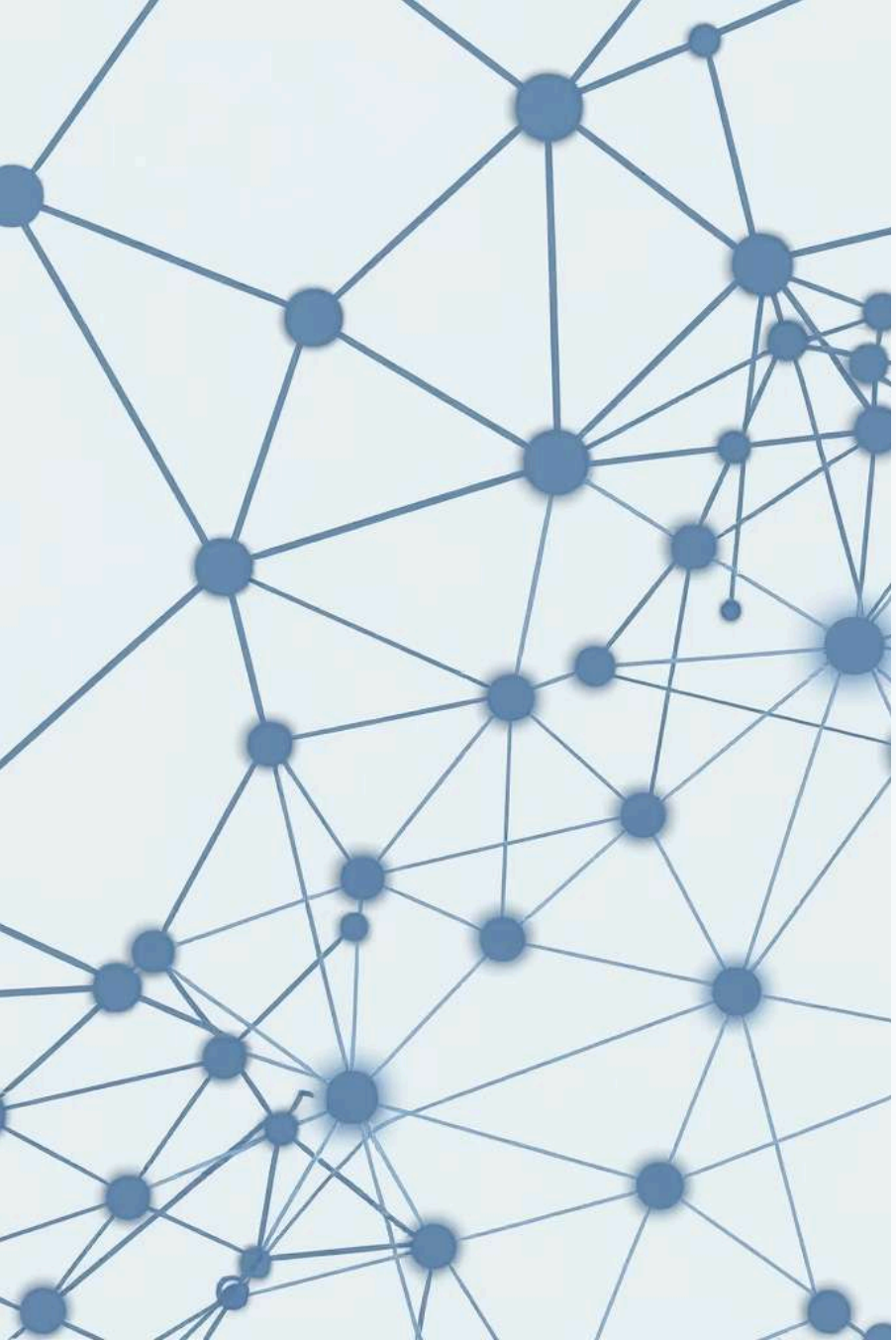


How Work–Life and Lifestyle Conditions Shape Burnout: A Cross-National Machine Learning Analysis

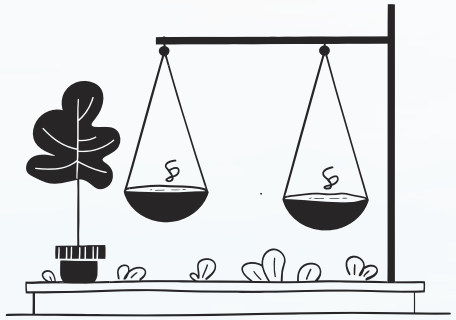
MSDSP 422B – Practical Machine Learning

Team Members: Anoushka, Biraj, Hunter, Samridhi

Understanding the Burnout Epidemic

Burnout Crisis

Recognized as a pressing global workplace crisis by WHO in 2019.



Causes

Extended work hours, imbalanced work-life ratios, and lifestyle stressors.

Effects

Decreased productivity, increased absenteeism & health-related risks.

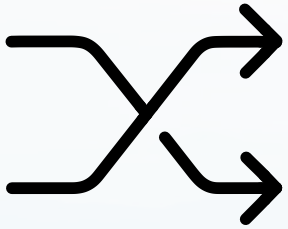
Solution

Overcome through data-driven analysis & preventive strategies.

Navigating the Challenges of Burnout Prediction

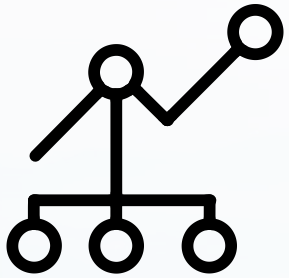
Widespread High Stress

Universally high



Blurred Boundaries

Overlap of everything



Typical Lifestyle

No significant deviations

Elusive Predictors

It changes through time

Research Questions

1

How do work–life conditions and lifestyle behaviors affect burnout?

Project Aim & Approach

Our goal is to uncover the key lifestyle and work factors that predict burnout

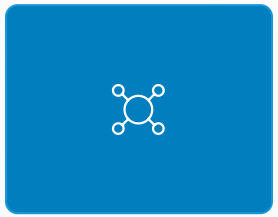
Analyze global records on stress, lifestyle, and work conditions

Define burnout using sustained high stress as the target variable

Apply machine learning models to identify strongest predictors

Provide insights for organizations and policymakers to address burnout

Literature Review: Key Findings



Multifactorial Nature

Burnout is influenced by a complex interplay of work hours, lifestyle, and mental health factors.

Extensive Research

Numerous studies on burnout exist across diverse professions and countries, highlighting its global prevalence.

Lifestyle's Role

Key lifestyle elements such as sleep, physical activity, and coping mechanisms are consistently emphasized in research.

ML Integration

Machine learning has emerged as a recent tool to analyze and understand the intricate complexity of burnout.

New Dataset & Insight

We were very suspicious of the dataset, so we compared our results to smaller studies to get further insights.

Case Study 1

Association between long working hours and sleep problems in white-collar workers

Case Study 2

The Effects of Sleep Quality and Resilience on Perceived Stress, Dietary Behaviors, and Alcohol Misuse: A Mediation-Moderation Analysis of Higher Education Students from Asia, Europe, and North America during the COVID-19 Pandemic

Case Study 3

Low Physical Activity and High Screen Time Can Increase the Risks of Mental Health Problems and Poor Sleep Quality among Chinese College Students

Case Study 4

Coping Styles as Predictors of Alcohol Consumption with Undergraduate College Students Perceiving Stress

Developed business assumptions, which improved our methods to identify burnout

Why this Dataset?

Trusted Source

From Zenodo, open & peer-reviewed research dataset (Oct 2024).

Extensive Coverage

ensuring broad representation

Rich Variables

Relevant to burnout

Research Focus

Specifically designed for mental health prediction

Exploratory Data Analysis: Initial Insights

Our EDA began with a comprehensive look at a robust, multi-country dataset, laying the groundwork for robust modeling.

- **Dataset Size:** ~50,000 individual-level records across multiple countries, ensuring diversity and representativeness.
- **Features:** 18 columns in total, comprising a balanced mix of:
  - **Demographic attributes (categorical):** Gender, Occupation, Country.
  - **Lifestyle habits (numerical):** Age, Sleep Hours, Work Hours, Physical Activity Hours, Social Media Usage.
  - **Lifestyle habits (categorical):** Diet Quality, Smoking Habit, Alcohol Consumption.
  - **Health-related (categorical):** Mental Health Condition, Consultation History, Medication Usage.
- **Target Variable:** Burnout, derived from self-reported stress level — High Stress = 1 (Burned Out), Low/Medium = 0 (Not Burned Out).
- **Data Quality:** Strong integrity overall, with minimal missingness.

Objective: Examine distributions and group differences to assess whether any single factor explains burnout, and to set the stage for robust multivariate modeling.

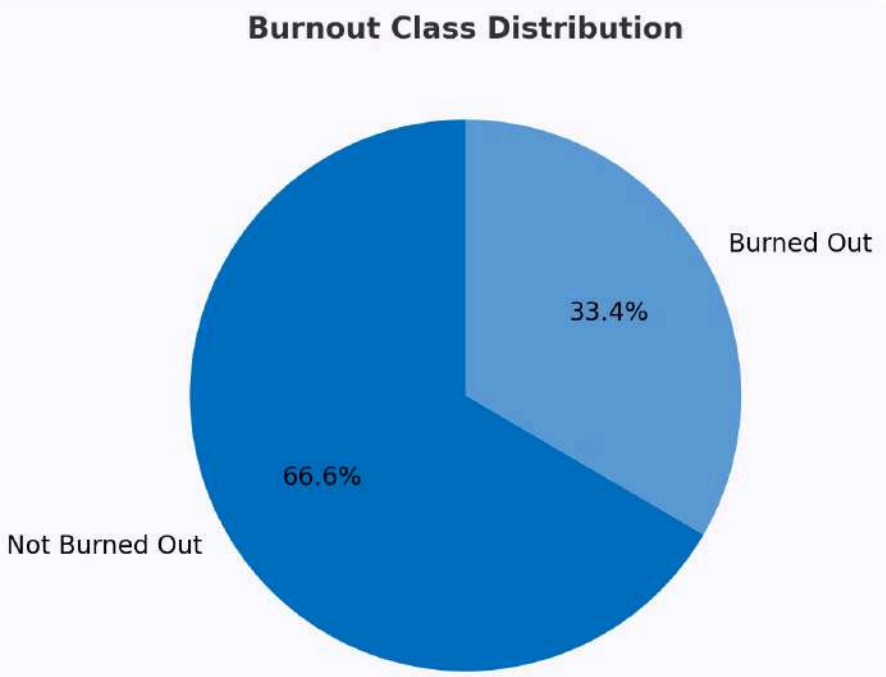
Target Definition & Class Balance

Burnout was rigorously defined to ensure a clear target variable for machine learning models.

Burnout Definition: Assigned as High Stress_Level = 1; while Low/Medium Stress_Level = 0.

Class Distribution:

- **Not Burned Out:** 33,293 individuals
- **Burned Out:** 16,707 individuals



The moderate imbalance observed is sufficient for robust and effective model training.

Target Definition & Class Balance

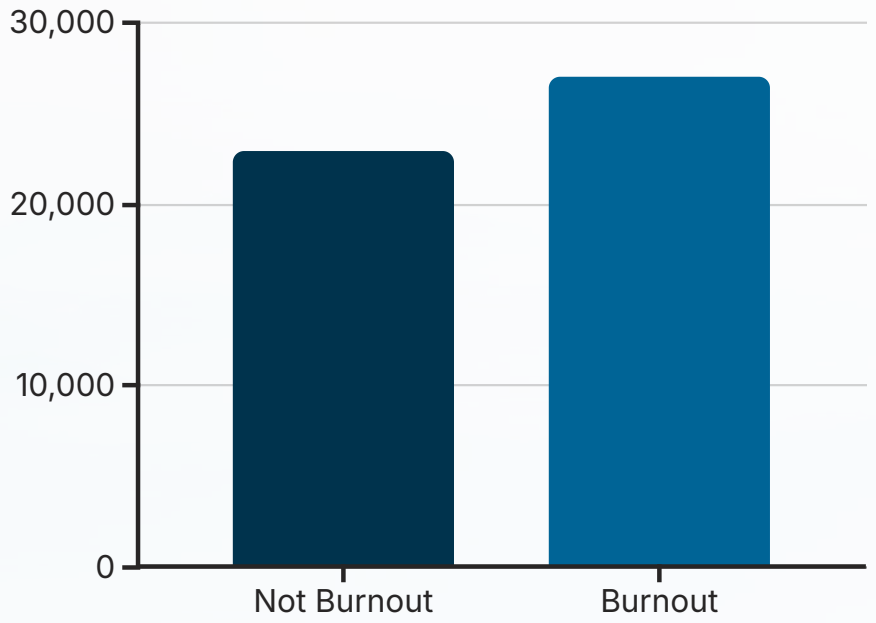
Burnout was rigorously defined to ensure a clear target variable for machine learning models.

Burnout Definition: Assigned as High Stress_Level = 1; while Low/Medium Stress_Level = 0.

Class Distribution:

- **Not Burned Out:** 22,945 individuals
- **Burned Out:** 27,055 individuals

The moderate imbalance observed is sufficient for robust and effective model training.

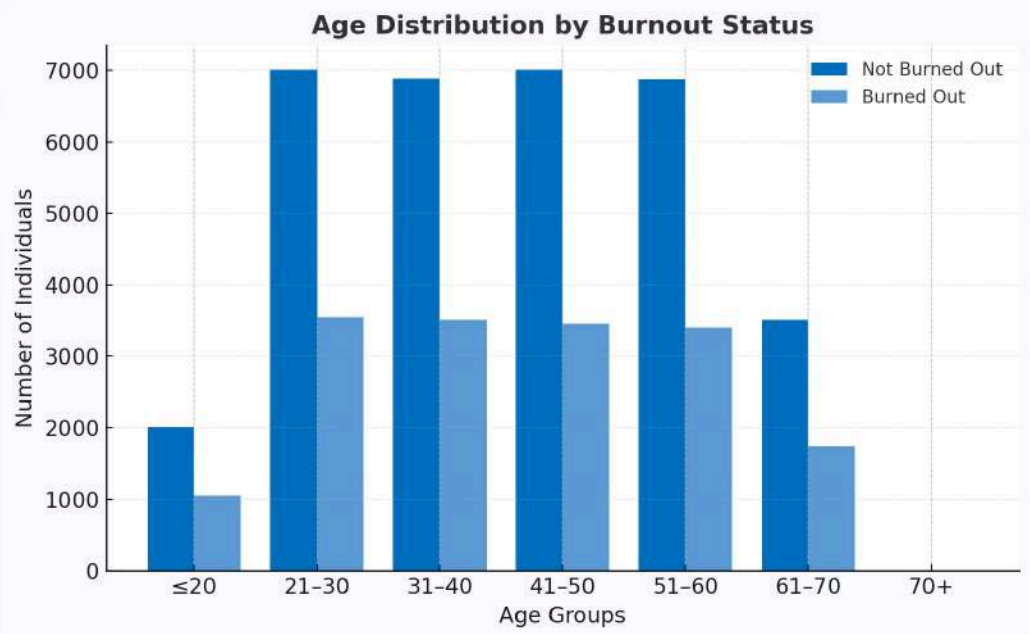


Exploratory Data Analysis

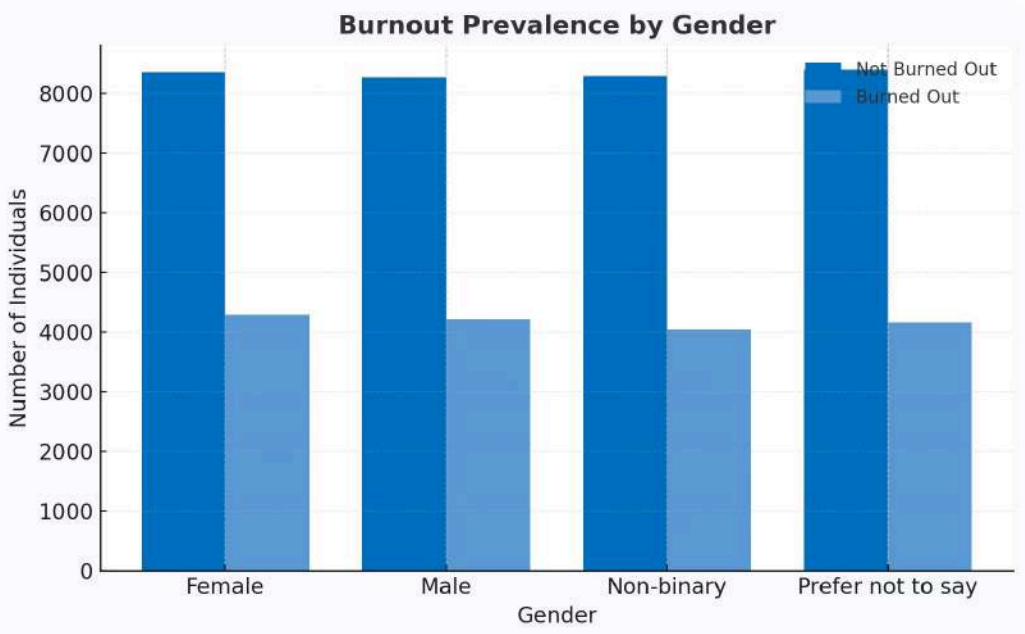
Demographic Analysis

Demographic factors such as age and gender did not reveal significant subgroup predispositions to burnout.

Age Distribution: While the dataset skewed towards the 21–40 age bracket, burnout was consistently present across all age groups.



Gender Parity: Burnout prevalence was nearly uniform across different gender groups, indicating no significant disparity.



Additional Demographics Insights:

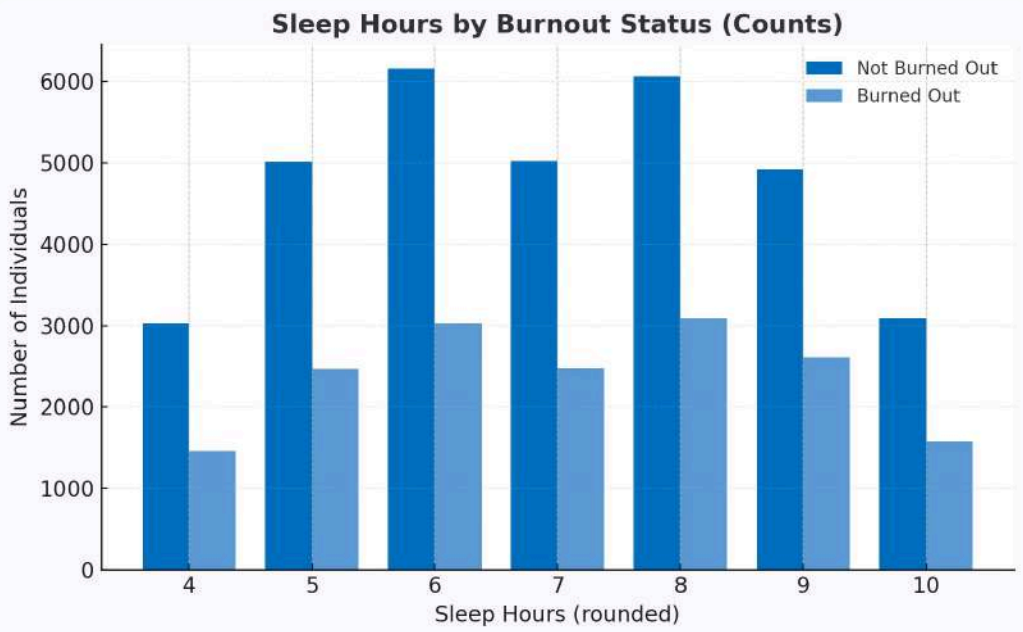
- **Country:** Across the top 10 most represented countries, burnout prevalence was consistently around **one-third of respondents**, showing no major country-specific differences.
- **Occupation:** While burnout rates were broadly similar across professions, some modest variation was observed:
  - **Highest:** Finance (~34.2%) and Sales (~33.6%) showed slightly higher burnout prevalence.
  - **Lowest:** Engineering (~32.9%) and Healthcare (~33.1%) were at the lower end of the range.

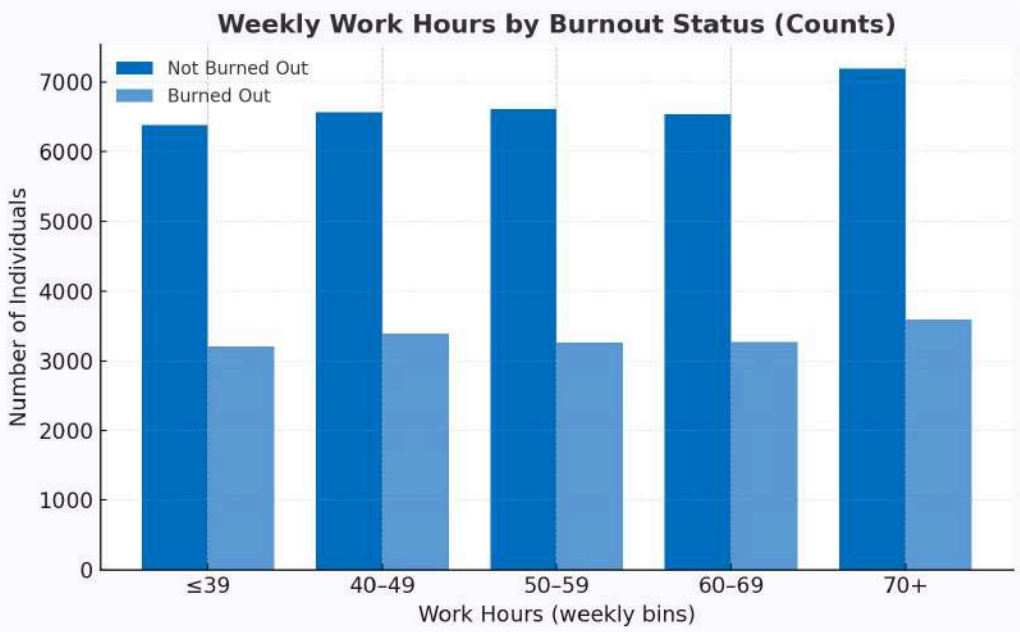
Sleep & Work Hours Analysis

Initial analysis of sleep and work hours indicates these factors alone do not singularly explain burnout.

Sleep Patterns: The majority of respondents reported 6–8 hours of sleep per night, with no clear separation observed between burnout classes.

Work Hours: The median work week consistently hovered around 55 hours, showing similar distributions across both burnout and non-burnout groups.



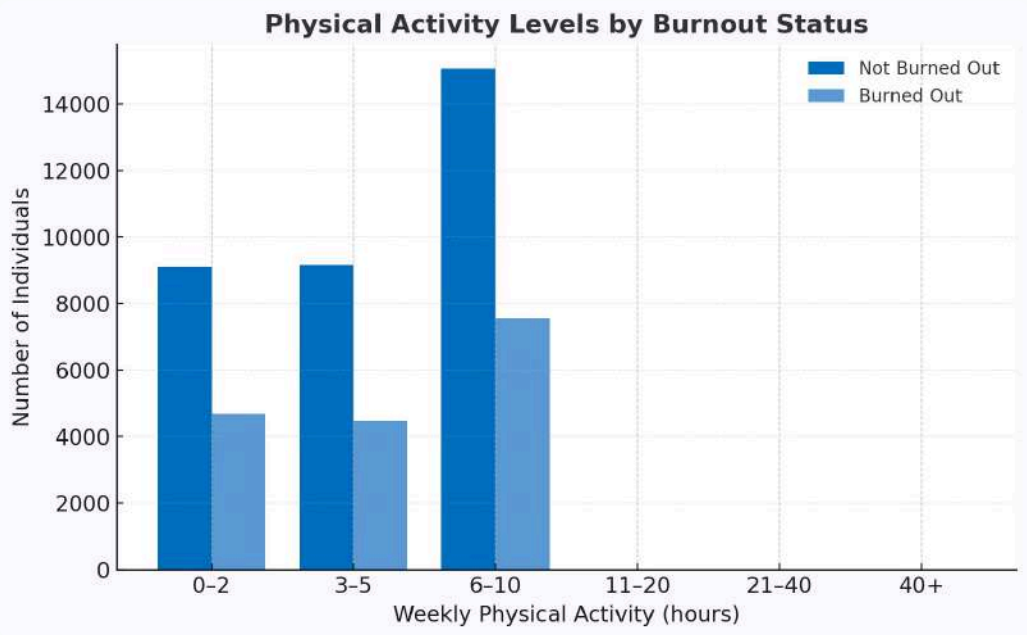


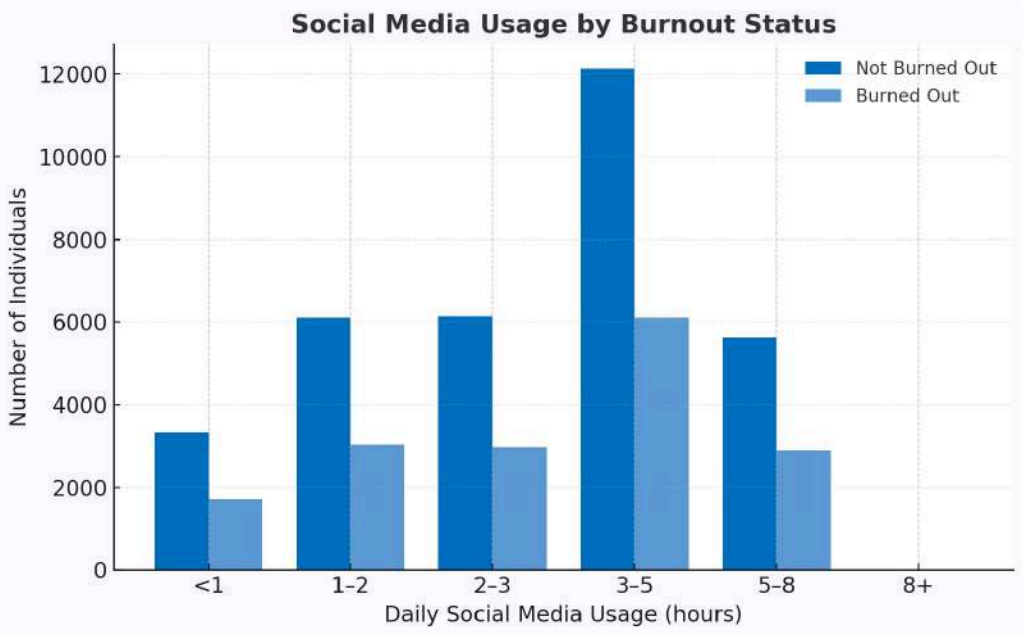
Lifestyle Factors

Physical activity and social media usage, when analyzed in isolation, showed limited direct explanatory power for burnout.

Physical Activity: Levels of physical activity were found to be largely comparable between individuals experiencing burnout and those not.

Social Media Usage: Social media usage exhibited an even distribution across the dataset, suggesting no strong or direct link to burnout on its own.





Business Rules and Burnout Hotspots

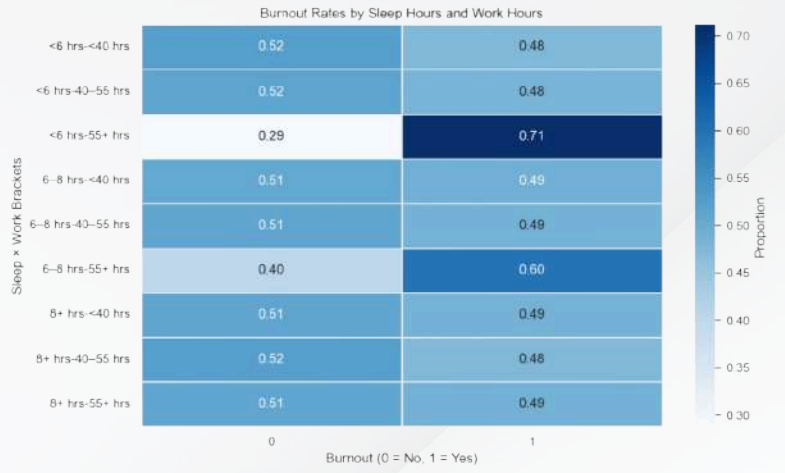
Work–Life Balance

The Impact of Sleep Hours and Work Hours on Burnout

Burnout rates are **consistently high** across most sleep and work hour combinations.

Individuals reporting **less than 6 hours of sleep and 55+ work hours per week** show the **highest burnout rate (~71%)**.

This highlights **chronic sleep deprivation combined with long work hours** as a critical risk factor, even within a widespread burnout landscape.



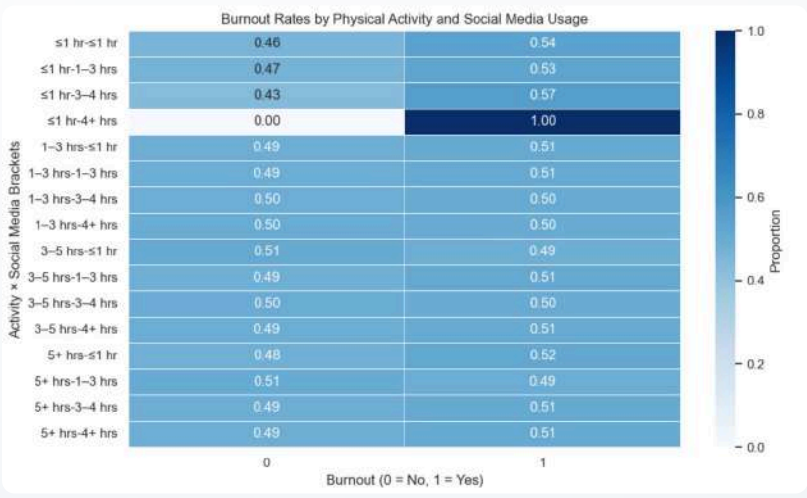
Lifestyle Balance

The Relationship Between Physical Activity and Social Media Usage on Burnout

Burnout risk is elevated for **sedentary individuals** engaging in **high social media use (≥4 hours/day)**.

While active lifestyles offer some protection, **burnout remains prevalent** across all groups.

This suggests that **low physical activity paired with excessive passive screen time** creates a vulnerable combination.



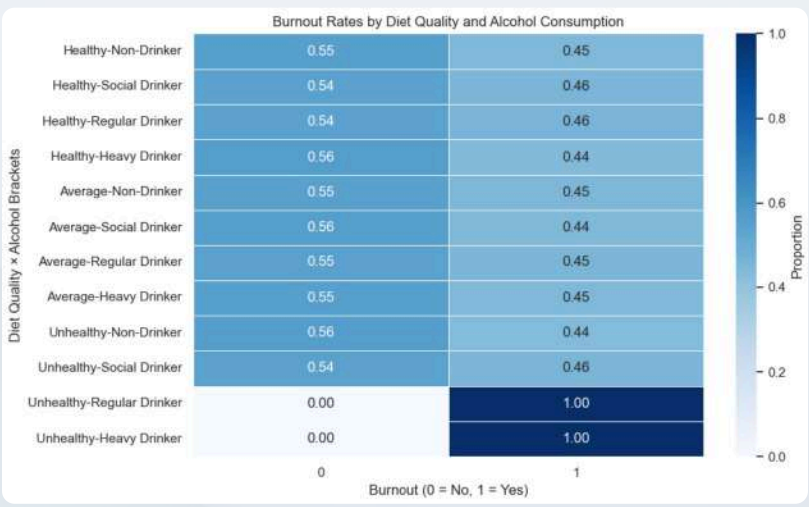
Coping Strategies

The Interplay of Diet Quality and Alcohol Consumption on Burnout

Burnout prevalence is **higher among individuals with unhealthy diets combined with regular/heavy alcohol consumption**.

This pattern indicates that **unhealthy coping mechanisms amplify vulnerability** to burnout.

Although overall burnout is systemic, these lifestyle choices, particularly with substance use, demonstrate **significant risk signals**.



Correlation Analysis

- Correlation matrix showed **weak pairwise relationships** between lifestyle and demographic variables.
- No single feature (e.g., sleep, work hours, social media) was strongly correlated with burnout.
- This supports the finding that burnout is **multifactorial** and not explained by any one variable.

Principal Component Analysis

- PCA was applied to reduce dimensionality and explore structure in the data.
- The first few principal components explained only a small portion of variance, meaning the dataset does not compress neatly into 1–2 drivers.
- This further confirms that burnout is influenced by a broad mix of interacting factors rather than a few dominant ones.

EDA Summary & Overall Findings

Initial analysis revealed burnout is widespread across all demographics, countries, and occupations, with no single subgroup disproportionately affected.

Individual lifestyle factors (sleep, work hours, diet, physical activity, social media) did not singularly explain burnout, as their distributions were similar between burned-out and non-burned-out groups.

Identified Burnout Hotspots

<6 hrs Sleep + 55+ Work Hours

Highest burnout rate (~71%).

Sedentary + 4+ hrs Social Media

Elevated burnout risk (~68%).

Unhealthy Diet + Regular/Heavy Drinking

Increased burnout vulnerability (~67–69%).

Key Message: Burnout is systemic and multifactorial. While specific risk zones exist, no isolated factor explains its prevalence, highlighting the need for advanced modeling to uncover hidden interactions.

Our EDA showed us that **simple explanations won't work**. To truly understand burnout, we need advanced machine learning models that can capture the complex interplay of factors.

Data Treatement

Before model training, our raw data underwent meticulous preprocessing to enhance its quality and predictive power.

Preventing Data Leakage

Ensured that information from the test set did not inadvertently influence the training process, maintaining model integrity.

Impute Missing Data

Simpleimputer to impute categorical and numeric data

Feature Scaling

Applied `StandardScaler` to continuous variables, ensuring all features contributed equally to model training.

Data Splitting (Stratified)

Divided the dataset into training and testing subsets (80/20) for robust model evaluation.

Data Encoding

Converted categorical variables into numerical representations using Label and One-Hot Encoding.

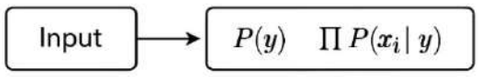
Feature Engineering

Derived new features including interaction effects, brackets for age, sleep, and work hours, enriching the dataset's predictive capacity.

Models Used

Benchmark Model: Naive Bayes

A family of **probabilistic classifiers** based on Bayes' theorem, making a "naive" assumption of feature independence. Best suited for classification tasks.

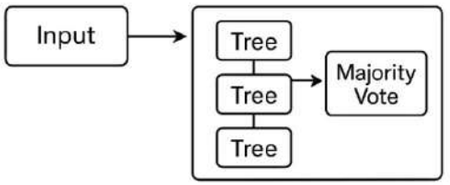


Logistic Regression

A **linear model** for classification (binary or multiclass), modelling the probability of class membership using a logistic (sigmoid/Softmax) function.

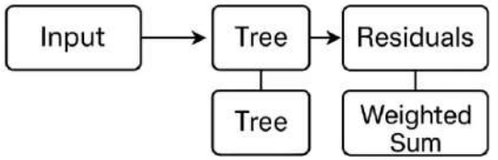
Random Forest

An **ensemble** of decision trees, trained with random feature and data sampling, with final predictions by majority vote (classification).



XGBoost

Type of gradient boosting algorithm that utilises decision trees as weak learners, combining them to create a strong predictive model.



Why these models and their Assumptions

Each model chosen for our analysis plays a distinct role, offering unique advantages and operating under specific assumptions that influence their suitability for burnout prediction.

Model	Why we have used?	Assumptions
Naive Bayes	Fast, effective, baseline	Features are conditionally independent given class.
Logistic Regression	Simple, interpretable	Linear relationship between features and log-odds; low multicollinearity; independent observations.
Random Forest	Handles non-linear, high-dimensional data; not impacted by noise and overfitting;	No missing values, low multicollinearity among features
XGBoost	State-of-the-art accuracy, handles missing data, scalable	Assumes additive model where new trees correct previous errors (gradient boosting principle).

Primary Model Fitness Indicator

To accurately assess the performance and reliability of our machine learning models in predicting burnout, we used the following key metrics:

1

Recall (Sensitivity)

Measures the proportion of actual positive cases that were correctly identified by the model.

2

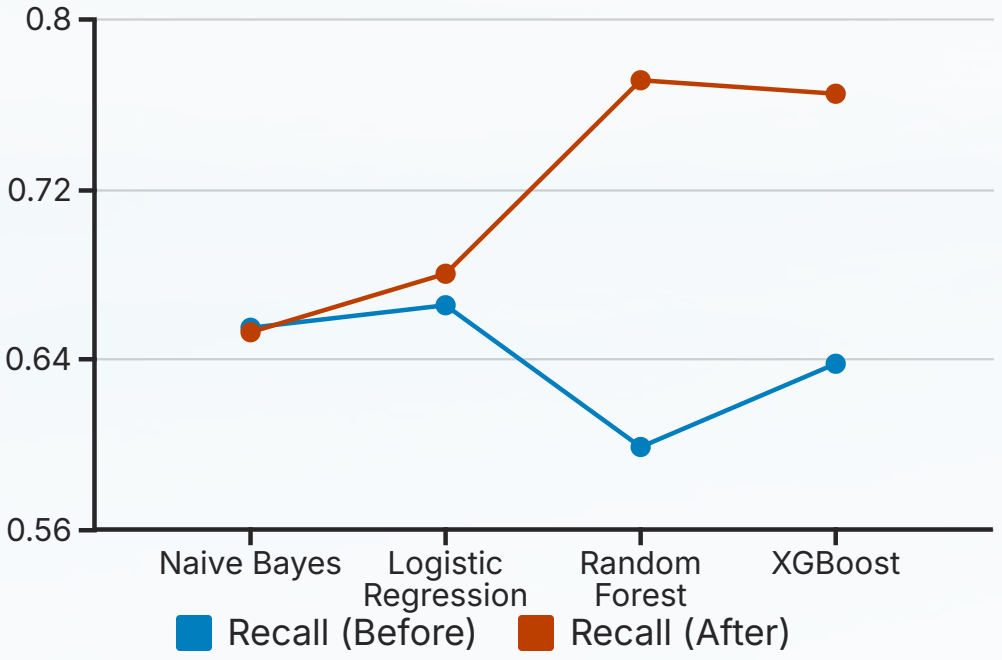
F1 Score

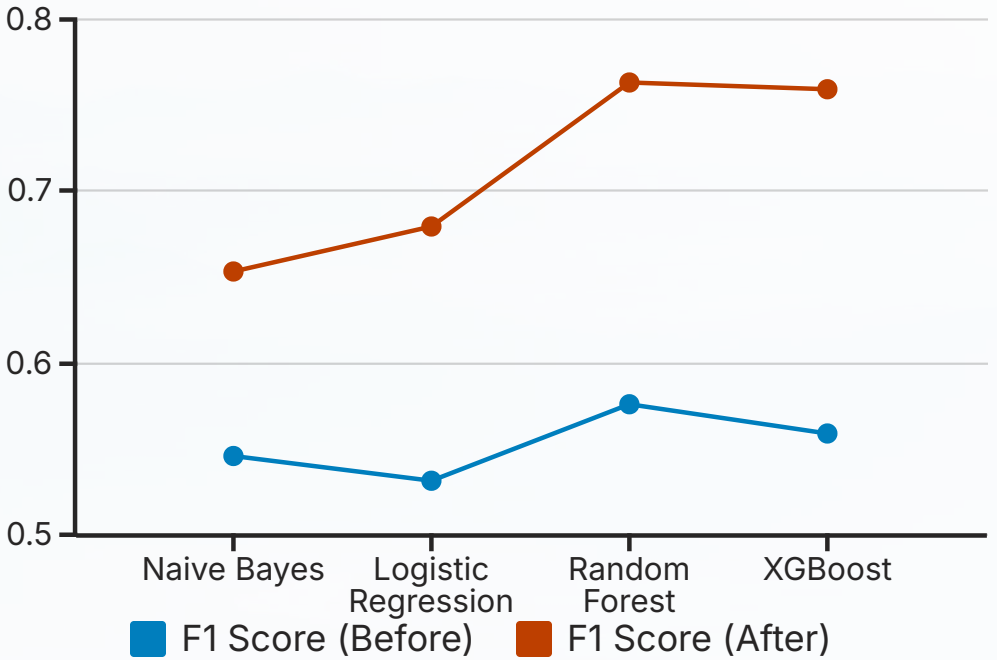
The harmonic mean of Precision and Recall

Primary Model Fitness Overview

We evaluated model performance both before and after applying business assumptions, focusing on Recall as a key metric for our business objectives. The chart below illustrates these results.

A comparison of Recall and F1 score Before and After Business assumption

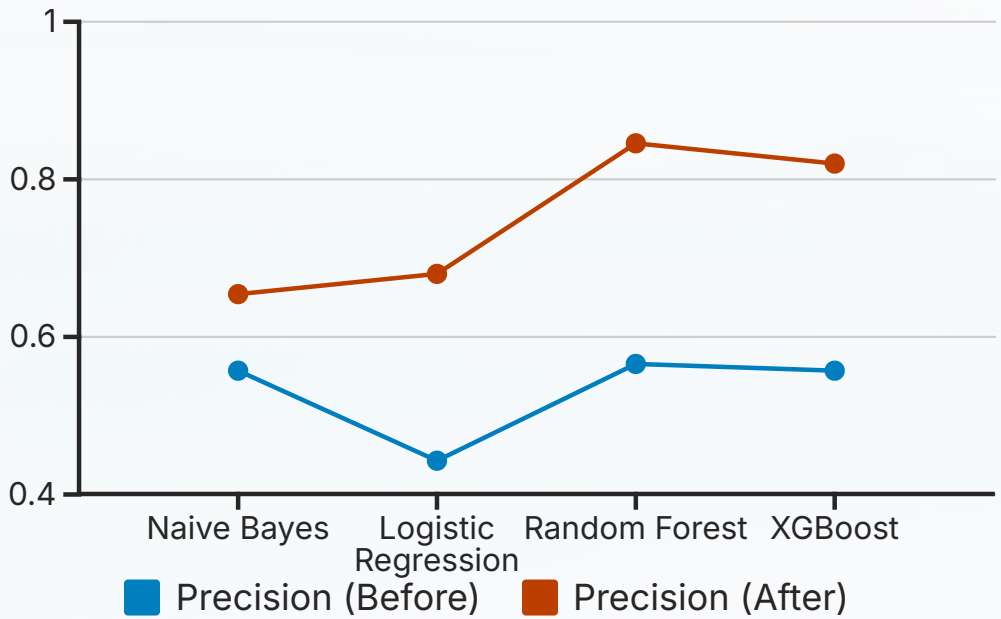




Secondary Model Fitness Overview

Beyond Recall and F1 Score, we also assessed model performance using Precision and Accuracy, both before and after applying business assumptions, to provide a comprehensive view of their predictive capabilities.

A comparison of Precision and Accuracy score Before and After Business assumption



How Hyperparameter Tuning Improves Models

Hyperparameter tuning is a crucial step in machine learning workflows, transforming base models into highly optimized predictors. It refines a model's underlying structure, leading to significant performance gains and more reliable predictions.

Improved Generalization

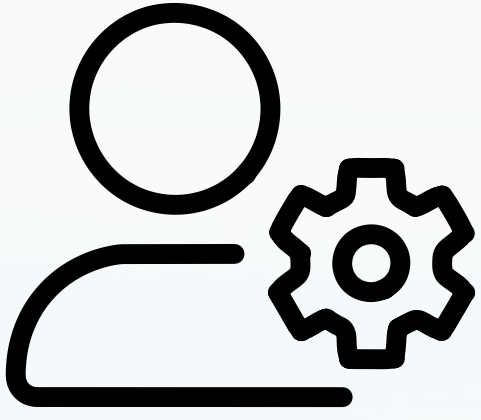
Ensures the model performs robustly on new, unseen data, reducing the risk of either overfitting or underfitting.

Reduced Bias & Variance

Balances the model's complexity, making it less prone to memorizing training noise (overfitting) or being too simplistic (underfitting).

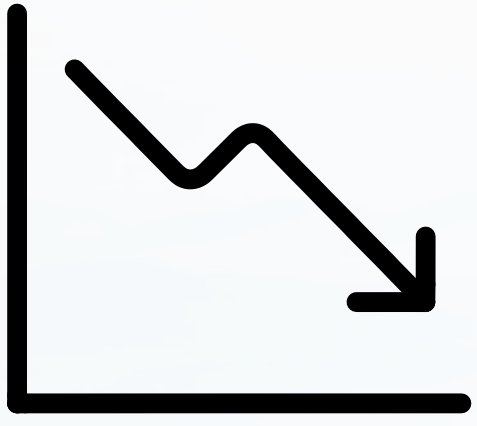
Hyperparameter Tuning with GridSearchCV

To maximize model performance and prevent overfitting, we systematically optimized each model's hyperparameters using `GridSearchCV`.



Defining Parameter Grids

For each algorithm (Logistic Regression, Random Forest, XGBoost), a comprehensive grid of potential hyperparameter values was defined and tested.



Performance Evaluation

Each parameter combination was evaluated based on predefined metrics (e.g., F1-score, accuracy) to identify the settings that got the best generalization performance.

Optimal Parameter Selection

The best performing set of hyperparameters for each model was then selected.

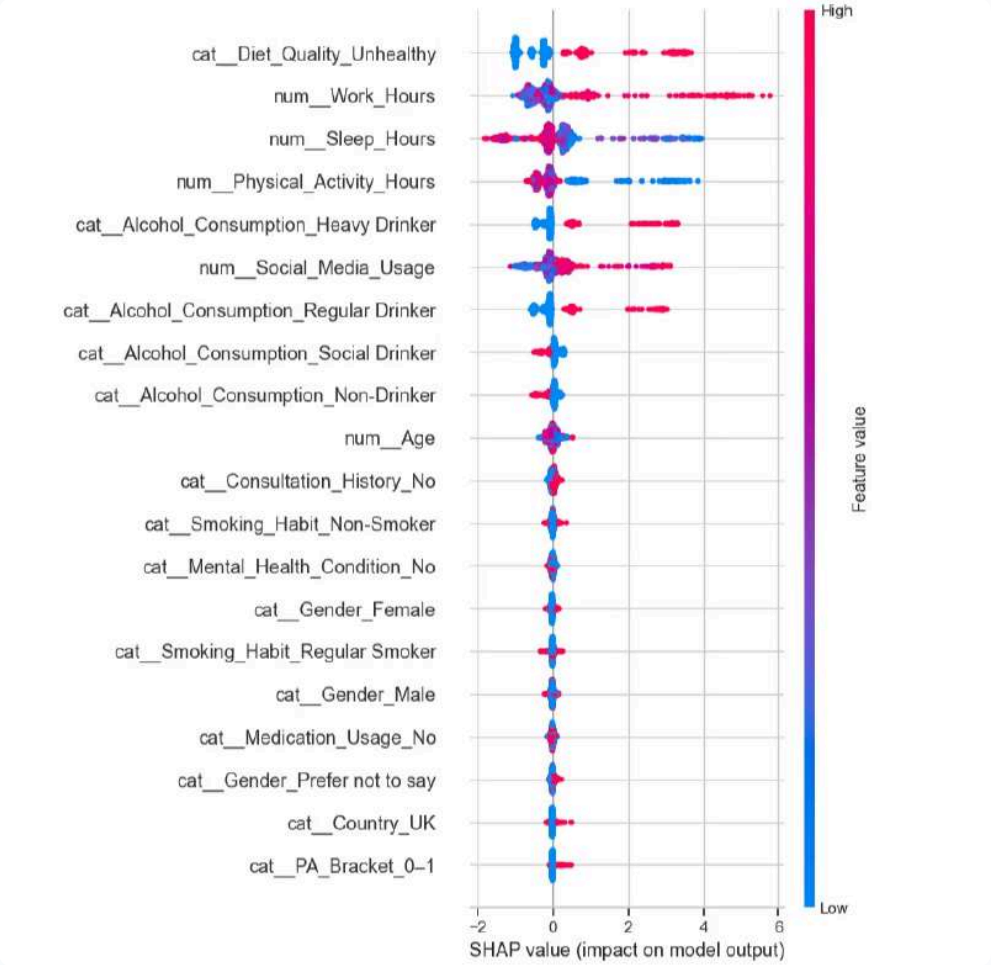
Hyperparameter Tuning with GridSearchCV

To optimize our models' performance and prevent overfitting, we systematically tuned their hyperparameters using GridSearchCV. This process refined each model to its peak predictive capability.

Model	Non-Tuned (Default)	Tuned Version	Key Differences / What Tuning Does
Logistic Regression	LogisticRegression()	LogisticRegression(solver="saga", penalty="l2", C=0.1, max_iter=2000, random_state=42)	Uses **SAGA solver** (handles large/sparse data); penalty="l2" applies ridge regularization; C=0.1 strengthens regularization (reduces overfitting); higher max_iter=2000 ensures convergence; reproducibility with random_state=42.
Random Forest	RandomForestClassifier()	RandomForestClassifier(n_estimators=300, max_depth=10, min_samples_split=5, min_samples_leaf=3, max_features="sqrt", class_weight="balanced", random_state=42)	More trees (n_estimators=300) improves stability; limits depth (max_depth=10) to prevent overfitting; min_samples_split=5 and min_samples_leaf=3 enforce node size for generalization; max_features="sqrt" reduces correlation between trees; class_weight="balanced" handles class imbalance.
Naive Bayes	GaussianNB()	GaussianNB(var_smoothing=1e-9)	var_smoothing=1e-9 prevents division by zero or instability when features have very low variance; improves numerical stability.
XGBoost	XGBClassifier()	XGBClassifier(eval_metric="mlogloss", random_state=42)	Uses mlogloss (multi-class log loss) for better classification evaluation; random_state=42 ensures reproducibility.

Model Interpretation - Key Drivers of Burnout

The Complex Web of Burnout Drivers



Factors that Increase Burnout Risk:

- Fewer sleep hours
- High social media usage
- Long work hours
- Little or no physical activity
- Poor/average diet quality
- Being a smoker

Translating Model Insights to Business Value

Proactive Risk Mitigation

Identify employees at high risk of burnout early, enabling timely interventions before severe symptoms emerge.

Targeted Interventions

Pinpoint the specific drivers of burnout within your organization to implement highly effective, personalized support strategies.

Improved Productivity & Retention

Foster a healthier, more engaged workforce, leading to increased productivity and reduced turnover costs.

Enhanced ROI on Wellness

Optimize resource allocation by focusing wellness initiatives where they will have the greatest impact on employee health and business outcomes.

Target Areas for Stakeholders

HR System Integration

Seamlessly integrate the prediction model into existing HR analytics dashboards and employee management systems for continuous monitoring.

Develop Intervention Programs

Design and deploy tailored support programs (e.g., stress management, flexible work policies) based on model insights.

Continuous Monitoring & Refinement

Establish a feedback loop for ongoing model validation and retraining, ensuring its accuracy adapts to evolving organizational dynamics.

Next Steps - Modelling Approach

Explore Deep Learning Approaches

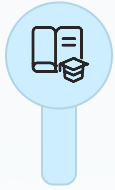
Investigate advanced neural network architectures, such as Recurrent Neural Networks (RNNs) or Transformers, to capture more complex temporal and contextual patterns in the data, potentially revealing hidden drivers of burnout.

Integrate Longitudinal Data

Incorporate time-series data to understand the progression of burnout and identify early warning signs, allowing for proactive interventions and personalized support strategies.

Expand Cross-National Analysis

Broaden the dataset to include more countries and diverse cultural contexts, enabling a more comprehensive understanding of how work-life and lifestyle factors contribute to burnout globally.



In-depth domain understanding

Continue building a strong understanding of the burnout phenomenon, its causes, and mitigation strategies.

Explore the Interactive Dashboard

Engage directly with our machine learning insights through a live, interactive dashboard. Visualize key findings, explore data relationships, and understand the model's predictions in real-time.

https://mentalhealthandburnoutdemo.streamlit.app/

Access the Project Code

Dive deeper into our methodology, data processing, and model implementations by exploring the complete code on our GitHub repository.

https://github.com/B1raj/Mental_Health_and_Burnout

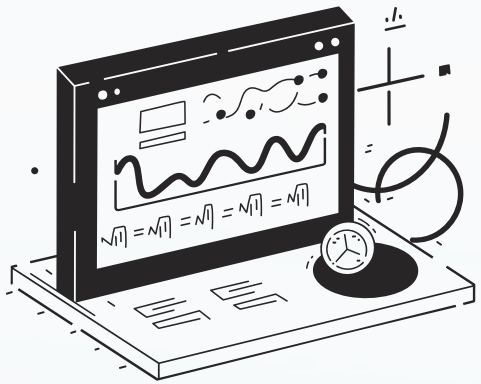
References

- **Shanafelt, T. D., S. Boone, L. Tan, L. N. Dyrbye, W. Sotile, D. Satele, ... and M. R. Oreskovich. 2015.** "Burnout and Satisfaction with Work-Life Balance among US Physicians Relative to the General US Population." *PLOS ONE* 10 (11): e0119607. https://doi.org/10.1371/journal.pone.0119607.
- **Sánchez-Oliva, D., J. J. Pulido-González, F. M. Leo, J. L. Chamorro, and T. García-Calvo. 2021.** "Low Physical Activity and High Screen Time Are Associated with Burnout and Mental Health Problems." *Nutrients* 13 (2): 442. https://doi.org/10.3390/nu13020442.
- **Rupp, A. 2014.** "Burnout, Stress, and Coping Mechanisms among Psychology Graduate Students." *PCOM Psychology Dissertations*, no.144. Accessed August 18, 2025. https://digitalcommons.pcom.edu/psychology_dissertations/144.
- **Folkman, S., and J. T. Moskowitz. 2000.** "Positive Affect and the Other Side of Coping." *American Psychologist* 55 (6): 647–54. https://pubmed.ncbi.nlm.nih.gov/20561174/.
- **Åkerstedt, T., and K. P. Wright. 2009.** "Sleep Loss and Fatigue in Shift Work and Shift Work Disorder." *Sleep Medicine Clinics* 4 (2): 257–71. https://pubmed.ncbi.nlm.nih.gov/19544749/.

Thank You!

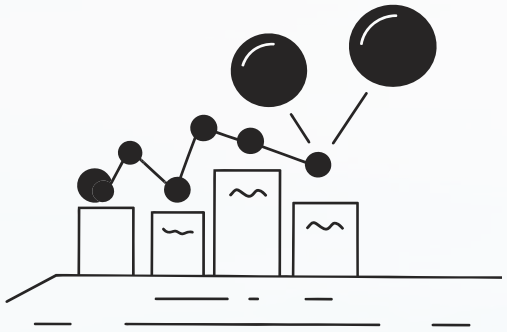
Extra Slides

Limitations



Limited Scope

Self-reported survey data → subject to bias

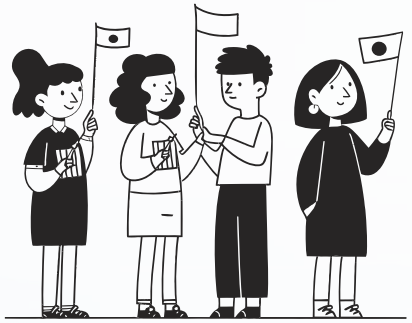


Descriptive, Not Predictive

One-time snapshot → no causal relationships

Isolated Factors

Missing organizational context



Missing Broader Analysis

Simplified burnout categories may miss nuance

Problem Statement

Burnout is widespread, but its root causes remain unclear and hard to isolate.

Unclear Causes

Burnout is widely acknowledged, but its **causes are unclear**

Overlapping Patterns

Lifestyle and work conditions often show **overlapping patterns** across groups

Hidden Signals

Traditional surveys highlight stress, yet **predictive signals remain hidden**

Data-Driven Need

Need to test whether **machine learning can uncover subtle drivers of burnout**

Complex Drivers

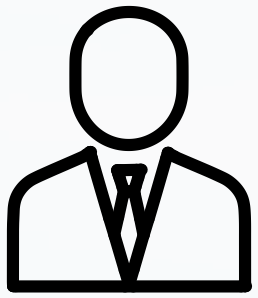
The challenge lies in the diverse and intricate factors contributing to burnout, spanning work hours, lifestyle choices, and demographic variations.

ML for Prediction

Our goal is to leverage machine learning to predict burnout instances and precisely identify the contributing factors at play.

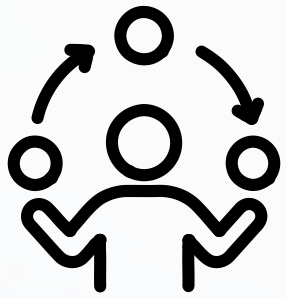
Burnout Drivers – Key Findings (3K Dataset)

Even with a smaller sample, the same story as the 50K dataset: burnout is driven by support, balance, growth, workload, and lifestyle not demographics.



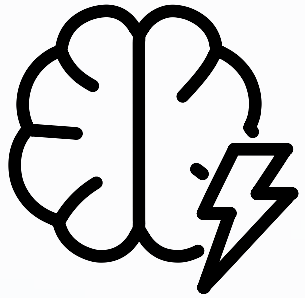
Manager Support

Low manager support strongly increases burnout #1 top driver



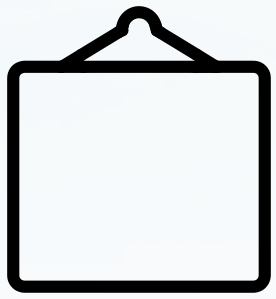
Work-Life & Job Satisfaction

Poor work-life balance, low job satisfaction, and limited career growth consistently raise burnout levels.



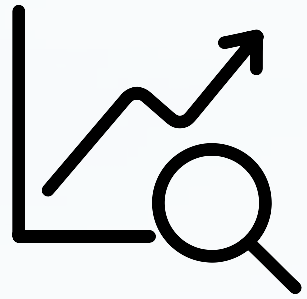
Stress & Lifestyle Factors

High stress, less sleep, low physical activity, and long commutes are significant contributors to higher burnout.



Structural Factors

Large teams, long work hours, and longer tenure within an organization also contribute to burnout.

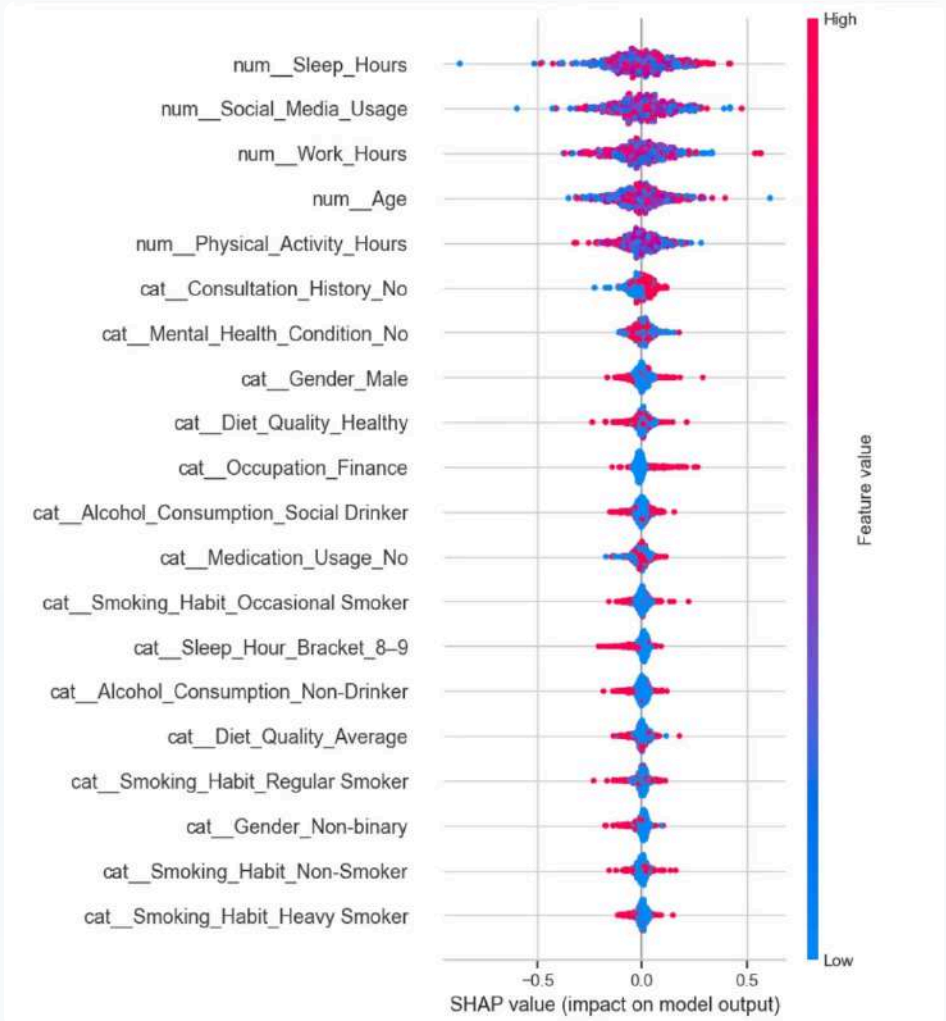


Demographics & Salary

These factors show minimal influence on burnout compared to the more impactful organizational and lifestyle drivers.

Model Interpretation - Key Drivers of Burnout

Understanding the Complex Web of Burnout Drivers



Factors that Increase Burnout Risk:

- Fewer sleep hours
- High social media usage
- Long work hours
- Little or no physical activity
- Poor/average diet quality
- Being a smoker