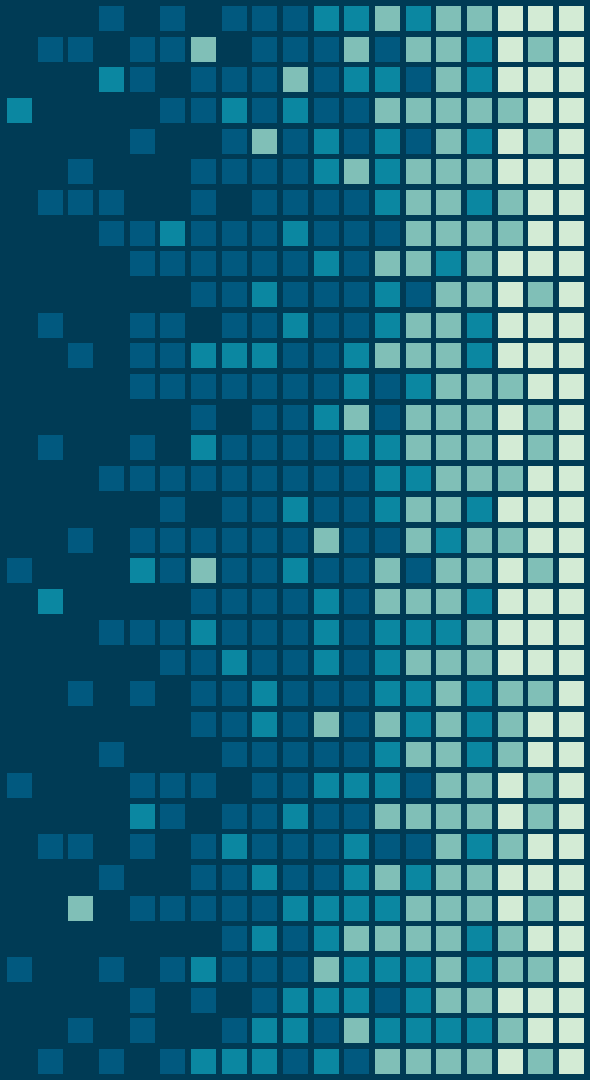
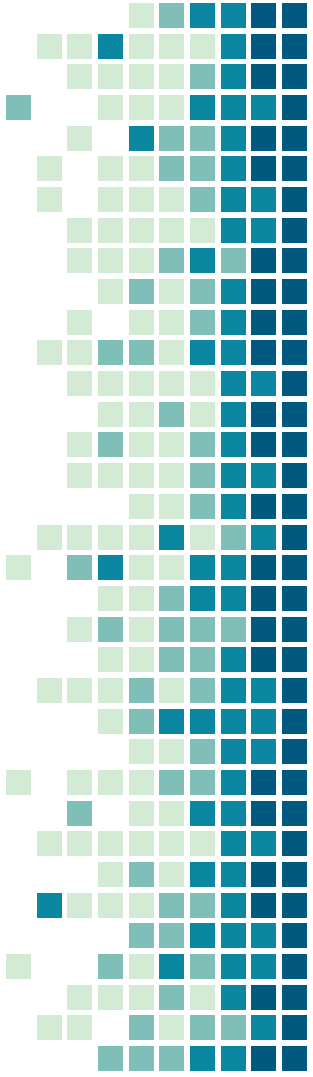
TEAM #1 GROUP PROJECT

By Lucas Allen, Kevin Ly, Olabanji Osifowokan, Aniruddha Pochimcherla, and Rishvita Vedartham



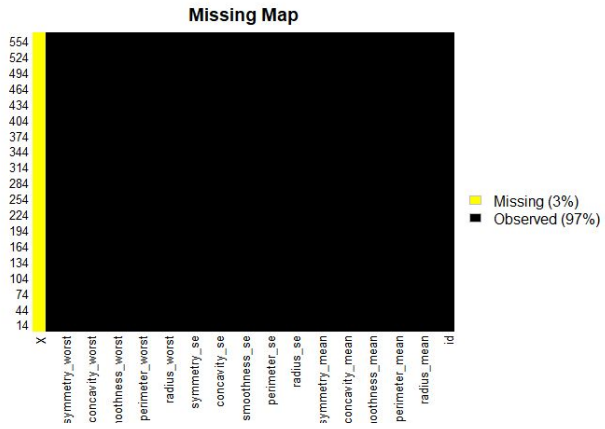
PROJECT INTRODUCTION

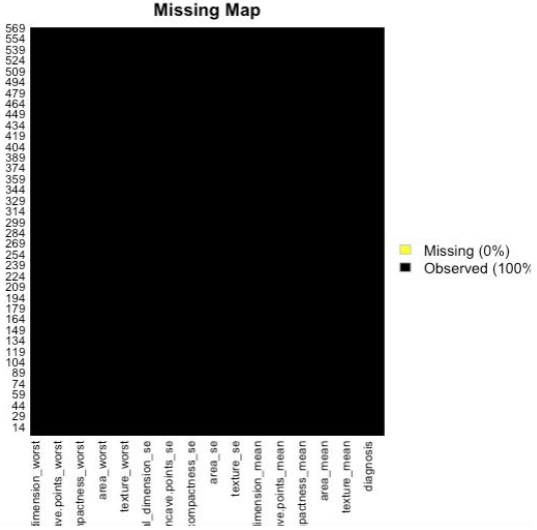


EXPLORATORY DATA ANALYSIS

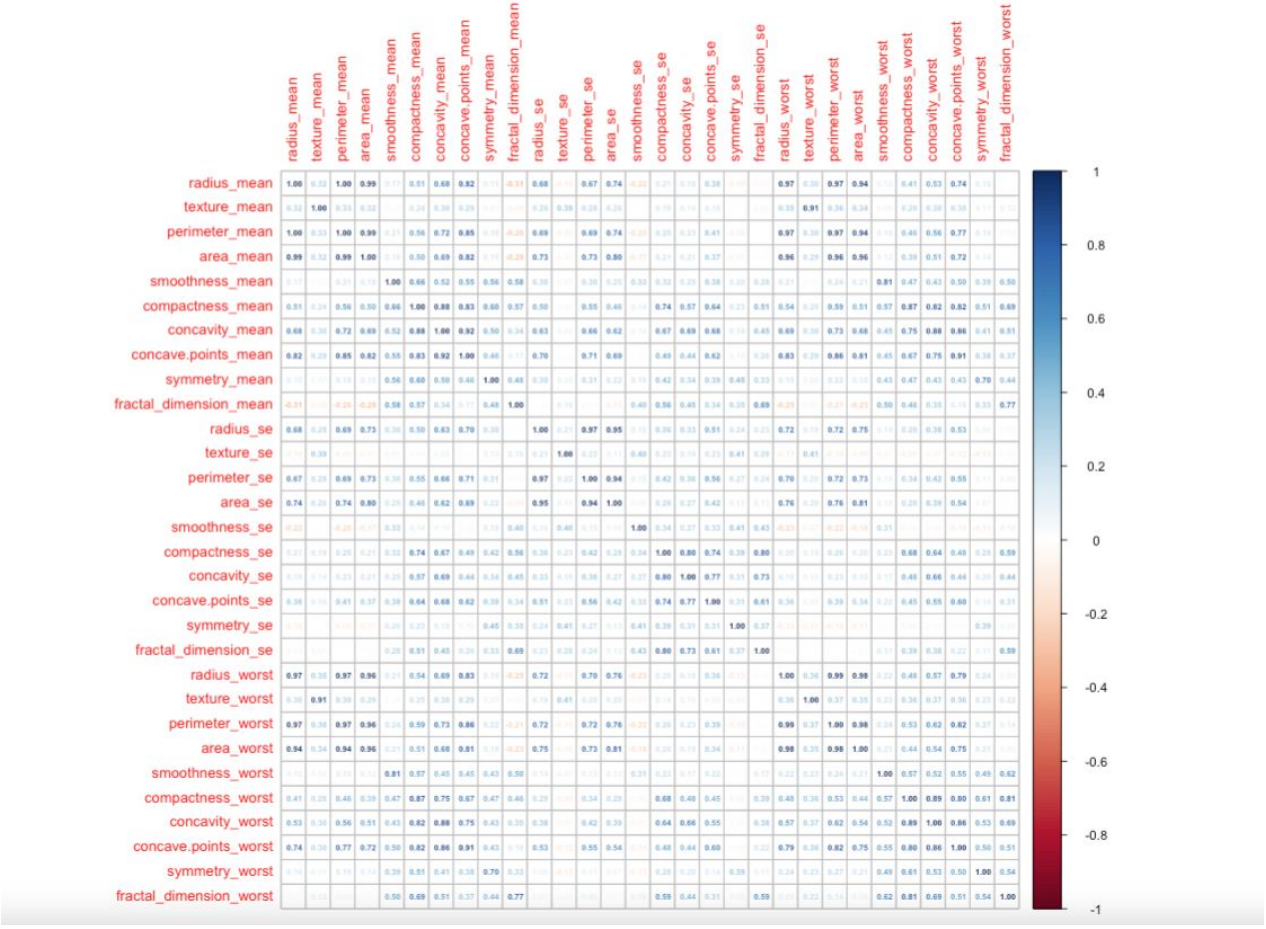
- **Correlation plot** - Made to understand the relationship between multiple variables
- **Box Plot** - Comparison of Mean Radius for Malignant and Benign Tumors
- **Histogram** - Distribution of Radius Mean by Diagnosis
- **Density Plots** - Smoother version of the Histogram

MISSING MAP

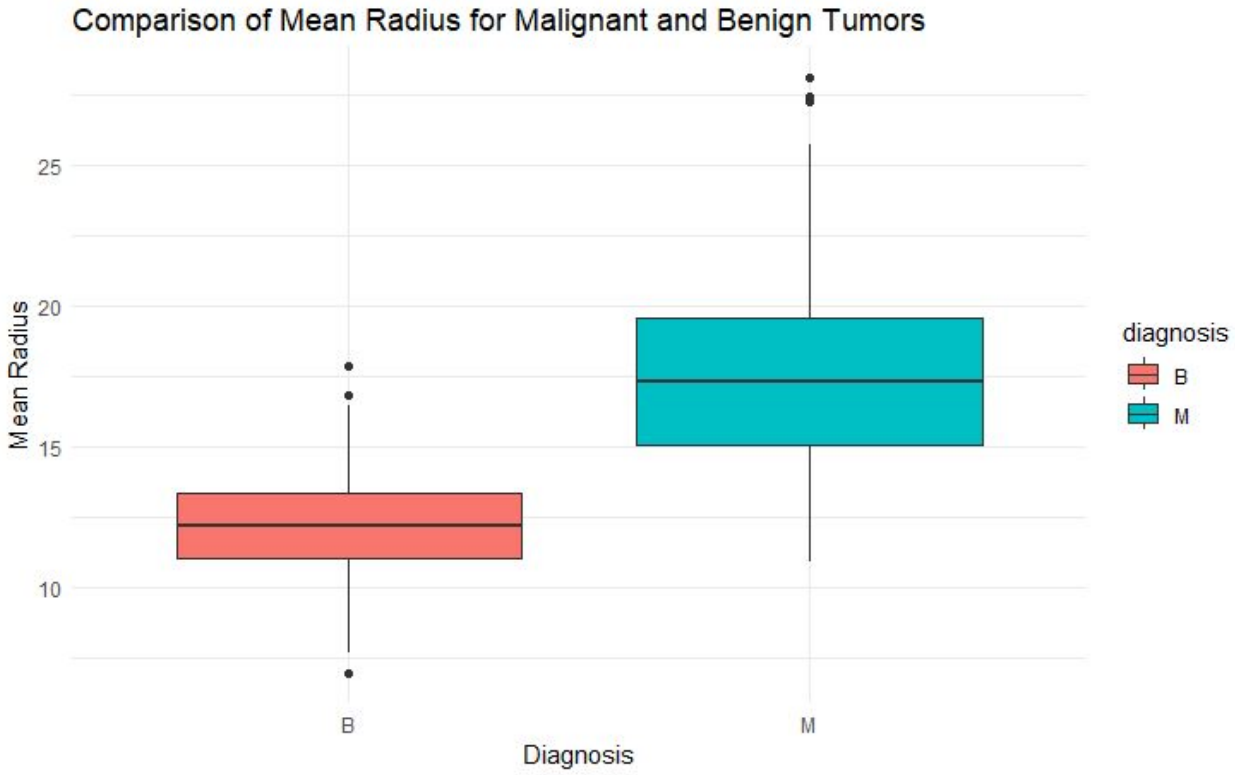

Missing Map
554
524
494
464
434
404
374
344
314
284
254
224
194
164
134
104
74
44
14
X
symmetry_worst
concavity_worst
perimeter_worst
radius_worst
symmetry_se
concavity_se
smoothness_se
perimeter_se
radius_se
symmetry_mean
concavity_mean
smoothness_mean
perimeter_mean
radius_mean
id
Missing (3%)
Observed (97%)


Missing Map
569
554
539
524
509
494
479
464
449
434
419
404
389
374
359
344
329
314
299
284
269
254
239
224
209
194
179
164
149
134
119
104
89
74
59
44
29
14
area_worst
texture_worst
area_se
texture_se
area_mean
texture_mean
diagnosis
Missing (0%)
Observed (100%

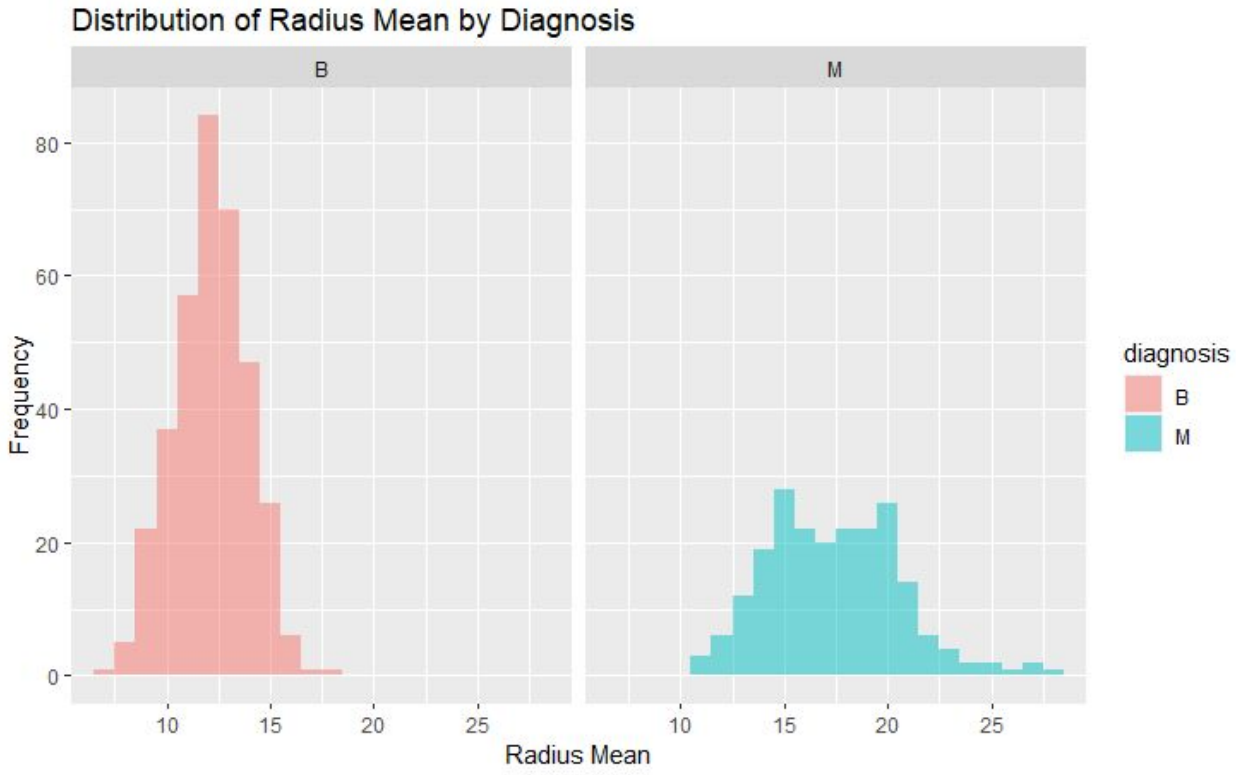
Correlation Plot



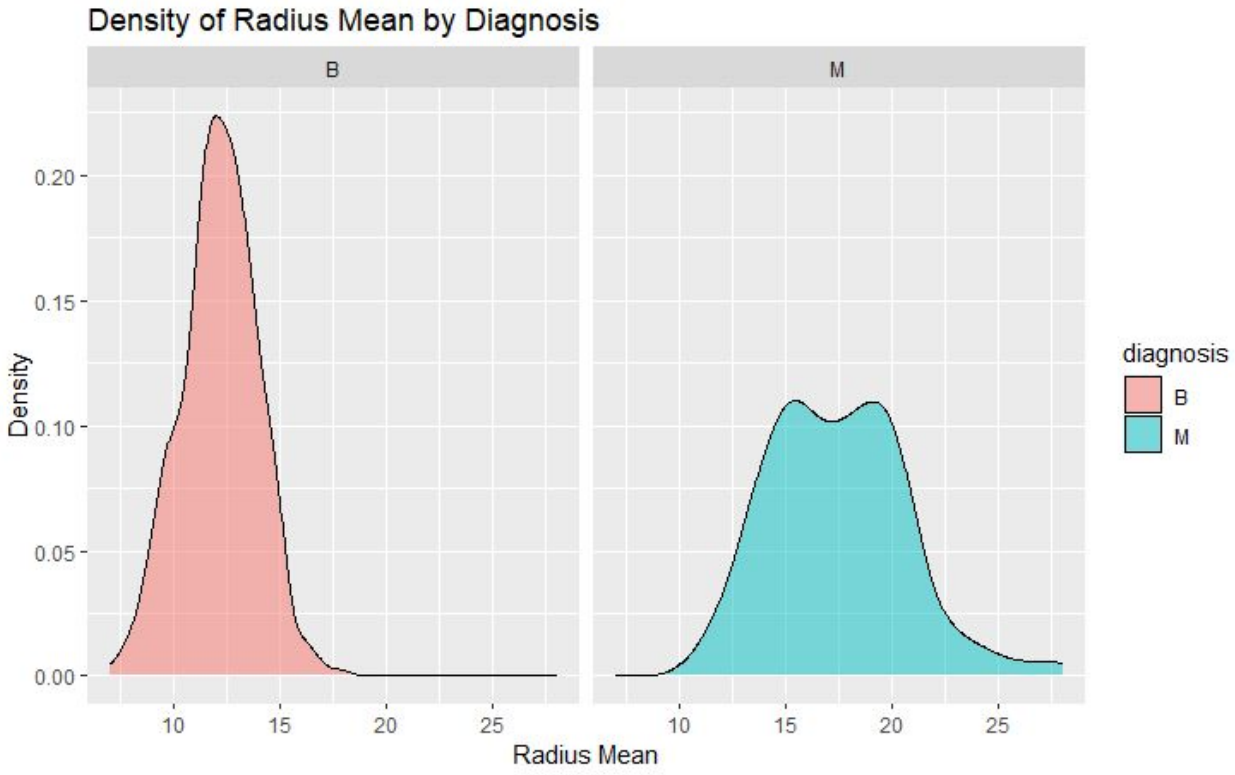
BOX PLOT


Comparison of Mean Radius for Malignant and Benign Tumors
Mean Radius
25
20
15
10
B
M
Diagnosis
diagnosis
B
M

HISTOGRAM


Distribution of Radius Mean by Diagnosis
B
M
Frequency
80
60
40
20
0
10
15
20
25
Radius Mean
diagnosis
B
M

DENSITY PLOTS



PROJECT ALGORITHMS

DECISION TREE

- A tree-like model that makes decisions by recursively splitting data into subsets based on significant features, leading to a set of rules for predicting a target variable

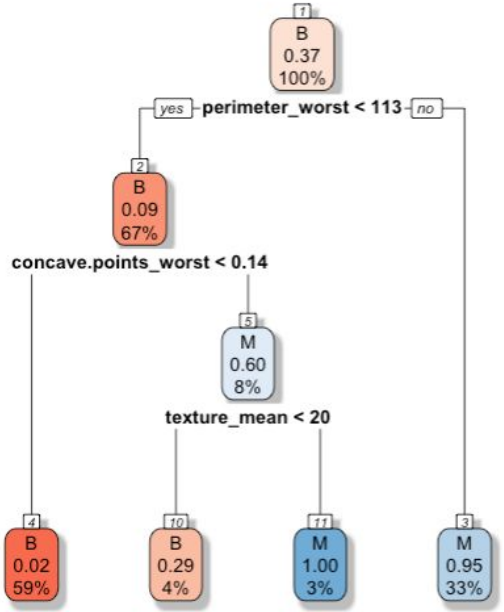
- Why Use Decision Trees?
  - Capable of handling lots of factors, which is present in our data
  - Identify patterns and make decisions based on those patterns
  - The model is similar to the diagnosis process used by physicians - Process of elimination starting using flowcharts.

DECISION TREE EVALUATIONS

	Benign	Malignant
Actual Benign	102	6
Actual Malignant	5	58

- Accuracy: 93.6%
- Precision: 94.4%
- Recall: 95.3%

DECISION TREE EVALUATIONS



K-NEAREST NEIGHBORS

- A model that classifies a data point by considering the majority class of its nearest k neighbors in the feature space.

- Why Use KNN?
  - Capable of handling multiple features in a dataset
  - Suitable for moderate-sized datasets.
  - Makes highly accurate predictions
  - Well-suited for binary classification tasks

K-Nearest Neighbors Evaluations

K =1	Benign	Malignant
Actual Benign	101	6
Actual Malignant	6	57

- Accuracy: 92.9%
- Precision: 90.4%
- Recall: 90.4%

K = 3	Benign	Malignant
Actual Benign	99	8
Actual Malignant	6	57

- Accuracy: 91.7%
- Precision: 87.6%
- Recall: 90.4%

K = 5	Benign	Malignant
Actual Benign	99	8
Actual Malignant	5	58

- Accuracy: 92.3%
- Precision: 87.8%
- Recall: 92.0%

THE ISSUE OF MISCLASSIFICATION

MISCLASSIFICATION

False Positives

- The model predicts a tumor as malignant when it is actually benign.
- It may lead to unnecessary treatments and increased cost of healthcare for patients who don't have cancer.

False Negatives

- The model predicts a tumor as benign when it is actually malignant.
- Delayed diagnosis and treatment for patients with cancer

SUPERVISED ALGORITHM COMPARISON

K-NEAREST NEIGHBOR VS DECISION TREE

KNN

- Best at classifying new observations, and is more accurate when trained on larger datasets
- Better at classifying large data sets, where the boundaries between neighborhoods seems arbitrary (higher variable independence)

Decision Tree

- Better overall for classification for smaller datasets
- This dataset was measuring mostly area and dimensions on images, so the dimensions would be somewhat dependent on each other

ETHICAL IMPLICATIONS

MACHINE LEARNING IN MEDICAL DIAGNOSIS

- Data Collection and Privacy
  - Securing patient consent for data collection and utilization is crucial
  - Being open with patients about how their data will be used and privacy measures taken to protect their personal information plays a significant role in gaining their trust
- Transparency and Accountability
  - Models with a higher level of clarity can help medical professionals better understand the criteria that is used to make predictions
  - Machine learning algorithms can streamline error detection and correction through an iterative process
- Regulatory Frameworks
  - Data systems must comply with legal obligations and follow ethical guidelines

FAIRNESS AND TRANSPARENCY IN PREDICTIONS

- Data Quality and Bias:
  - Reliable data will result in more accurate predictions and can be used with confidence to analyze and make decisions
  - Unfairness in data collection, sometimes even unintentionally, can skew the data in a particular way
  - This leads to unreliable data which may lead to inaccurate predictions and costly decisions
  -
- Ensuring Fairness and Transparency in our Model:
  - Removed the column named 'X' due to all the missing values as shown on the Missing Map ('missmap')

CONCLUSION

- Project Description/Introduction
- Exploratory Data Analysis
- Models & Evaluations
- Issues of Misclassification
- Comparison of Models
- Ethical Implications:
    - ML in Medical Diagnosis &
    - Fairness and Transparency

THANK YOU FOR LISTENING!