Predicting Health Risks in Students Relative to Physiological and Psychological Metrics

Christopher Guzman, Dereck Lin, Haris Nioulikos, Samiha Uddin

Goals and Motivation

- Studies have shown that student physiological and psychological health can be predicted from lifestyle and habit, ie:
  - Chowdhury et al. (2025)
  - Zhang et al. (2024)
- Predict student’s health risk levels given demographic, physiological, psychological, occupational, and lifestyle data

Data

- n = 1000
- Imbalanced dataset
- Simulated data
- Label:
  - Health_Risk_Level
- Features
  - demographic, physiological, psychological, occupational, and lifestyle data

Processing

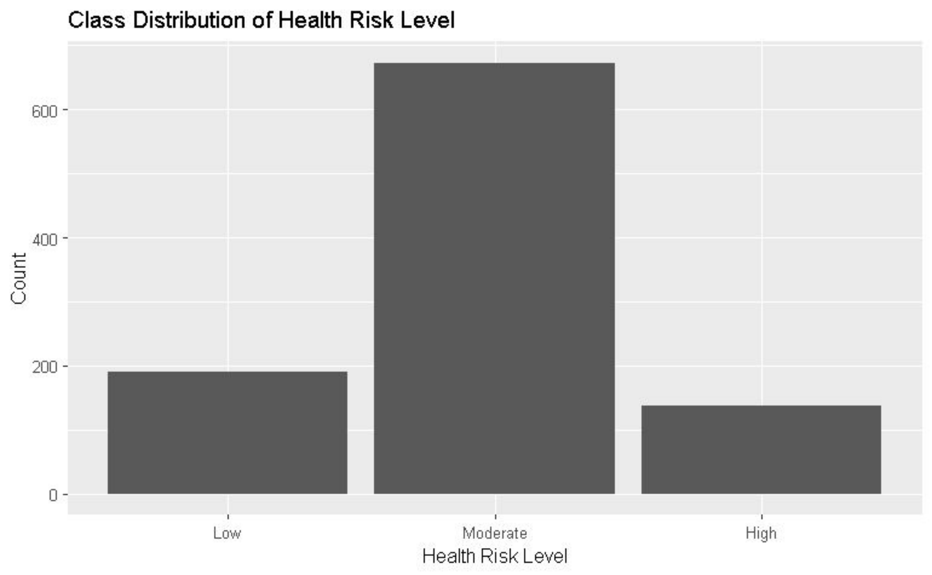
- Encoded response label using One-Hot encoding
- Removed the unique identifier Student_ID
- Encoded categorical features
    - Features that were ordinal in nature were encoded as 0, 1, 2
    - Features that were binary in nature were encoded as 0, 1

Data Splits

- Test, 20%
- Train, 80%
    - Fold1
    - Fold2
    - Fold3
    - Fold4
    - Fold5

Baseline

- Majority rule
  - Always predict the most frequent class
  - Overall accuracy on test set: 67.33%



Methodology

- For each model:
  - Tune model hyperparameters using cross validation on previously defined folds
  - Fit each model on entire train set using tuned hyperparameters
  - Predict on test set
- Evaluated models using metrics such as log loss, accuracy, and sensitivity

Logistic regression

- One vs. Rest strategy to transform multiclass problem into a series of binary problems
- Independently trained a series of binary classifiers for each class label following our methodology
- Normalized the resultant vector of probabilities to make a prediction on the hard class label

Logistic regression results

- Overall accuracy of 85.93%
- Sensitivity on Class = Moderate: 97.01%
- Sensitivity on Class = High: 37.037%
- Strong predictors for High by magnitude:
    - Physical_Activity
    - Sleep_Quality

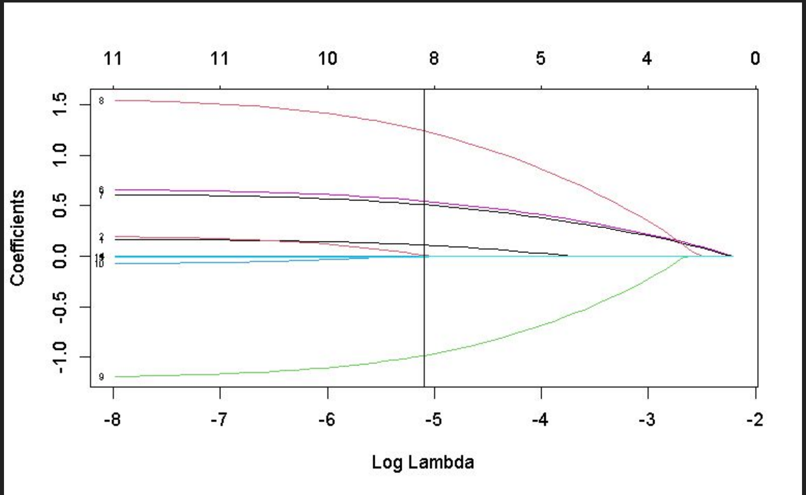
	Actual		
Predicted	Low	Moderate	High
Low	31	1	3
Moderate	7	130	14
High	0	3	10

LASSO

- Technique that simplifies model
  - uses Lambda, λ, L1 a penalty to minimize weak predictors
- Same process as our Logistic Regression model
  - Employed One vs. Rest Strategy
  - Normalized probabilities and generated hard class predictions

Lasso results

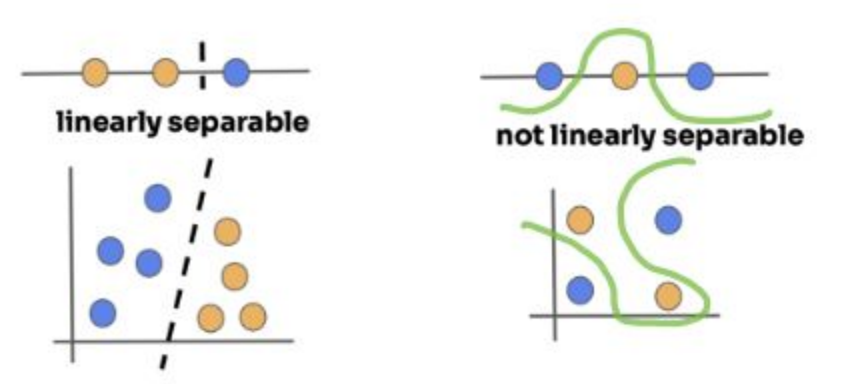
- Optimal λ for Class = High: 0.06117
- Overall accuracy of 83.93%
- Sensitivity on Class = Moderate: 98.51%
- Sensitivity on Class = High: 22.22%
- Strong predictors for High by magnitude:
  - Physical_Activity
  - Sleep_Quality
- Strongest Predictors for High by solution path:
  - Stress_Level_Self_Report
  - Stress_Level_Bio_Sensor



	Actual		
Predicted	Low	Moderate	High
Low	28	1	3
Moderate	10	132	18
High	0	1	6

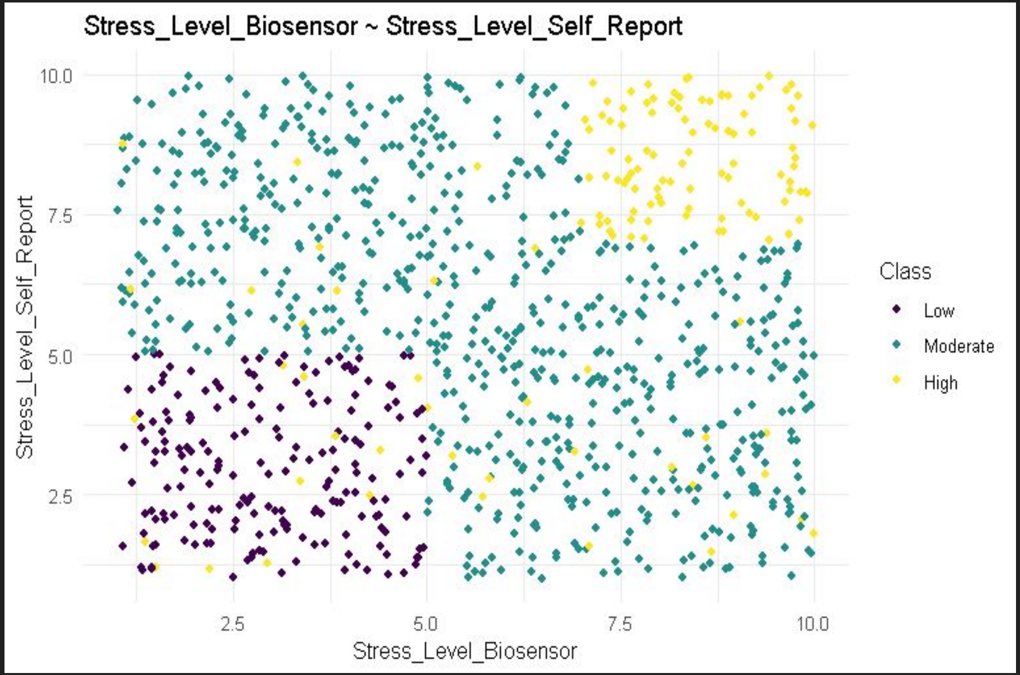
Analysis

- Overall accuracy decreased
- Sensitivity on Class = High decreased
- Simplifying the model made it more biased towards the majority class
- Model’s linear nature may not be able to capture the minority class



Analysis cont.

- Linear classifier fails because predicting High increases overall loss
- Next steps:
  - Adjust weights to penalize misclassifying High
  - Choose a more fitting model
    - Decision tree

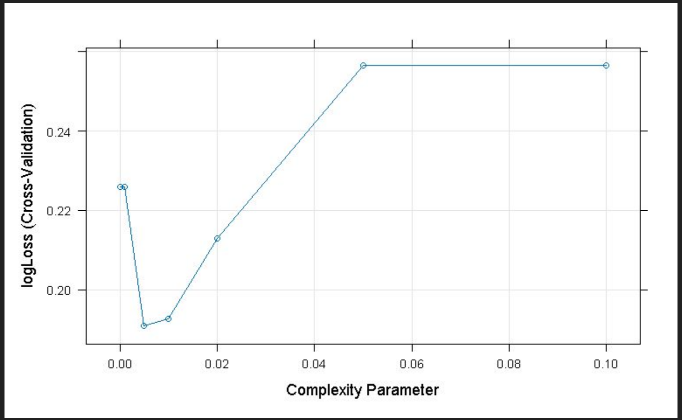


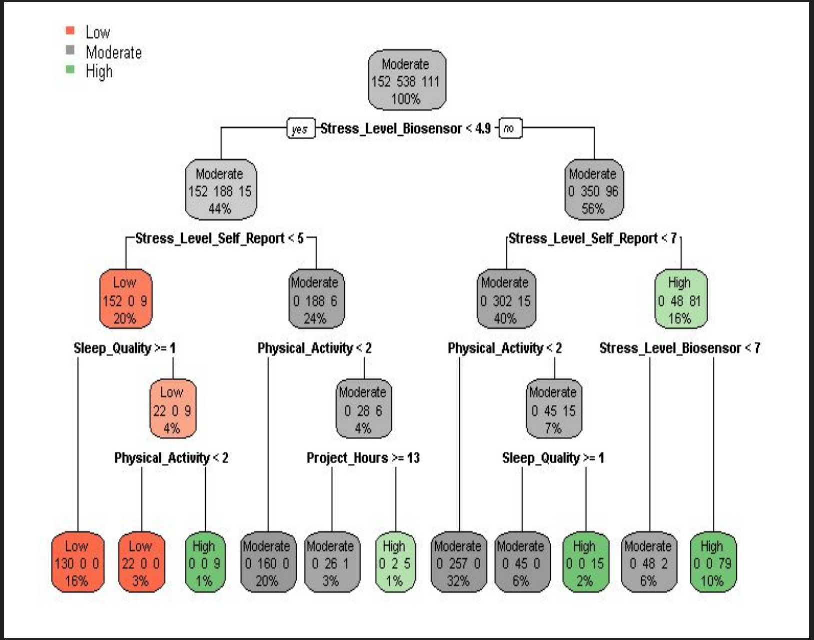
Decision Tree

- Constructs binary splits using a singular feature
  - More splits can approximate a better decision boundary for our target
- Naturally multiclass, no need for One vs. Rest
- Used cv to tune the complexity parameter

Decision Tree results

- Optimal complexity parameter: 0.005
- Overall accuracy: 95.98%
- Sensitivity on Class = High: 96.3%



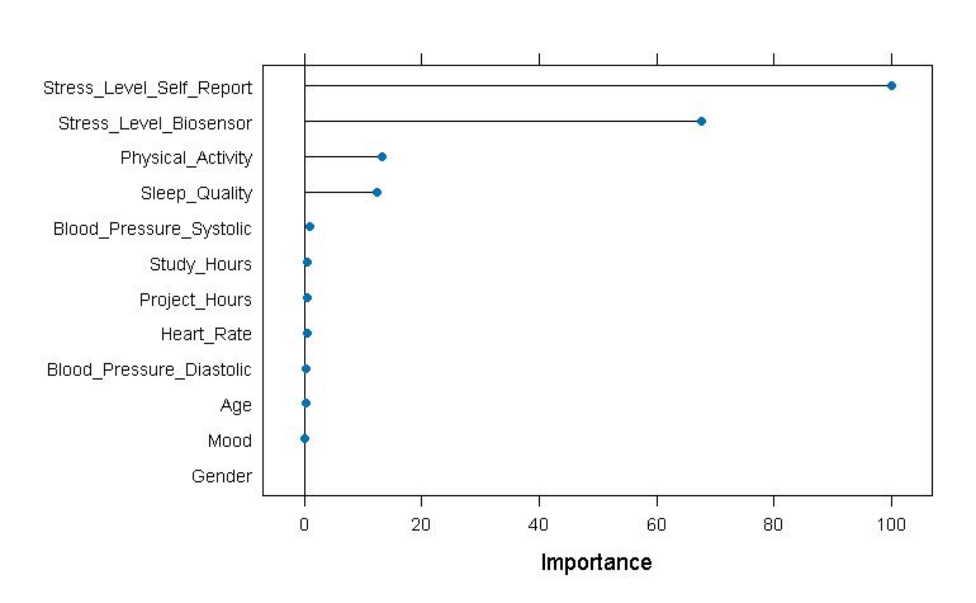


Random forest

- Use more trees to average a prediction
  - M = 500
- Used cv to tune mtry, the number of features considered in each split

Decision tree results

- mtry = 12
- Overall accuracy: 98.99%
- Sensitivity on High: 100%
- Stress_Levels were the strongest predictor; consistent with lasso & decision tree



Final Results

	Baseline	Logistic Regression	LASSO	Decision Tree	Random Forest
Accuracy	67.33%	85.93	83.42%	95.88%	98.99%
Sensitivity (High)	0%	37.037%	22.22%	96.3%	100%

Conclusion

- Success
  - Models better than baseline
  - New model outperformed previous model
- Failure
  - Does not fit our goal of preventative care or early detection
  - Similar concerns as with Chowdhury et al; poor generalization