EFFECTS ON RINGS OF ABALONE

By Megan Vaughn, Sierra Sesto, and Alexandria Lacoursière

Big Data Analytics Fall 2023

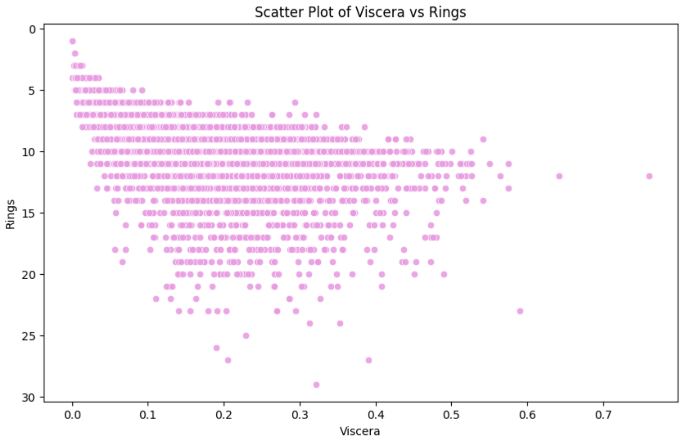
BACKGROUND ON ABALONE

- A marine mollusk commonly found in Australia, New Zealand, South Africa, and Japan.
- Abalone are harvested as a source of food and as decorative items.
- Related to snails, but their appearance is more similar to an oyster.

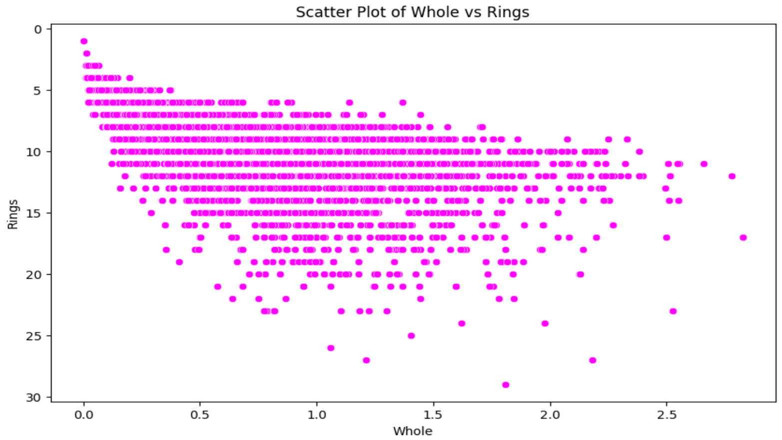
LINEAR REGRESSION

- The chosen dataset's variables are gender, length, diameter, height, whole, shucked, viscera, and shell.
- The objective is to use these variables to predict the rings on an Abalone, the rings can be used to determine how old the Abalone is.
- Gender was omitted from the data due to it being a categorical variable.
- All of the variables were statistically significant except for length.

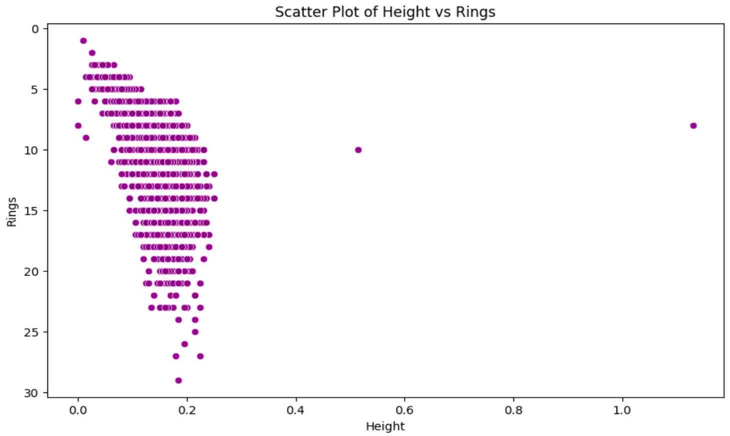
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  rings   R-squared:                       0.528
Model:                            OLS   Adj. R-squared:                  0.527
Method:                 Least Squares   F-statistic:                     665.2
Date:                Sun, 26 Nov 2023   Prob (F-statistic):               0.00
Time:                        17:12:09   Log-Likelihood:                -9250.0
No. Observations:                4177   AIC:                         1.852e+04
Df Residuals:                    4169   BIC:                         1.857e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.9852      0.269     11.092      0.000       2.458       3.513
length        -1.5719      1.825     -0.861      0.389      -5.149       2.006
diameter      13.3609      2.237      5.972      0.000       8.975      17.747
height        11.8261      1.548      7.639      0.000       8.791      14.861
whole          9.2474      0.733     12.622      0.000       7.811      10.684
shucked      -20.2139      0.823    -24.552      0.000     -21.828     -18.600
viscera       -9.8297      1.304     -7.538      0.000     -12.386      -7.273
shell          8.5762      1.137      7.545      0.000       6.348      10.805
==============================================================================
Omnibus:                      933.799   Durbin-Watson:                   1.387
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2602.745
...
Mean Squared Error (MSE): 5.055541144299392
```



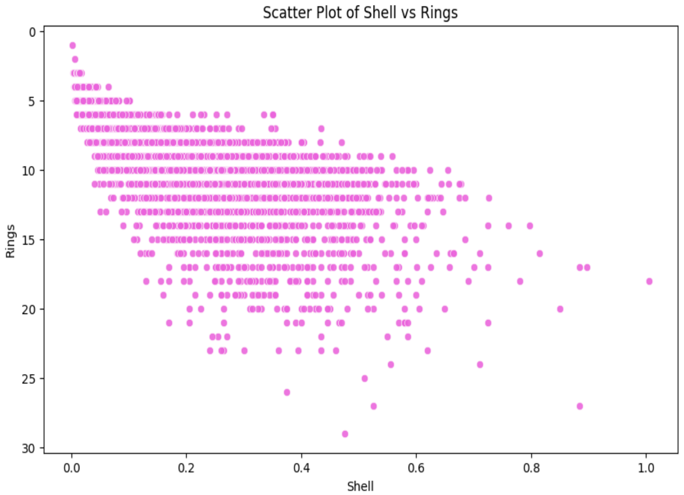
The scatterplot of viscera and rings has a moderate positive trend. Abalone who have more viscera tend to have more rings.



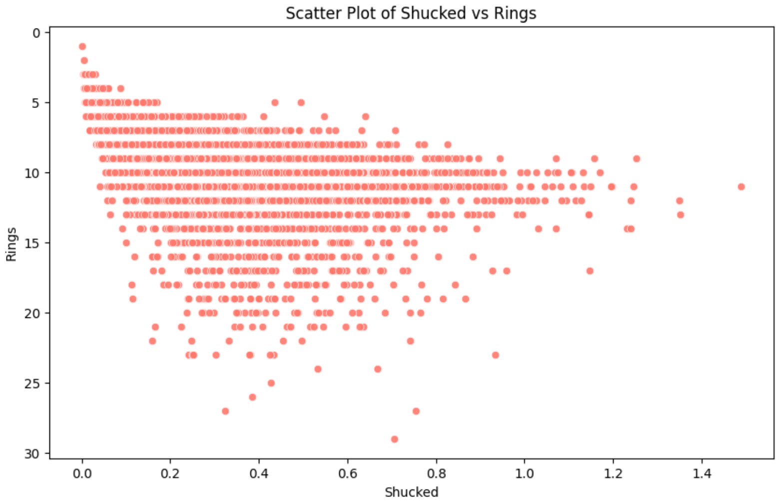
The scatterplot of whole and rings has a moderate positive trend. Abalone who are whole tend to have more rings.



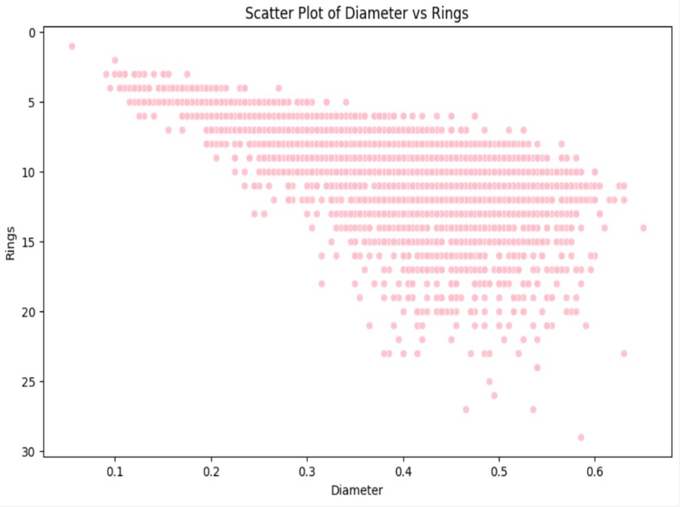
The scatterplot of height and rings has a strong positive trend. As height increases, the number of rings increases. Abalone who are taller may live longer lives.



The scatterplot of shell and rings has a moderate positive trend. Abalone who have shells tend to have more rings.



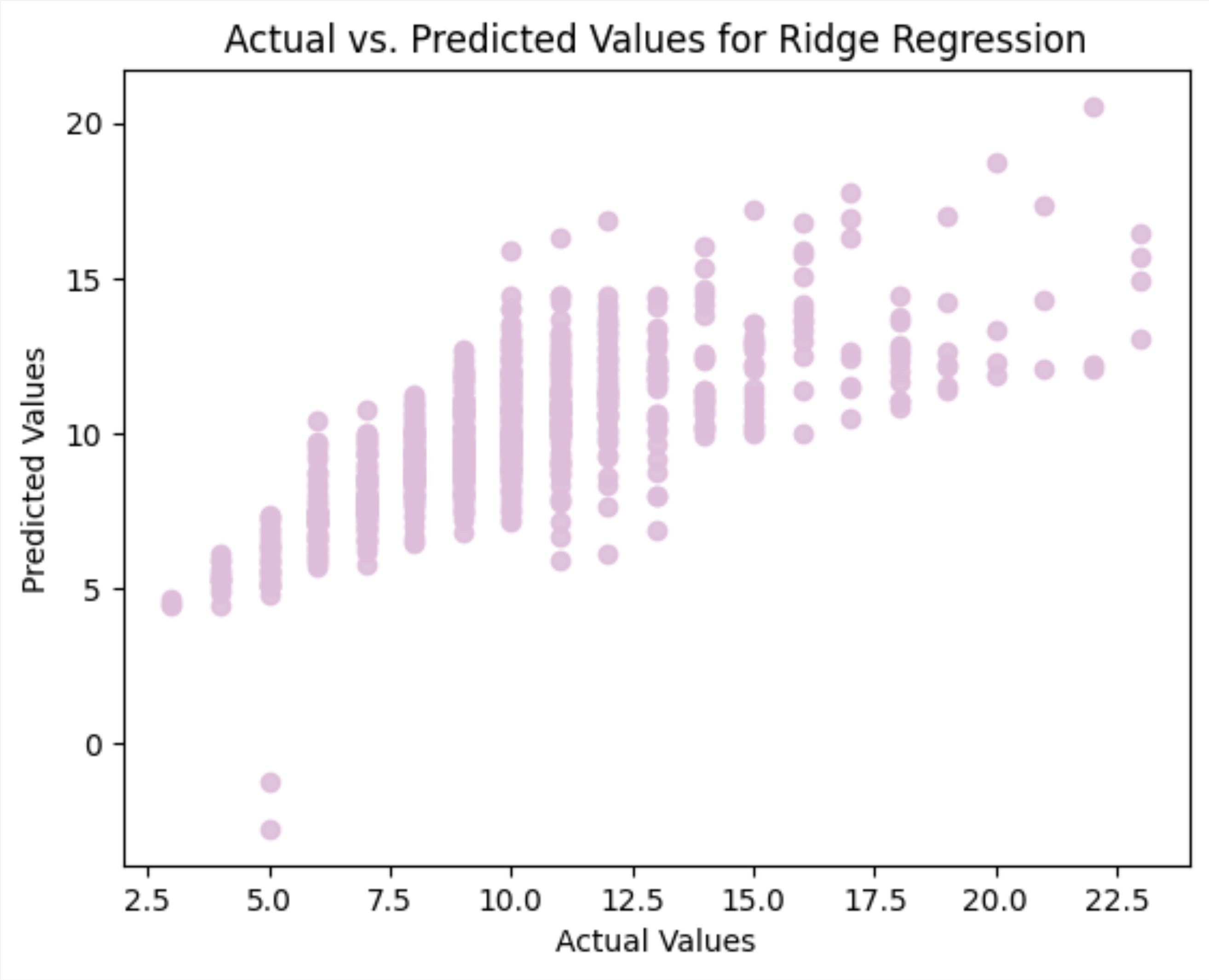
The scatterplot of shucked and rings has a moderate positive trend. Abalone who are shucked tend to have more rings.

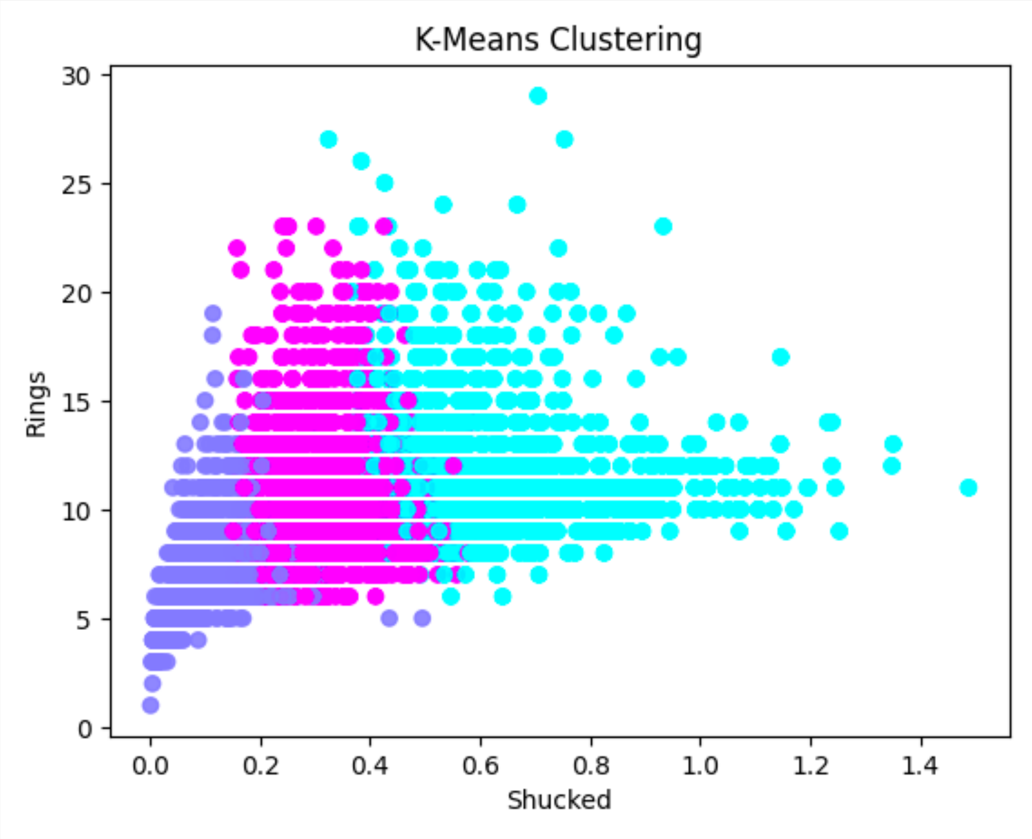


The scatterplot of diameter and rings has a moderate downward trend. Abalone with a larger diameter may live shorter lives.

RIDGE REGRESSION

- *Ridge regression plot is moderately strong, but the plot should have a straight line at 45 degrees, with a few stray points, therefore this is not the best model to fit the data.*
- *Ridge regression model would be more favorable in terms of overfitting.*






```
Mean Rings Value within Each Cluster:
cluster
0    11.673339
1     7.328755
2    10.422575
```

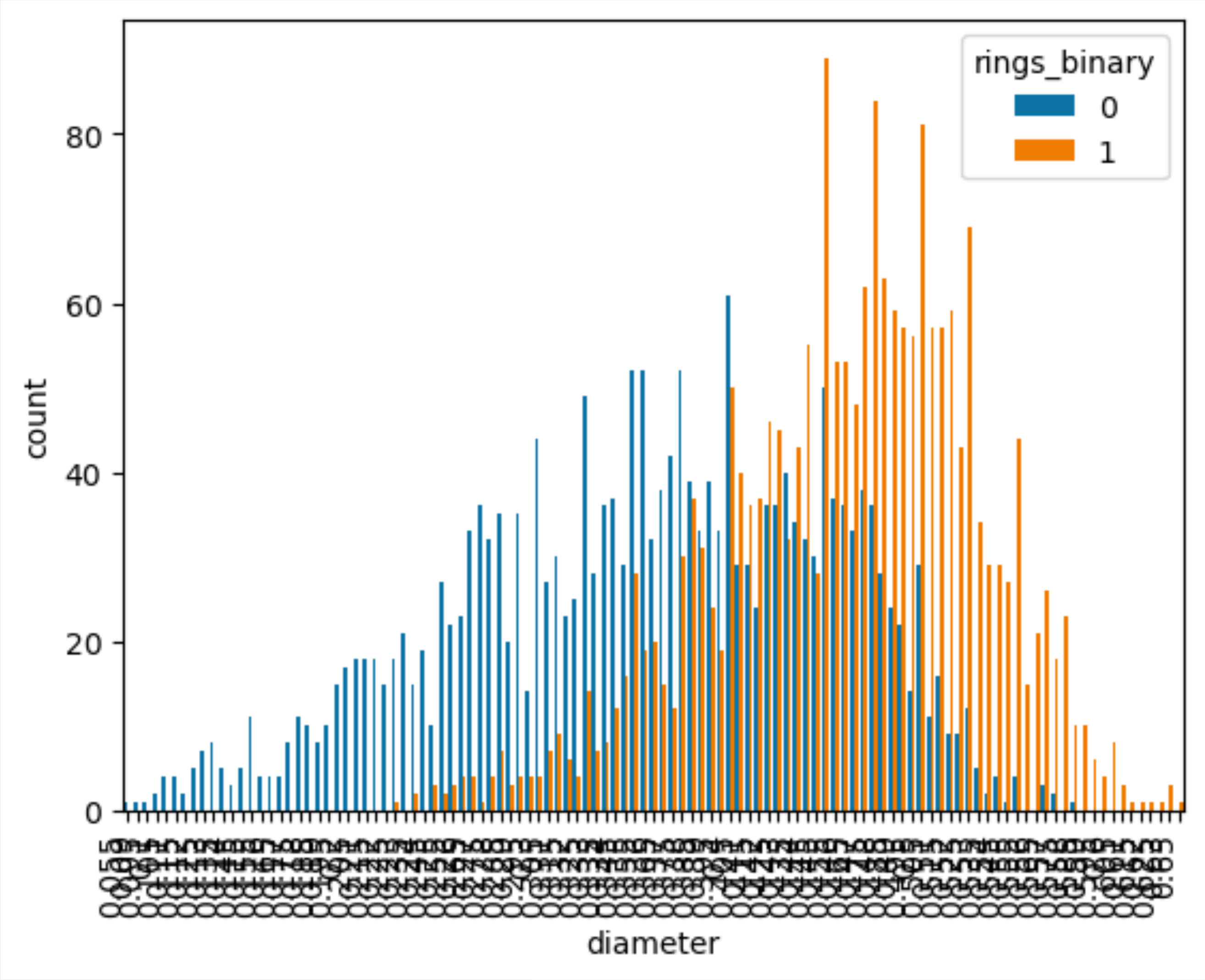
CLUSTERING

- *The variable shucked was chosen due to it having the largest |t-value| out of all the other variables in the linear regression model.*
- *There are three clusters. The clusters are not well-separated, often overlapping. The clusters overall have a moderate positive trend.*
- *Cluster 1 has a strong positive trend, with few stray points.*
- *Cluster 2 has a moderate positive trend, with more stray points than Cluster 1.*
- *Cluster 3 has a moderate negative trend, with more stray points than both Cluster 1 and Cluster 2.*
- *The number of rings an abalone has and whether it was shucked are closely related relationship.*

NAÏVE BAYES

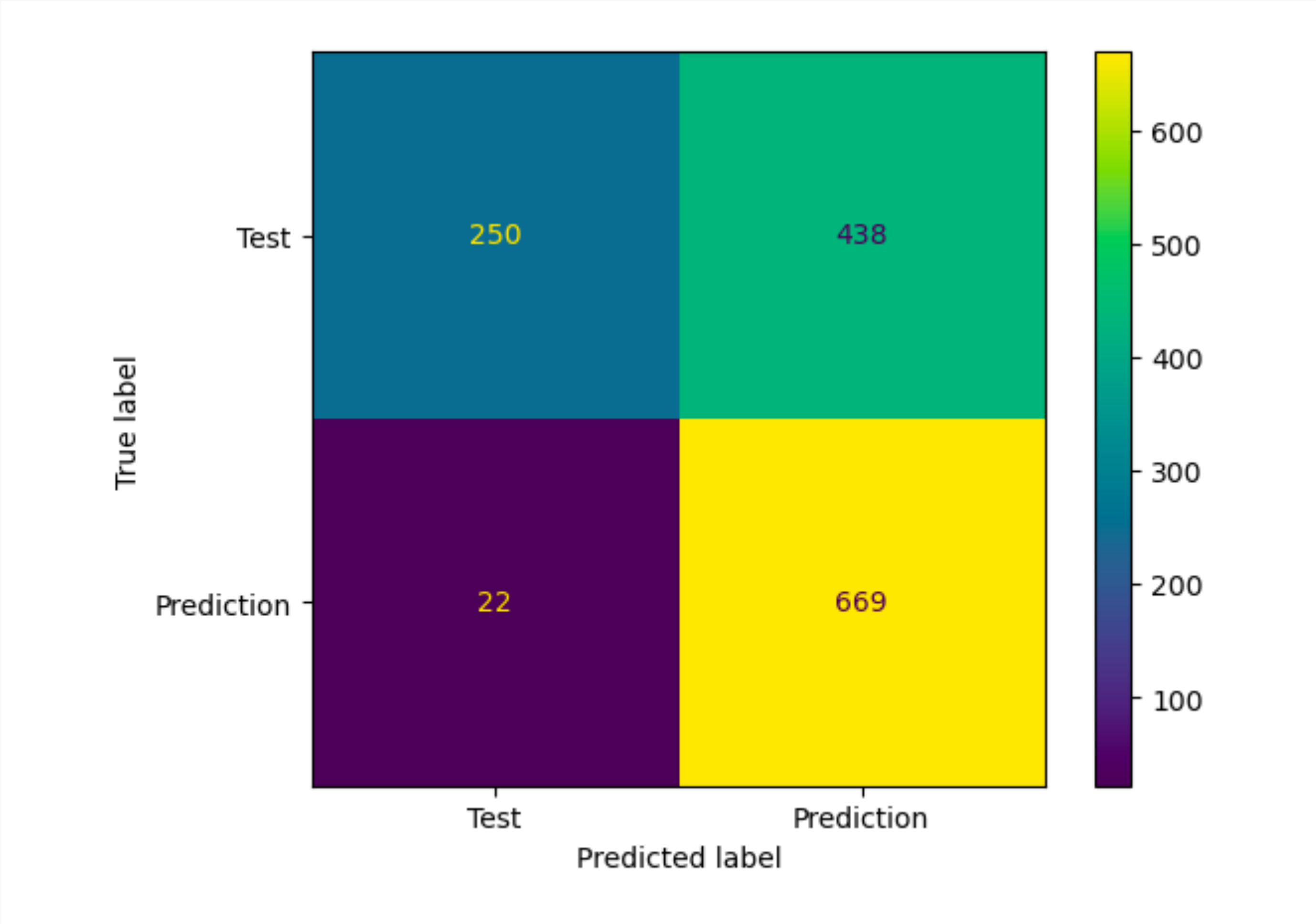
- We can see that in this plot, abalone that have less than 10 rings have smaller diameters while those with rings that are greater or equal to 10 have greater diameters, but this is not 100% accurate since it mixes in the middle and on out.

- The accuracy with this model is low, but at least it’s accurate a little over half the time.



Accuracy: 0.6664249456127629
F1 Score: 0.7001102129699612

NAÏVE BAYES



- As we see in this confusion matrix, our true negatives (669 yellow) are high which is good but so is our false negatives (438 green). In return, our true positives (250 blue) and false positives (22 purple) are very low.
- In conclusion, this is not a good model in predicting the number of rings an abalone may have.

K-NEAREST NEIGHBOR ANALYSIS

```python
import pandas as pd

cols = ['sex','l','d','h','ww','sw','vw','slw','r']
labs = ['1','2','3','4','5']

data = pd.read_csv('abalone.data', header = None, names = cols)
print(data)

data = pd.DataFrame(data)

data = data.drop('sex', axis = 1)

data['r'] = pd.qcut(data['r'], q=[0, 0.2, 0.4, 0.6, 0.8, 1.0], labels=labs)

print(data)
```

K-NEAREST NEIGHBOR ANALYSIS

```python
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
import numpy as np
from sklearn.model_selection import cross_val_score

X = data[['l','d','h','ww','sw','vw','slw']]
y = np.ravel(data[['r']])

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state = 120123)

k_values = [3, 5, 7, 9, 11]
for k in k_values:
    knn_classifier = KNeighborsClassifier(n_neighbors=k)
    scores = cross_val_score(knn_classifier, X_train_scaled, y_train, cv=5)
    print(f'k={k}, Mean Accuracy: {scores.mean()}')
```

K-NEAREST NEIGHBOR ANALYSIS

```
k=3, Mean Accuracy: 0.44507084485737O5
k=5, Mean Accuracy: 0.46392595973971334
k=7, Mean Accuracy: 0.4797937756773448
k=9, Mean Accuracy: 0.49176221547935517
k=11, Mean Accuracy: 0.4902710274518227 7
```

K-NEAREST NEIGHBOR ANALYSIS

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

K-NEAREST NEIGHBOR ANALYSIS

```python
from sklearn.model_selection import GridSearchCV

grid = {'n_neighbors': [3, 5, 7, 9, 11], 'weights': ['uniform', 'distance']}
search = GridSearchCV(KNeighborsClassifier(), grid, cv = 5)
search.fit(X_train_scaled, y_train)

params = search.best_params_
knn = search.best_estimator_

print(params)
print(knn)
```

K-NEAREST NEIGHBOR ANALYSIS

```python
from sklearn.ensemble import BaggingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

knn_classifier = KNeighborsClassifier(n_neighbors = 9)

bagging_classifier = BaggingClassifier(estimator = knn_classifier, n_estimators = 10, random_state = 120123)

bagging_classifier.fit(X_train, y_train)

y_pred = bagging_classifier.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f"Bagging Accuracy: {accuracy * 100:.2f}%")
```

REFERENCES

- Nash,Warwick, Sellers,Tracy, Talbot,Simon, Cawthorn,Andrew, and Ford,Wes. (1995). Abalone. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W.