Machine learning midterm report: -

Aim(abstract): -

The project aims to create machine learning models using information taken from digitalized fine needle aspirate (FNA) pictures to predict whether breast cancer is benign or malignant and will choose the best ML model for this dataset. Based on the given feature set, our results demonstrate how well the machine learning models were able to predict the malignancy of breast cancer. Over the years, it has been shown that patients with similar cancer types exhibit hidden patterns. Deep learning networks have also been shown to be particularly good at modelling patterns that are not immediately obvious. Thus, the project's goal is to accurately classify images of breast lesions as malignant, benign and normal.

Introduction: -

Any nationality might be affected by the cancer epidemic. Since cancer is such a widespread issue, many individuals have either personally experienced cancer themselves or know someone who has. Moreover, the accurate diagnosis of cancer requires time and a qualified specialist who can misdiagnose the patient or not arrive in time to assist. The individual's life expectancy and quality are directly impacted by all of this. Because therapy may begin sooner and the prognosis for the sickness is better, the earlier a malignant tumour is discovered. Machine learning is especially well-suited to assist in this area.

The UCI Machine Learning Repository provided the dataset, which consists of 10 real-valued features (such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension) that are computed for each cell nucleus. Each image has 30 features, and these features were calculated using the mean, standard error, and "worst" values. I aim to differentiate between benign and malignant instances within the 3-dimensional space specified by these characteristics, as outlined in Bennett and Mangasarian's work. There are 212 cases of malignancy and 357 benign cases in the sample.

The dataset used for this project is taken from the UCI machine learning repository and can be accessed via Kaggle.

Dataset: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data

UCI: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Background: -

KNeighborsClassifier: -
A machine learning method called KNeighborsClassifier is used for classification tasks, such determining if breast cancer is benign or malignant. It works on the tenet that cases with comparable labels in a multidimensional feature space should exist. Using a majority voting method, the algorithm determines a new instance's k-nearest

neighbours and forecasts its label. The number of neighbours that are taken into account is set by the n_neighbors argument. The method can capture intricate, non-linear decision boundaries and is adaptable and non-parametric. Its performance depends on careful evaluation of distance measures and proper feature scaling. The KNeighborsClassifier is useful in image recognition, medical diagnosis, and other fields since it uses instance similarity to make precise predictions.

Linear regression: -
By fitting a linear equation to observed data, linear regression attempts to describe the connection between a dependent variable (target) and one or more independent variables (features). In this instance, a quantitative indicator of the breast cancer's severity can be the dependent variable. This might be a continuous number related to the disease's severity or course. The different traits or measures linked to cancer cells that might affect the severity of breast cancer would be considered independent variables. These characteristics may consist of mitotic rate, cell size, shape, or other pertinent information gleaned via diagnostic testing.

To help researchers and clinicians understand how these features contribute to the overall severity of the disease, linear regression is used in the context of breast cancer prediction. It attempts to establish a linear relationship between measurable features of cancer cells and a quantitative measure of cancer severity.

Bayesian classification: -
A labelled dataset with examples of both benign and malignant cases is used to train the algorithm. It picks up on the conditional odds of seeing particular attributes that are provided in every class. Naive Bayes uses the probabilities it learned during training to determine the likelihood that a new instance with a given collection of attributes will belong to the benign or malignant class. The outcome that is anticipated is the class with the highest probability. The features in the context of breast cancer may include different traits of cells seen during diagnostic examinations. To predict whether a tumour is likely to be benign or malignant, the Naive Bayes method uses the probability of these characteristics given the class labels.

Logistic regression: -
Logistic regression is a better option for binary classification, such as the prediction of benign or malignant conditions. Based on linear combinations of input characteristics, Logistic Regression estimates the likelihood that a given instance falls into a specific class (e.g., benign or malignant). After that, the linear combination is transformed into a number between 0 and 1, which represents the likelihood, using the logistic function (sigmoid function). Next, a threshold (usually 0.5) is established, and cases with projected probabilities higher than this threshold are assigned to one class, while those with lower probabilities are assigned to the other.

Because they can produce distinct binary outcomes, logistic regression and other classification models are frequently used in the medical field for tasks like cancer prediction.

Methodology: -

For methodology, I used: -

Training test splitting: -

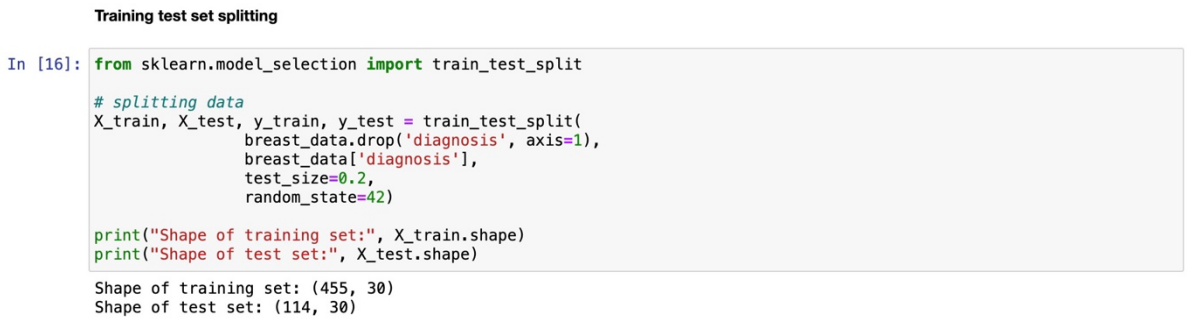

Fig1: Training test splitting

The machine learning model is trained on one part of the dataset, and its performance is tested on the other. By doing this, it is made sure the model is assessed on data that it did not observe during training, which results in a more accurate evaluation of its capacity for prediction.

- The training set (X_train and y_train) is used to train the machine learning model to learn the patterns and relationships within the data.

- The test set (X_test and y_test) is then used to evaluate the model's performance by making predictions on unseen data. The difference between the predicted labels and the actual labels in the test set indicates how well the model generalizes to new, unseen instances.

By splitting up the data, the machine learning model is better equipped to predict outcomes accurately in new and untested scenarios. The model is trained on the training set, and its performance and capacity for generalisation are evaluated on the test set.

Data scaling: -

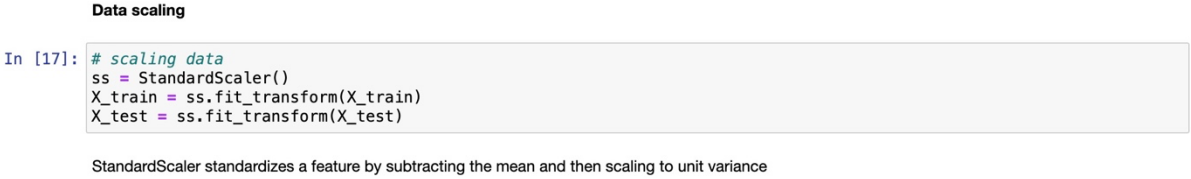

Fig2: Data scaling

A critical pre-processing step that improves the efficacy and dependability of machine learning models is scaling the data using StandardScaler. This is especially true for tasks like determining if breast cancer is benign or malignant. It guarantees that the model can generalise effectively to a variety of datasets, improves performance, and helps to consistent model behaviour.

Classifier evaluation metrics: -

```
In [194]: print(confusion_matrix(y_test, prediction1))
          print("\n")
          print(classification_report(y_test, prediction1))
```

```
[[70  1]
 [ 4 39]]


              precision    recall  f1-score   support

           B       0.95      0.99      0.97        71
           M       0.97      0.91      0.94        43

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Fig3: Classifier evaluation metrics

The classification report, one may learn more about the model's advantages and disadvantages in terms of foretelling benign and malignant situations. It offers thorough insights into how effectively a machine learning model works, particularly in terms of predicting cases of breast cancer that are benign or malignant. In the medical profession, where timely and precise interventions depend on accurately diagnosing malignant situations, it serves as a decision-making aid.

Cross-validation: -

```
In [89]: from sklearn.model_selection import cross_val_score

         LogisticRegression_cross_val = cross_val_score(LogisticRegression(),X_train,y_train)
         print("Cross validation score of Logistic Regression Model:")

         count = 0
         for i in LogisticRegression_cross_val:
             count+=1
             print(f"{count}) {round(i*100, ndigits = 2)} %")
```

```
Cross validation score of Logistic Regression Model:
1) 97.8 %
2) 96.7 %
3) 100.0 %
4) 97.8 %
5) 94.51 %
```

Fig4: Cross-validation

Cross-validation gives a more accurate estimation of the model's performance and makes sure it is not overfitting to a particular train-test split. This is critical for healthcare applications since the model's generalisation capacity and reliability are critical for producing precise predictions on new, unseen patient data.

Results: -

The aim was to find the best ML model for this dataset and we did so using the results from our 4 machine learning models.

The best ML model for this dataset will be determined by each models accuracy.

```
Accuracy of linear regression model: 95
Accuracy of naive bayes model: 95
Accuracy of the Logistic Regression Model is: 98
Accuracy of K Neighbors Classifier Model is:  95
```

Hence, the best model for this dataset is logistic regression models.

Fig5: Results

Evaluation: -

A bigger dataset with more detailed findings would also contribute to the accuracy of the prediction models. Since I don't have access to a better dataset, there isn't much I can do right now to enhance the model. If the dataset is larger and consists of more details this method can be done with different programming languages, using other ML libraries. We could improve by exploring additional features and considering more advanced deep learning architectures. Nevertheless, by demonstrating the efficacy of logistic regression in predicting breast cancer malignancy based on FNA pictures, our effort significantly advances the field.

Conclusions: -

It became evident that the logistic regression model emerged as the most effective and reliable solution for this dataset. The model demonstrated superior accuracy and precision in predicting whether breast lesions were malignant or benign, aligning well with the project's overarching goal.

The utilization of digitalized FNA images provided valuable insights into hidden patterns within breast cancer types. While deep learning networks were considered for their ability to capture intricate patterns, our findings highlight the success of logistic regression in this specific context.

The results underscore the importance of model selection in medical image classification tasks. Logistic regression proved to be a robust choice for distinguishing between malignant and benign cases of breast cancer. This model's accuracy and precision are important factors that make it appropriate for use in practical healthcare applications.