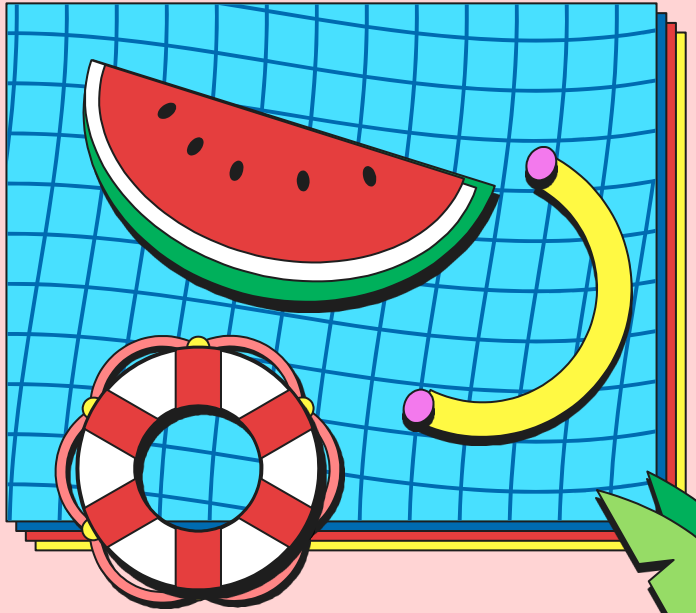
WATERMELON QUALITY PREDICTION

COGS 109 group 19: Junkee Shin, Odemuno Ogelohwohor, Janet Lin



MOTIVATION

Purpose

- Judging watermelon quality based on its apparent properties such as texture or color is difficult.
- Scientific concern in agriculture and food retail

Objective

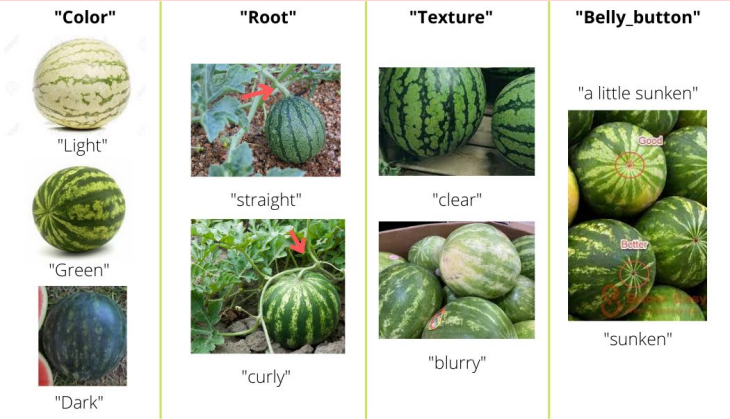
- Predict the watermelon quality using Logistic Regression and K-Nearest Neighbors regression (KNN)

Hypothesis

- Will belly button, sound and sugar rate be the best predictors for the quality of a watermelon?

Relevant data (from Kaggle)

- 209 data points
- 6 categorical & 2 numerical predictors
- 1 categorical outcome label (Good / Bad)



Categorical: color, root, texture, belly_button, sound, touch
Numerical: density, sugar_rate

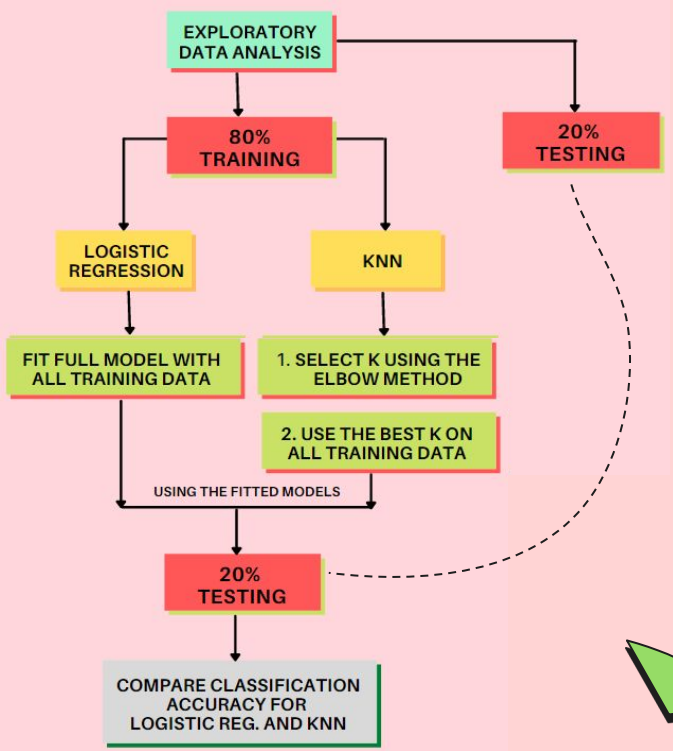
METHODS

Logistic Regression

- Uses a parametric algorithm
- Categorical outcome
- Applies to binary classification
- Tests the significance for individual regression coefficients

K-Nearest Neighbors

- Uses a non-parametric algorithm
- More flexible
- Does not need to be normal distributed
- Calculates the distance to all points in our data based on the K closest points



RESULTS & INTERPRETATION

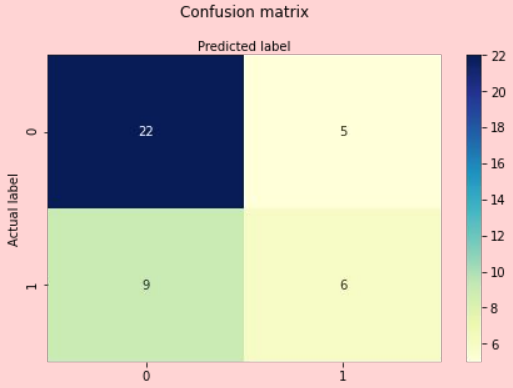
```python
import pandas as pd # data processing, input CSV file
import matplotlib.pyplot as plt # plotting data
import statsmodels.api as sm # linear & logistic regression
import numpy as np # linear algebra
import seaborn as sns # confusion matrix heatmap
import statsmodels.formula.api as smf # log-odds ratio
```

```python
## Data Exploration
df['GB'].value_counts()
```

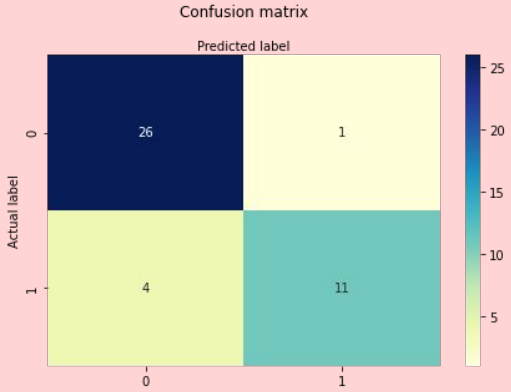
```
No     130
Yes     79
```

```
Train set has:  103 bad watermelons
Train set has:  64 good watermelons
```

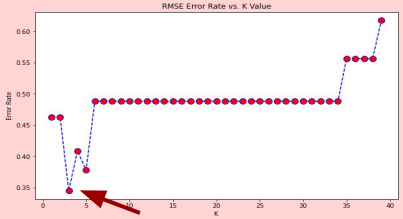
```
Test set has:  27 bad watermelons
Test set has:  15 good watermelons
```



Logistic



KNN

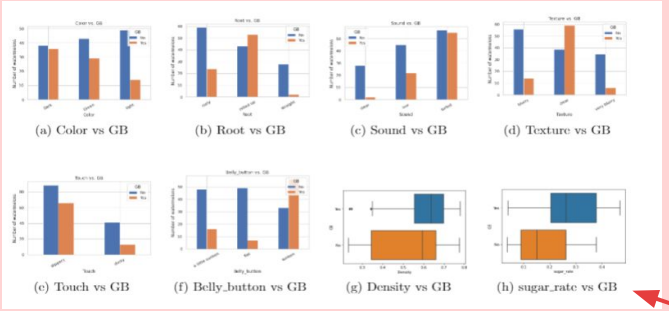


Plot RMSE values to find best K (k=3)

From the confusion matrix:

	Logistic	KNN
Accuracy	66.66%	88.09%
Precision	54.54%	91.66%
Error rate	33.33%	11.90%

RESULTS & INTERPRETATION



Logit Regression Results

Dep. Variable:	GB	No. Observations:	209
Model:	Logit	Df Residuals:	200
Method:	MLE	Df Model:	8
Date:	Sun, 18 Jul 2021	Pseudo R-squ.:	0.2839
Time:	21:50:47	Log-Likelihood:	-99.236
converged:	True	LL-Null:	-138.58
Covariance Type:	nonrobust	LLR p-value:	8.958e-14

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.3908	1.648	-3.271	0.001	-8.621	-2.161
Color	-0.2767	0.349	-0.793	0.428	-0.961	0.407
Root	0.4966	0.541	0.917	0.359	-0.565	1.558
Sound	1.1058	0.447	2.472	0.013	[illegible]	1.982
Texture	0.2569	0.468	0.549	0.583	-0.659	1.173
Belly_button	0.3981	0.338	1.179	0.238	-0.264	1.060
Touch	-0.4919	0.820	-0.600	0.549	-2.099	1.116
Density	2.1697	2.322	0.934	0.350	-2.382	6.721
sugar_rate	6.4586	2.468	2.617	0.009	[illegible]	[illegible]

```
print(np.exp(fitted.params))

Intercept          0.004558
Color              0.758312
Root               1.643137
Sound              3.021667
Texture            1.292895
Belly_button       1.488933
Touch              0.611481
Density            8.755790
sugar_rate       638.171086
dtype: float64
```

- In choosing the best model, KNN gives the best accuracy and lowest error rate, but does not tell us about the predictor significance
- We chose logistic regression for better interpretability for predictors using coefficient table
- **Only sound and sugar rate** have significant influence (p-values < 0.05) on predicting the qualities of the watermelons
- There is minimal overlap for sugar rate, so this proves our hypothesis for this predictor

Each unit increase in sound increases the log odds of having good watermelons by 3.02

REFLECTION

- Logistic regression gives us what predictors are the most important, which **confirms** our hypothesis that sugar rate and sound are the best predictors of watermelon
- However, our model **denies** our hypothesis that belly button is an important predictor
- **Limitation**: we do not know certainly how sensory data like color, sound, texture, and touch are measured; they may vary based on the observer. For "Texture", it is subjective to differentiate "blurry" from "very blurry"
- **If you want to choose a sweet watermelon, the sound heard when tapping it should be clear as opposed to turbid!** *(trust the science & your guts)*

THANK YOU!

Any questions?