1. Machine Learning Final Project Proposal

2. Student Health Analysis

2.1 Introduction

For our project, we have selected the Student Health Data dataset from Kaggle, which is available at Student Health Data - Kaggle. The dataset provides a comprehensive simulation of physiological, psychological, and academic data for college students. The primary focus is on assessing health risks in high-stress environments such as academic settings. Our objective is to leverage this dataset to develop machine learning models that can predict and evaluate health risks, based on numerous factors like sleep deprivation, stress, physical, and mental health. The dataset includes various health indicators such as heart rate, blood pressure, stress levels (both biosensor and self-reported), along with behavioral factors like study hours and physical activity, making it ideal for developing a robust health risk prediction system.

2.2 Approach (Supervised vs Unsupervised)

We decided to divide our group into two further subgroups so that we could tackle different questions using different target variables– Student Health Risk and Self Reported Stress Levels.

Target Variable (Health Risk Level):
We are addressing a supervised learning problem since we have a clearly defined target variable (Health_Risk_Level) with labeled categories (Low, Moderate, High), and our goal is to predict these health risk levels using various input features. This is specifically a multi-class classification problem where we utilize labeled training data to train our models to recognize patterns and relationships between the input features (such as stress levels, sleep quality, physical activity, and physiological measurements) and the corresponding health risk outcomes. The supervised learning approach is appropriate for our objective as we aim to learn from known health risk classifications to predict future risk levels for students based on their health indicators.

Target Variable: (Self-Reported Stress Levels):
Our analysis addresses a supervised learning problem where the model is trained on a labeled dataset to predict the target variable, which in this case is self-reported stress levels. This is specifically a regression problem, where the goal is to understand the relationships between various input features (e.g., physical activity, sleep quality, mood, and physiological metrics) and the continuous target variable of stress levels. By leveraging labeled data, the supervised

approach enables the model to identify patterns and relationships to accurately predict stress levels for students based on their health and behavioral indicators.

3. Source and Background of the Data

3.1 Dataset Overview

This dataset has been sourced from Kaggle and is titled Student Health Data. It is designed to support the development and testing of machine learning models for predicting health risks in high-pressure environments. Each row of the dataset represents a unique student's health, workload, and lifestyle characteristics.

3.2 Key Features of the Dataset

Demographic Information:

- Age and gender data are provided, along with unique student identifiers.

Physiological Data:

- Includes real-time biosensor metrics such as heart rate and blood pressure (systolic and diastolic).
- Provides stress levels derived from biosensor measurements to assess physical health.

Psychological Data:

- Self-reported stress levels and mood states offer insights into students' mental and emotional well-being.

Academic and Entrepreneurial Activity:

- Captures the hours students spend on academic tasks and entrepreneurial projects, reflecting workload and time management skills.

Physical Activity and Sleep Quality:

- Categorized data on daily physical activity and sleep quality, which are critical factors in overall health.

Health Risk Level (Target Variable):

- This target variable indicates the health risk level (low, moderate, or high), derived from combinations of physiological and psychological metrics.

4. Problems Addressed by our ML model

- Can we predict a student's health risk category based on various health-related features (e.g., age, stress, physical activity, sleep quality)?
- What are the most important factors (features) contributing to health risk prediction, and how can they be ranked effectively?
- How can we visualize the distribution of independent variables across different health risk levels to identify patterns?
- Which machine learning algorithms (KNN, Bagging, Decision Tree, Random Forest) provide the best performance for predicting health risk, and how do they compare in terms of accuracy, precision, recall, and F1 score?
- What is the relationship between objective (biosensor) and subjective (self-reported) stress measurements in predicting student health risks?
- To what extent can ensemble methods improve prediction accuracy compared to individual models?
- What are the most significant factors influencing a student's self-reported stress levels?
- How do lifestyle choices, such as sleep quality and physical activity, impact a student's self-reported stress?
- What physiological indicators, like heart rate or blood pressure, are associated with stress?
- What behavioral and academic factors contribute to a student's self-reported stress levels?
- Can at-risk students be identified based on their health and lifestyle data?
- How accurately can a machine learning model predict a student's subjective (self-reported) stress levels using the given features?

5. Group Contribution & Techniques Used

5.1 Group Members

- Ritika Dawadi
- Prashant Shah
- Samman Bhetwal
- Sumit Shiwakoti

5.2 Ritika Dawadi & Prashant Shah

Target Variable: Student Health
Type: Supervised Machine Learning Problem as the model is trained on labeled dataset, and the model is used to predict a target variable which is student health in this case.

5.2.1 Exploratory Data Analysis (EDA) and Insights Generation:

Conducted thorough EDA to understand dataset structure and characteristics:

- Generated distribution plots for numerical and categorical variables.
- Created stacked histograms with dynamically calculated subplots to visualize independent variable distributions, excluding the target column and hiding unused plots for clarity.
- Analyzed feature distributions across different risk levels and computed detailed statistical summaries.
- Examined basic dataset information, identified missing values, and analyzed gender-based differences in health metrics.

5.2.2 Data Preprocessing for Optimal Model Performance

Cleaned and Prepared Data Through Comprehensive Preprocessing:

- Removed unnecessary identifiers (e.g., Student_ID).
- Converted quantitative variables (e.g., Physical Health, Sleep, Mood) into integer dummy variables (-1, 0, 1).
- Standardized numerical features using StandardScaler and normalized data to ensure consistency.
- Split data into training (80%) and test (20%) sets to maintain balanced representation.

5.2.3 Relationship Analysis Between Features

Investigated Relationships and Dependencies Among Variables:

- Computed and visualized a correlation matrix heatmap to identify feature relationships, with a focus on correlations with the target variable (Health_Risk_Level).
- Explored stress-related relationships and specialized health metric visualizations.
- Analyzed key variable relationships across different health risk levels.

5.2.4 Implementation and Evaluation of Classifiers

Deployed and Evaluated Multiple Classifier Algorithms:

- **Logistic Regression**: *(PS)*
- **K-Nearest Neighbors (KNN)**: *(RD)*
- **Decision Tree** *(RD & PS)*
- **Random Forest** *(RD & PS)*
- **Stochastic Gradient Descent (SGD)** *(RD)*
- **Support Vector Machine (SVM)** *(PS)*

- **Bagging** *(RD)*
- **Voting Classifier** *(PS)*

5.2.5 Performance Metrics and Analysis

Computed and analyzed key performance metrics for all models:

- Accuracy, Precision, Recall, and F1 Scores.
- Cross-validation scores for robust evaluation.
- Confusion matrices and heatmaps to assess classification performance.
- ROC curves and AUC scores for model comparison.
- Learning curves to understand model behavior over training iterations.

5.2.6 Comprehensive Visualization & Analyzation Techniques

Created visualizations to interpret data and model results effectively:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R^2 Score
- Feature importance plots for tree-based models.
- Model comparison plots to highlight strengths and weaknesses.
- Learning curve plots and ROC curves for performance comparison.
- Confusion matrix heatmaps for detailed error analysis.
- Distribution plots for all features.

5.2.7 Detailed Model Comparison and Feature Analysis

Conducted in-depth comparison and evaluation across models:

- Compared accuracy and other metrics to identify the optimal model.
- Assessed the effectiveness of ensemble methods and model-specific advantages.
- Generated feature importance rankings:
    - Analyzed key predictors of health risk.
    - Ranked features using various importance evaluation methods across models.
- Provided actionable insights into the most influential factors driving health risk predictions.

5.3 Sumit Shiwakoti & Samman Bhetwal

Target Variable: Self-Reported Stress Levels
Type: Supervised Machine Learning Problem as the model is trained on a labeled dataset and used to predict the target variable, which is stress levels in this case.

5.3.1 Exploratory Data Analysis (EDA) and Insights Generation

Conducted thorough EDA to uncover patterns, trends, and relationships within the dataset:

- Displayed dataset structure, data types, and descriptive statistics for initial insights.
- Investigated missing values and confirmed their appropriate handling or absence.
- Generated visualizations to explore feature distributions and relationships:
    - Histograms: Displayed distributions for individual features such as Age, Heart Rate, Blood Pressure, and Stress Levels.
    - Scatterplots: Explored relationships between continuous variables (e.g., Study Hours, Sleep Quality) and the target variable, Stress_Level_Self_Report.
    - Correlation Heatmaps: Assessed dependencies and multicollinearity among features.
- Analyzed feature distributions across different health and stress levels and computed detailed statistical summaries.

5.3.2 Data Preprocessing for Optimal Model Performance

Cleaned and prepared data through comprehensive preprocessing steps:

- Handled missing values and encoded categorical variables using one-hot encoding and logical point mappings for features like Gender, Physical Activity, Sleep Quality, and Mood.
- Removed unnecessary identifiers such as Student_ID to streamline data analysis.
- Scaled numerical features using StandardScaler to normalize data for consistent input across models.
- Engineered features by converting categorical variables (e.g., Mood, Physical Activity) into interpretable numeric representations.
- Split the dataset into training (80%) and testing (20%) sets for balanced representation.

5.3.3 Relationship Analysis Between Features

Investigated relationships and dependencies among variables:

- Computed and visualized a correlation matrix heatmap to identify feature relationships, focusing on correlations with the target variable (Stress_Level_Self_Report).
- Explored stress-related relationships through scatter plots (e.g., Blood Pressure vs. Stress) and boxplots (e.g., Mood vs. Stress Levels).
- Analyzed multifactorial relationships, such as combining Sleep Quality and Physical Activity to understand their combined effect on stress levels.

5.3.4 Implementation and Evaluation of Regression Models

Deployed and evaluated multiple regression algorithms:

- **Linear Regression** *(SB)*
- **Random Forest Regressor** *(SB & SS)*
- **Decision Tree Regressor** *(SB)*
- **Support Vector Regressor** *(SS)*
- **Bagging Regressor** *(SS)*
- **Voting Regressors** *(SS)*

5.3.5 Performance Metrics and Analysis

Computed and analyzed key performance metrics for all models:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R^2 Score

Generated graphical representations for result analysis:

- Residual Plots: Evaluated model fit by analyzing residual distributions.
- Prediction vs. Actual Plots: Compared predictions to actual values for intuitive performance evaluation.
- Error Distribution Plots: Analyzed prediction error spread to detect biases or inconsistencies.

5.3.6 Comprehensive Visualization and Analysis Techniques

Created detailed visualizations to interpret data and model results effectively:

- Feature Importance Plots: Illustrated rankings of feature contributions to stress levels using Random Forest and Gradient Boosting models.
- Scatter Plots: Visualized relationships between key features and stress levels.
- Boxplots and Count Plots: Highlighted trends and distributions for categorical variables.
- Heatmaps: Displayed dependencies and multicollinearity.

5.3.7 Detailed Model Comparison and Feature Analysis

Conducted in-depth evaluation across models:

- Compared accuracy and performance metrics to identify the best-performing model.
- Assessed the effectiveness of ensemble methods and specific model advantages.
- Generated feature importance rankings to identify key predictors of stress levels:
    - Analyzed the impact of behavioral and academic factors, such as Physical Activity, Sleep Quality, and Study Hours, on self-reported stress levels.
- Provided actionable insights into the most influential factors driving stress predictions.