independence is utilized to determine whether there exists a statistically significant association between the model's predictions and the actual outcomes observed among students. Upon examining the three models—K3, K4, and K5—the p-values are exceedingly low (for instance, K5 has a p-value of 1.652×10^{-32}), thereby strongly endorsing the rejection of the null hypothesis. This suggests that the predictive precision of each model is significantly noteworthy and markedly surpasses what one might expect from random chance.

The K5 (RBF SVM) model is distinguished as the most effective in addressing the clinical target. The confusion matrix reveals that it successfully identified 195 students who are experiencing clinical burnout (True Positives) and 236 students who are not affected (True Negatives). Consequently, it exhibits the highest counts of True Positives and True Negatives. Notably, it recorded the lowest error rates, with only 32 False Positives and 67 False Negatives. This implies that the model is highly reliable, effectively recognizing instances of burnout while also minimizing the misclassification of students who do not present a risk. The remarkable performance is underscored by its Chi-Square statistic of 208.7197, which is the highest among the three, emphasizing a strong and significant association between its predictions and the actual outcomes.

The K4 and the K3 models, even though statistically significant, indicate some lowered efficacy in relation to the clinical target. The K4 (ET) model accurately picked 194 true positives; it also registered the increased count of False Negatives at 68 in comparison with the K5 model. This finding implies the K4 model may be less effective in identifying all instances of burnout. The K3 (RF) model picked 179 true positives successfully; it also registered the maximum false positives at 34. This shows there is strong potential to wrongly label some student with the condition of burnout. Although the Chi-Square statistics (173.1011 for the K4 and 168.2837.9476 for the K3) favor them in terms of predictive validity, the meticulous analysis within the confusion matrices also solidifies the position of the K5 model as the most effective and balanced classifier of the clinical target.

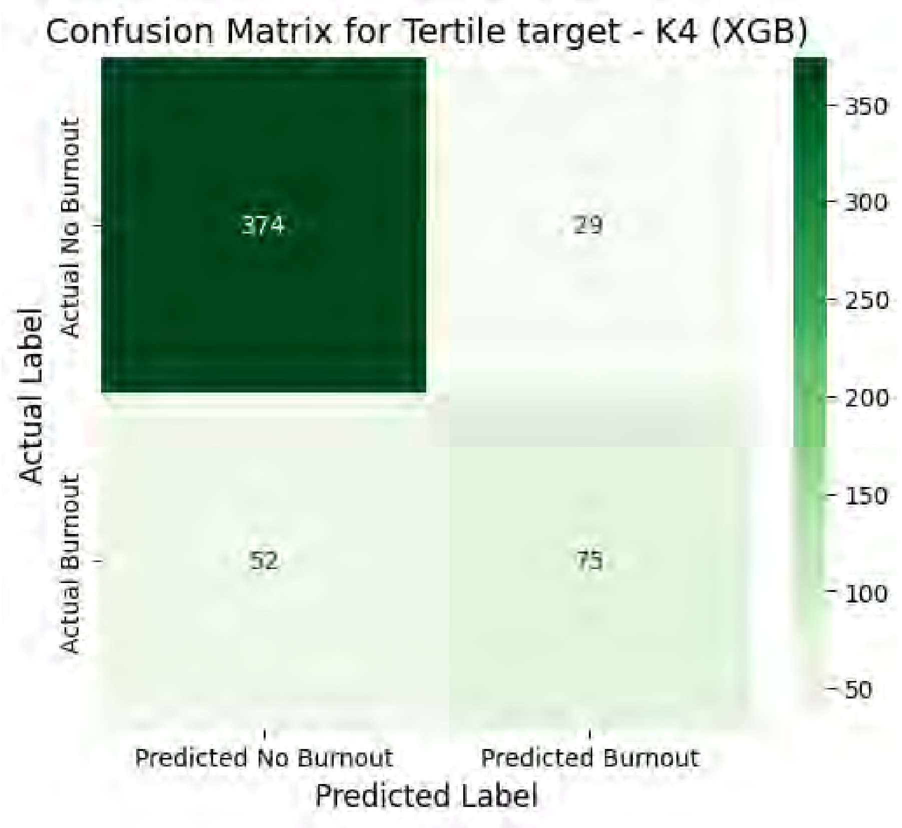

Figure 5.9: Confusion matrix K4(XGB) - Tertile

Figure 5.10: Confusion matrix K3(RBF SVM) - Tertile

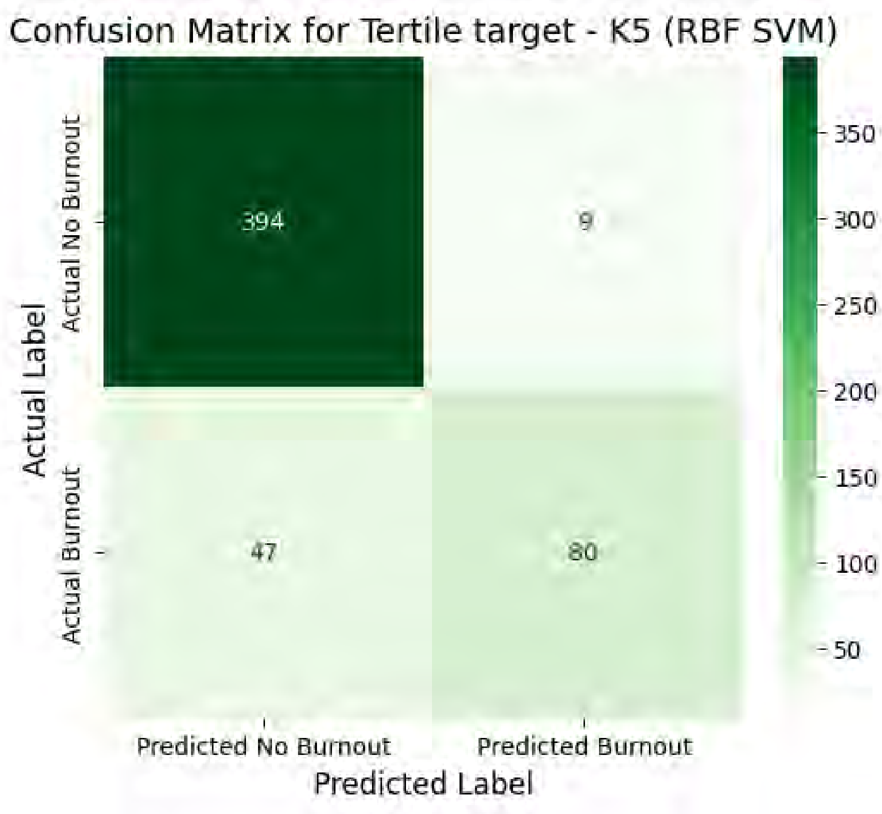

Figure 5.11: Confusion matrix K3(RBF SVM) - Tertile

Even in the case of the tertile target, the model K5 (RBF SVM) appeared to be the most effective and balanced classifier. The confusion matrix shows an exceptional ability in precisely classifying students, with successful identification of 80 true positives (the students at the risk of experiencing burnout) and 394 true negatives (the students who do not face such risk). This result is significant in that it reveals the lowest error rates registered, made up of just 47 false negatives which represent missed instances, and 9 false positives which represent wrongful alerts. The significant degree of accuracy is testified to by its Chi-Square statistic of 250.8083 which is the highest among the three models tested and indicates a strong positive correlation between its output and the actual outcomes.

The performance of the K4 and the K3 models remains statistically significant; the trend shows progressive decline with the decrease in the number of features. The K4 (XGB) model registered 52 False Negatives and 29 False Positives, which indicates increased frequency of errors. The K3 (RBF SVM) model reflected the similar trend with 62 missed instances and 23 false alarms. The latter increase in misclassification is reflected in the related Chi-Square statistics at 161.3897 for K4 and 140.9476 for K3. Although the metrics suggest the predictive models to be strong in robustness, comprehensive error analysis computed through the confusion matrices shows the 5-item feature set to be the best and most accurate model in determining students based on the definition of the risk tertile.

5.5 Comparisons and Relationships

When inspecting the analysis of label systems, we discovered that the clinical cut-offs led to the prevalence of tertiles in our sample being by an order of two. Also, the confusion matrix showed that there are many instances of clinical-only positive students whose high score on the EX and a high score on the CY or low score on the PE but are not in the highest tertile of the sample distribution. This is naturally expected, because tertiles are relative measures; a group that is usually under a high amount of stress will still lead to only one-third of them being considered as highs.

5.6 Discussions

The project shows it is possible to be loyal to the discovered measurement model and decision rule and construct an operational system for early detection. The micro-screener makes no attempt at redefining burnout; it collapses the information required to re-create labels already justified. This position reduces potential risk of conceptual drift. The prominent limitation is the availability of just one cross-sectional data set, the lack of tests for multi-group invariance because demographics are lacking, and the dependence on sole self-reports. A potential research agenda would be to test invariance of measurement across gender and year, calibrate on new cohorts to track drift, and, where ethical and practical, introduce low-burden digital-trace signals like learning-management logins to supplement the limited self-reports.

The CFA outputs become the anchor for the rest of the thesis: EX, CY, and PE remain unique in this multinational student sample, in keeping with the initial MBI-SS theory. The dual-label method is an important methodological contribution. The established percentile convention is successfully operationalized with a clinically verified Exhaustion+1 rule and also tracks the effect of this decision on model choice, capacity planning, and calibration. A brief five-item micro-screener successfully obtains strong discrimination and shortens survey length by about 67% or more than the full measure, thus making campus-wide assessments more practical and feasible.

Chapter 6

Conclusion

6.1 Summary of Findings

We successfully created a secondary substitute for the MBI-SS, which may not be as accurate as MBI-SS but can be used for weekly monitoring. We also devecompared two labeling systems (tertiles versus clinical Exhaustion+1), validated the three-factor measurement model through confirmatory factor analysis, and evaluated a range of classifiers across three feature regimes (all 15 items, k-item micro-screener). The three-factor CFA demonstrated a significant advantage over the one-factor alternative when evaluated against widely recognized SEM criteria. Clinically anchored labels identified a significantly greater number of students compared to tertiles and were more accurately represented by a five-item micro-screener that aligns with the $EX + CY/PE$ logic established in the validation literature.

6.2 Contributions to the Field

Through empirical analysis, we present a reproducible comparison of label systems that are frequently addressed implicitly in the literature. Our findings indicate that a brief micro-screener can effectively approximate a clinically validated burnout flag, exhibiting strong discrimination capabilities. In our approach, we combine SEM-based construct validation with contemporary, calibrated classifiers and decision-oriented metrics that are appropriate for student-support environments with limited resources. The choices regarding the micro-screener conceptually reflect the three-factor structure that serves as the foundation for the MBI-SS.

6.3 Recommendations for Future Work

To strengthen the pipeline for production, institutions should reassess calibration every semester and monitor subgroup performance when demographic data are accessible. Measurement invariance across groups should be checked to ensure that a score means the same thing for all students. When privacy rules allow it, ethically gathered behavioral data can be included to help identify sudden stress changes that self-reports might miss. Ultimately, the translational work should produce a straightforward guide for operators and a collection of clear, student-friendly phrases. This way, the micro-screener can be easily used by many.

In the future, the micro-screener should be tested prospectively against clinical interviews in the local setting. Multi-group invariance testing (female, year, country) should also be considered to assess stability across subpopulations, a next step often recommended in SEM practice. If schools begin to collect digital-trace variables such as LMS activity, hybrid behavioral + micro-screener models may help identify problems earlier while remaining easy to use.

Bibliography

[1] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data.," *Biometrics*, vol. 33 1, pp. 159–74, 1977. [Online]. Available: https://api.semanticscholar.org/CorpusID:11077516

[2] A. Feinstein and D. Cicchetti, "Feinstein ar, cicchetti dvhigh agreement but low kappa: I. the problems of two paradoxes. j clin epidemiol 43:543-549," *Journal of clinical epidemiology*, vol. 43, pp. 543–9, Feb. 1990. DOI: 10.1016/0895-4356(90)90158-L

[3] L.-t. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 6, no. 1, pp. 1–55, 1999. DOI: 10.1080/10705519909540118 eprint: https://doi.org/10.1080/10705519909540118. [Online]. Available: https://doi.org/10.1080/10705519909540118

[4] W. Schaufeli, I. Martínez, A. Marques-Pinto, and M. Salanova, "Burnout and engagement in university studentsa cross-national study," *Journal of Cross-cultural Psychology - J CROSS-CULT PSYCHOL*, vol. 33, pp. 464–481, Sep. 2002. DOI: 10.1177/0022022102033005003

[5] W. Schaufeli, M. Salanova, and V. González-Romá, "The measurement of engagement and burnout: A two sample confirmatory factor analytic approach," *Journal of Happiness Studies*, vol. 3, pp. 71–92, Feb. 2002. DOI: 10.1023/A:1015630930326

[6] D. Flora and P. Curran, "An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data," *Psychological Methods*, vol. 9, pp. 466–491, Dec. 2004. DOI: 10.1037/1082-989X.9.4.466

[7] M. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, Oct. 2012. DOI: 10.11613/BM.2012.031

[8] A. Meade and B. Craig, "Identifying careless responses in survey data," *Psychological Methods*, vol. 17, pp. 437–455, Apr. 2012. DOI: 10.1037/a0028085

[9] M. Rhemtulla, P. Brosseau-Liard, and V. Savalei, "When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions," *Psychological Methods*, vol. 17, pp. 354–373, Jul. 2012. DOI: 10.1037/a0029315

[10] C.-H. Li, "Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares," *Behavior research methods*, vol. 48, Jul. 2015. DOI: 10.3758/s13428-015-0619-7

[11] R. L. S. Xu Zhao and H. Haste, “Academic stress in chinese schools and a proposed preventive intervention program,” *Cogent Education*, vol. 2, no. 1, M. Cheng, Ed., p. 1 000 477, 2015. DOI: 10.1080/2331186X.2014.1000477 [Online]. Available: https://doi.org/10.1080/2331186X.2014.1000477

[12] N. D. Wickramasinghe, D. S. Dissanayake, and G. S. Abeywardena, “Clinical validity and diagnostic accuracy of the maslach burnout Inventory-Student survey in sri lanka,” en, *Health Qual. Life Outcomes*, vol. 16, no. 1, p. 220, Nov. 2018.

[13] World Health Organization, *Burn-out an occupational phenomenon: International classification of diseases*, https://www.who.int/news/item/28-05-2019-burn-out-an-occupational-phenomenon-in-the-international-classification-of-diseases, News release, May 2019. Accessed: Oct. 8, 2025.

[14] J. Marôco et al., “Predictors of academic efficacy and dropout intention in university students: Can engagement suppress burnout?” *PLOS ONE*, vol. 15, pp. 1–26, Oct. 2020. DOI: 10.1371/journal.pone.0239816 [Online]. Available: https://doi.org/10.1371/journal.pone.0239816

[15] F. T Serrano et al., “Síndrome de burnout, según el año de estudio, en estudiantes de medicina de ocho países de latinoamérica, 2016-2017,” *Ciencia e Investigación Medico Estudiantil Latinoamericana*, vol. 26, pp. 22–29, Jul. 2020. DOI: 10.23961/cimel.v26i1.1237

[16] S. Bd and S. Baniadamdizaj, “Prediction of iranian efl teachers’ burnout level using machine learning algorithms and maslach burnout inventory,” *Iran Journal of Computer Science*, vol. 6, Aug. 2022. DOI: 10.1007/s42044-022-00112-x

[17] F. T. Serrano, *Dataset and analysis conducted for the study: Burnout syndrome according to the year of study in medical students from eight countries in latin america, 2016–2017*, version 2, CC BY 4.0, Mendeley Data, Apr. 2022. DOI: 10.17632/2dhbz56s6h.2 [Online]. Available: https://data.mendeley.com/datasets/2dhbz56s6h/2

[18] X. Gao, “Academic stress and academic burnout in adolescents: A moderated mediating model,” *Frontiers in Psychology*, vol. 14, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259065806

[19] G. Larraín, N. Rojas-Morales, N. Gonzalez, and D. Olcay, “Using a classifier ensemble for preventing burnout in university students: A study case in valparaíso,” *IEEE Access*, vol. PP, pp. 1–1, Jan. 2023. DOI: 10.1109/ACCESS.2023.3325785

[20] A. Abraham et al., “Burnout increased among university students during the covid-19 pandemic: A systematic review and meta-analysis,” *Scientific Reports*, vol. 14, Jan. 2024. DOI: 10.1038/s41598-024-52923-6

[21] A. Boone et al., “Burnout and predictive factors among medical students: A cross-sectional survey,” *BMC Medical Education*, vol. 24, Jul. 2024. DOI: 10.1186/s12909-024-05792-6

[22] R. De Filippis and A. Foysal, “Comprehensive analysis of stress factors affecting students: A machine learning approach,” *Discover Artificial Intelligence*, vol. 4, Sep. 2024. DOI: 10.1007/s44163-024-00169-6

[23] K. Oluwadiya, O. Owoeye, and A. Adeoti, “Evaluating the factor structure, reliability and validity of the copenhagen burnout inventory-student survey (cbi-ss) among faculty of arts students of ekiti state university, ado-ekiti, nigeria,” *Scientific Reports*, vol. 14, May 2024. DOI: 10.1038/s41598-024-61310-0

[24] L. Z. Sannyuya Liu Yunhan Zhang and Z. Liu, “Academic stress detection based on multisource data: A systematic review from 2012 to 2024,” *Interactive Learning Environments*, vol. 0, no. 0, pp. 1–27, 2024. DOI: 10.1080/10494820.2024.2387744 [Online]. Available: https://doi.org/10.1080/10494820.2024.2387744

[25] H. Kiss and B. F. Pikó, “Risk and protective factors of student burnout among medical students: A multivariate analysis,” *BMC Medical Education*, vol. 25, Article 386, 2025, Open Access. DOI: 10.1186/s12909-025-06956-8 [Online]. Available: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-025-06956-8

[26] P. Lialiou and I. Maglogiannis, “Students’ burnout symptoms detection using smartwatch wearable devices: A systematic literature review,” *AI Sensors*, vol. 1, p. 2, May 2025. DOI: 10.3390/aisens1010002

[27] R. Tapio, “Predicting burnout in college students: A machine learning approach using decision tree and psychometric data,” *Asian Journal of Probability and Statistics*, vol. 27, pp. 50–60, May 2025. DOI: 10.9734/ajpas/2025/v27i5754