Capstone 3 - Project Report

Predicting breast cancer using Machine learning models

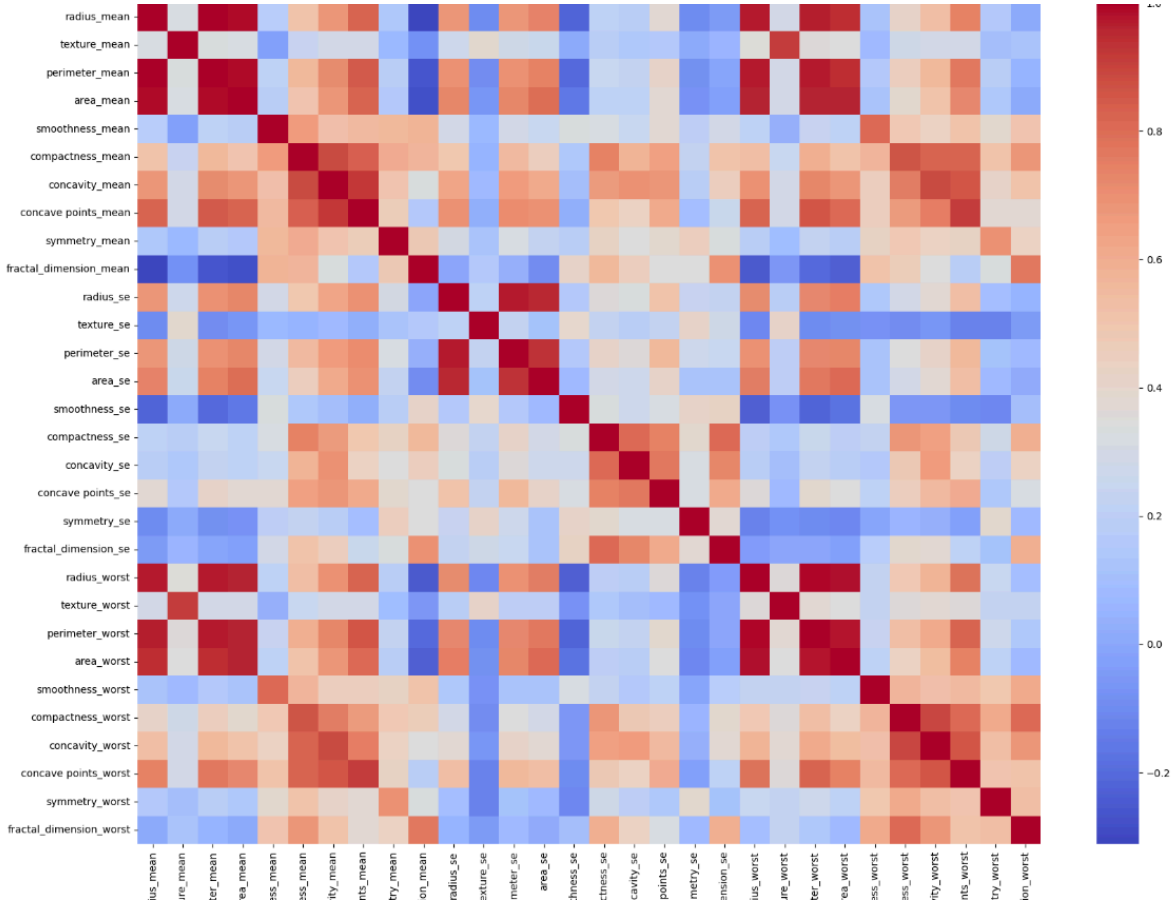
By Lakshika Jain

Objective: Building a Machine Learning model to identify if image features of a cell are for malignant or benign tumors in breast cancer.

Data: The morphological and dimensional data of the tumor was procured from the kaggle website. This data had 32 columns and about 600 rows. There is - ID and unnamed:32 columns that I dropped as they were either nulls or unique identifiers that didn't contribute to the analysis. The target variable was Diagnosis - Which was either M for malignant or B for benign tumors. Other 29 column provided the texture/size/shape of the tumors that were imaged during breast cancer screening.
Link to data: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data

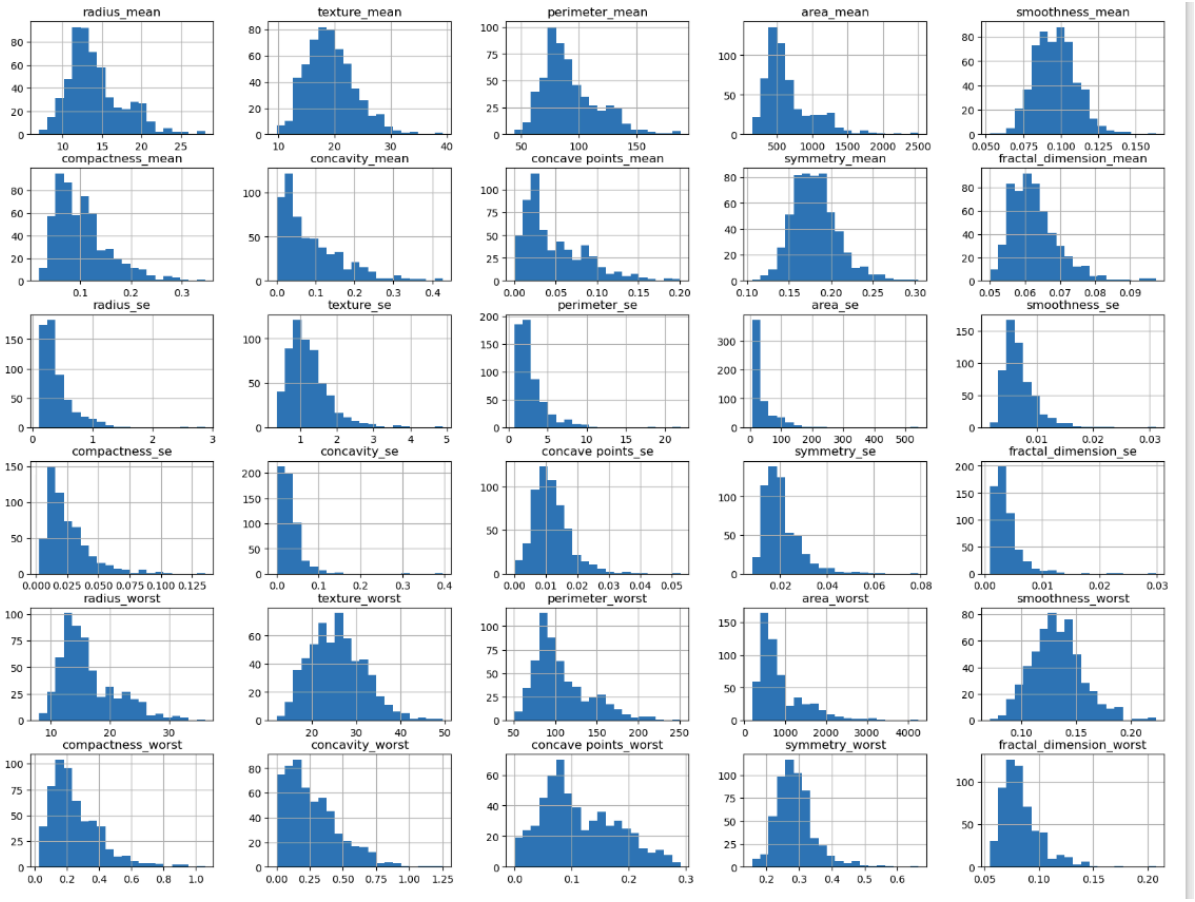
Data Wrangling & Exploratory Data Analysis:

- After dropping the ID and unnamed:32 columns from the dataset, I looked if there were any other columns that had any nulls.
- I also looked for the datatypes of these variables, the Diagnosis column was an object and everything else was a float64 field.
- The data had 357 rows of benign tumor diagnosis and 212 rows of malignant tumor diagnosis.
- There were no features with 0 standard deviation that needed to be dropped.
- Correlation heat map of this data was very interesting, there were many features which were highly correlated to each other.

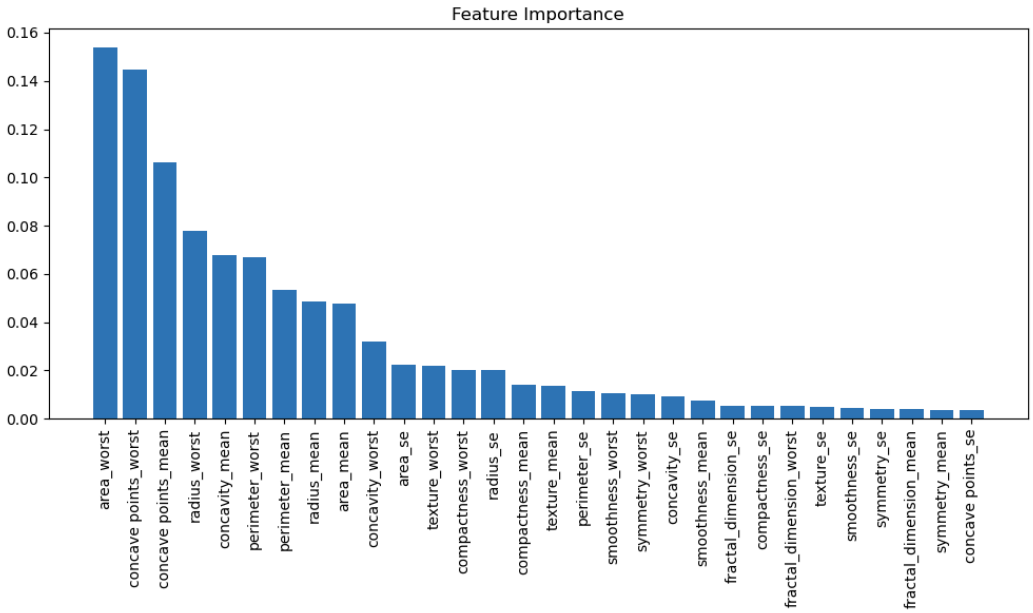


1. **High Correlations Between Size-Related Features**: Features like **radius_mean**, **area_mean**, and **perimeter_mean** are highly correlated, suggesting that these size-related metrics are strongly linked
   Multicollinearity: There is significant correlation between several features, especially the "mean" and "worst" values of the same measurements.

- Upon plotting the histogram, following was deduced:
  Skewed distributions: Many features are right-skewed, suggesting that most samples have lower values for these features, but some samples with much higher values (potential outliers) exist.
  Outliers: There are potential outliers, especially in features like area_mean and perimeter_mean, which could indicate certain samples with very large tumors.
  Symmetry and Spread: Some features, like texture_mean and smoothness_mean, are more symmetrically distributed, showing a more uniform distribution of values across samples.
  High Range in "Worst" Features: The "worst" measurements typically exhibit a larger range compared to the means, as expected, and are useful in identifying the extreme or most severe cases.



- Since many ‘mean’ and ‘worst’ values were highly correlated to each other, performing feature importance and removing the features with low importance can help with reducing the noise in the dataset.
    1. **Area_worst**, **concave points_worst**, and **concave points_mean** are the most influential features, with size-related and shape-related characteristics having the greatest impact on predicting whether a tumor is benign or malignant.
    2. Features related to the **worst-case measurements** (extreme values) tend to dominate the model's decisions, indicating that the most aggressive features of the tumor are crucial for the prediction task.
    3. Less important features, such as **fractal dimension** and **symmetry**, seem to contribute less to the model, suggesting that they may be less helpful for tumor classification in this dataset.




```
area_worst: 0.1538923646320539
concave points_worst: 0.14466326620735526
concave points_mean: 0.10620998844591636
radius_worst: 0.07798687515738044
concavity_mean: 0.06800084191430111
perimeter_worst: 0.06711483267839193
perimeter_mean: 0.05326974612817967
radius_mean: 0.04870337173775524
area_mean: 0.047555008860185516
concavity_worst: 0.031801595740040434
area_se: 0.02240696016045847
texture_worst: 0.021749011006763203
compactness_worst: 0.020266035899623565
radius_se: 0.020138917194191527
compactness_mean: 0.013944325074050483
```

- Dropped the following features from the dataset:

```
[45]: ['fractal_dimension_worst',
       'radius_mean',
       'symmetry_mean',
       'radius_se',
       'compactness_mean',
       'smoothness_mean',
       'area_mean',
       'texture_mean',
       'compactness_se',
       'perimeter_worst',
       'concavity_mean',
       'concave_points_se',
       'perimeter_se',
       'compactness_worst',
       'radius_worst']
```

Modelling:

- To prepare the data for modeling, I normalized the features and split the data into training and validation sets. The split was 70:30.
- Since this is a binary classification problem, the goal is to classify whether a tumor is benign or malignant.
- I built 4 models to test which model performed the best. Logistic Regression, SVM, KNN, Neural Networks.

Logistic Regression :

Logistic Regression:

	precision	recall	f1-score	support
0	0.97	0.98	0.98	108
1	0.97	0.95	0.96	63
accuracy			0.97	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.97	0.97	0.97	171

SVM:

SVM:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	108
1	0.98	0.87	0.92	63
accuracy			0.95	171
macro avg	0.96	0.93	0.94	171
weighted avg	0.95	0.95	0.95	171

KNN:

KNN:

	precision	recall	f1-score	support
0	0.94	0.94	0.94	108
1	0.90	0.89	0.90	63
accuracy			0.92	171
macro avg	0.92	0.92	0.92	171
weighted avg	0.92	0.92	0.92	171

Neural Network:

Neural Network:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	108
1	0.97	0.97	0.97	63
accuracy			0.98	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.98	0.98	0.98	171

Logistic Regression performed very well overall, with a balanced performance across both classes. The recall for malignant cases is 95%, meaning it detects 95% of the actual malignant cases, which is important in this context.

SVM has a high precision for detecting malignant cases (98%) but a lower recall (87%), meaning it misses some malignant cases. This may be critical in medical diagnosis, where missing malignant cases could have serious consequences.

KNN performs reasonably well, but its recall and precision for malignant cases are lower compared to Logistic Regression and Neural Networks. This suggests that KNN may not be the best choice for this task.

The Neural Network achieves the highest accuracy (98%) with balanced precision, recall, and F1-scores for both classes. This suggests it performs well at both detecting and correctly identifying benign and malignant cases.

Neural network:

The architecture consisted of:

Input Layer: Matching the number of features in the dataset.

Hidden Layers: Two fully connected (dense) layers, where the number of neurons and the activation functions were tuned.

Output Layer: A single neuron with a sigmoid activation for binary classification (outputting a probability between 0 and 1).

Compilation: The model was compiled using the Adam optimizer, the binary crossentropy loss function (appropriate for binary classification), and accuracy as the metric for evaluation.

Step 3: Hyperparameter Tuning

Keras Tuner was used to optimize the model. The hyperparameters that were tuned included:

- Number of neurons in the hidden layers.
- Dropout rate to prevent overfitting.
- Learning rate to control the step size during optimization.
- Batch size and epochs for controlling the training process.

The best hyperparameters found during tuning were:

Units in the first layer: 224 neurons.

Units in the second layer: 512 neurons.

Dropout rate: 0.2 (20% of the neurons dropped randomly during training to reduce overfitting).

Learning rate: 0.01, which worked well in balancing the training speed and stability.

Model Training and Validation

The model was trained for 10 epochs with the best hyperparameters found during the tuning process.

Throughout training, the training accuracy steadily improved, reaching 88.58% by the final epoch.

The validation accuracy showed strong performance, reaching 94.74% by the final epoch, indicating that the model was generalizing well to unseen data.

The validation loss decreased over time, indicating good convergence and no signs of overfitting.

After tuning, the model achieved a validation accuracy of 94.93%, which is very strong for this task.
The model's final test accuracy after retraining with the best hyperparameters was approximately 94.74%, indicating good generalization on unseen data.

Cross validation Results:
SVM:
Mean Accuracy: 96.48%
Standard Deviation: 2.13%
SVM performs well, showing high accuracy and low variation across folds. However, in one fold, the accuracy dropped to 92.31%, which is lower than the others. This may indicate some instability in certain parts of the data.

KNN:
Mean Accuracy: 91.87%
Standard Deviation: 3.30%
KNN has lower accuracy compared to SVM, with more variation across folds. The standard deviation is relatively higher, suggesting that KNN's performance is more sensitive to how the data is split.

Logistic Regression:
Mean Accuracy: 95.82%
Standard Deviation: 2.13%
Logistic Regression performs quite well, similar to SVM. The standard deviation indicates stable performance across different folds.

Neural Network:
Mean Accuracy: 87.70%
Standard Deviation: 4.82%
The neural network has the lowest mean accuracy and the highest standard deviation, indicating that its performance is less stable across folds compared to other models. This may suggest overfitting or underfitting, depending on the network's architecture or hyperparameters.

Key Findings:

Best overall model - The Neural Network model emerged as the top performer with the best balance between precision, recall and overall accuracy especially after hyperparameter tuning.
Logistic Regression was also a high performing model with the benefit of being simpler and more interpretable compared to the other models.

SVM had high accuracy but it's recall for malignant cases was lower which makes it less suitable for tasks where false negatives are so important
KNN performed reasonably well but not as good as the other models.

Conclusion:

Neural Networks showed the best overall performance, especially for recall, which is crucial for medical diagnoses like cancer detection. The model achieved an accuracy of 97.66% after fine-tuning.
Logistic Regression offers a strong alternative with high accuracy and interpretability, making it suitable for situations where model transparency is critical.
SVM had high precision but slightly lower recall, which may not be ideal in a medical context.
KNN was less robust but could still be useful as part of an ensemble or with further tuning.