ODUKOYA ADEWALE

DATA ANALYSIS/ML PROJECT ON:

BREAST CANCER PREDICTION

APTECH FINAL PROJECT ONE(1)

Breast Cancer Prediction Report Using Machine Learning

ABSTRACT

Breast cancer is one of the most deadly cancers in the world. It is important to detect breast cancer at an early stage because it raises the likelihood of a person being healed. This breast cancer predictor also aids in the prevention of the disease's relapse. As a result, I have predominantly concentrated on breast cancer prediction using various machine learning algorithms such as Decision Tree, Logistic Regression, Random Forest Classifier, and SVM. These algorithms are used to train the pre-processed data, enabling the model to be predicted with greater accuracy.

Logistic Regression Classification has the highest accuracy of all models. Python is the programming language used, and we have also imported libraries such as NumPy, pandas, sklearn, matplotlib, and joblib. My primary goal is to find the best machine learning model by doing a comparative study of different machine learning approaches that can have good accuracy when evaluated on a large dataset. Women are at significant risk of high morbidity and mortality due to breast cancer.

The lack of robust prognostic models makes it impossible for the psychiatrist to formulate a rehabilitation strategy which will take years to survive. Thus, it is essential to establish a methodology that gives minimal accuracy error. Future studies can be done to forecast other parameters and research into breast cancer can rely on other parameters categories.

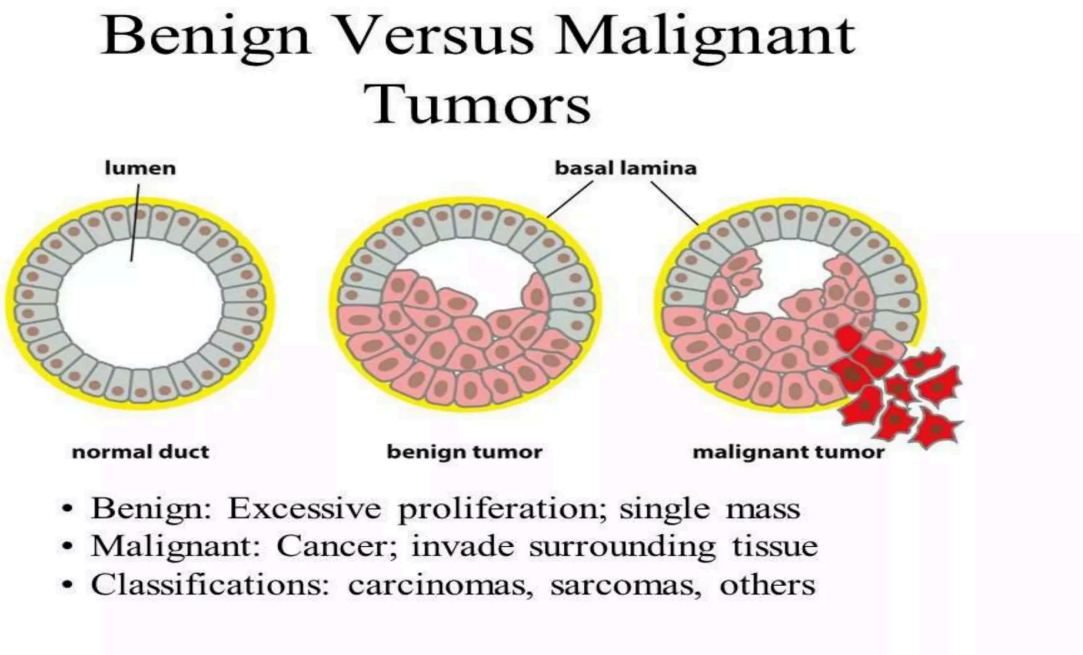
CHAPTER 1

INTRODUCTION

Breast cancer is the most common cancer among women and the second most dangerous cancer after lung cancer, according to the World Health Organization. According to the research, a total of 670,000 women died from breast cancer in 2018, accounting for 20% of all cancer deaths among women. People should see an oncologist if they experience any symptoms. Breast ultrasound, diagnostic mammogram, magnetic resonance imaging (MRI), and biopsies can all help doctors identify breast cancer. Doctors may recommend additional tests or treatment based on the results of these tests. Early detection of breast cancer is crucial. If cancer is predicted at an early stage, the patient's chances of recovery may keep improving. Another method for predicting breast cancer is to use machine learning algorithms to predict abnormal tumours. As a result, research has been undertaken in order to properly diagnose and classify patients into malignant and benign groups.

The three types of tumours are as follows:

a. Benign tumours are not cancerous, cannot spread, and develop slowly. Even if they are removed by surgeons, they cannot damage the human body.
b. Premalignant tumours are not cancerous, but they have the ability to become so.
c. Malignant tumours are cancerous and can grow quickly across the body.



OBJECTIVE

Cancer classification can be performed using benign or malignant cells in machine learning, and while this may greatly boost our ability to predict prognosis in cancer patients, no progress has been made in their use in clinics. However, before gene expression profiling can be used in medical care, wider data sets and more thorough evaluation are needed. As a result, our goal is to create a predictive framework that can determine the occurrence of breast cancer based on large data sets.

OVERVIEW OF THE PROJECT

A number of classification algorithms are available and breast cancer results are predicted. The current study describes a contrast among SVM, Random Forest Classifier, Decision Tree, KNN, Logistic Regression which are among the most influential Machine Learning algorithms in the research community. Our aim is to test the performance, sensitivity, specificity and precision of all these algorithms.

CHAPTER 2

ANALYSIS AND DESIGN

DATA DESCRIPTION

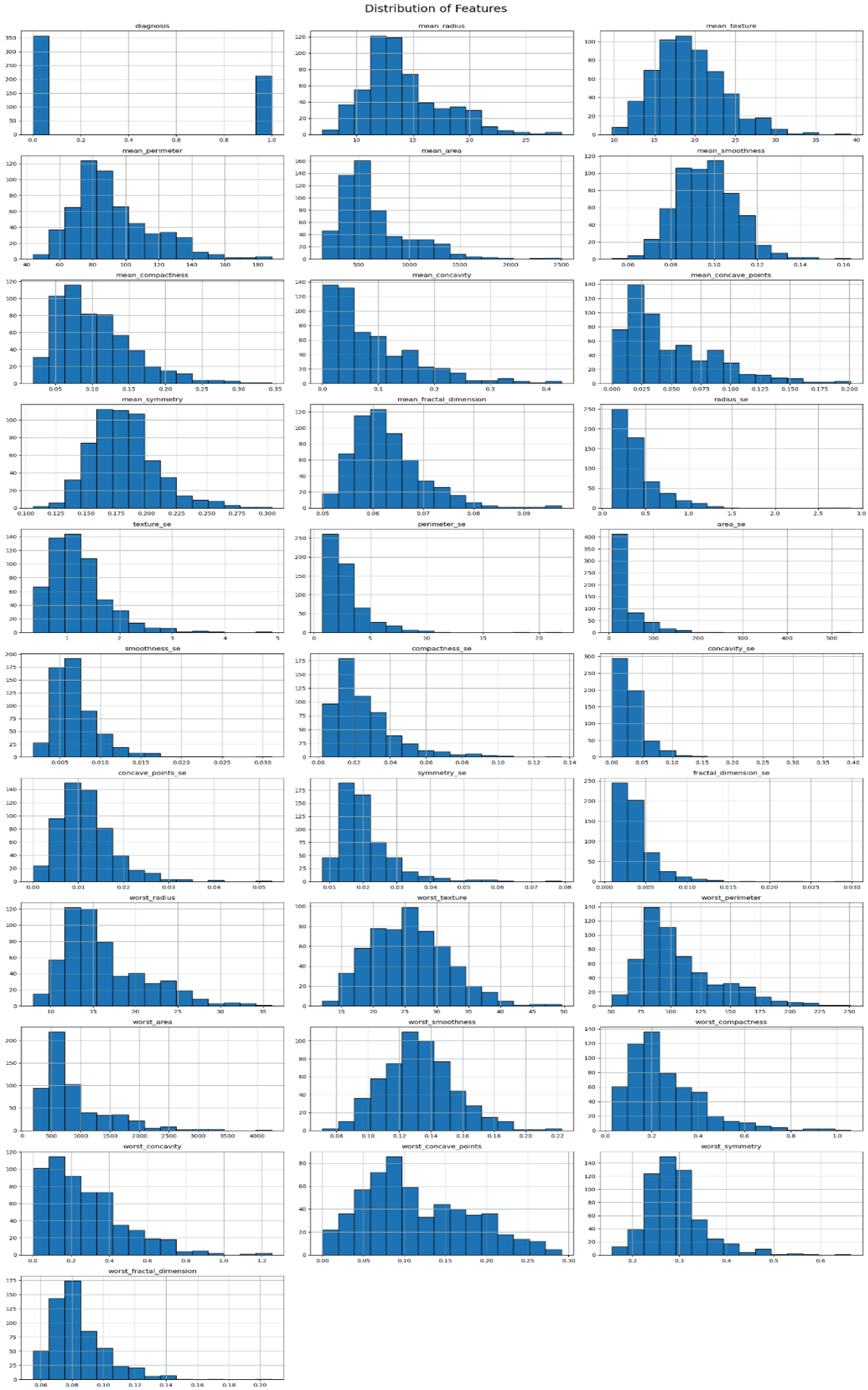
The breast cancer data and its description was taken from the UCI machine learning repository. This data has 569 instances and 32 attributes. And this data notably has no missing values. There is a target (diagnosis) attribute which contains cases where there are either malignant or benign. The class distribution for the following data is given as 357 (62.75%) are benign and 212 (37.25%) are malignant. The benign cases are classified as a negative class (Class 0), and the malignant cases are classified as a positive class (Class 1).

IMPORTING LIBRARIES AND THE DATASET

Firstly, I have imported the libraries and the breast cancer data. Then, I downloaded the data from UCI, not the dataset. Therefore, I have converted the data to a csv file using the data.to_csv function. Now that I have the dataset, I can use it for further EDA and modelling data.

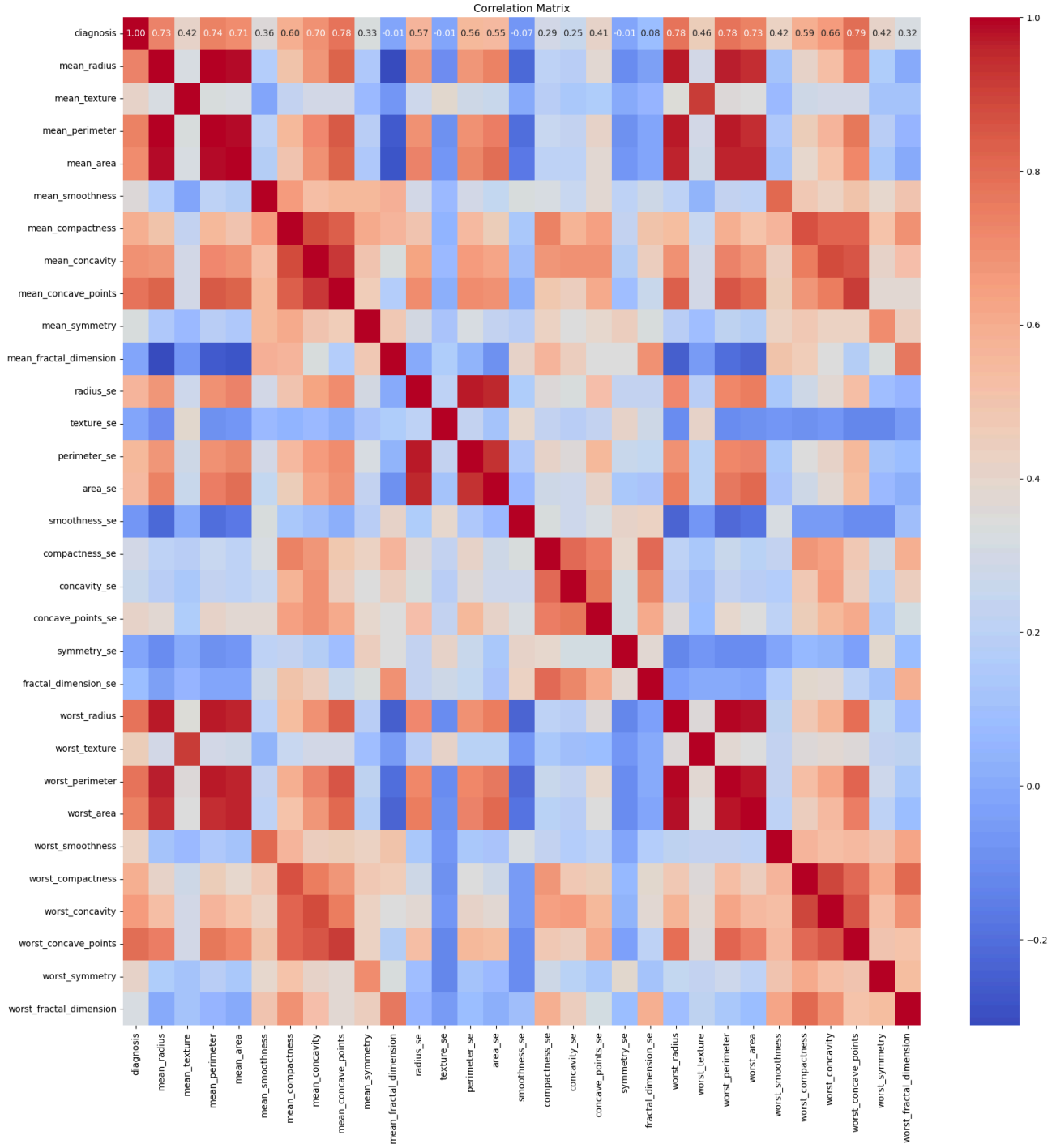
DATA VISUALIZATION

Data visualisation is an important component of data science. It aids in the comprehension of data as well as the explanation of data to others. The visualisation of each value from the dataset as shown below:


Distribution of Features
diagnosis
mean_radius
mean_texture
mean_perimeter
mean_area
mean_smoothness
mean_compactness
mean_concavity
mean_concave_points
mean_symmetry
mean_fractal_dimension
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave_points_se
symmetry_se
fractal_dimension_se
worst_radius
worst_texture
worst_perimeter
worst_area
worst_smoothness
worst_compactness
worst_concavity
worst_concave_points
worst_symmetry
worst_fractal_dimension

VISUALIZING CORRELATION OF ALL THE FEATURES USING HEATMAP

Correlation is a metric that determines how often two variables pass in relation to one another. We will be visualising the correlation between all of the features in this part.



DROPPING COLUMNS WITH LOW CORRELATION TO TARGET ('diagnosis')

After setting a benchmark of 0.49 from the correlation matrix, I dropped all features with correlation less than 0.49 according to matrix leaving me with 16 features to work with for further EDA and modelling data.

CHAPTER 3

IMPLEMENTATION

FEATURE AND LABELS

Many of the columns of data that we include as input to a model are referred to as functions. Label, on the other hand, is the intended output column that will be given to the model. We can make predictions using features, and we can classify raw data using labels (target). In this case, we'll use all 15 attributes as features and the diagnosis as the label.

```
X = df1.drop(columns=['diagnosis'], axis = 1)
y = df1['diagnosis']
```

SPLITTING THE DATASET FOR TESTING

The training phase extracts features from the dataset, and the testing phase determines how practical the model performs for prediction. The dataset is split into two sections: training and testing. We have trained 80% of the data and tested 20% of the data.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```

TRAINING THE MODEL

The given dataset has been pre-processed and visualised earlier. Now we will train our data using the following machine learning techniques:

Classifiers used: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and K-Nearest Neighbors.

```python
classifiers = {
    'Logistic Regression': LogisticRegression(),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'Random Forest': RandomForestClassifier(random_state=42),
    'Gradient Boosting': GradientBoostingClassifier(random_state=42),
    'Support Vector Machine': SVC(random_state=42),
    'K-Nearest Neighbors': KNeighborsClassifier()
}

# evaluation metrics
metrics = {
    'Accuracy': accuracy_score,
    'Precision': precision_score,
    'Recall': recall_score,
    'F1 Score': f1_score,
    'ROC AUC': roc_auc_score
}

results = []

# Train and evaluate each classifier using the pipeline
for clasf_name, clasf in classifiers.items():
    pipeline = Pipeline(steps=[
        ('scaler', StandardScaler()),
        ('classifier', clasf)
    ])

    # Fitting the models
    pipeline.fit(X_train, y_train)

    # Predict on test dataset
    y_pred = pipeline.predict(X_test)

    # Calculate evaluation metrics
    scores = {}
    for metric_name, metric_func in metrics.items():
        scores[metric_name] = metric_func(y_test, y_pred)

    # Append results as a dictionary to the list
    results.append({
        'Model': clasf_name,
        'Accuracy': scores['Accuracy'],
        'Precision': scores['Precision'],
        'Recall': scores['Recall'],
        'F1 Score': scores['F1 Score'],
        'ROC AUC': scores['ROC AUC']
    })

## puts output into easily readable format
results_df = pd.DataFrame(results)

print(results_df)
```

CHAPTER 4

TEST RESULTS

TESTING

We used six machine learning techniques, namely Logistic Regression, Decision Tree, Random Forest Classification, Gradient Boosting, SVM and K-Nearest Neighbors, and to predict whether a cell is benign or malignant. Scikit-learn, an open-source machine learning library written in Python, is used. For each technique, the confusion matrix is computed. We used 455 of the 569 instances in the dataset to train with all six strategies, accounting for 80% of the total data. We studied Logistic Regression Classification with a maximum precision of 96.49% over all other machine teaching techniques, while SVM, Gradient Boosting, Knn, Random Forest and Decision Tree obtained 95.7% of the next highest exact accuracy. In short, theLogistic Regression Classification has shown its reliability and efficiency in terms of precision and recall. Following are the output of the algorithms:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
0	Logistic Regression	0.964912	0.953488	0.953488	0.953488	0.962660
1	Decision Tree	0.938596	0.950000	0.883721	0.915663	0.927776
2	Random Forest	0.956140	0.952381	0.930233	0.941176	0.951032
3	Gradient Boosting	0.956140	0.952381	0.930233	0.941176	0.951032
4	Support Vector Machine	0.956140	0.975000	0.906977	0.939759	0.946446
5	K-Nearest Neighbors	0.964912	0.975610	0.930233	0.952381	0.958074

DEPLOYMENT

Deployment is a process by which a machine learning algorithm is integrated into an existing production system to make realistic data-based business decisions. It is one of the last steps in the life cycle of the system and can be one of the most difficult.

LOADING THE SAVED PICKLE FILES

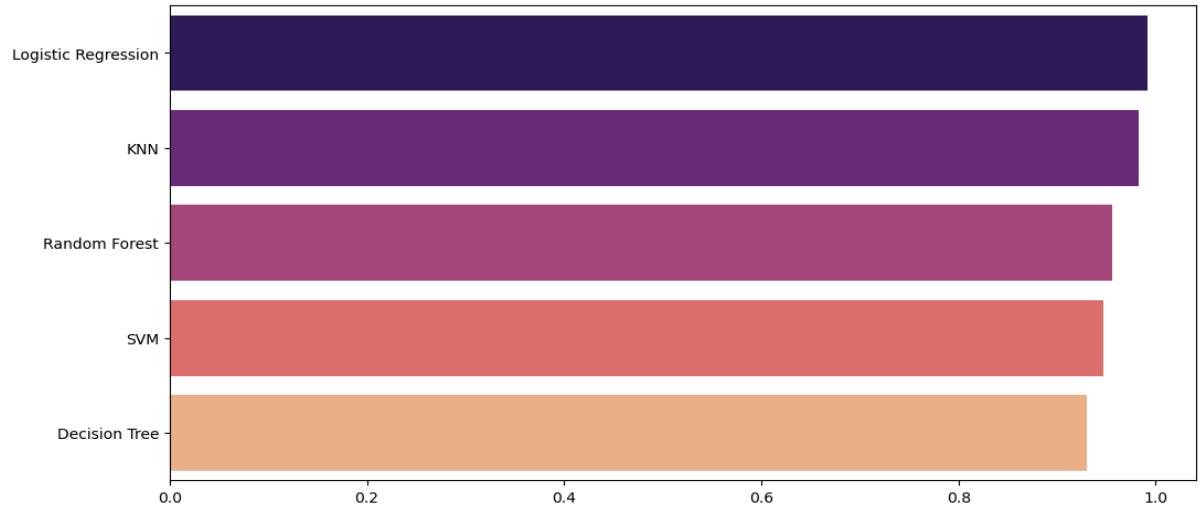
Firstly, we need to load the pickle file which we have saved during the training process. This step is very important as a pickle file is a trained model of a corresponding algorithm which can be used for further calculation

DESIGNING THE WEB APP

Here, we need to design the webpage to make it look even more attractive and catchy. So I have used html and added some background colour, padding and text alignment. I would've added a slide bar option where one can choose a specific model from the list of models specified(but that'll be pursued personally). It would have helped us to check various models for sample input at the same time. Also I have created some number input functions which helps in selecting the input.

COMPARATIVE ANALYSIS

We used four machine learning techniques, namely Decision Tree, Random Forest Classification, SVM, and Logistic Regression, to predict whether a cell is benign or malignant. We studied Random Forest Classification with a maximum accuracy of 98.57% over all other machine teaching techniques, while SVM, Logistic Regression, and Decision Tree obtained 95.7% of the next highest exact accuracy. In short, the Random Forest Classification has shown its reliability and efficiency in terms of precision and recall. We compared the results achieved in this study.



The image above gives us a perfect visualisation of the accuracies of all the algorithms; Support Vector Machine, Logistic Regression, Decision Tree, K-nearest Neighbor and Random Forest Classifier. Logistic Regression

Classifier gives the maximum accuracy and overtakes all other algorithms.

CHAPTER 5

CONCLUSION

This project reviews various models and calculates their accuracy and compares them to allow doctors to use the best model for cancer detection, which is comparatively quicker in real life than previous approaches when it comes to providing a diagnosis on breast tumours/cancers. The preceding research suggested that the Logistic Regression Classification algorithm is more proficient and efficient for identifying breast cancer than the Decision Tree, SVM, and Random Forest Regression algorithms. Various machine learning approaches are used for analysing medical evidence.

Building effective and computationally reliable classifiers for medical applications is a significant challenge in machine learning. On the Wisconsin Breast Cancer (original) dataset, primarily I used five algorithms: SVM, LR, RF, K-nearest and DT. To find the best classification accuracy,then compare the performance and efficacy of those algorithms in terms of accuracy, precision, Recall, and F1 Score. The precision of LR outperforms all other algorithms. A breast cancer predictive system will allow the doctor to determine optimally, correctly and helps to minimise the potential costs of healthcare. For the implementation of the model, different classifiers were used but the Logistic Regression Classifier has been found to provide the best accuracy in classification when used for the most predictive variables.

The proposed method significantly reduces treatment costs and increases the standard of living through early prediction of breast cancer. More information on datasets and more important results will be investigated in future work. It will help make disease prediction and diagnostic systems more efficient and accurate, helping to build a healthier healthcare system by lowering costs, duration and mortality rate. To conclude, Logistic Regression Classifier has demonstrated its prediction and evaluation efficiency in breast cancer and achieves maximum results with regard to accuracy and low error rate.