CS M148 Project

Max Deng, Aaron Isara, Aaron Shi, Ryan Yeo, Jaden Lee

The Problem

Predict: whether a prostate has malignant cancer or if it is benign.

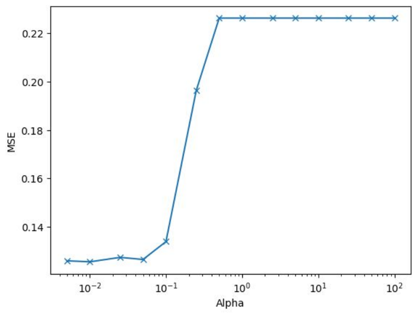
Dataset

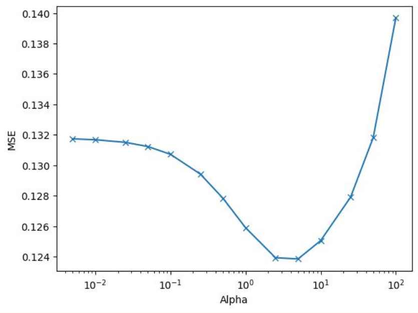
- Prostate Cancer Dataset
- 100 observations and 10 variables
  - Smaller dataset size, but clean data
- id, radius, texture, perimeter, area, smoothness, compactness, diagnosis_result, symmetry, and fractal dimension
- Supervised learning binary classification task

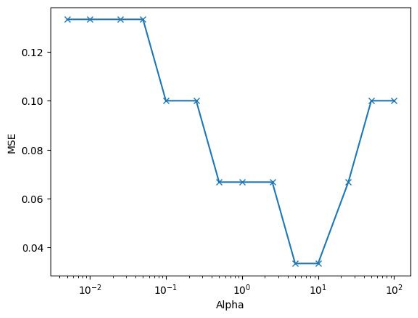
Models

Linear Regression, LASSO and Ridge Regularization

- Optimization: Hyperparameters were determined as follows:
  Alpha = 0.01 for LASSO regression,
  Alpha = 5 for Ridge regression,
  Alpha = 5 for Ridge classifier.

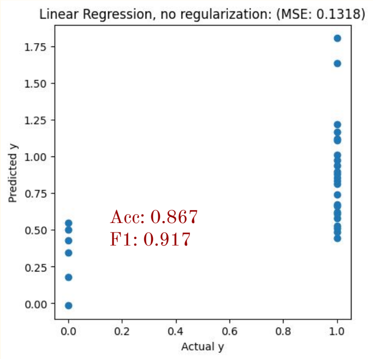




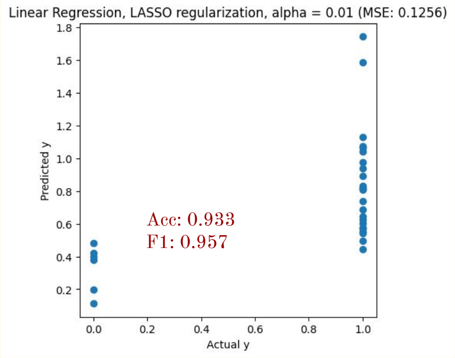


Linear Regression, LASSO and Ridge Regularization

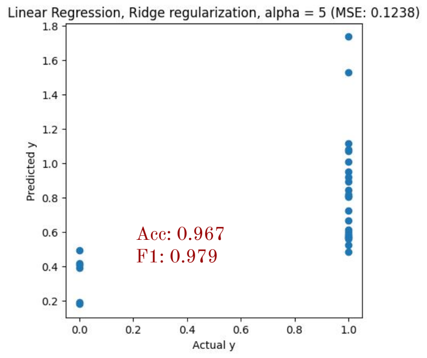
- Results: Linear regression was not expected to perform well because of the binary nature of the output, but the results exceeded expectations, with an MSE of 0.1238 for Ridge regularization and 0.0333 for Ridge classifier.


Linear Regression, no regularization: (MSE: 0.1318)

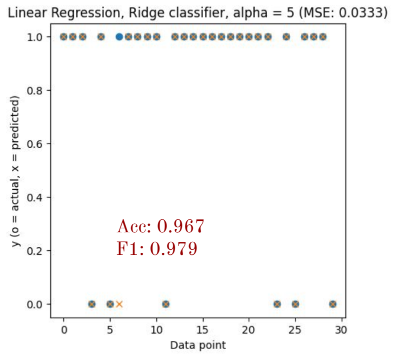
Acc: 0.867
F1: 0.917


Linear Regression, LASSO regularization, alpha = 0.01 (MSE: 0.1256)

Acc: 0.933
F1: 0.957


Linear Regression, Ridge regularization, alpha = 5 (MSE: 0.1238)

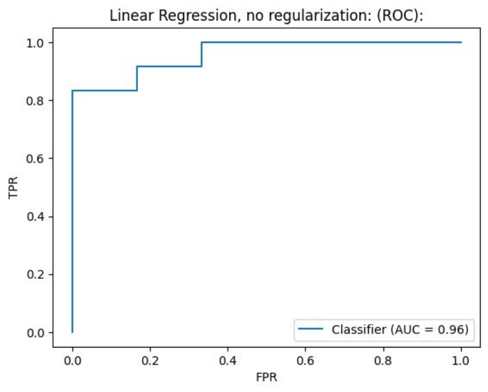
Acc: 0.967
F1: 0.979


Linear Regression, Ridge classifier, alpha = 5 (MSE: 0.0333)

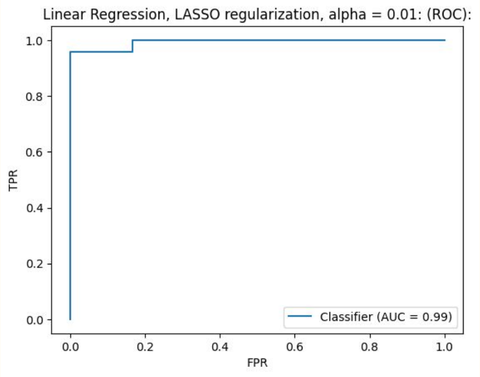
Acc: 0.967
F1: 0.979

Linear Regression, LASSO and Ridge Regularization

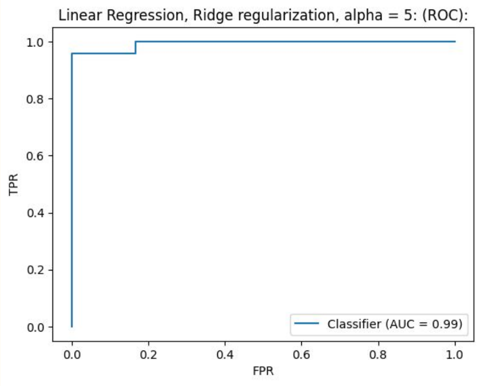
- The following ROC curves indicate we can set the positive threshold fairly high after regularization to minimize ratio of FPR to TPR



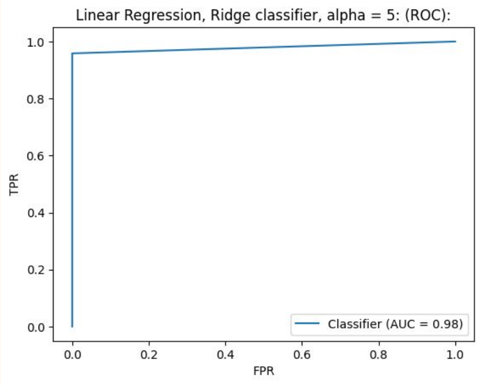
Linear Regression, no regularization: (ROC):



Linear Regression, LASSO regularization, alpha = 0.01: (ROC):



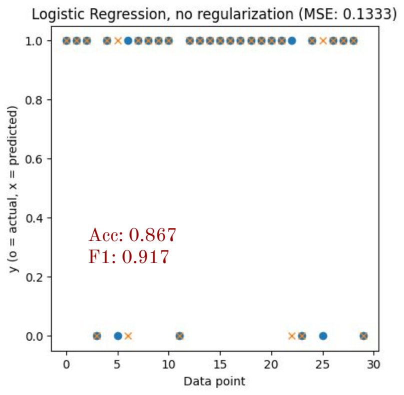
Linear Regression, Ridge regularization, alpha = 5: (ROC):

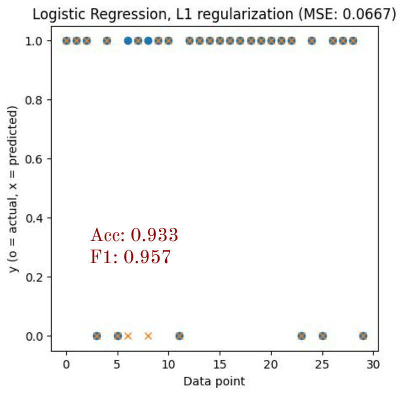


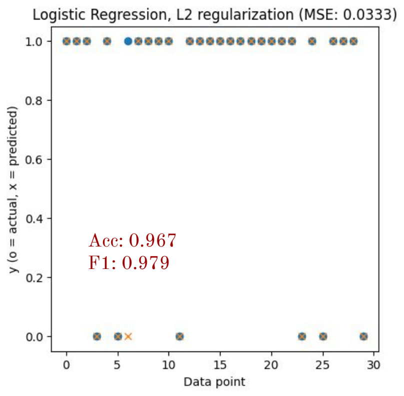
Linear Regression, Ridge classifier, alpha = 5: (ROC):

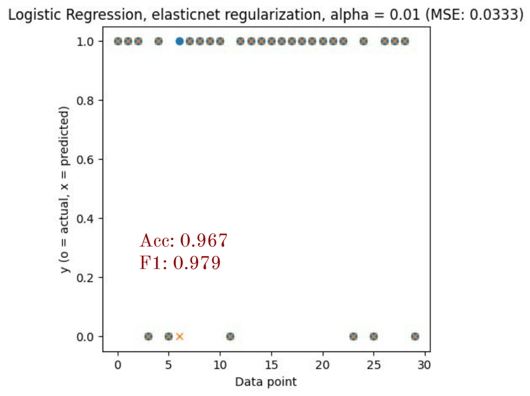
Logistic Regression: L1, L2, Elasticnet Regularization

- Optimization: L1 ratio = 0.01 was determined to be the best hyperparameter for Logistic elasticnet (L1 + L2) regression, basically relying entirely on L2.
- Results: Performed worse than expected without regularization, but achieved good performance with L1 regularization and excellent performance with L2 regularization.



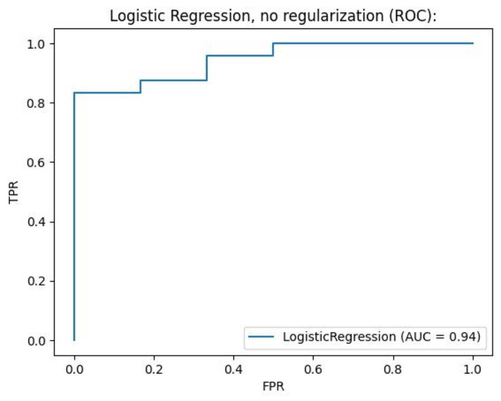


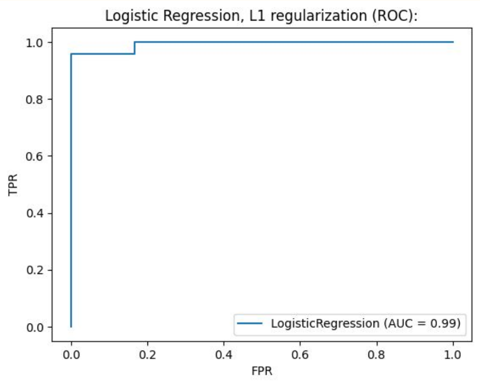


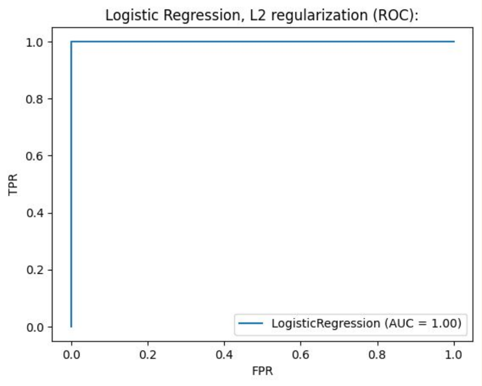


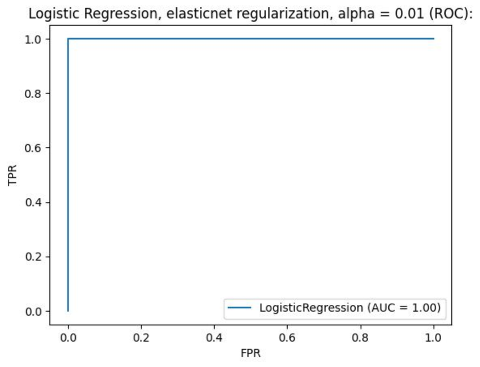
Logistic Regression: L1, L2, Elasticnet Regularization

- Very similar results to linear regression, surprisingly
- The following ROC curves indicate we can set the positive threshold fairly high after regularization to minimize ratio of FPR to TPR
- These results seem almost too good and may be an indication of overfitting







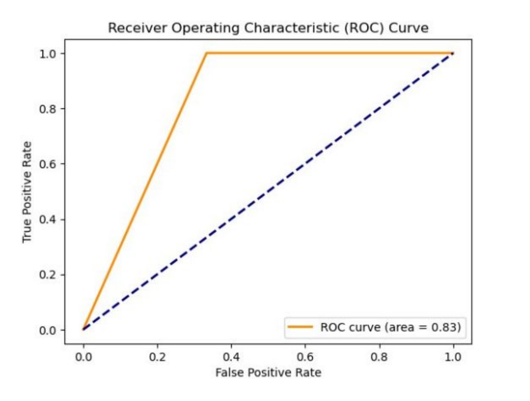


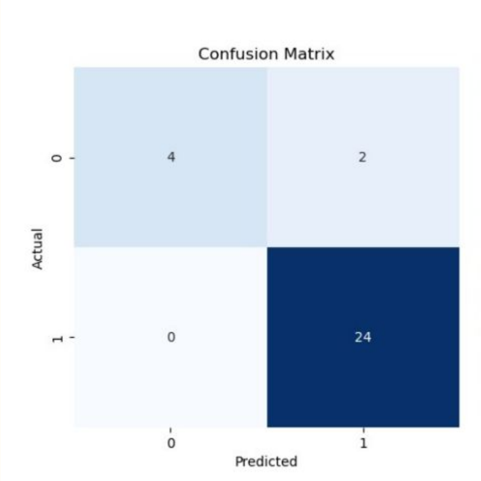
Support Vector Machine

- As part of data preprocessing, our target variable (diagnosis_result) was converted from M/B to 1/0
- Using grid search for Support Vector Classification (SVC), optimal hyperparameters were determined from the following values:
  - Regularization Parameter: 0.1, 1, 10, 100, 1000
  - Kernel: Linear, RBF(Radial Basis Function)
    - Polynomial Kernel was too computationally expensive to add among the hyperparameters
  - Kernel Coefficient: 0.0001, 0.001, 0.01, 0.1, 1
- Set random state to 0 to improve consistency and compare results among different algorithms

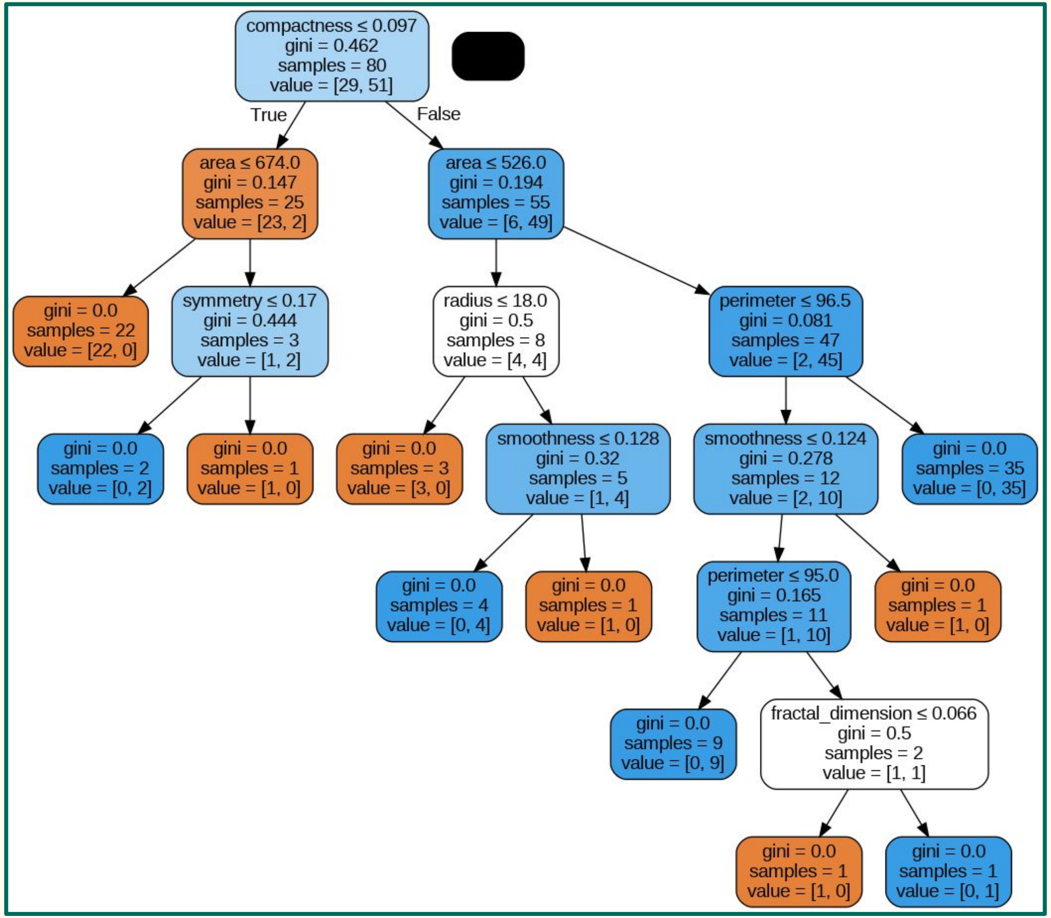
Support Vector Machine

- Optimization:
  - Regularization: 1000
  - Kernel: Linear
  - Kernel Coefficient: 0.0001
- Results: With proper data preprocessing and optimization, SVM demonstrated great performance with a MSE of 0.067 and a ROC curve area of 0.83.

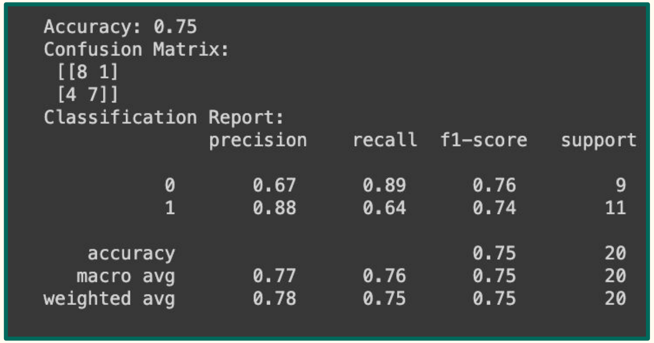




Decision Tree Pre-Optimization

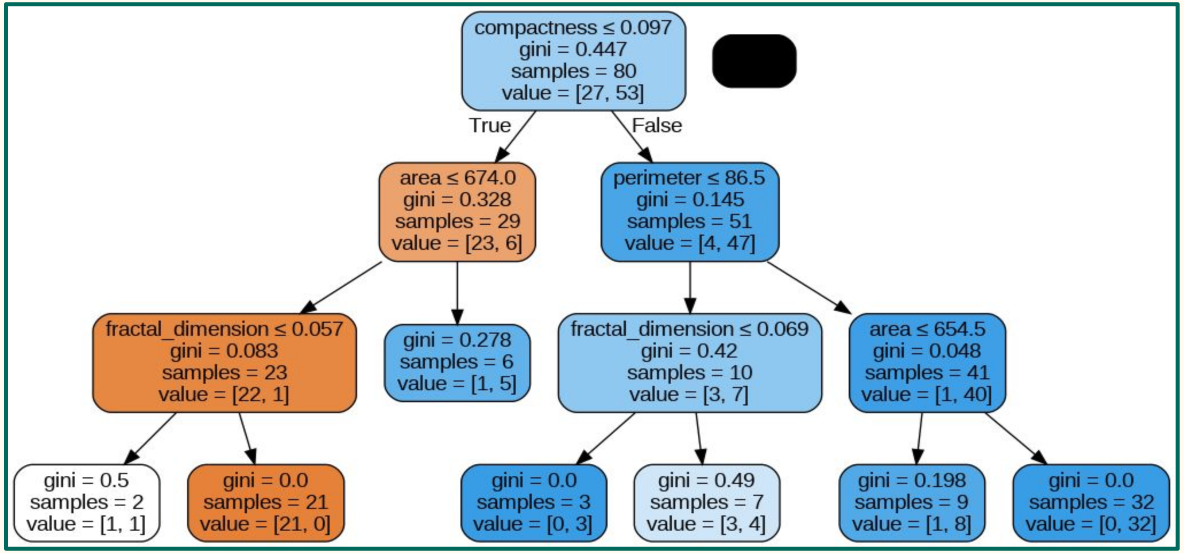


Accuracy of **0.75**

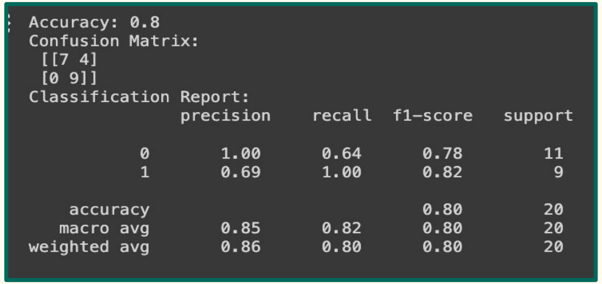


```
Accuracy: 0.75
Confusion Matrix:
 [[8 1]
 [4 7]]
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.89      0.76         9
           1       0.88      0.64      0.74        11

    accuracy                           0.75        20
   macro avg       0.77      0.76      0.75        20
weighted avg       0.78      0.75      0.75        20
```

Decision Tree Post Optimization

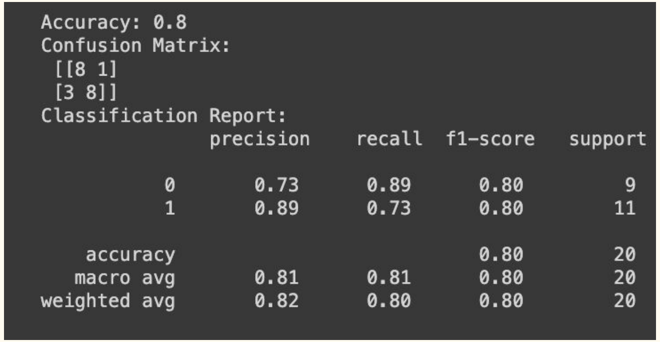


Accuracy of **0.8**



```
Accuracy: 0.8
Confusion Matrix:
 [[7 4]
 [0 9]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.64      0.78        11
           1       0.69      1.00      0.82         9

    accuracy                           0.80        20
   macro avg       0.85      0.82      0.80        20
weighted avg       0.86      0.80      0.80        20
```

Minor increase in accuracy, but much more interpretable and less likely to overfit.



```
Accuracy: 0.8
Confusion Matrix:
 [[8 1]
 [3 8]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.89      0.80         9
           1       0.89      0.73      0.80        11

    accuracy                           0.80        20
   macro avg       0.81      0.81      0.80        20
weighted avg       0.82      0.80      0.80        20
```

Random forest performed similarly to the optimized decision tree with an accuracy of 0.8

Conclusions

Data Selection

- In hindsight, the dataset was a little **too small** to produce the most accurate results.
- Prioritizing clean useable data over raw amount made the process **more amateur-friendly**, but made the results **harder to compare.**
- With additional time, we would have liked to look for similar prostate cancer datasets online to further validate our models on new data.

Model Selection

- With the data provided, we selected reasonable data models and produced good results.
- The optimized **decision tree** is the **most interpretable**, with a clear visual representation and low likelihood of overfitting on other data, but had a **slightly lower accuracy score** then the linear models.
- The **linear models** performed extremely well, perhaps in part due to **overfitting**.

Data Selection

- In hindsight, the dataset was a little **too small** to produce the most accurate results.
- Prioritizing clean useable data over raw amount made the process **more amateur-friendly**, but made the results **harder to compare.**
- With additional time, we would have liked to look for similar prostate cancer datasets online to further validate our models on new data.

Model Selection

- With the data provided, we selected reasonable data models and produced good results.
- The optimized **decision tree** is the **most interpretable**, with a clear visual representation and low likelihood of overfitting on other data, but had a **slightly lower accuracy score** then the linear models.
- The **linear models** performed extremely well, perhaps in part due to **overfitting**.

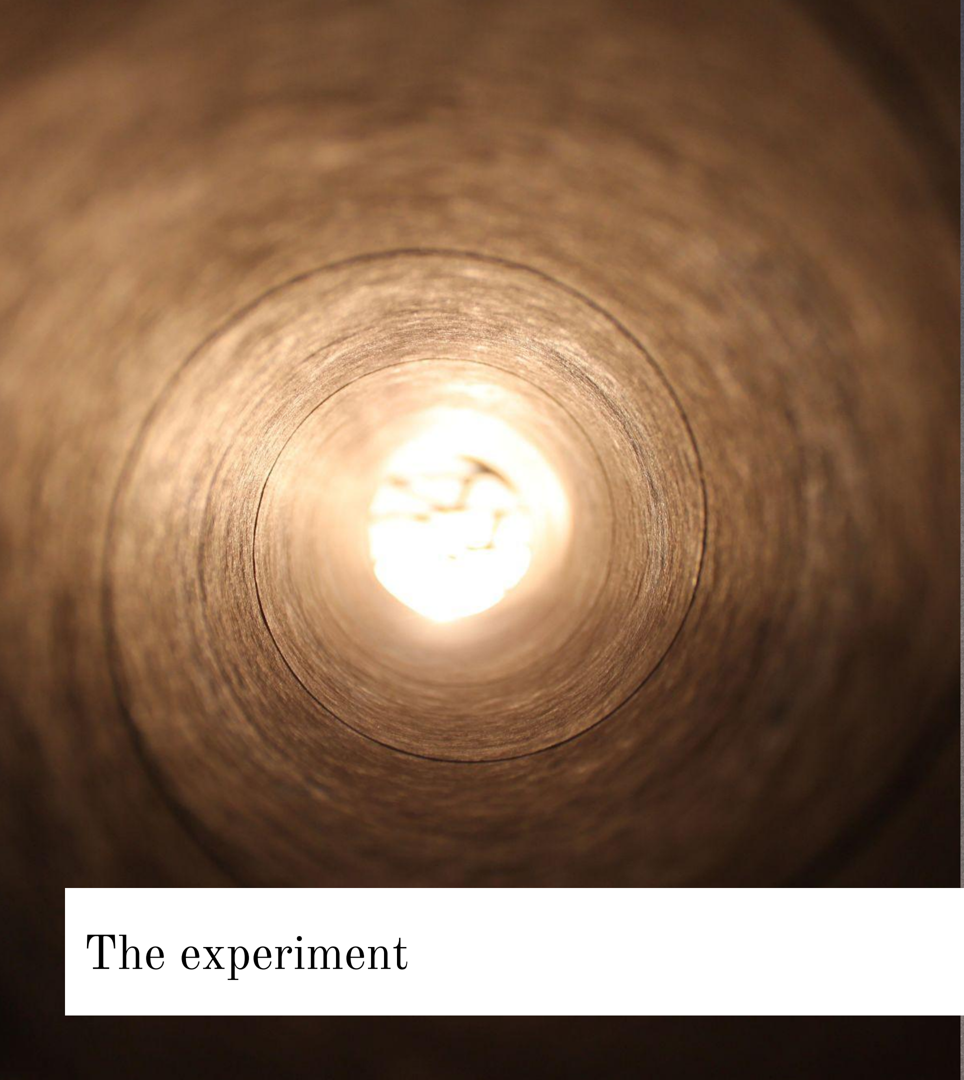
Hypothesis support

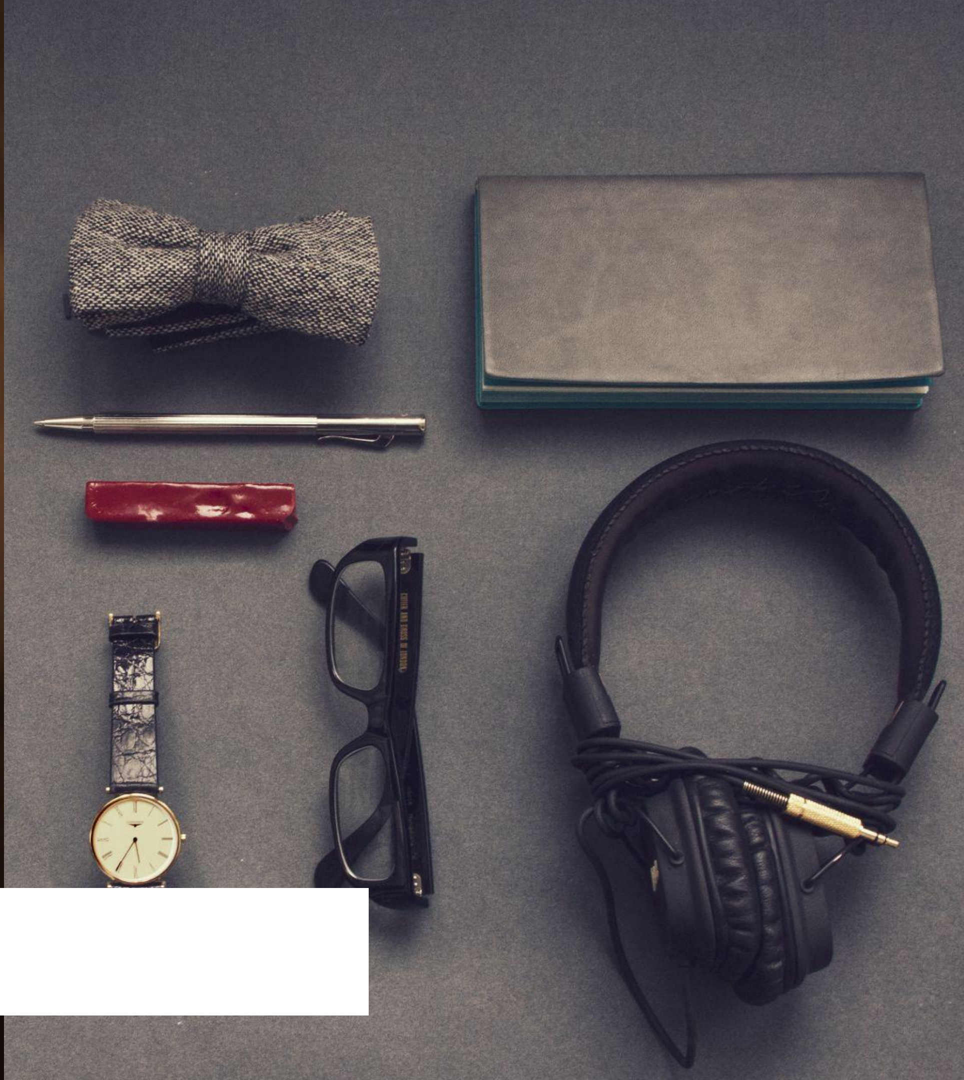
I think this is what's going to happen because...

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip.

Variables that may affect the outcome...

- Lorem ipsum dolor sit amet, consectetur adipiscing elit
- Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua





The experiment

Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip.