Predicting Obesity Status

STATS 101C GROUP 4P

Gavin Cardeno, Gerardo Munoz, Luning Ding, Eric Jiang

Table of Content

01

Introduction

Obesity context and data set overview

Predicting Obesity Status Using Machine Learning

Obesity is a growing global health issue, associated with numerous health risks such as heart disease and diabetes. Understanding factors that contribute to obesity and accurately predicting obesity status is essential for public health interventions.

Machine learning models can help in identifying patterns in large datasets that may be difficult for traditional methods to uncover. By predicting an individual’s obesity status, these models can help with early intervention.

Obesity Data Set

42686

Observations

Each observation refers to an individual entry or record in the dataset whose obesity status is being predicted

30

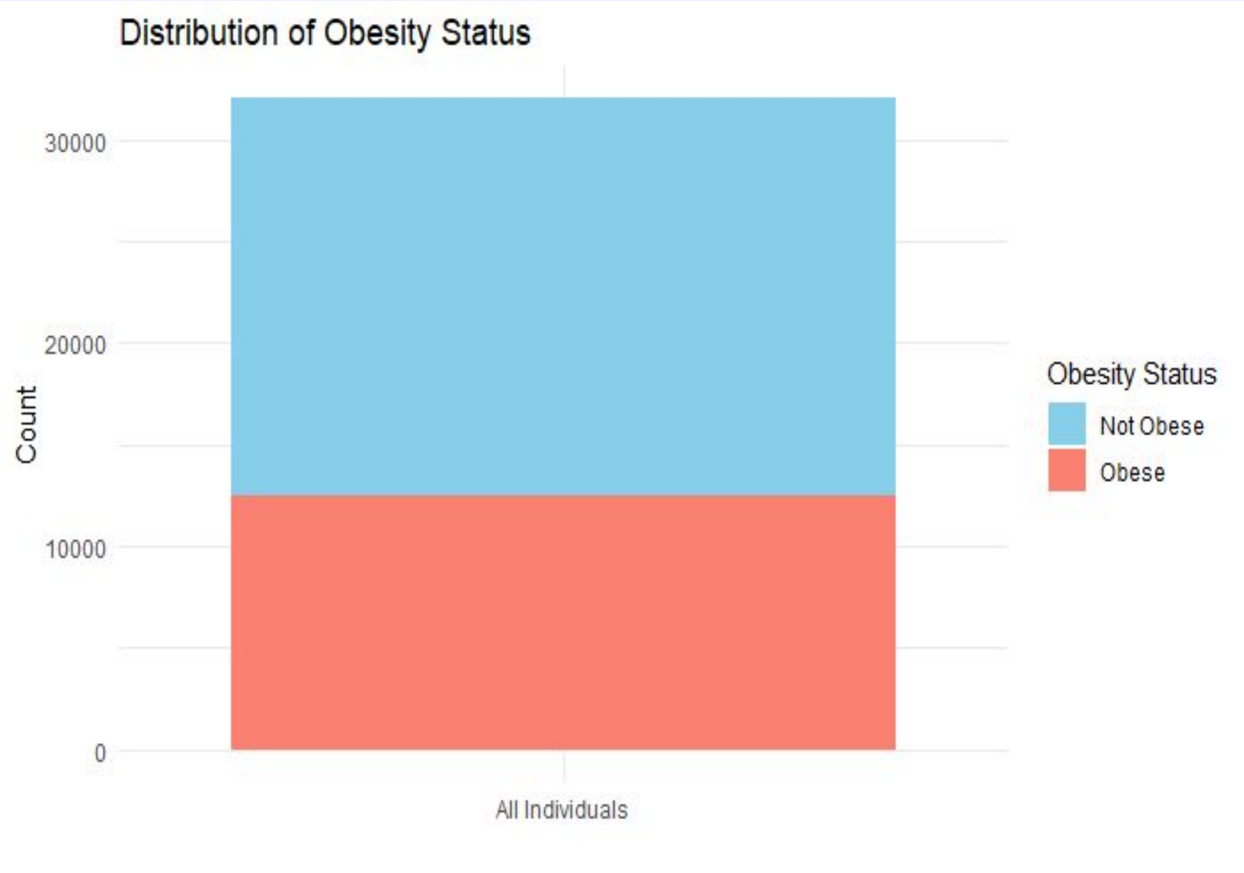
Variables

The features or attributes that describe each person (such as age, gender, BMI, etc.)

02

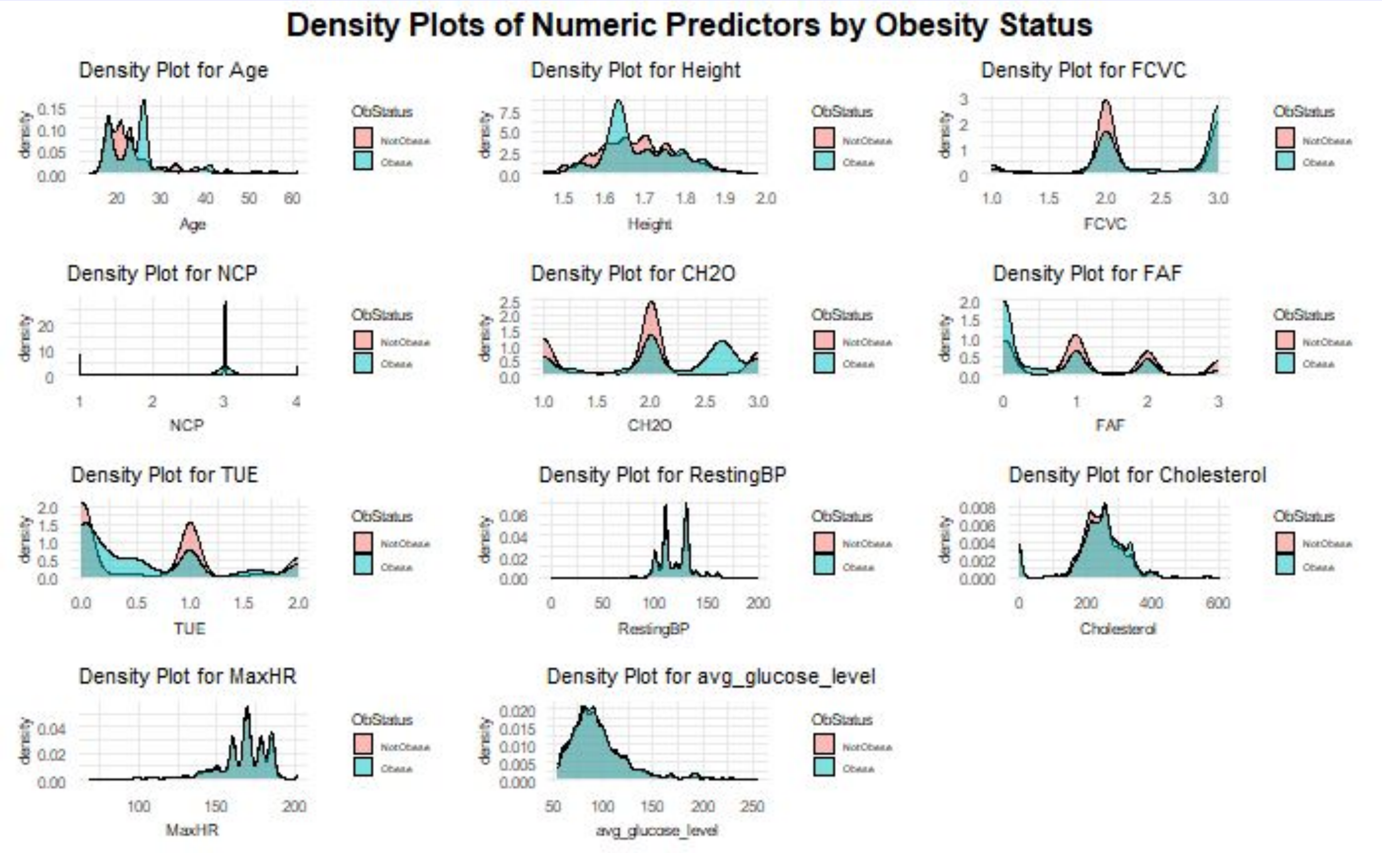
Preprocessing

Data visualization and imputation

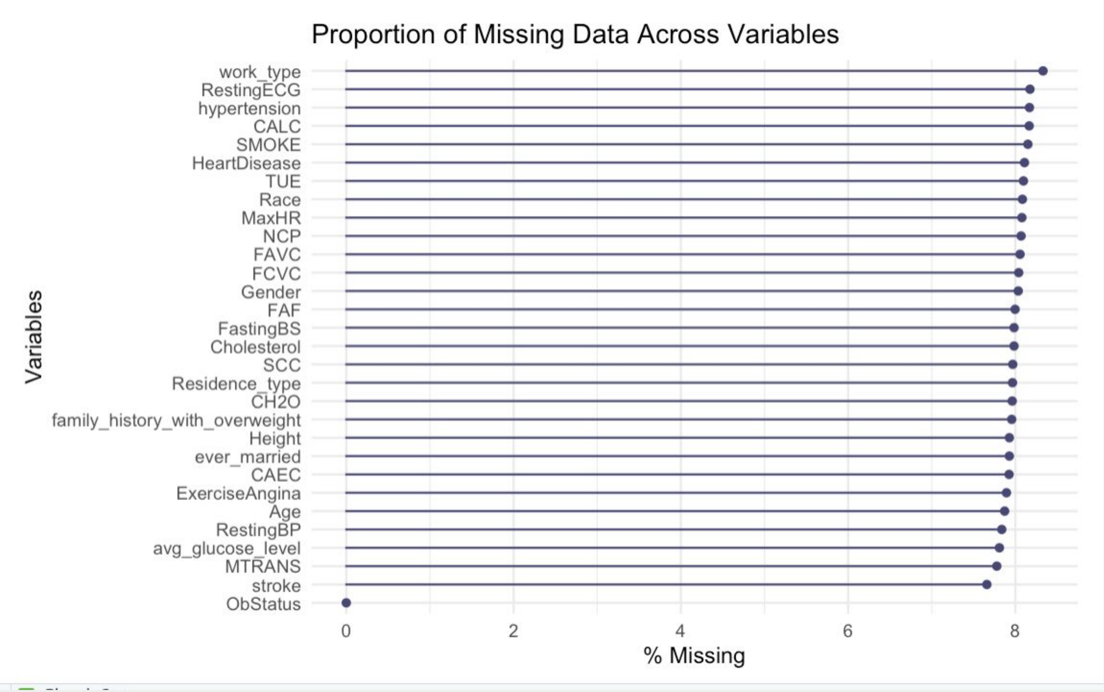


Distribution Status: Obese vs. Not Obese

- There are significantly more Not Obese individuals than Obese individuals (approximately 3:1)
- Imbalance can lead to a biased model that predicts "Not Obese" more frequently.


Density Plots of Numeric Predictors by Obesity Status
Density Plot for Age
Density Plot for Height
Density Plot for FCVC
Density Plot for NCP
Density Plot for CH2O
Density Plot for FAF
Density Plot for TUE
Density Plot for RestingBP
Density Plot for Cholesterol
Density Plot for MaxHR
Density Plot for avg_glucose_level
density
Age
Height
FCVC
NCP
CH2O
FAF
TUE
RestingBP
Cholesterol
MaxHR
avg_glucose_level
ObStatus
NotObese
Obese

Data imputation



Libraries Used:

- missForest (for numerical data)
- FactoMineR (for categorical data)

Data imputation

1. **Load dataset and extract numeric and categorical columns**

2. **Impute missing values in numeric columns using missForest( )**
   - Leverages the predictive power of Random Forests to handle complex relationships.
   - Handles nonlinear interactions and works for large datasets.

3. **Impute missing values in categorical columns using imputeMCA( ), which is a function in "FactoMineR" package.**
   - is an extension of PCA and uses Multiple Correspondence Analysis (MCA) to impute missing categorical data by identifying relationships between variable categories and projecting them into a reduced latent space.

03

Methodology

Modeling

Model Comparison

	Accuracy on Test Dataset
GLM	0.7535
LDA	0.7399
QDA	0.7439
Random Forest	0.9653
XGB	0.9736

Model Comparison

GLM

- assumes linear relationships, unsuitable for nonlinear patterns
- Sensitive to outliers

LDA/QDA

- Normality Assumption: Assumes features are normally distributed within each class
- LDA assumes equal variances across classes
- Sensitive to Multicollinearity
- Sensitive to outliers

Random Forest

- Suitable for both linear and nonlinear data
- Not sensitive to multicollinearity
- Computationally intensive, tuning required
- Handles outlier and noise well

04

Results & Final Discussion

Final Model Assessment

Why we chose XGB..

- Dataset Complexity & Preprocessing
  - Over 40,000 observations with a mix of categorical and numerical variables.
  - Presence of NA values necessitated data imputation and mutation for accurate modeling.
- Model Selection & Performance
  - Achieved 96% accuracy using a Random Forest model with 5-fold cross-validation.
  - Transitioned to XGBoost for enhanced classification performance within the ensemble framework.
- Advantages of XGBoost over random forest
  - XGBoost optimizes and assesses decision trees sequentially, correcting errors from previous trees.
  - This boosting methodology reduces residual errors, improving the gap between actual and predicted values.

Benefits of XGB

01

XGB Model Advantages:

- Ridge and Lasso penalties prevent overfitting
- Reduces noise through sequential error correction
- Effective handling of large datasets with variable context

02

Model Performance

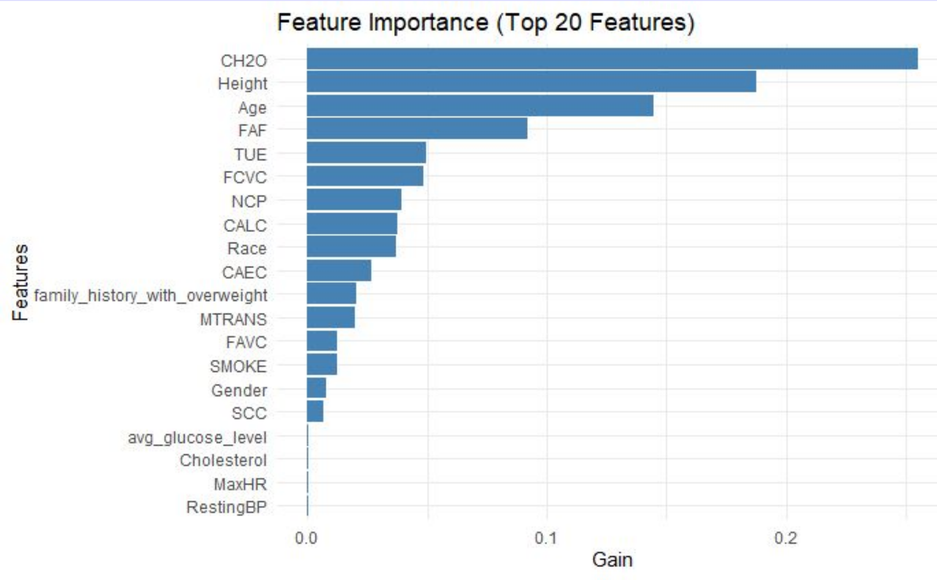
- 97.36% accuracy with a misclassification rate of 0.02643
- Imputed data enabled robust obesity classification predictions

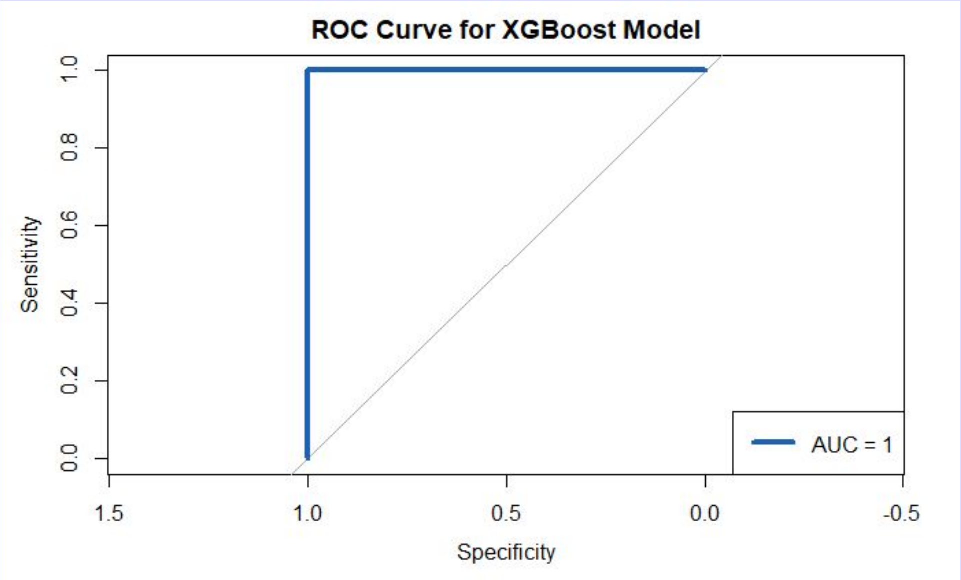
03

Key Features & Efficiency

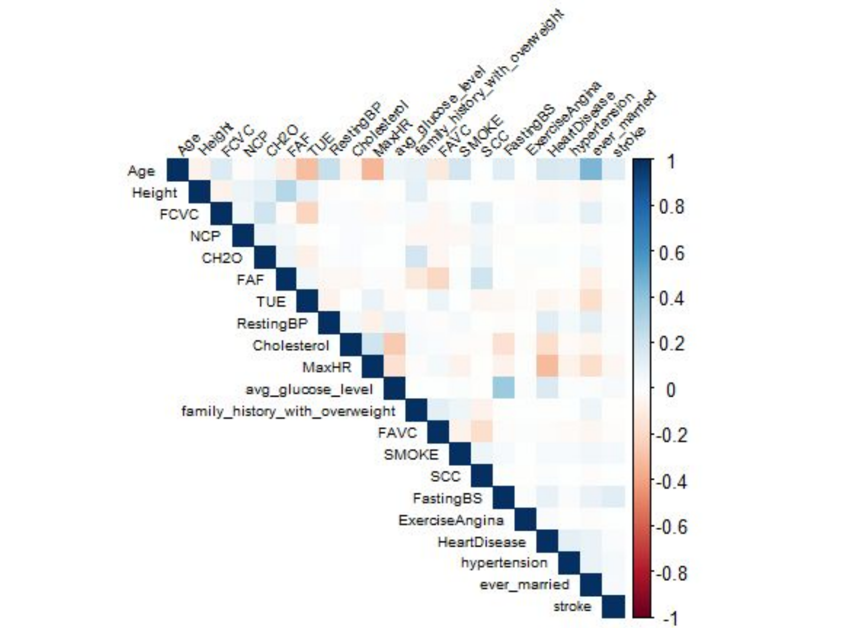
- Identifies impactful variables for better interpretability
- Manages extreme values efficiently, optimizing workload
- High accuracy with reduced time costs despite cross-validation challenges

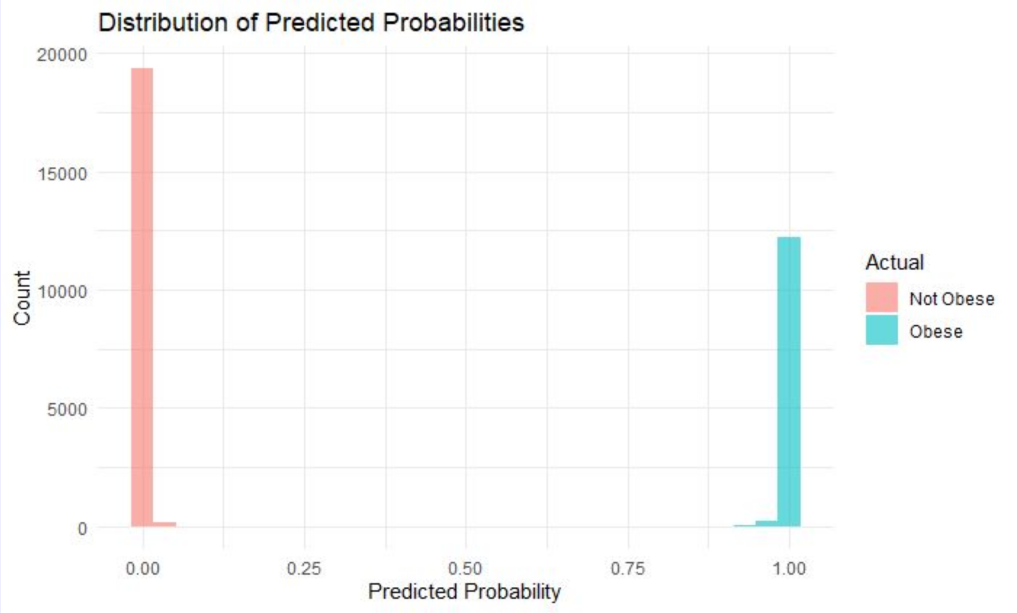
Visualizations of XGB



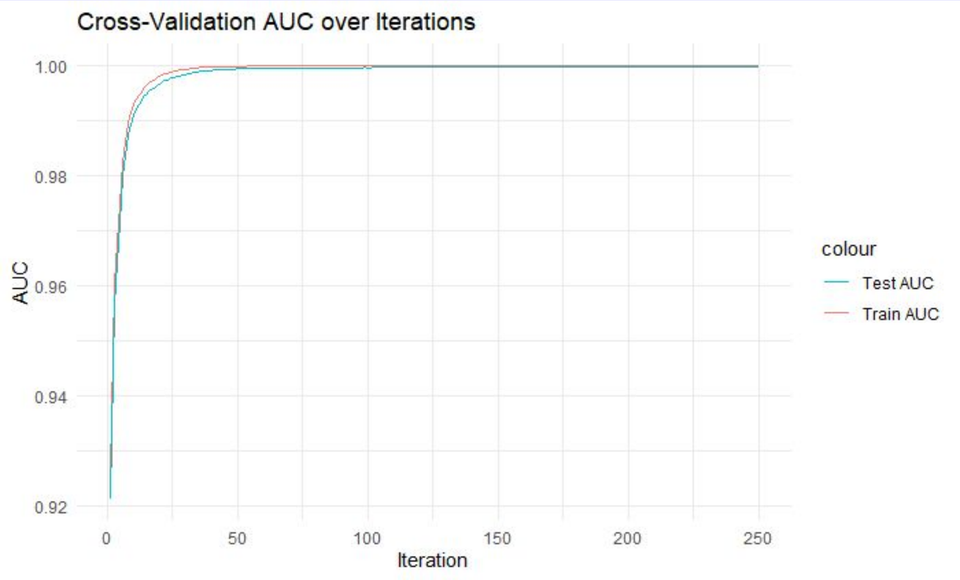


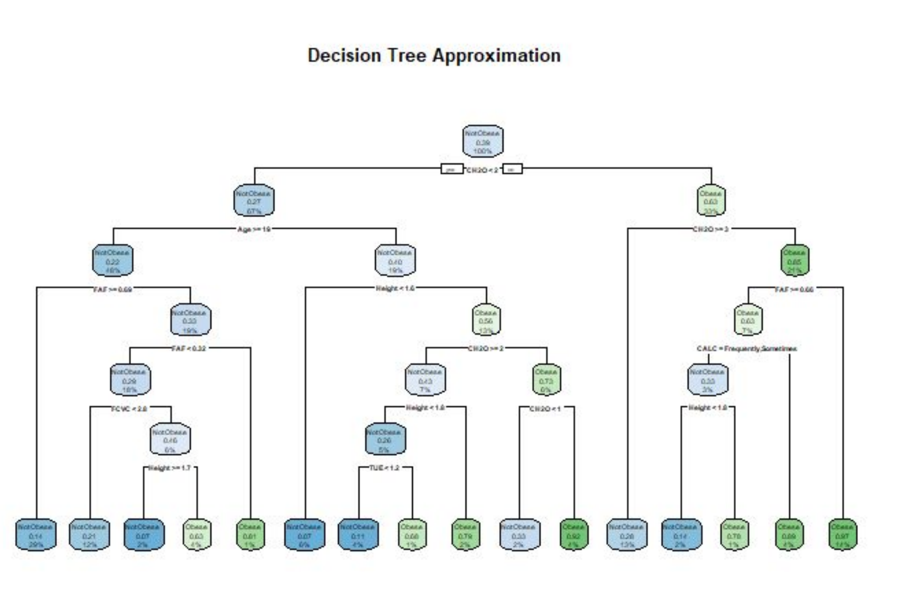
Visualizations of XGB





Visualizations of XGB





05

Limitations & Final Words

Setbacks and Assumptions

Limitations

1. Some weaker variables may have been included, potentially impacting model efficiency.

2. Both Random Forest and XGBoost are ensemble models, making them harder to interpret.

3. Both models are computationally intensive requiring significant processing power and time for over 40,000 observations.

Final thoughts ✓

Overall we achieved a good prediction with our model.

Our project shows that machine learning is a powerful tool in solving complex health challenges and guiding data-driven decisions.