Supporting Student Wellness at Northeastern: A Machine Learning Approach

By: Madison Rodriguez and Sai Prashanth Vaditha

Executive Summary

Business Opportunity

- Use **machine learning** to identify obesity risk levels among students, supporting Northeastern's wellness program in creating **targeted** health interventions.

Project Objective

- Develop and compare **logistic regression** and **decision tree** models to predict obesity categories based on lifestyle data.

Background

Context

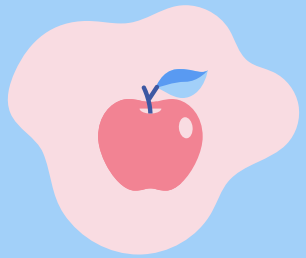
As students transition to college life, they face **challenges** with maintaining a balanced diet, regular physical activity, and healthy habits for the first time.

Importance

By understanding factors that contribute to obesity, Northeastern can provide **targeted** resources and support to students in adopting healthier habits.

Obesity: increased risk of heart disease, diabetes, and mental health disorders
Underweight: weakened immune systems, nutrient deficiencies, and decreased energy levels

Core Team & Stakeholders

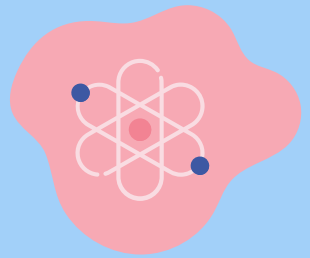


Core Team

Data science and health analytics team within Northeastern.

Internal Stakeholders

University wellness programs, campus dining, fitness teams, and mental health services.



External Stakeholders

Students and potential partners in health-focused initiatives.

Scope & Objectives

Scope

- **Predict** obesity risk categories (i.e. Insufficient Weight, Normal Weight, Obesity Types I-III).
- **Assess** feature importance to inform wellness programming (i.e. physical activity and diet).

Exclusions

- Personalized Health Recommendations
- Medical Advice

Project Development Timeline

1

Research

Explore databases & create a business idea

2

Data Exploration

Clean, preprocess, & explore data to find patterns/correlations.

3

Model Development

Train logistic regression & decision tree models.

4

Validation

Evaluate model performance with metrics like accuracy, precision, and recall

5

Conclusions

Select the “best” model and summarize key insights

Development Strategy

Logistic Regression

- **Interpretability**: Provides correlation coefficients which shows how each feature affects obesity risk.
- **Multiclass Classification**: Handles multiple obesity categories well.

Decision Tree

- **Non-Linear Relationships**: Captures complex feature interactions.
- **Feature Importance**: Identifies key factors influencing obesity.
- **Visual Representation**: Clearly shows how factors determine categories.

Data Dictionary

Gender	Male/Female
Age	Age of the individual
Height	(meters)
Weight	(kgs)
Family_history_with_overweight	Whether the individual has a family history of being overweight (Yes/No)
FAVC	Frequent consumption of high-calorie food (Yes/No)
FCVC	Frequency of vegetable consumption (1-3 scale)
CALC	Alcohol consumption frequency (Sometimes, Frequently, Always, etc.)

MTRANS	Mode of transportation (e.g., Walking, Public Transport)
NCP	Number of meals per day
CAEC	Consumption of food between meals (Sometimes, Frequently, Always, etc.)
SMOKE	Whether the individual smokes (Yes/No)
CH2O	Daily water consumption (in liters)
SCC	Monitoring calorie consumption (Yes/No)
FAF	Physical activity frequency (number of days per week)
TUE	Time spent using technology (hours per day)

NObeyesdad	Obesity Level

Development Details

Data Overview

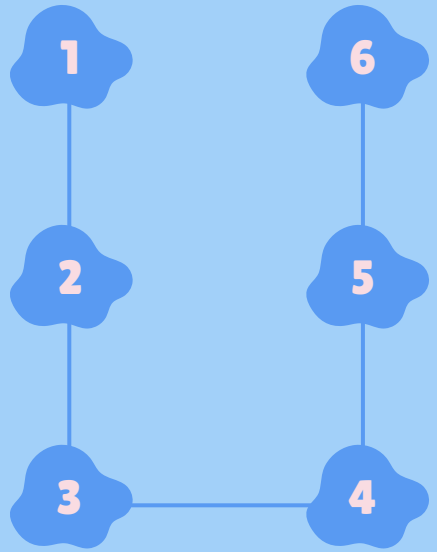
Reviewed dataset structure, statistics, and unique values.

Exploratory Data Analysis

Used pie charts, heatmaps, and histograms to explore patterns and correlations.

Data Preparation

Rounded fields, removed missing values/duplicates, encoded categories, and normalized data.



Performance Evaluation & Model Comparison

Calculated accuracy and recall for all classes to select the best model.

Model Building

Built Logistic Regression and Decision Tree models for obesity prediction.

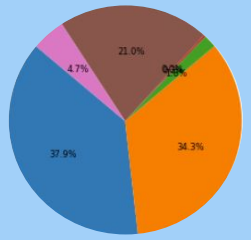
Data Partitioning

Split data into 70% training and 30% testing for model evaluation.

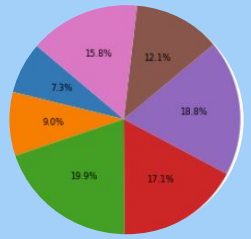
Exploratory Data Analysis

Family_history_with_overweight

No



Yes

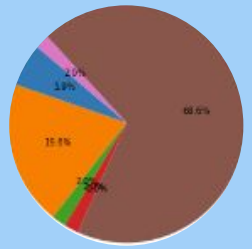


CAEC

"Snacking" between meals

Always

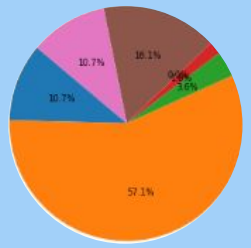
No



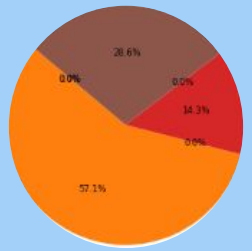
MTRANS

Mode of Transportation

Walking



Bike



Public Transaction

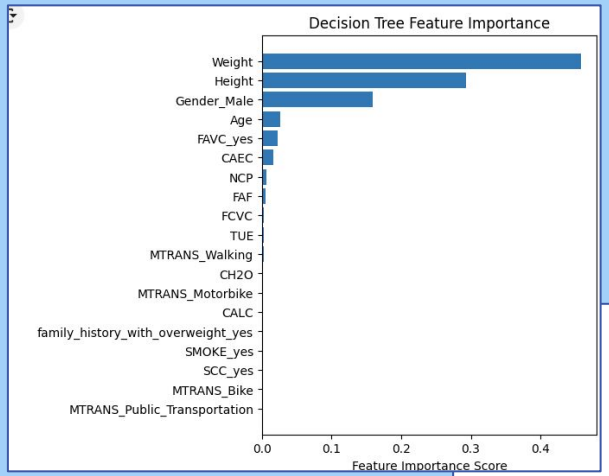
Logistic Regression Coef.

	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Weight	-11.158957	-6.283290	4.310116	8.872693	7.342089	-2.746405	-0.336246
FCVC	0.108355	-0.251114	-0.489973	-0.217877	1.467257	-0.373335	-0.243313
NCP	-0.049033	-0.132080	-0.340058	-0.147464	1.025320	-0.055239	-0.301445
family_history_with_overweight_yes	-0.530057	-0.465896	0.225055	0.074220	0.860473	-0.382094	0.218299
CALC	-0.158852	-0.119775	-0.193405	-0.317404	0.766906	0.269299	-0.246769
MTRANS_Public_Transportation	-0.200386	-0.124611	-0.274926	0.342917	0.510016	-0.257480	0.004470
FAVC_yes	-0.169249	-0.211777	0.535561	-0.395574	0.379265	0.267454	-0.405680
CAEC	0.237601	0.397113	-0.140824	-0.307008	0.246612	-0.276186	-0.157308
CH2O	0.048152	-0.026635	0.189308	-0.472328	0.238529	-0.013279	0.036254
TUE	-0.006980	-0.141505	0.102654	-0.186613	0.176285	-0.157083	0.213243
SMOKE_yes	-0.420035	0.097527	0.183198	0.015723	0.032301	0.027497	0.063789
Age	-0.848809	-0.264630	-0.150824	1.116658	0.016648	-0.268863	0.399820
MTRANS_Bike	-0.060909	0.107671	-0.033272	-0.006770	-0.013146	0.185720	-0.179293
MTRANS_Motorbike	-0.159564	0.153767	0.153159	-0.017547	-0.030611	-0.264659	0.165455
MTRANS_Walking	0.006589	0.227964	-0.122839	-0.051135	-0.079266	0.101059	-0.082373
SCC_yes	0.037369	0.134012	-0.548350	0.128548	-0.094791	0.342830	0.000382
FAF	0.141419	0.326934	0.283081	-0.185630	-0.760163	0.166569	0.027791
Height	2.964133	1.476501	-1.629586	-2.411090	-1.410664	0.909666	0.101039
Gender_Male	-0.131720	0.406356	-0.053470	1.899093	-2.605933	-0.132017	0.617690

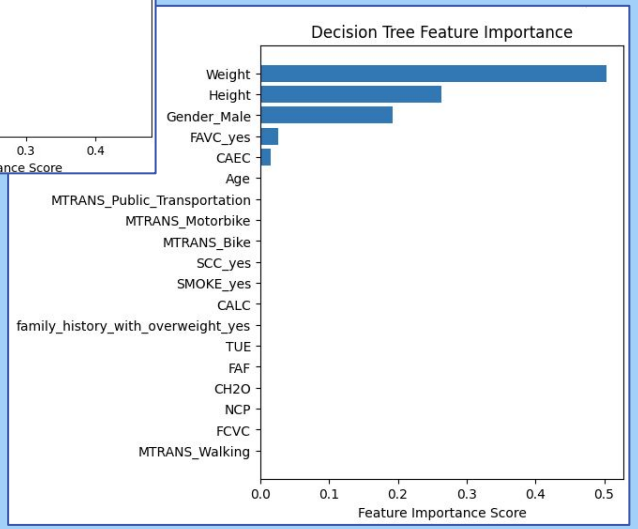
- **Height and Weight:** Both height and weight are primary predictors.
- **Family History:** Individuals without a family history of being overweight have a higher likelihood of being underweight or normal weight.
- **Gender:** Males are more likely in Obesity Type II and females more likely in Obesity Type III.
- **Alcohol:** Higher alcohol consumption shows a correlation with Obesity Type III.

Decision Tree Feature Importance

Un-Pruned

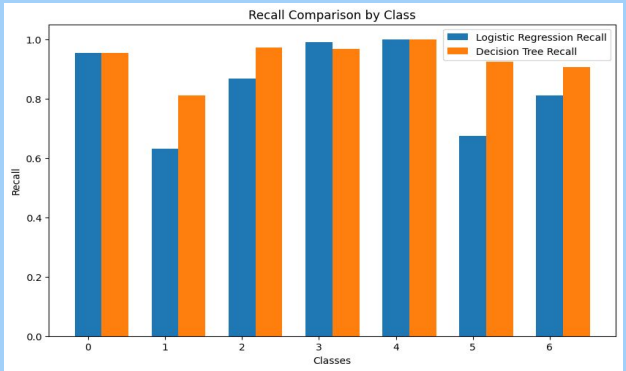


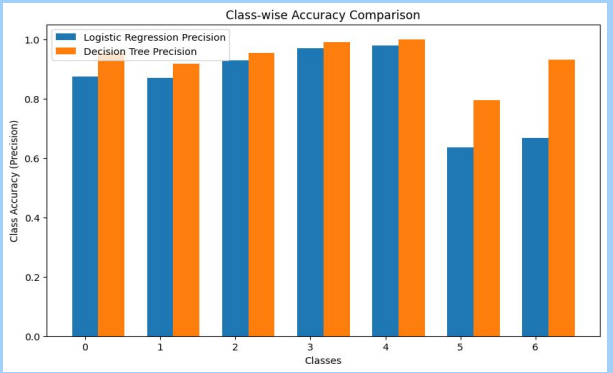
Pruned



- **Weight and Height**: Significant indicators of obesity levels across all categories.
- **Gender**: Gender has a high impact on obesity risk
- **Dietary Habits**: Frequent consumption of high-calorie foods (FAVC) and snacking between meals (CAEC) are important factors influencing obesity risk.

Results





Logistic Regression

Precision: Good precision across classes; slightly lower in classes 5 and 6.
Recall: Lower recall in classes 1, 2, and 5, missing some high-risk cases.

Decision Tree

Precision: Higher precision in classes 1, 5, and 6, capturing more high-risk cases.
Recall: Strong recall across most classes, especially in high-risk areas.

Conclusion: : The Decision Tree model demonstrates higher precision and strong recall for high-risk categories, making it the better choice for our dataset.

Value & Next Steps

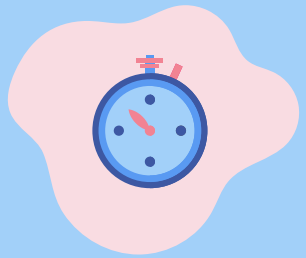
Value

- This model provides actionable insights for targeting high-risk students
- It also supports resource allocation for wellness programs

Next Steps

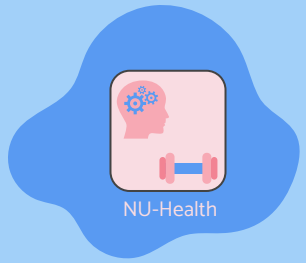
- **Wellness Initiatives:** Introduce programs such as Blue-Bike discounts, a NU walking club, nutrition counseling, and free healthy snack options.
- **Student-Facing App**: Develop an app to centralize wellness resources, giving students easy access to all available programs and health support options.

Socialization & Distribution



Regular Reporting

Monthly reports for internal wellness staff to monitor trends in obesity risk factors.



Student–Facing App

Implement the model in a wellness app where students can sign-up for health related activities on campus.

Interactive Dashboard

Provide dashboards for stakeholders to track overall student wellness trends and respond proactively.

Any Questions?

Thanks for listening :)