Binary Classification of Benign 😊 and Malignant 😳 Cancer 🎗

Cheeto Team (32) - Nathan, Jonathan, Nguyen, Mukta

01. Introduction and Addressing the Problem

2 million people will be diagnosed with cancer and 609,000 will die this year. Patient survival is correlated to early detection and identification of the cancer, but some types of cancer can be hard to correctly diagnose. This is where our project aims to fill the gaps.

02. Dataset and Methods

Dataset Information:

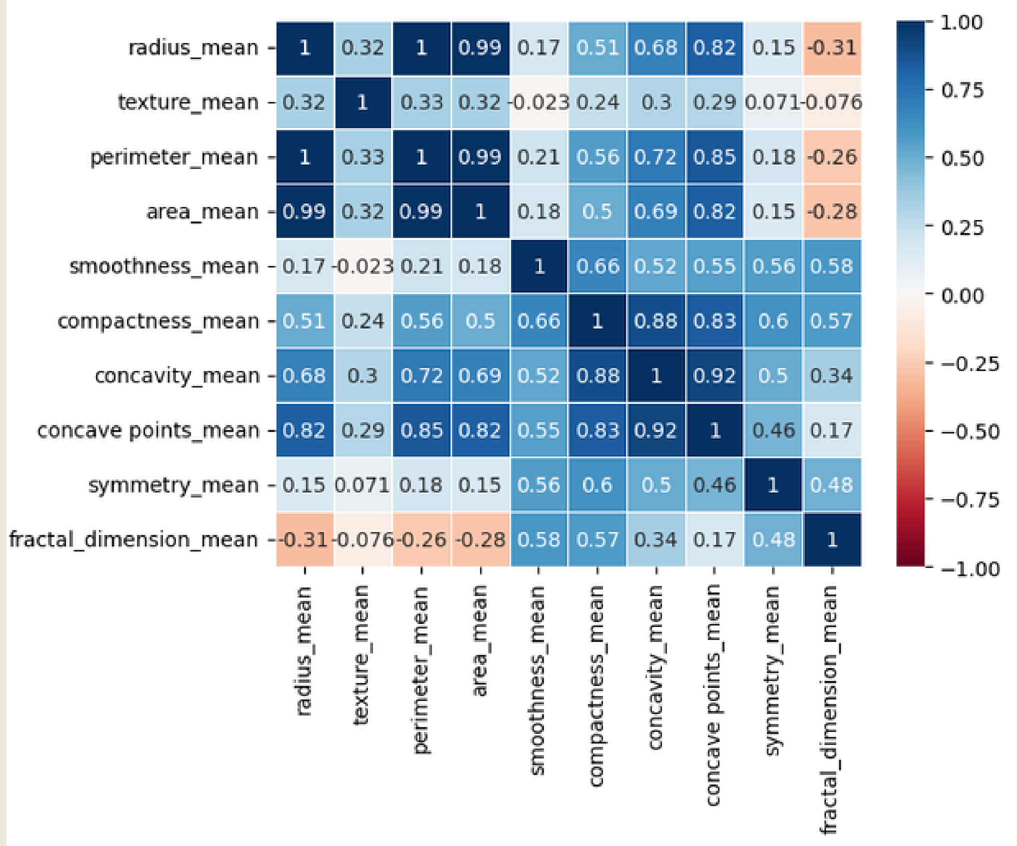
1. The dataset includes 570 records of tumors
2. Each record includes up to 30 attributes describing the appearance of the tumor with characteristics
3. 63% of dataset is benign tumors and 37% for malignant tumors

Dataset Exploratory Analysis:

1. Dropped redundant or unnecessary columns (e.g., id column).
2. Generated correlation matrix to identify closely correlated columns (e.g., radius, perimeter, and area) and either dropped or combined them to avoid multicollinearity issues.
3. Normalized the data to bring extreme values between 0 and 1, reducing prediction model errors.
4. Binarized the M and B values of the prediction column (diagnosis column) as 1 and 0 for analysis.

Analysis Algorithms used:

1. Logistic regression
2. Feed-forward neural network
3. Support Vector Machine



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
radius_mean	1	0.32	1	0.99	0.17	0.51	0.68	0.82	0.15	-0.31
texture_mean	0.32	1	0.33	0.32	-0.023	0.24	0.3	0.29	0.071	-0.076
perimeter_mean	1	0.33	1	0.99	0.21	0.56	0.72	0.85	0.18	-0.26
area_mean	0.99	0.32	0.99	1	0.18	0.5	0.69	0.82	0.15	-0.28
smoothness_mean	0.17	-0.023	0.21	0.18	1	0.66	0.52	0.55	0.56	0.58
compactness_mean	0.51	0.24	0.56	0.5	0.66	1	0.88	0.83	0.6	0.57
concavity_mean	0.68	0.3	0.72	0.69	0.52	0.88	1	0.92	0.5	0.34
concave points_mean	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1	0.46	0.17
symmetry_mean	0.15	0.071	0.18	0.15	0.56	0.6	0.5	0.46	1	0.48
fractal_dimension_mean	-0.31	-0.076	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1

03. Experimental Results and Findings

We judged our models on accuracy, F-Score, and mean-squared error. We chose to consider F-Score because our dataset has slightly more benign samples than malignant. However, in most cases the F-Score was close to the accuracy metric, so we believe that the sample imbalance was not a major factor in this case.

1 We classified the data points into categories which gave us a probability estimate of our result. **Logistic Regression**

2 The nodes in the feed-forward neural network were processed with a sigmoid/logistic activation function. **Feed Forward NN**

3 We learned that the real support vector machines were the friends we made along the way. **Support Vector Machine**

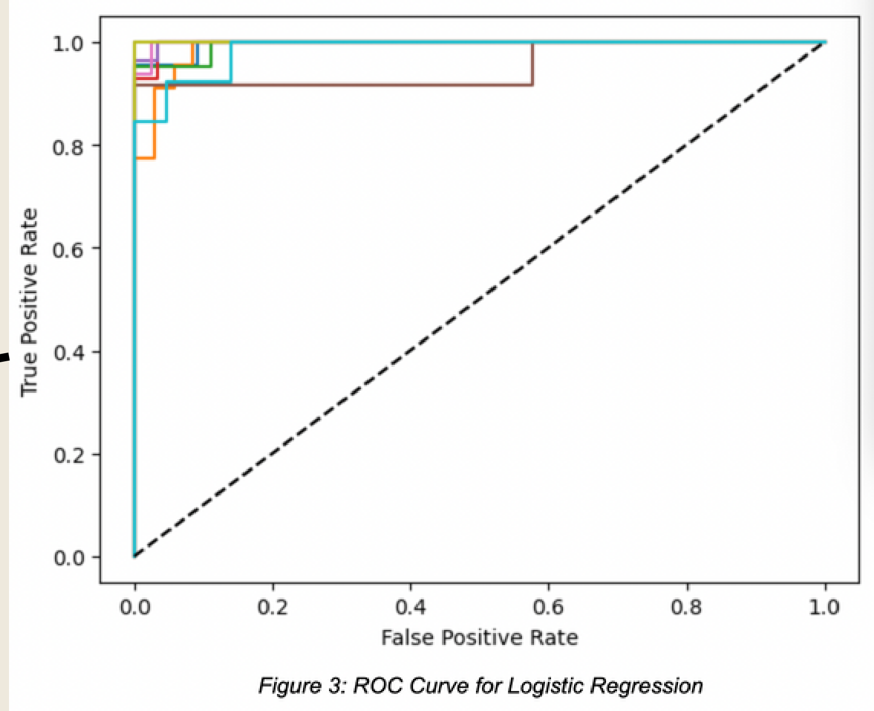


Figure 3: ROC Curve for Logistic Regression

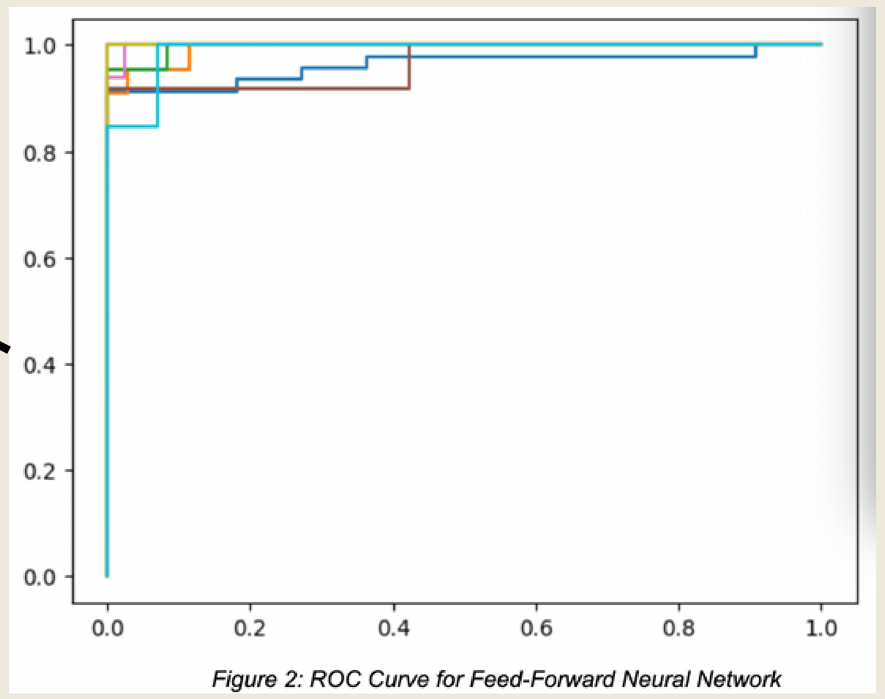
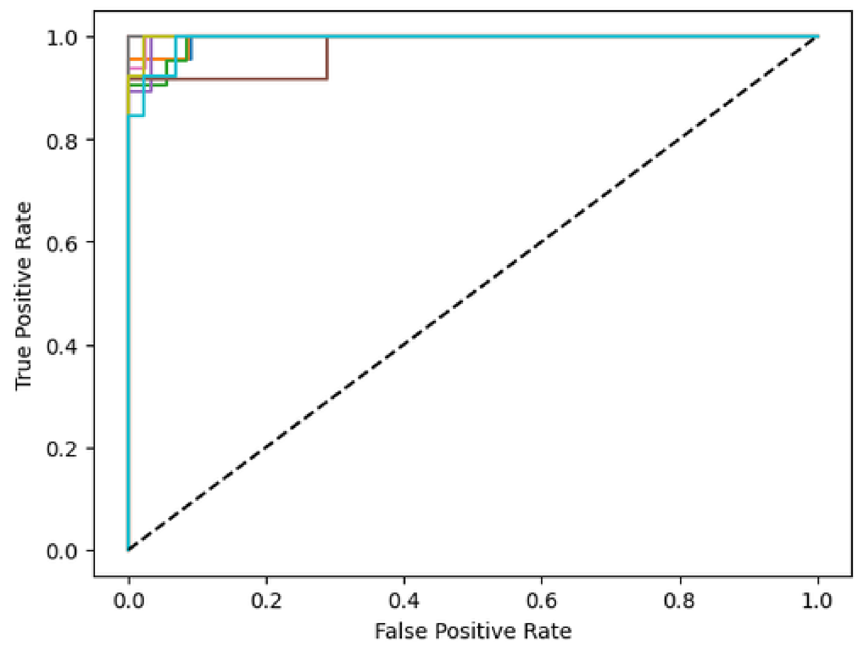


Figure 2: ROC Curve for Feed-Forward Neural Network



Model	Accuracy	MSE	F-Score
Logistic Regression	0.9578	0.0422	0.9470
Feed-Forward Neural Network	0.8875	0.1124	0.8520
Support Vector Machine	0.9560	0.0439	0.9380

04. Conclusion and Key Takeaways

In conclusion, we found that the logistic regression model was the most accurate in classifying tumors.

We performed k-fold cross-validation to assess the generalization of the logistic regression model and yielded a very high average accuracy for this model. This ensures that the model performs consistently on unseen data and guards against overfitting or under-fitting issues. Interestingly enough, the best possible model produced for the Feed-Forward Neural Network and the Logistic Regression models had 100% accuracy. This seems oddly unrealistic and may point to some faults in the dataset or model overfitting.

In the future, we could improve our model by using hyper-parameter tuning to optimize the performance of our logistic regression model. We could also extend our model to identify pictures of tumors and automatically categorize them as malignant or benign.

05. References

https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf

https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm

https://seer.cancer.gov/statfacts/html/common.html

https://canvas.ucdavis.edu/courses/786744