Group 3:
Wanshan Mao,
Muhetaer Mayila,
Yankun Song,
Yu Xiao,
Yue Wang,
Duygu Torun

WHITEWINE QUALITY PREDICTION

Data Mining for Business

(BYGB-7967-V01)

Instructor: Professor Lin Hao

FORDHAM | Gabelli School of Business
THE JESUIT UNIVERSITY OF NEW YORK

TABLE OF CONTENTS

I. Introduction

II. DIDA Framework

III. Challenges / Insights

IV. Python Implementation Of Prediction Model:

   i. Logistic Regression

   ii. Classification Tree& English Rules

   iii. Neural Network

   iv. k-NN

V. Prediction Model Comparison and Result

VI. The “Perfect Wine”- Scenario

VII. Conclusion and Limitations

INTRODUCTION

Is it wine o'clock yet?

- **Goal:** Examine expectation of wine quality to decide market launch
- Wine quality can vary from **1 to 10**
- Segmentation for good wine starts with a **score above 6**
- Target **high-end** white wine market
- **Physicochemical variables**: alcohol, sulfates, volatile acidity, density, etc.
- Dataset source: Kaggle

DIDA FRAMEWORK

Data	● Dataset contains **chemical attributes** of white wines ● Each row represents a specific wine → **individual-level** and **historical data** ● **5** predictors & **4899** observations → ensuring **portrait-shape** ● DV: "***Quality***" (1: good or 0: not good → binary)
Insights	● **Probability**: How likely wine will be perceived as good on the market?
Decision	● Whether to **launch** the white wine to the high-end market or not ● If **probability > 50%**, wine is considered “good” (score > 6.0) → company launches wine to high-end market
Advantage	● **Cost-cutting** → no need for sample preparation or send-out to wine experts/customers for rating purposes ● Increase of high-end wine **market share & presence** ● **Profit maximization** & **brand awareness reinforcement** as high-end wine producer

CHALLENGES / INSIGHTS

Accuracy or AUC?

Highly unbalanced problem, a very skewed sample distribution
we care the “one”

	Num	%
0	3838	78.4%
1	1060	21.6%

LOGISTIC REGRESSION

Characters	Coefficients
Fixed Acidity	0.469865
Volatile Acidity	-0.390498
Citric Acid	-0.095696
Residual Sugar	1.525274
Chlorides	-0.339358
Free Sulfur Dioxide	0.150111
Total Sulfur Dioxide	0.007699
Density	-1.979349
PH	0.501027
Sulphates	0.231539
Alcohol	0.160923
Intercept	-1.717265

Top 5 predictors:

1. Density
2. Residual Sugar
3. PH
4. Fixed Acidity
5. Volatile Acidity

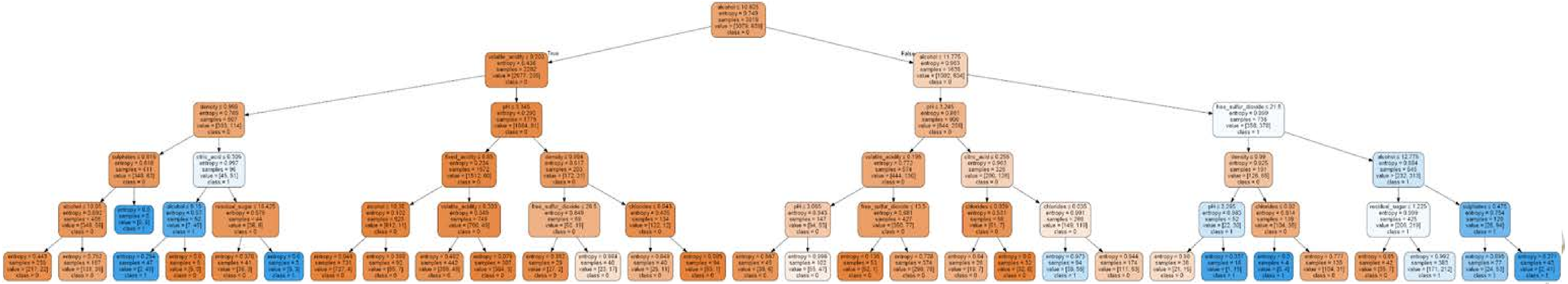
Accuracy	AUC
0.7887	0.8007

CLASSIFICATION TREE

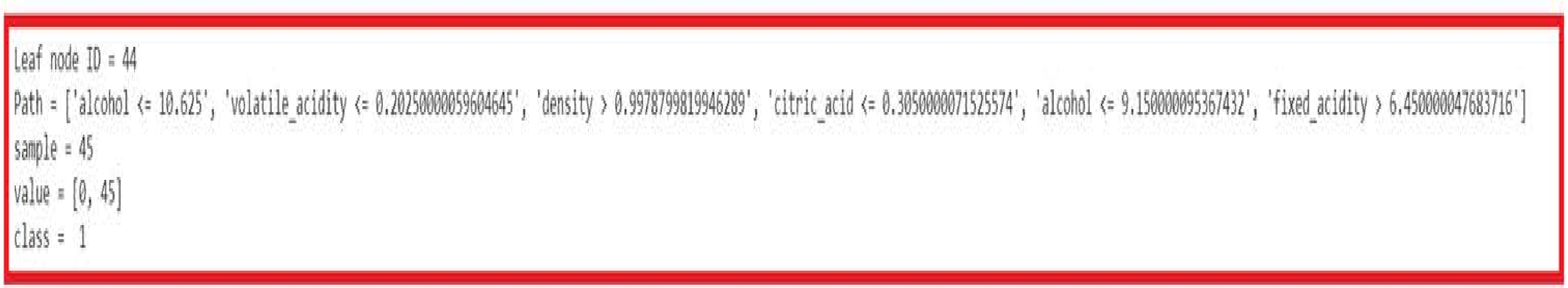
The order of the predictors appear from the root:

1. Alcohol
2. Volatile Acidity
3. Density
4. PH
5. Free Sulfur Dioxide

Accuracy	AUC
0.8285	0.8249



ENGLISH RULES



```
Leaf node ID = 44
Path = ['alcohol <= 10.625', 'volatile_acidity <= 0.20250000059604645', 'density > 0.9978799819946289', 'citric_acid <= 0.30500000071525574', 'alcohol <= 9.150000095367432', 'fixed_acidity > 6.450000047683716']
sample = 45
value = [0, 45]
class =  1
```

- The predicted probability given by a leaf node: 100%

- IF alcohol <= 9.15and volatile_acidity <= 0.2025and density > 0.998 and citric_acid <= 0.305andfixed_acidity > 6.45, , THEN it is high quality wine.

NEURAL NETWORK

- drop the quality column
- set the testpart size to 0.2
- set alpha level to 0.1, the hidden levels to 3
- get the weight for 11 predictors and 3 levels for each predictor

Accuracy	AUC
0.8204	0.8421

K-NN

- measure the similarity between the new data and sample data
- measure the distance between each data with the “euclidean” function
- set the n_neighbors = 5as pre-specify k to get the AUC score

Accuracy	AUC
0.8500	0.8379

PREDICTION MODEL COMPARISONS

Techniques	Accuracy	AUC
Logistic Regression	0.7887	0.8007
Classification Tree	0.8285	0.8249
kNN	0.8500	0.8379
Neural Network	0.8204	0.8421

THE "PERFECT WINE"-SCENARIO

fixed_acidity(+)	volatile_acidity(-)	citric_acid(-)	residual_sugar(+)
9.1	0.24	0.29	10.6
chlorides(-)	free_sulfur_dioxide (+)	total_sulfur_dioxide (+)	density(-)
0.018	57	139	0.98965
pH(+)	sulphates(+)	alcohol(+)	
3.41	0.61	12.9	

CONCLUSION AND LIMITATIONS

Summaries:

- Developed **4 prediction models**
- Each achieves a good performance around **80%**
- offers company **reliable** results, **lessen** costs & **time**, **grow profits** & business

Limitations:

- Accuracy vs. Interpretability
- **Need**: both **high accuracy** & **high interpretability**
- **Actual**: kNN & Neutral Network , **high AUC**, **low interpretability**

Thank you!

*Please do NOT **drink** and drive. ... Please **drink responsibly** and with moderation. We do not, under any circumstances, accept **responsibility** for any damages that result to yourself or anyone else due to the consumption of alcoholic beverages or the use of this site and any materials located on it.