MULTIPLE DISEASE PREDICTION

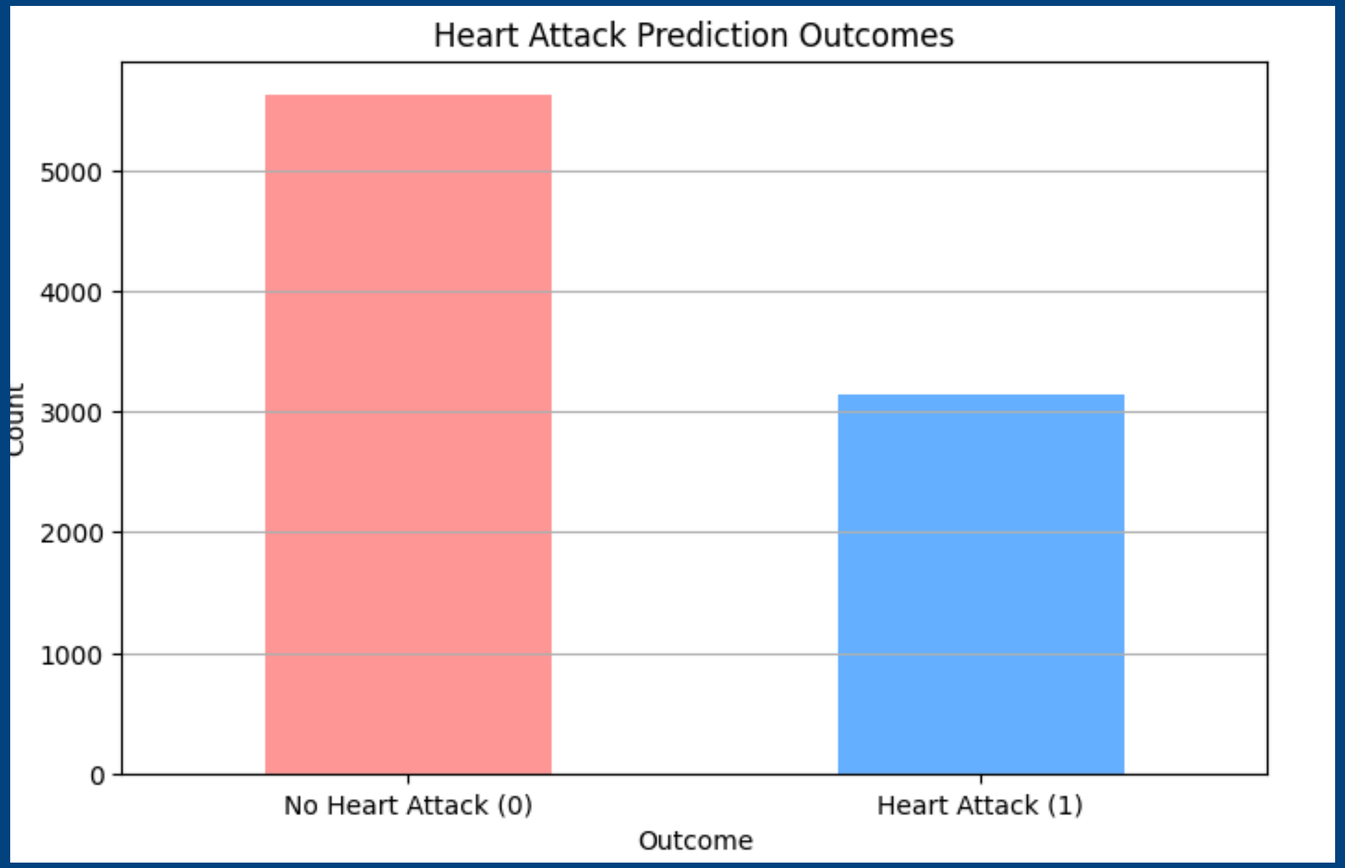
Using Machine Learning

Sakshi Shinde - 324066
Isha Varade - 324071
Pratiksha Naik - 324058

HEART ATTACK PREDICTION

Data Preprocessing & Train-Test Split

- **Data Preprocessing:** Applied techniques like scaling (RobustScaler) to ensure features are on the same scale.
-
- **Model Training:** Split the dataset into training (70%) and testing (30%) sets to evaluate model performance.



$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}}$$

MODELS USED

- Logistic Regression
- Gaussian Naive Bayes (GNB)
- Decision Tree
- Random Forest

EVALUATION METRICS

- **Accuracy** :The percentage of correct predictions overall.
- **Precision:** The accuracy of positive predictions.
- **Recall :**The ability to find all actual positives.
- **F1-Score:**The balance between precision and recall.

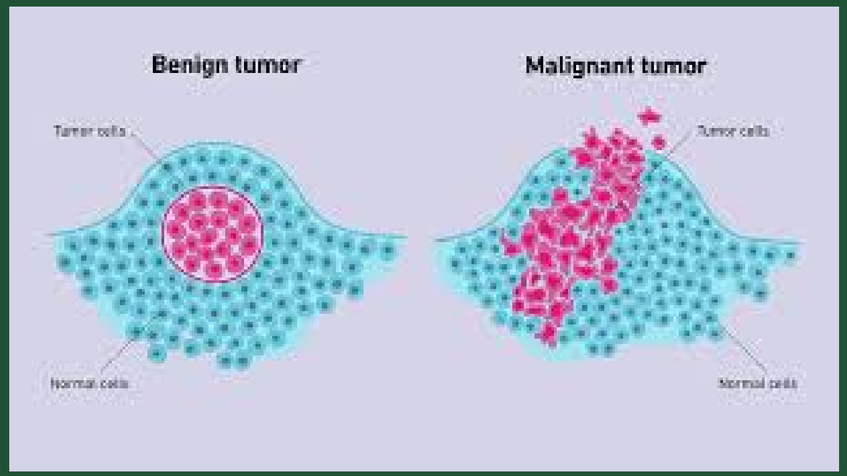
MODEL COMPARISON

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	71.97%	99.86%	42.66%	59.78%
GaussianNB	70.19%	87.94%	45.15%	59.66%
Decision Tree	61.42%	60.15%	62.20%	61.16%
Random Forest	68.83%	78.27%	50.06%	61.07%

CONCLUSION

In summary, Logistic Regression stands out due to its combination of high accuracy and very high precision, making it a suitable choice

BREAST CANCER DETECTION



Benign Tumor:

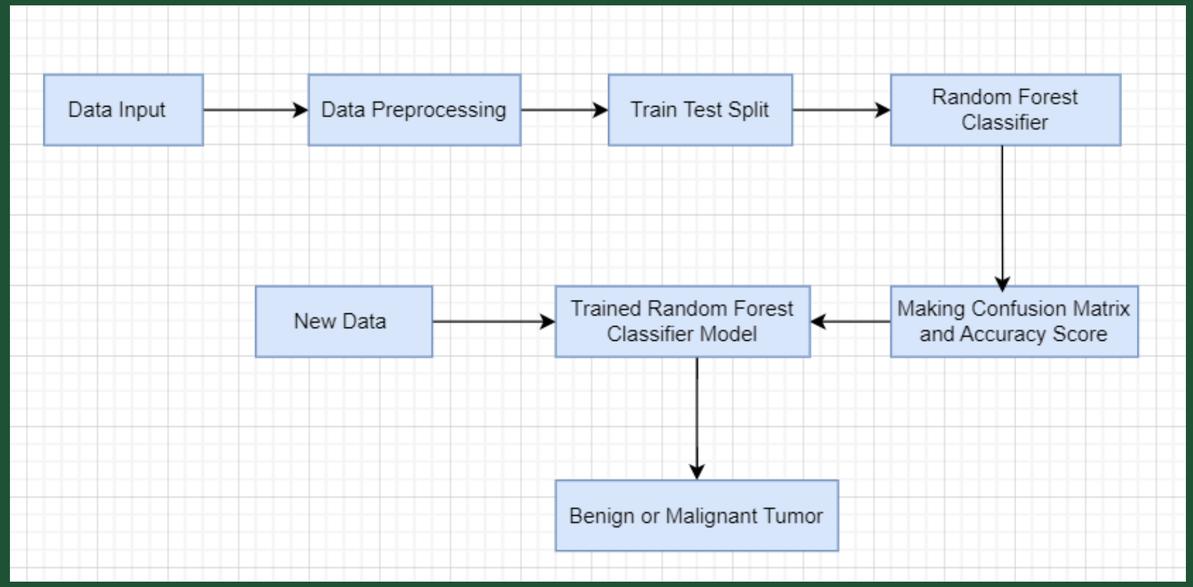
- Nature: Non-cancerous.
- Spread: Does not invade nearby tissues or spread to other parts of the body (non-metastatic).

Malignant Tumor:

- Nature: Cancerous.
- Spread: Can metastasize (spread) to other parts of the body, such as the lymph nodes, lungs, liver, or bones.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	842302	0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
1	842517	0	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
2	84300903	0	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
3	84348301	0	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	84358402	0	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

5 rows × 32 columns



Data Preprocessing:

- Handling Missing Data: Filling in missing values if present.
- Feature Scaling: Scaling features to bring them to a similar range using StandardScaler.
- Label Encoding: Converting categorical labels (binary classifivcation) into numerical format.

Machine Learning Algorithms Used :

1)Logistic Regression:

- Logistic Regression is a linear model used to classify the tumor as malignant or benign based on the features.

2)K-Nearest Neighbors (KNN):

- KNN uses the proximity of a data point to its neighbors to predict the class. K=5 was used in this model.

3)Random Forest Classifier:

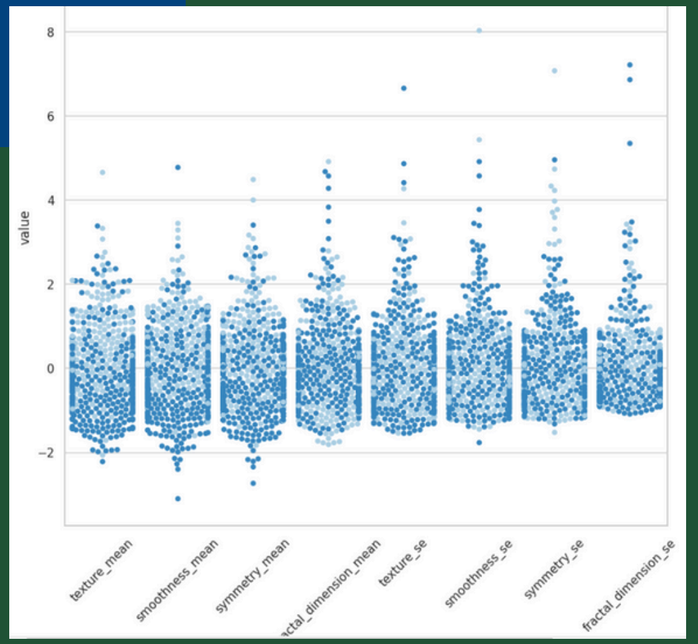
- Random Forest is an ensemble learning method that combines multiple decision trees to improve classification performance.

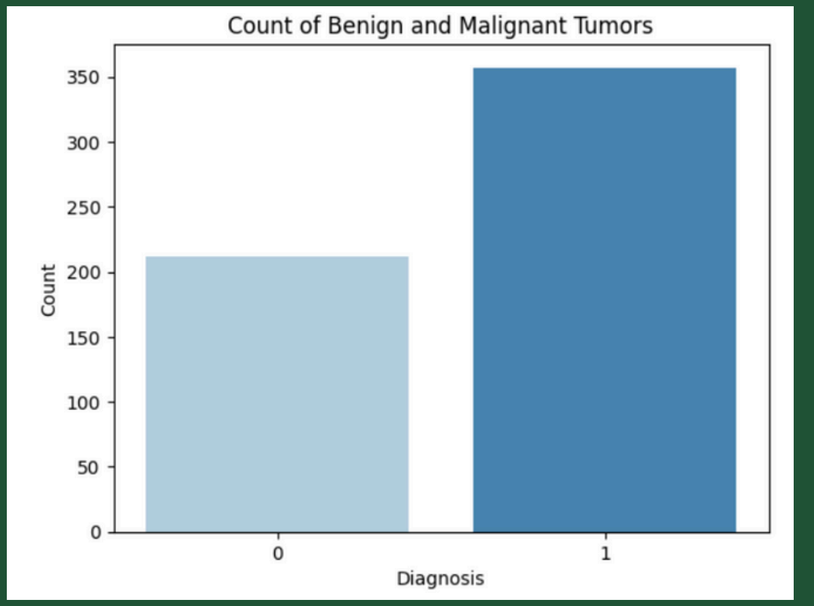
4)Decision Tree Classifier:

- A Decision Tree splits the data into smaller subsets based on feature values, ultimately predicting the class of the tumor.

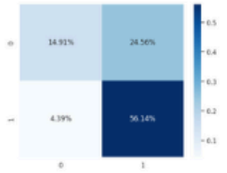
5)Naive Bayes:

- Naive Bayes uses the Bayes theorem for classification, assuming that features are independent of each other.

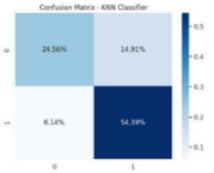




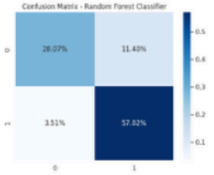
Logistic Regression



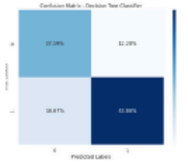
K-Nearest Neighbors



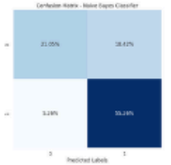
Random Forest Classifier



Decision Tree Classifier



Naive Bayes Classifier



Model	Training Accuracy	Testing Accuracy
Logistic Regression	76%	77%
K-Nearest Neighbors	85%	77%
Random-forest Classifier	100%	80%
Decision-Tree Classifier	100%	77%
Gaussian Naïve Bayes	83%	74%

Accuracy Comparison:

The Random Forest Classifier performed the best on the test data with an accuracy of 80%, followed by KNN and Logistic Regression with 77%. The Decision Tree also achieved 77%, while Naive Bayes had a slightly lower accuracy of 74%.

KIDNEY DISEASE PREDICTION

Kidney Disease refers to conditions that impair kidney function, potentially leading to life-threatening complications like waste buildup, electrolyte imbalances, or kidney failure. Early detection through medical tests is crucial for effective treatment and prevention of further damage.

Data Preprocessing :

- Handling Missing Data
- Feature Scaling
- Encoding Categorical Variables: Select relevant features such as age, blood pressure, and blood tests that are critical to kidney disease diagnosis, reducing irrelevant data to improve model performance.

Machine Learning Algorithms Used :

- **Logistic Regression** is often used for binary classification, helping predict the likelihood of disease presence.
- **Random Forest and Decision Trees** help identify key features and classify patients based on risk factors.
- **K-Nearest Neighbors (KNN)** can be used to classify patients by comparing them to similar historical cases based on their medical features.

THANK YOU

Learn More

@reallygreatsite

www.reallygreatsite.com

+123-456-7890