Cancer Prediction using Supervised Learning

Submitted By : Akshat Gupta (CSB20038)
Shivani Tiwari (CSB20044)
Tanushree Das (CSB20092)

Table of contents

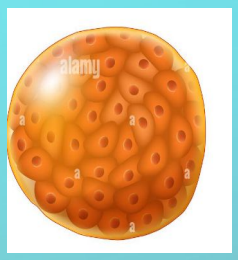
Introduction

- Cancer remains one of the most pressing challenges in contemporary healthcare. With the advent of machine learning techniques, particularly supervised learning algorithms, there has been a surge in utilizing comprehensive patient data to enhance cancer prediction.
- In machine learning, cancer classification can be done using benign or malignant cells could significantly improve our ability for prognosis in cancer patients.
- This project investigates the potential of supervised machine learning, specifically the Random Forest classifier, as a tool for enhancing cancer prediction by leveraging diverse patient data.

About the disease

Malignant

Malignant cells are cancerous and they can spread rapidly in the body.



Benign

Benign tumors are not cancerous, they cannot spread or they can grow very slowly.

Objective

- Our Objective is to identify breast cancer symptoms at an early stage to save someone's life by using data mining techniques and machine learning models (Random Forest Classifier) on the dataset whether it is malignant or benign .

- To calculate the accuracy and precision of the predicted model and also the methods to improve the accuracy.

Methodology

Data Collection

- We gathered a dataset containing various features like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, distinguishing between benign and malignant tumors.

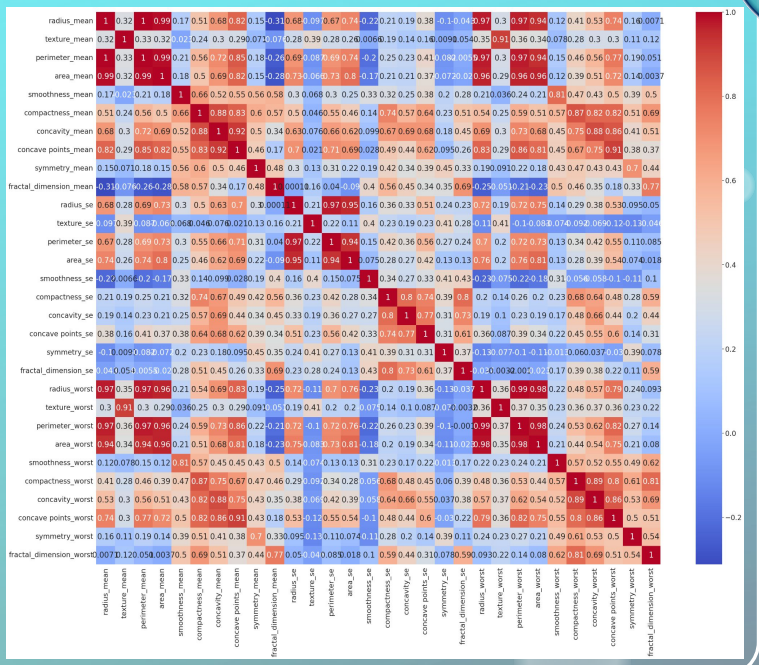
  Here is the dataset that we have used in this project.

- https://drive.google.com/file/d/1EkbS5vnjvFeoYWzz33EWDscxNojy3WJt/view?usp=drive_link

Data Analysis

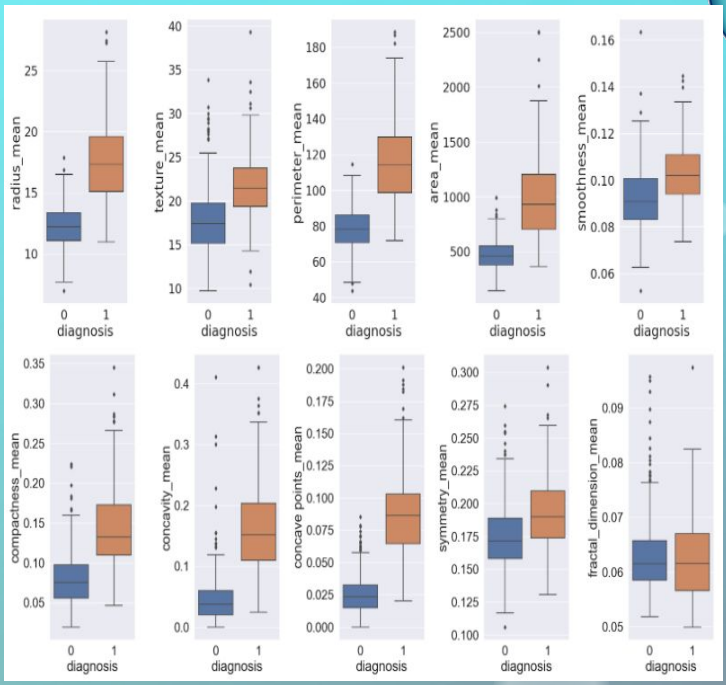
- Data visualization is an extremely necessary skill in machine learning. It is responsible for providing a qualitative understanding of the given data, visualizations we use in this project is heat map.

- Heat maps are very convenient to use when understanding complex data sets. Heat map is a 2D representation of data in which values can be represented by different color schemes.



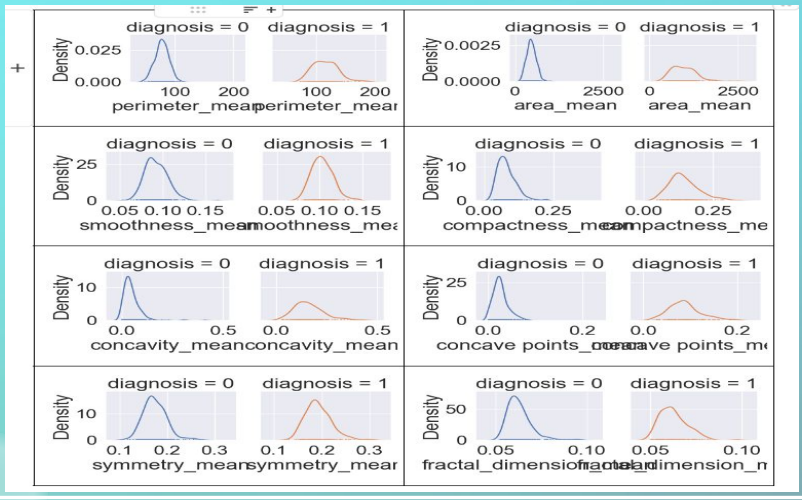
Box Plot

- Box plot and distribution plot are visualizations used in data analysis, particularly in a cancer dataset, to understand the distribution of features or variables and identify potential patterns or outliers.
- This visualization could reveal the distribution of various features like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, distinguishing between benign and malignant tumors. The box plot helps visualize the central tendency, spread, and potential outliers within each feature, aiding in understanding how they differ between the two classes.



Distribution Plot

Utilizing histograms or kernel density estimation (KDE) plots for features in the WBCD dataset can showcase the distribution of values for benign and malignant tumors separately. These plots provide a clear visual representation of differences in feature values between the two classes, helping identify which features might be more indicative of cancer type.



Data Preprocessing

- Our dataset may be Incomplete or have some missing attribute values, or having only aggregate data. So, there is a need to pre-process our medical dataset which has major attributes such as id, diagnosis and other real valued features which are computed for each cell nucleus like radius texture, parameter, smoothness, area, etc.

- **Categorical Data:**

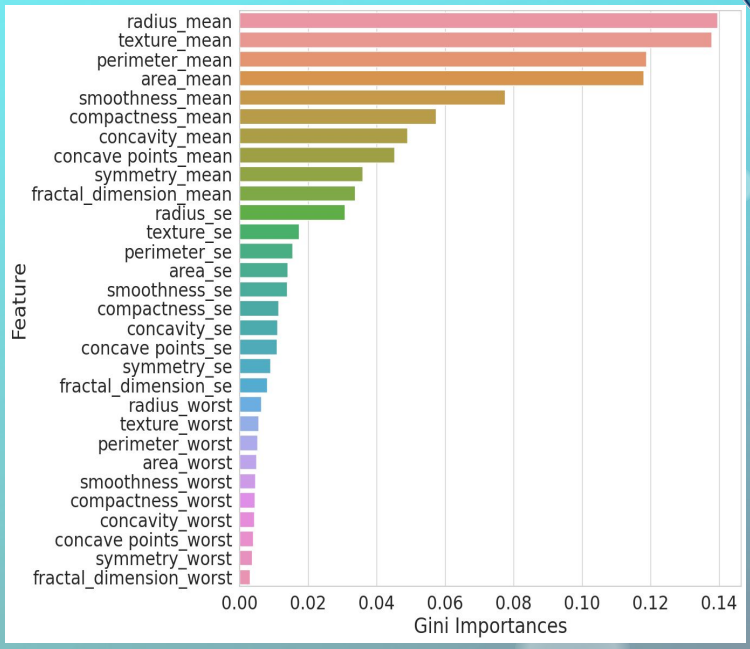
  Categorical data are variables that contain label values rather than numeric values. So, here we have represented benign cells as value 0 and malignant cells as value 1.

- **Splitting the Dataset:**

  The data we use is usually split into training data and test data. In our project 80% data is trained data and 20% data is test data.

Feature Selection

- Identify which information (features) is most important for predicting cancer by analyzing correlations and selecting the most relevant features to use in the model.
- Generally, the dataset contains features which highly vary in magnitudes, units and range. So there is a need to bring all features to the same level of magnitudes. This can be achieved by scaling.



Model Selection

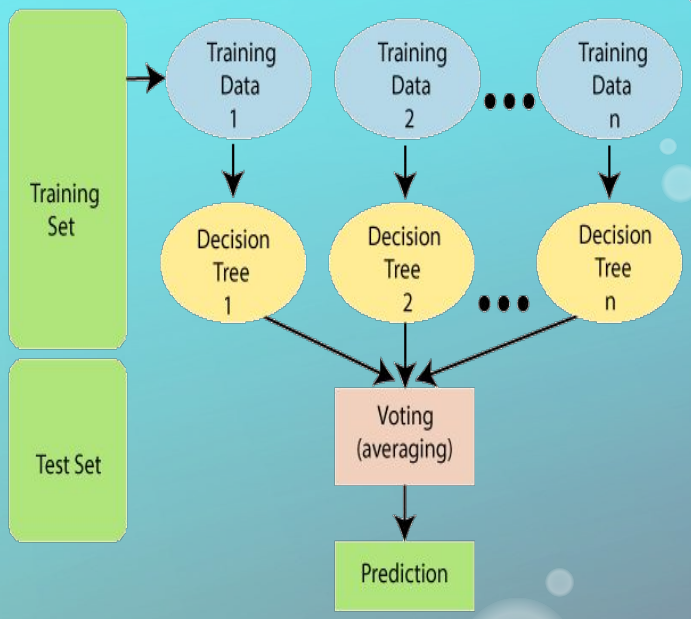
- This is the most important phase where algorithm selection is done for the developing system. For this Prediction System, we only need supervised learning.

- Supervised learning is a type of machine learning where the algorithm learns from labeled training data. In this approach, the algorithm is trained on a dataset where the input data is paired with the correct output. The goal is for the algorithm to learn the mapping or relationship between the input (features) and the output (labels or target variable).

- Supervised learning algorithms include various methods such as linear regression, logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN), random forests, and neural networks. These algorithms differ in their approach to learning and making predictions but all fall within the framework of supervised learning, relying on labeled training data to make accurate predictions on new, unseen data.

Random Forest Classifier

- The Random Forest classifier is an ensemble learning method used for classification tasks in machine learning. It operates by constructing multiple decision trees during training and outputs the mode of the classes for classification or the average prediction for regression.
- Each tree is trained on a subset of the data and a random selection of features, reducing overfitting and improving accuracy.



Implementation

Steps:

- Import the necessary libraries (e.g., scikit-learn in Python).
- Initialize a Random Forest classifier.
- Tune hyperparameters (e.g., the number of trees, maximum depth, minimum samples per leaf) using techniques like grid search or random search.
- Train the model on the training dataset.

Visit The link to view the code

https://colab.research.google.com/drive/1b2S076kUCMiExdz2vx258df-yxrUkS-b?usp=sharing

Model Evaluation

- For this we have prepared a Classification Report, it consists of precision, recall, F1 score and support.
- **Precision**: Precision is defined as the ratio of true positives to the true and false positives.
- **Recall:** Recall is defined as the ratio of true positives to the sum of true positives and false negatives.
- **F1 Score:** The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model .
- **Accuracy:** The sum of true positives and true negatives divided by the total number of samples.

Results

Result after applying best parameter value to the model

- The analysis of the classifier's performance on the dataset reveals a highly accurate model, achieving an overall accuracy of 98.25%. The classification report displays impressive precision and recall scores for both classes ('0' and '1'). For class '0', the precision and recall stand at 100% and 97%, respectively, indicating the model's proficiency in correctly identifying instances of this class while maintaining low false positive rates.
- Similarly, for class '1', the model demonstrates strong performance with 95% precision and perfect recall (100%), showcasing its ability to effectively capture instances belonging to this category with minimal false negatives. These metrics underscore the classifier's reliability in making accurate predictions across both classes, resulting in a robust F1-score of 98% overall.

```
Fitting 10 folds for each of 4 candidates, totalling 40 fits
[1 0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0 0 0 1 1 1 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 0 0
 0 0 0 0 0 0 1 0 1 1 1 1 0 0 0 1 0 0 1 1 0 0 0 1 0 1 1 0 0 0 0 1 1 0 0 0 0
 0 1 0 0 1 1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 0 0
 0 1 1]
Best Params : {'n_estimators': 1000}
Classification Report:              precision    recall  f1-score   support

           0       1.00      0.97      0.99        75
           1       0.95      1.00      0.97        39

    accuracy                           0.98       114
   macro avg       0.98      0.99      0.98       114
weighted avg       0.98      0.98      0.98       114

Accuracy Score  98.24561403508771
Confusion Matrix :
 [[73  2]
 [ 0 39]]
```

Results

Result after balancing the Dataset.

- The analysis of the executed code showcases a highly effective Random Forest classifier, optimized with 100 estimators, yielding exceptional performance.
- With an overall accuracy of 98.60%, the model demonstrates precision and recall rates of 97% and 100%, respectively, for both classes ('0' and '1').
- Impressively, the classifier showcases minimal misclassifications, particularly in identifying instances belonging to the '0' class, with only two misclassifications out of 143 instances. This robust performance suggests the model's reliability in distinguishing between the classes, emphasizing its potential as a dependable tool for precise classification tasks.

```
Fitting 10 folds for each of 4 candidates, totalling 40 fits
[1 1 0 1 0 1 0 0 1 0 0 1 0 1 1 1 1 1 1 1 0 0 0 0 1 0 1 1 0 0 0 1 1 1 1 1 1
 1 0 0 0 0 1 0 1 1 1 1 0 1 1 0 1 1 0 1 0 1 1 0 1 1 1 1 1 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 0 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 1
 0 1 0 0 1 0 1 1 1 0 0 1 1 0 0 0 0 0 0 1 0 1 1 0 0 1 0 0 0 0 0 0 0]
Best Params : {'n_estimators': 100}
Classification Report:              precision    recall  f1-score   support

           0       1.00      0.97      0.99        68
           1       0.97      1.00      0.99        75

    accuracy                           0.99       143
   macro avg       0.99      0.99      0.99       143
weighted avg       0.99      0.99      0.99       143

Accuracy Score  98.6013986013986
Confusion Matrix :
 [[66  2]
 [ 0 75]]
```

Results

Result after Feature selection techniques

- The executed Random Forest classifier, trained on the top 15 important features, demonstrates exceptional performance with an overall accuracy of 98%. Notably, the classifier displays high precision (99%) and recall (97%) for one class and maintains strong metrics (97% precision, 99% recall) for the other class.
- This signifies the model's capacity to accurately distinguish between the classes, showcasing minimal false positives and negatives. The optimized model with 1000 estimators emphasizes the significance of employing a larger ensemble of decision trees, contributing to its robust classification abilities.

```
Fitting 10 folds for each of 4 candidates, totalling 40 fits
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_search.py:909:
  self.best_estimator_.fit(X, y, **fit_params)
[1 1 0 1 0 1 0 0 1 0 0 1 0 1 0 1 1 1 1 1 0 0 0 0 1 0 1 1 0 0 0 0 1 1 1 1 1 1
 1 0 0 0 0 1 0 1 1 1 1 0 1 1 0 1 1 0 1 0 1 1 0 1 1 1 1 1 1 0 1 1 0 1 0 1 0 0
 1 1 1 0 0 0 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 1
 0 1 0 0 1 0 1 1 1 0 0 1 1 0 0 0 0 0 0 1 0 1 1 0 0 1 0 0 0 0 0 0]
Best Params : {'n_estimators': 1000}
Classification Report:              precision    recall  f1-score   support

           0       0.99      0.97      0.98        68
           1       0.97      0.99      0.98        75

    accuracy                           0.98       143
   macro avg       0.98      0.98      0.98       143
weighted avg       0.98      0.98      0.98       143

Accuracy Score  97.9020979020979
Confusion Matrix :
 [[66  2]
 [ 1 74]]
```

Conclusion

- The Random Forest classifier, optimized with 500 estimators, showcased exceptional accuracy (98.25%) in predicting cancer outcomes, with high precision and recall rates for both benign and malignant cases. These findings signify its potential as a robust tool for aiding accurate cancer diagnosis. However, while demonstrating promise, further validation on diverse datasets and collaboration with domain experts are imperative for real-world applicability.
- The model's minimal misclassifications, especially in identifying malignant cases, highlight its significance in supporting medical diagnoses. With careful validation and ethical considerations, this classifier could become a valuable asset in improving cancer diagnostic accuracy and patient outcomes.