Mental Health and Burnout Prediction: A Cross-National Machine Learning Analysis

Abstract

This research investigates burnout prediction using a comprehensive dataset of 50,000 individuals across diverse demographics, occupations, and lifestyle factors. Through extensive exploratory data analysis and machine learning modeling, we identify key patterns in workplace stress and develop predictive models for burnout risk assessment. The study reveals that burnout is a multifactorial phenomenon affecting all demographics equally, with complex interactions between lifestyle, work patterns, and mental health outcomes. Our Random Forest model achieves 77.12% accuracy in predicting burnout risk, providing actionable insights for organizational interventions and workplace wellness programs.

Table of Contents

Introduction

Burnout has emerged as a critical workplace mental health concern, recognized by the World Health Organization (WHO) as a pressing global workplace crisis in 2019. This research aims to:

- Identify key predictors of workplace burnout using machine learning techniques
- Develop predictive models for early burnout risk assessment
- Analyze demographic and lifestyle factors contributing to stress levels
- Provide evidence-based recommendations for organizational interventions

The study addresses the challenge that burnout is widespread across demographics and occupations, with no single factor providing clear separation between at-risk and healthy populations. This necessitates sophisticated multivariate modeling approaches to uncover the complex interactions that drive burnout risk.

Research Questions

Primary Research Question: How do work-life conditions and lifestyle behaviors affect burnout across different demographics and occupational groups?

Secondary Questions:

1. Which lifestyle factors (sleep, physical activity, social media usage, diet) are most predictive of burnout risk?
2. How do work hours and sleep patterns interact to influence burnout probability?
3. Can machine learning models effectively predict burnout risk for early intervention?
4. What are the key business applications for burnout prediction models in organizational settings?

Dataset Overview

Data Source

- Source: Zenodo open research dataset (October 2024)
- Sample Size: 50,000 individuals

- Geographic Coverage: 7 countries (USA, India, Germany, Canada, Australia, UK, Other)
- Occupational Diversity: 7 major sectors (Finance, IT, Healthcare, Education, Engineering, Sales, Other)

Key Variables

- Demographics: Age (18-65), Gender (balanced distribution), Country, Occupation
- Work Patterns: Work hours (30-80 hrs/week), Consultation history
- Lifestyle: Sleep hours (4-10 hrs/night), Physical activity (0-10 hrs/week), Social media usage
- Health Factors: Diet quality, Smoking habits, Alcohol consumption, Medication usage
- Mental Health: Mental health condition status, Severity levels
- Target Variable: Stress level categorized as Low, Medium, and High (High = Burnout)

Data Quality

- Completeness: Strong data integrity with minimal missing values
- Missing Values: 50% missing values in Severity column handled through median imputation
- Class Balance: 33.3% burnout cases (High stress), 66.7% non-burnout cases

Methodology

1. Data Preprocessing

- Target Variable Creation: High stress level mapped to Burnout (1), Low/Medium to No Burnout (0)
- Missing Value Treatment: SimpleImputer for categorical and numeric data
- Feature Engineering: Created categorical brackets for continuous variables, interaction effects
- Encoding: One-hot encoding for categorical variables, standardization for numerical features

- Train-Test Split: 80-20 stratified split maintaining class balance

2. Exploratory Data Analysis

Comprehensive analysis including:

- Univariate Analysis: Distribution analysis of all variables
- Bivariate Analysis: Cross-tabulations and correlation analysis between variables and target
- Multivariate Analysis: Principal Component Analysis (PCA) for dimensionality reduction
- Business Rules Development: Identification of burnout hotspots through interaction analysis

3. Machine Learning Models

Four models evaluated with hyperparameter tuning:

Random Forest (Best Performer)

```
RandomForestClassifier(
    n_estimators=300,
    max_depth=10,
    min_samples_split=5,
    min_samples_leaf=3,
    max_features="sqrt",
    class_weight="balanced",
    random_state=42
)
```

XGBoost (Second Best)

```
XGBClassifier(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=6,
    min_child_weight=3,
    subsample=0.8,
    colsample_bytree=0.8,
    reg_alpha=0.1,
    reg_lambda=1.5,
    random_state=42
```

```
)
```

Logistic Regression

```
LogisticRegression(
    solver="saga",
    penalty="l2",
    C=0.1,
    max_iter=2000,
    random_state=42
)
```

Naive Bayes

```
GaussianNB(var_smoothing=1e-9)
```

4. Model Interpretability

- SHAP Analysis: SHapley Additive exPlanations for feature importance and model interpretability
- Feature Importance: Analysis of top predictive features across models
- Business Rules: Development of actionable insights from model predictions

Key Findings

Demographic Insights

- Gender Distribution: Balanced across all categories (Male: 24.9%, Female: 25.3%, Non-binary: 24.7%, Prefer not to say: 25.1%)
- Age Distribution: Concentrated in 21-40 age range (peak working years)
- Geographic Distribution: Relatively uniform across countries
- No Demographic Bias: Burnout affects all groups equally, indicating systemic rather than demographic-specific issues

Stress Level Distribution

- High Stress (Burnout): 27,055 cases (54.1%)
- Medium Stress: 11,612 cases (23.2%)
- Low Stress: 11,333 cases (22.7%)
- Business Rules Impact: 13,173 cases reclassified to High stress based on lifestyle risk factors

Lifestyle Pattern Analysis

Sleep Patterns

- Majority sleep 6-8 hours (normal distribution)
- Median sleep: 7.0 hours across all activity levels
- No strong correlation between physical activity and sleep duration

Work Patterns

- Work hours range: 30-80 hours per week
- Median: 55 hours per week across all groups
- Right-skewed distribution with extreme cases (80+ hours)

Physical Activity

- Range: 0-10 hours per week
- Median: 5 hours per week
- Balanced distribution across activity brackets

Burnout Hotspots Identified

1. Sleep Deprivation + Long Work Hours: <6 hours sleep + 55+ work hours per week (Highest burnout rate: ~71%)
2. Sedentary + High Social Media: Low physical activity + 4+ hours daily social media usage (~68% burnout rate)
3. Unhealthy Coping: Poor diet quality + regular/heavy alcohol consumption (~67-69% burnout rate)

Multivariate Analysis Results

- Correlation Analysis: Weak correlations between continuous variables (all <0.2)
- PCA Analysis: No clear separation between burnout and non-burnout groups
- Interaction Analysis: No significant hotspots in simple variable combinations
- Conclusion: Burnout is a complex, multifactorial phenomenon requiring advanced modeling

Model Performance

Performance Metrics (Weighted Average)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	77.12%	84.46%	77.12%	76.37%
XGBoost	76.51%	82.06%	76.51%	75.95%
Logistic Regression	68.01%	67.99%	68.01%	68.00%
Naive Bayes	65.29%	65.53%	65.29%	65.35%

Model Selection Rationale

Random Forest was selected as the final model due to:

- Highest overall performance metrics across all evaluation criteria
- Robust handling of mixed data types (categorical and numerical)
- Excellent interpretability through feature importance analysis
- Resistance to overfitting through ensemble approach
- Balanced performance across precision and recall metrics

SHAP Analysis and Feature Importance

Top Predictive Features

SHAP analysis revealed the following key predictors consistently across models:

1. Age-related factors: Complex non-linear relationships with burnout risk
2. Work hours: Particularly extreme values (>60 hours/week)
3. Sleep hours: Both insufficient (<6 hours) and excessive (>9 hours) sleep
4. Occupation-specific factors: Certain industries showing higher risk
5. Social media usage patterns: High usage (>4 hours/day) correlating with risk
6. Physical activity levels: Low activity (<2 hours/week) increasing risk
7. Diet quality indicators: Poor diet quality compounding other risk factors

Key Insights from SHAP Analysis

- Non-linear relationships: Tree-based models capture complex interactions better than linear models
- Age effects: Younger professionals (21-30) show different risk patterns than older workers
- Work-life balance: Extreme work hours combined with poor sleep create highest risk
- Lifestyle multiplicative effects: Poor diet + low activity + high social media usage compound risk

Business Applications

Organizational Interventions

Based on model insights, organizations should focus on:

Proactive Risk Mitigation

- Early Warning Systems: Implement model-based screening for high-risk employees

- Personalized Interventions: Target specific risk factors identified by the model
- Regular Monitoring: Continuous assessment of lifestyle and work pattern changes

Targeted Wellness Programs

- Work Hour Management: Monitor and limit excessive work hours (>60 hours/week)
- Sleep Hygiene Programs: Promote healthy sleep patterns (7-8 hours)
- Physical Activity Initiatives: Encourage regular exercise programs
- Mental Health Support: Proactive consultation and support services
- Digital Wellness: Manage social media usage and screen time

Policy Recommendations

- Implement maximum work hour policies
- Promote flexible working arrangements
- Invest in comprehensive employee wellness programs
- Regular mental health assessments and check-ins
- Create supportive, inclusive work cultures

ROI and Business Value

- Improved Productivity: Healthier, more engaged workforce
- Reduced Turnover: Lower recruitment and training costs
- Enhanced Retention: Better employee satisfaction and loyalty
- Optimized Resource Allocation: Focus wellness initiatives where they have greatest impact
- Risk Mitigation: Prevent burnout before it becomes costly

Interactive Dashboard

Experience our research findings through an interactive Streamlit dashboard that allows real-time exploration of:

- Model predictions and feature importance

- Demographic and lifestyle factor analysis
- Burnout risk assessment tools
- Data visualizations and insights

🔗 Live Dashboard: https://mentalhealthandburnoutdemo.streamlit.app/

Dashboard Features

- Individual Risk Assessment: Input personal data to get burnout risk prediction
- Feature Impact Analysis: Understand how different factors affect your risk score
- Population Analytics: Explore patterns across different demographic groups
- Interactive Visualizations: Dynamic charts and plots for data exploration

Installation & Usage

Prerequisites

- Python 3.8 or higher
- Required packages listed in `requirements.txt`

Installation Steps

1. Clone the Repository

```
git clone https://github.com/B1raj/Mental_Health_and_Burnout.git
cd Mental_Health_and_Burnout
```

2. Set up Virtual Environment

```
python -m venv venv
source venv/bin/activate  # On Windows: venv\Scripts\activate
```

3. Install Dependencies

```
pip install -r requirements.txt
```

Usage Examples

Running the Jupyter Notebook Analysis

```
jupyter notebook Mental_Health_Workplace_Survey.ipynb
```

Using the Trained Model

```
import pickle
import pandas as pd
import numpy as np

# Load the trained model
with open('trained_model.pkl', 'rb') as f:
    model = pickle.load(f)

# Prepare your data (ensure it matches the training format)
# Make predictions
predictions = model.predict(your_data)
probabilities = model.predict_proba(your_data)
```

Running the Streamlit Dashboard Locally

```
cd streamlit
streamlit run app.py
```

Project Structure

```
Mental_Health_and_Burnout/
├── README.md                                    # This comprehensive guide
├── LICENSE                                      # Project license
├── requirements.txt                             # Python dependencies
├── Mental_Health_Workplace_Survey.ipynb         # Main analysis notebook
├── trained_model.pkl                            # Trained Random Forest model
├── data/
│   └── mental_health_data_final_data.csv        # Primary dataset
├── streamlit/                                   # Interactive dashboard
│   ├── app.py                                   # Main dashboard application
│   ├── dashboard.py                             # Dashboard components
│   ├── mental_health_demo.py                    # Demo functionality
│   ├── requirements.txt                         # Dashboard dependencies
│   └── trained_model.pkl                        # Model for dashboard
├── archieve/                                    # Previous versions and
```

Technical Implementation Details

Data Pipeline Architecture

```
# Preprocessing Pipeline
numeric_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])

categorical_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(handle_unknown='ignore'))
])

preprocessor = ColumnTransformer([
    ('num', numeric_pipeline, numeric_columns),
    ('cat', categorical_pipeline, categorical_columns)
])
```

Model Training Configuration

```
# Cross-validation strategy
cv_strategy = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Hyperparameter tuning
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 3, 5]
}

grid_search = GridSearchCV(
    estimator=RandomForestClassifier(random_state=42),
    param_grid=param_grid,
    cv=cv_strategy,
    scoring='f1_weighted',
    n_jobs=-1
)
```

Evaluation Metrics

```
def evaluate_model(model, X_test, y_test):
```

```
    predictions = model.predict(X_test)
    probabilities = model.predict_proba(X_test)[:, 1]

    metrics = {
        'accuracy': accuracy_score(y_test, predictions),
        'precision': precision_score(y_test, predictions, average='weighted'),
        'recall': recall_score(y_test, predictions, average='weighted'),
        'f1_score': f1_score(y_test, predictions, average='weighted'),
        'roc_auc': roc_auc_score(y_test, probabilities)
    }

    return metrics
```

Limitations and Future Work

Study Limitations

- Cross-sectional Design: Limits causal inference capabilities
- Self-reported Data: Potential for response bias and social desirability effects
- Missing Data: 50% missing severity data required imputation strategies
- Temporal Constraints: Limited longitudinal data for trend analysis
- Confounding Variables: Potential unmeasured factors affecting burnout risk

Future Research Directions

Short-term Enhancements

- Industry-specific Models: Develop tailored models for different sectors
- Temporal Analysis: Incorporate time-series data for trend prediction
- External Validation: Test model performance on independent datasets
- Feature Engineering: Explore advanced feature interactions and transformations

Long-term Research Goals

- Longitudinal Studies: Establish causal relationships through time-series analysis
- Intervention Studies: Measure effectiveness of model-guided interventions
- Cross-cultural Analysis: Expand to more diverse geographic and cultural contexts

- Integration with Wearables: Incorporate objective physiological measures
- Real-time Monitoring: Develop continuous assessment and alert systems

Advanced Technical Approaches

- Deep Learning Models: Explore neural network architectures for complex pattern detection
- Ensemble Methods: Combine multiple model types for improved performance
- Federated Learning: Enable privacy-preserving multi-organizational studies
- Explainable AI: Develop more intuitive model interpretation tools

Contributing

We welcome contributions to improve this research and expand its applications. Please see our contribution guidelines:

How to Contribute

1. Fork the Repository: Create your own fork of the project
2. Create Feature Branch: `git checkout -b feature/AmazingFeature`
3. Make Changes: Implement your improvements or fixes
4. Add Tests: Ensure your changes include appropriate tests
5. Commit Changes: `git commit -m 'Add AmazingFeature'`
6. Push to Branch: `git push origin feature/AmazingFeature`
7. Open Pull Request: Submit your changes for review

Areas for Contribution

- Model Improvements: Enhanced algorithms or feature engineering
- Data Analysis: Additional EDA or statistical insights
- Visualization: Improved charts, plots, or dashboard features
- Documentation: Better explanations, tutorials, or examples
- Testing: Unit tests, integration tests, or validation studies

Research Collaboration

For academic collaborations or research partnerships, please contact the team through GitHub issues or direct communication.

License

This project is licensed under the MIT License - see the LICENSE file for details.

Data Usage Rights

- The dataset is sourced from Zenodo under open research data policies
- Commercial use is permitted under the MIT license terms
- Attribution to original research is required for derivative works
- Modifications and distributions must retain license notices

Acknowledgments

Research Team

MSDSP 422B – Practical Machine Learning Team

- Anoushka
- Biraj
- Hunter
- Samridhi

Data Sources

- Primary Dataset: Zenodo Open Research Platform
- Validation Studies: Multiple peer-reviewed sources for comparative analysis
- Industry Benchmarks: WHO guidelines and occupational health standards

Technical Infrastructure

- Computing Resources: Northwestern University HPC facilities
- Development Tools: Python ecosystem (pandas, scikit-learn, XGBoost, SHAP)
- Deployment Platform: Streamlit Cloud for interactive dashboard

Citations and References

Primary References

1. Shanafelt, T. D., et al. (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
2. Sánchez-Oliva, D., et al. (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
3. Rupp, A. (2014). "Burnout, Stress, and Coping Mechanisms among Psychology Graduate Students." *PCOM Psychology Dissertations*, 144. https://digitalcommons.pcom.edu/psychology_dissertations/144
4. Folkman, S., & Moskowitz, J. T. (2000). "Positive Affect and the Other Side of Coping." *American Psychologist*, 55(6), 647-54. https://pubmed.ncbi.nlm.nih.gov/20561174/
5. Åkerstedt, T., & Wright, K. P. (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/

Dataset Citation

```
@dataset{mental_health_2024,
    title={Mental Health and Lifestyle Dataset for Burnout Prediction},
    author={Anonymous Contributors},
    year={2024},
    publisher={Zenodo},
    version={1.0},
    url={https://zenodo.org/}
```

```
}
```

Project Citation

```
@article{burnout_ml_2024,
    title={Mental Health and Burnout Prediction: A Cross-National Machine
Learning Analysis},
    author={Anoushka and Biraj and Hunter and Samridhi},
    year={2024},
    journal={MSDSP 422B Final Project},
    institution={Northwestern University},
    url={https://github.com/B1raj/Mental_Health_and_Burnout}
}
```

Contact Information

Project Repository

🔗 GitHub: https://github.com/B1raj/Mental_Health_and_Burnout

Interactive Dashboard

🔗 Live Demo: https://mentalhealthandburnoutdemo.streamlit.app/

Research Documentation

For detailed technical documentation, methodology, and extended results, refer to the complete academic report included in the repository.

This research contributes to the growing understanding of workplace mental health and provides actionable insights for creating healthier, more sustainable work environments through evidence-based machine learning approaches.