Voice-Based Health Classification: Capstone Project Report

Date: November 9, 2025

Author: D. Saif

1. Introduction

1.1 Project Overview

This capstone project aims to analyze audio-derived features from voice recordings to classify individuals as either "Healthy" or "Unhealthy." Voice-based health monitoring is an emerging area of biomedical signal processing and artificial intelligence, providing non-invasive, accessible diagnostic support through speech analysis.

1.2 Objective

The primary objectives are:

- Perform comprehensive exploratory data analysis (EDA)
- Visualize insights through an interactive Power BI dashboard
- Build and compare predictive machine learning models

1.3 Dataset Description

- File: VowelA_High_latest.csv
- Total Records: 2,035 samples
- Key Features: Signal_Energy, Spectral_Brightness, Spectral_Spread, Spectral_Rolloff, Zero_Crossing_Rate, MFCC_1–20, Age, Gender
- Target Variable: Health_Status (Healthy / Unhealthy)

2. Exploratory Data Analysis (EDA)

EDA revealed **no missing values or duplicate records**.

2.1 Categorical Distributions

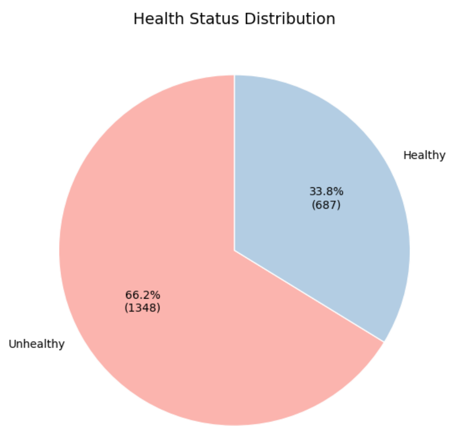
1. **Health Status:** The number of **Unhealthy** individuals exceeds that of **Healthy** ones → **66.8% Unhealthy**, **33.2% Healthy** (moderate class imbalance).

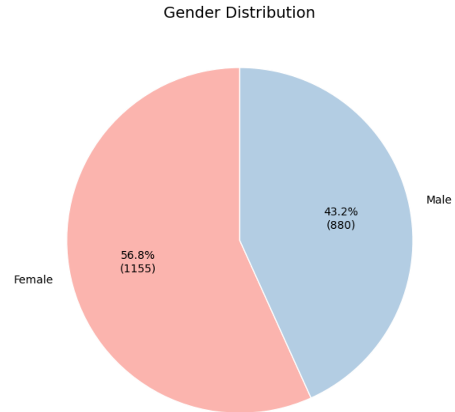
2. **Gender: Female individuals are more frequent** (56.8%) than **Male** (43.2%).

3. **Age Group: Most individuals fall within the 21–40 years age group**.

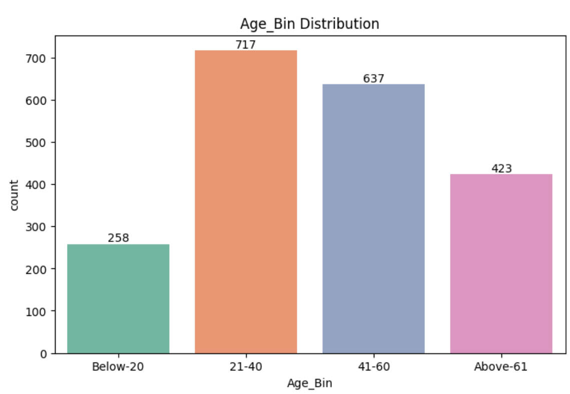
    - Median age: **42 years**

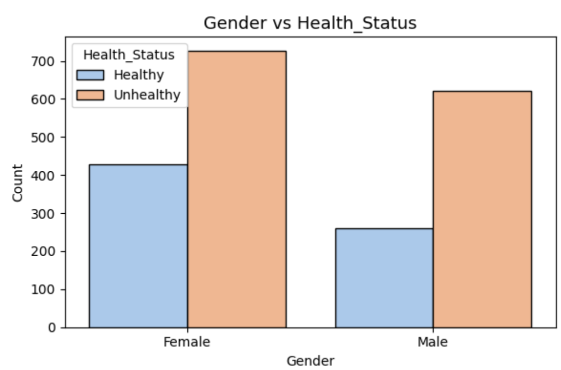
    - Age range: 9–94 years





Gender vs Health Insight: Although females are more represented overall, **the proportion of unhealthy males (70.6%) is notably higher** than unhealthy females (63.9%). This suggests males exhibit a greater prevalence of unhealthy vocal conditions.



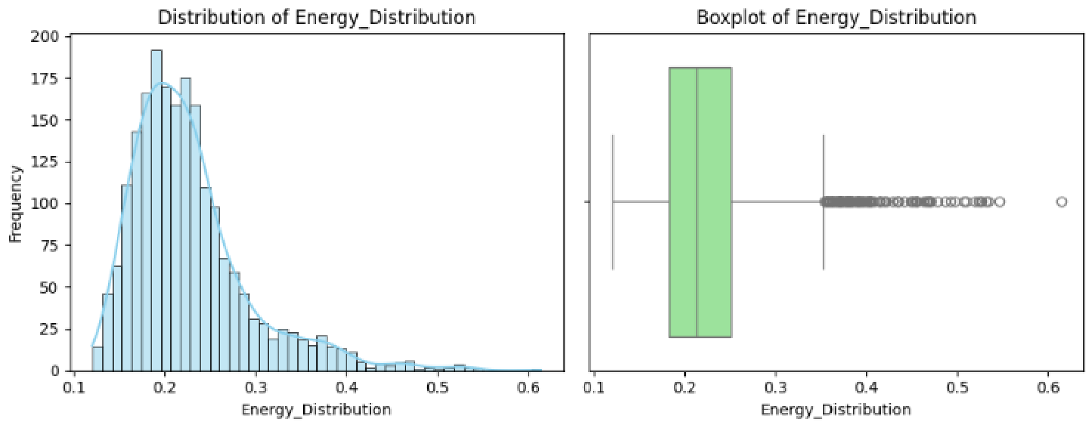


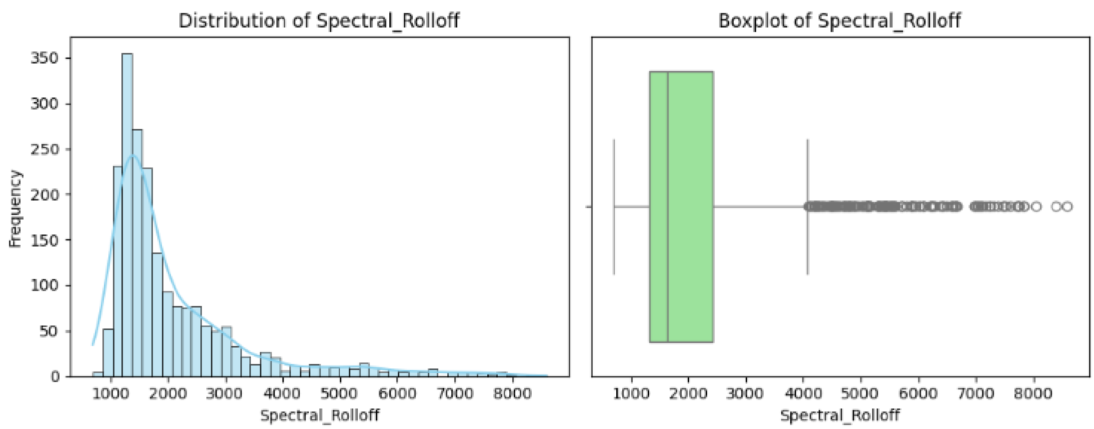
2.2 Numerical Feature Distributions

The numerical features were examined for skewness and outliers:

- **Positively skewed with outliers:** Energy_Distribution, Spectral_Brightness, Spectral_Rolloff, Zero_Crossing_Rate, MFCC_4, MFCC_19, MFCC_20
- **Symmetric & well-behaved:** Signal_Energy, MFCC_2, MFCC_3, MFCC_5, MFCC_7

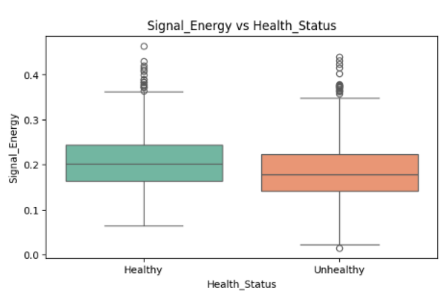
Recommendation: Apply scaling (StandardScaler) and consider log-transformation for skewed features during modeling.

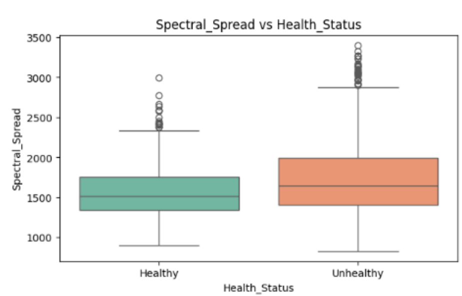




2.3 Bivariate Analysis: Features vs Health Status

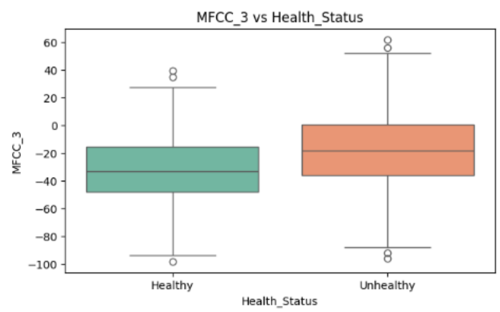
1. **Signal & Spectral Features**
   - **Signal_Energy** is **higher in Healthy** individuals → stronger, stable phonation.
   - **Energy_Distribution, Spectral_Brightness, Spectral_Spread, Spectral_Rolloff** are **wider and higher in Unhealthy** → greater spectral noise and variability.
   - **Zero_Crossing_Rate** shows many outliers in Unhealthy → irregular signal fluctuations.

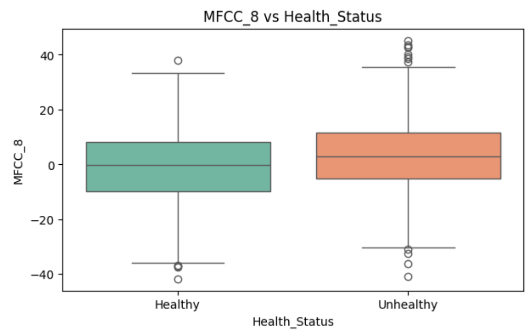




2. **MFCCs**

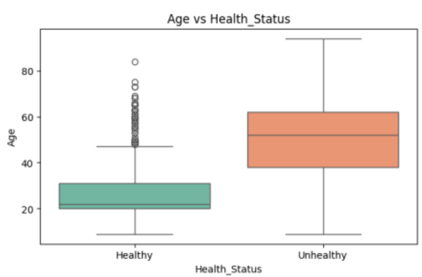
   - Most MFCCs similar across groups.
   - **MFCC_1, MFCC_3, MFCC_8** slightly **higher** in Unhealthy.
   - **MFCC_8 and MFCC_15** slightly **lower** in Unhealthy.





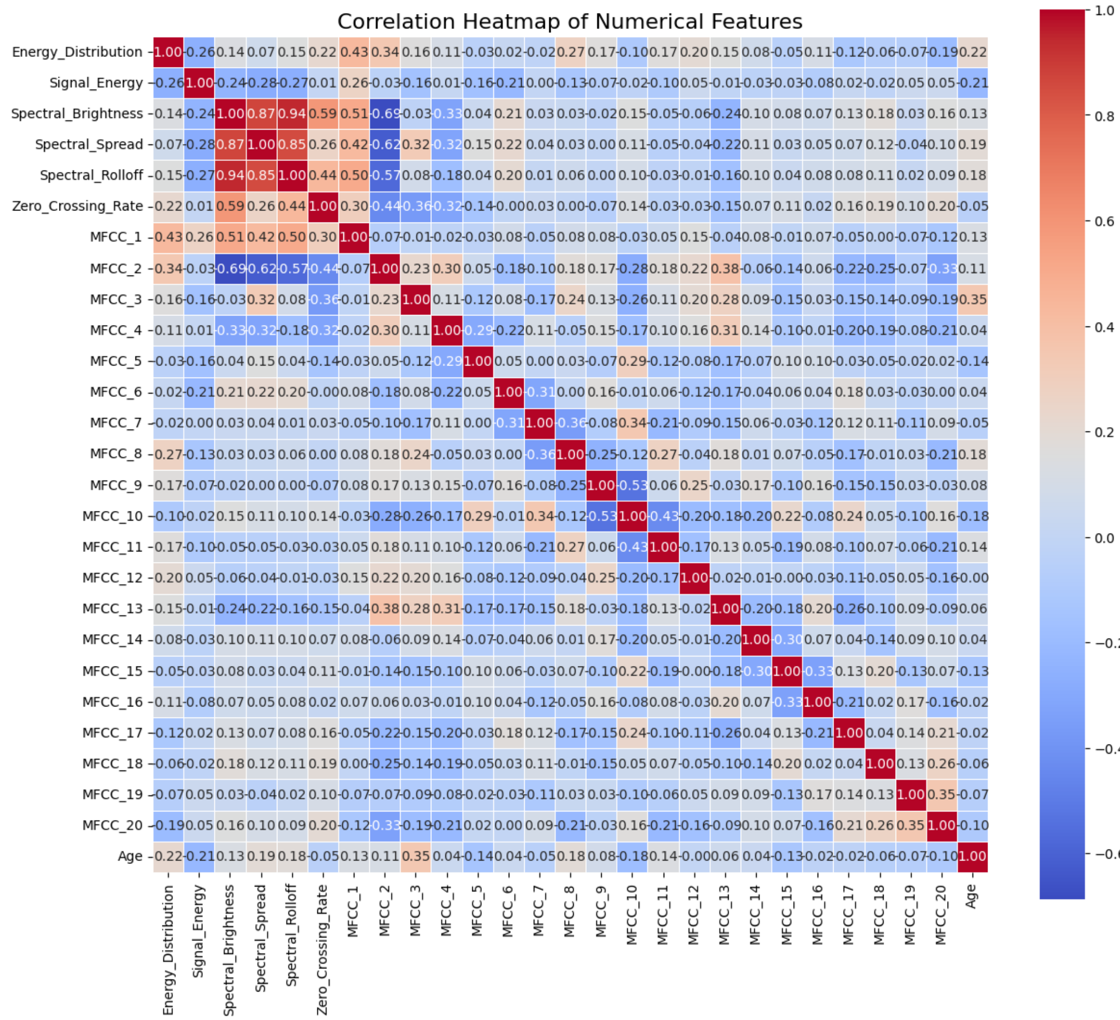
3. **Age**

   - **Unhealthy individuals tend to be older** → aligns with real-world health deterioration trends.

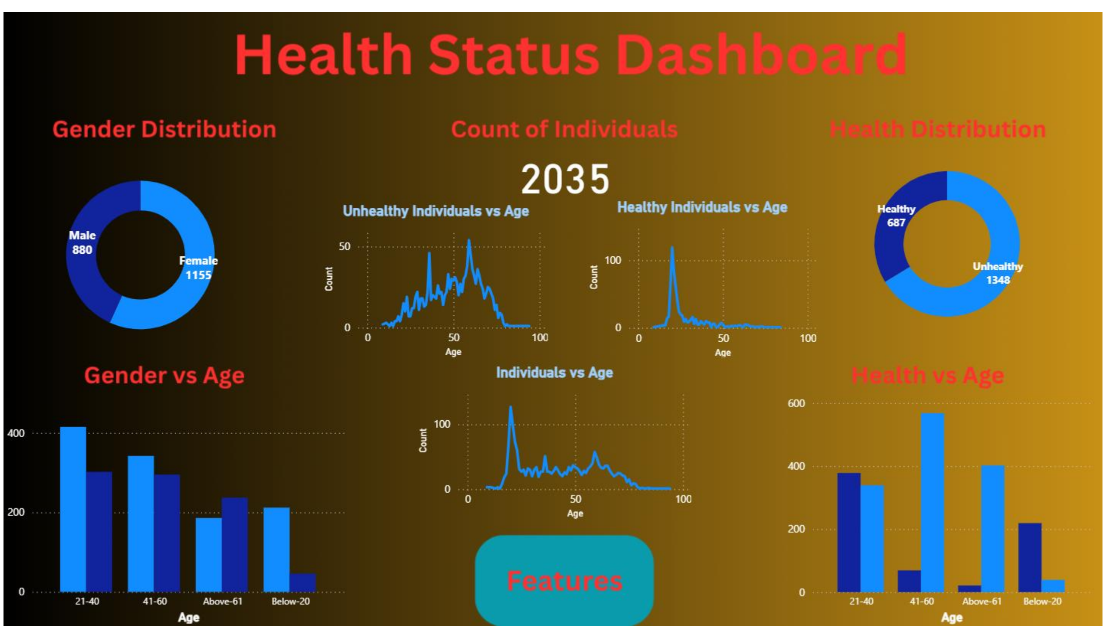


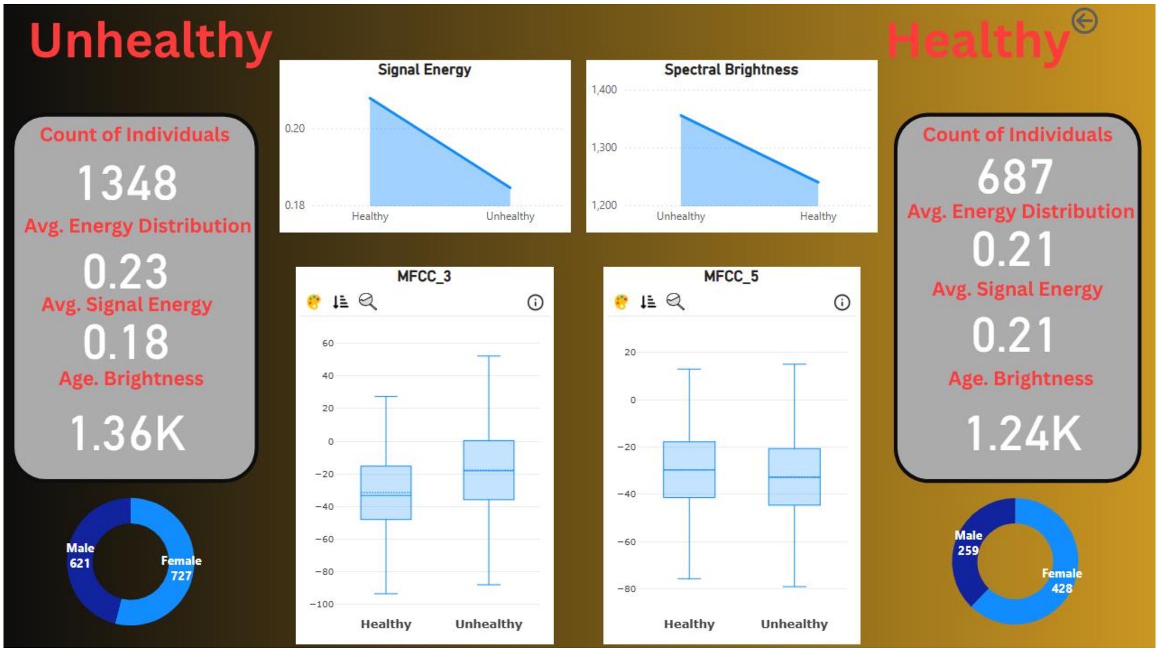
3. Correlation Analysis

- **Strong multicollinearity** among spectral features:
  - Strong multicollinearity among spectral features:
  - Spectral_Brightness ↔ Spectral_Rolloff (r = 0.94)
  - Spectral_Brightness ↔ Spectral_Spread (r = 0.87)
  - Spectral_Spread ↔ Spectral_Rolloff (r = 0.85)
- MFCC_2 inversely related to spectral measures → valuable contrasting feature.
- Signal_Energy and Age show low correlation with others → independent predictors.



4. Power BI Dashboard



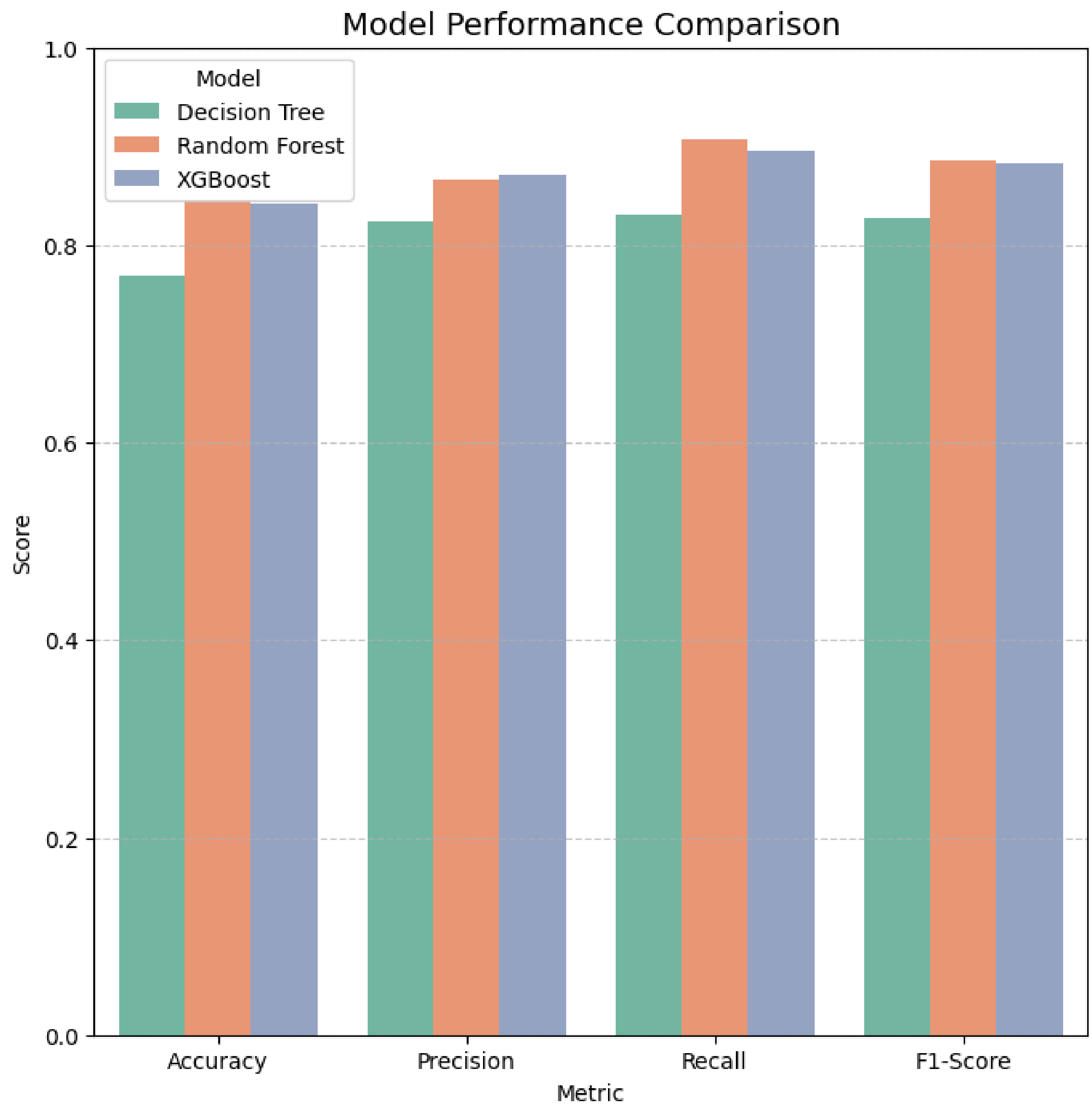


5. Predictive Modeling

Modeling workflow summary:

- Encode categorical variables (Gender), standardize numerical features using StandardScaler, split data (80/20 stratified).
- Models evaluated: Decision Tree, Random Forest, XGBoost.
- Use evaluation metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC, and confusion matrices.

Model	Accuracy	Precision	Recall	F1-Score	Insights
Decision Tree	0.7690	0.8240	0.8300	0.8270	Lowest performance; serves as a baseline.
Random Forest	**0.8452**	0.8657	**0.9074**	**0.8861**	**Best overall balance, highest Recall**.
XGBoost	0.8428	**0.8705**	0.8963	0.8832	Competitive;



6. Conclusion

This project successfully demonstrates that voice acoustic features can reliably distinguish Healthy from Unhealthy individuals, achieving up to 84.5% accuracy with Random Forest.

Key Findings:

- Unhealthy voices exhibit higher spectral noise, lower signal energy, and greater variability.
- Age and gender significantly influence health status.
- Ensemble models are robust to multicollinearity and class imbalance.

Recommended Model for Deployment: Random Forest (best recall + interpretability).

8. Deliverables Summary

- EDA & Modeling Notebook: Capstone_CodeCademy (3).ipynb
- Power BI Dashboard: .pbix file with interactive filters
- Cleaned Datasets: Vowel_A_clean.csv, VowelA_PBI.csv
- Final Report: This document (PDF/Word)