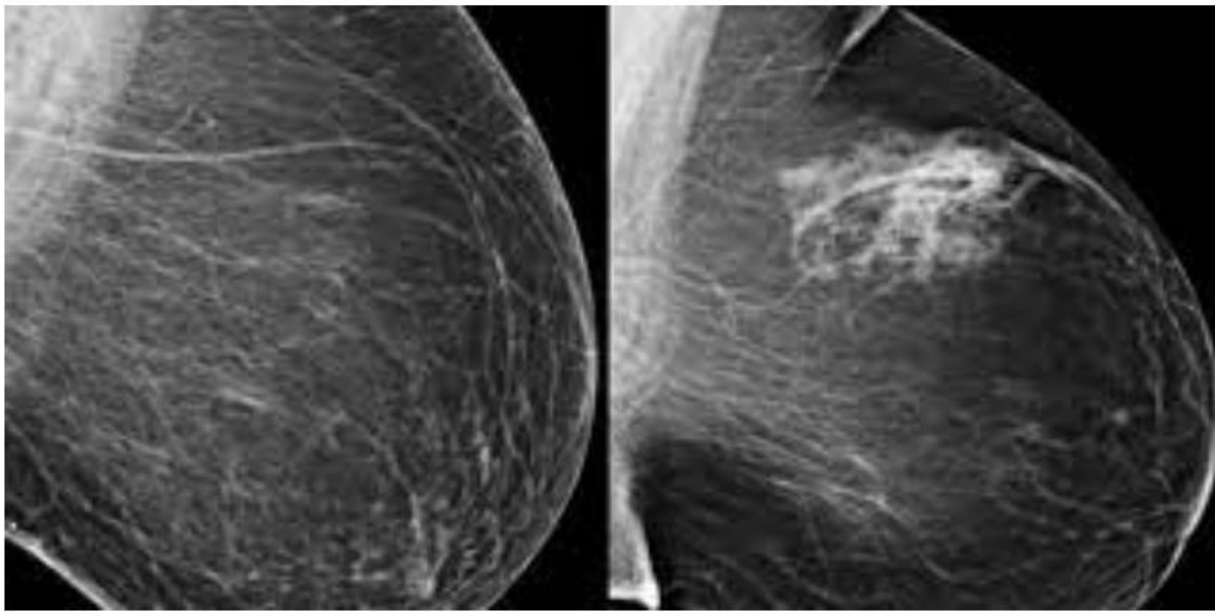


1 -) Research Data

This section of the report presents a comprehensive review of the collection process, reliability and intended use of the "Breast Cancer Wisconsin (Original)" dataset, which is frequently used in classification studies.

Collection of data set

This data set was collected for use in cancer research. From January 1989 to July 1992, the data set was collected periodically by Dr Wolberg. The data set has a total of 699 samples and consists of different cases from different dates. In addition, the data consists of clinical cases coming to Dr Wolberg. The purpose of using this dataset in our project is to find out whether this tumour is benign or malignant according to the characteristics of the given cells. Benign tumours respond to treatment and do not spread to other parts of the body. Malignant tumours, on the contrary, spread throughout the body and sometimes do not respond to treatment.

Reliability of the Data Set and Intended Use in Other Articles

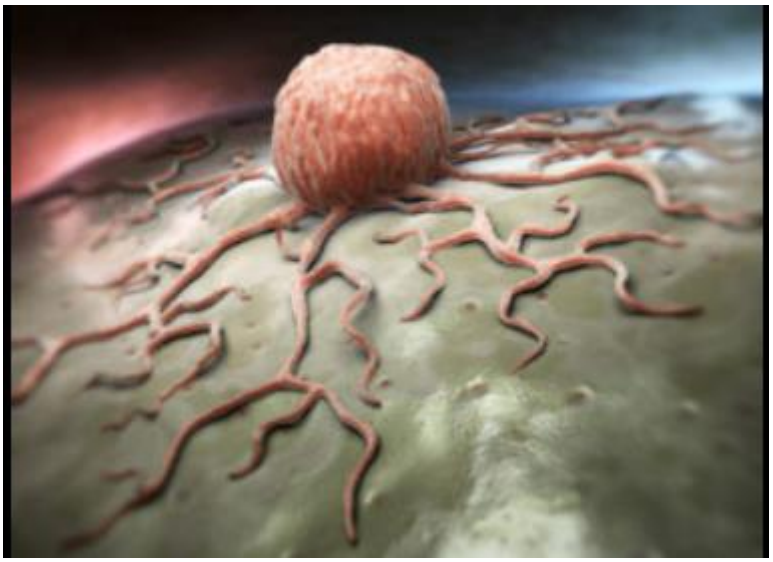
The Breast Cancer Wisconsin (Original) dataset is an important resource used in breast cancer research and diagnosis. The reliability of the dataset is based on many years of solid history and a comprehensive data collection process. In particular, it contains real clinical data from studies on breast cancer screening and diagnosis. Therefore, it can be said that this dataset is reliable.

In the fields of machine learning and data analysis, this dataset can be used as a testbed for the development and testing of classification algorithms. In particular, it can be used to evaluate the performance of machine learning models used to accurately classify breast cancer types.

In papers, this dataset has often been used as a tool to study the effect of different characteristics used in breast cancer diagnosis. For example, the impact of various clinical features such as tumour size, shape, rate of cell division on the risk of cancer spread has been investigated. The dataset has also been used to compare the performance of different classification and prediction models used in breast cancer diagnosis.

2 -) Examining Data (EDA)

Now, first of all, since I do not know how breast cancer cells are no different from normal cells, I did research about it on the internet. Let's examine the features in our data set given one by one.



Clump Thickness: This refers to the thickness of the given tumour groups. If it is high, it is more likely to be malignant.

Uniformity of Cell Size: This examines the consistency of the size of the given cells. If the cells are more different from each other, the cell group is more likely to be malignant.

Uniformity of Cell Shape: This is when the shape of the given cells does not resemble each other. This is more likely to occur in malignant tumours.

Marginal Adhesion: This is the ability of cells to adhere. This is high in healthy cells, but malignant cells tend to adhere less.

Single Epithelial Cell Size: Single epithelial cell size. If there is a very enlarged cell, it is likely to be a malignant cell**.**

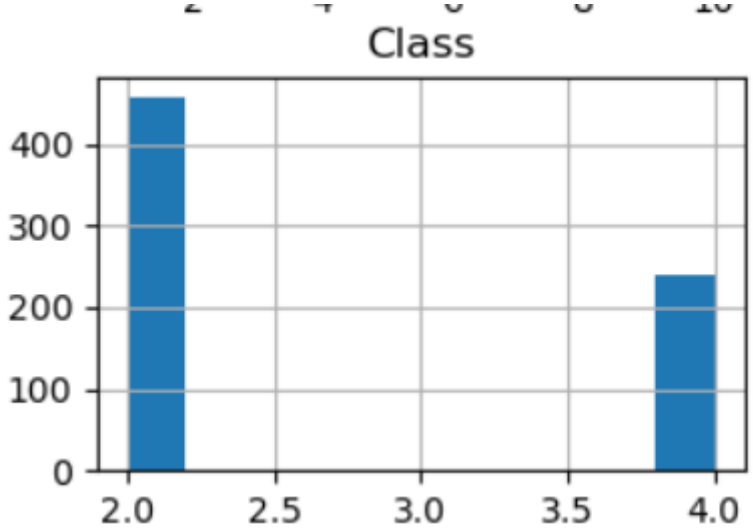
Bare Nuclei: This feature indicates the presence or absence of bare nuclei of cells. The presence of bare nuclei is associated with the aggressiveness of cancer cells and their tendency to divide rapidly

Bland Chromatin: It is the score whether the chromatin is homogenous. Benign tumours are generally more homogeneous.

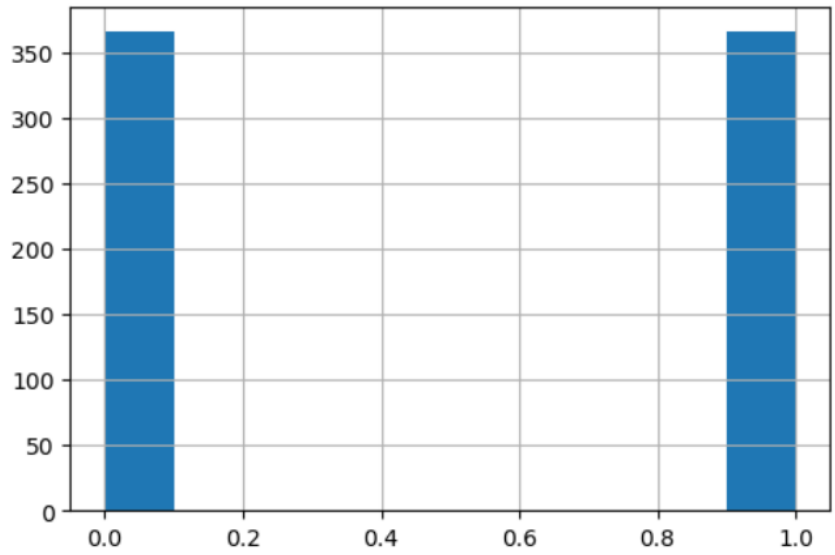
Normal Nucleoli: Nucleoli are very small in normal cells but become very large in malignant cells.

Mitoses The number of mitoses the cell undergoes. The more it is, the more likely it is to be malignant.

Firstly, I analysed the data and converted the Bare_Nuclei column, which is of object data type, to float type. Then I looked at how many null data there were. I preferred to fill them with the average instead of discarding them. Because there was already little data and I didn't want to lose information. Also, by experimenting later, I saw that my model performance decreased when I discarded the data. I examined the correlation between them and decided to discard the '**Uniformity_of_cell_size**' column since '**Uniformity_of_cell_size' and 'Uniformity_of_cell_shape'** are very compatible with each other. In this way, I reduced the complexity of the model and prevented it from memorising the data a bit more. In addition, I decided to discard the '**Mitoses'** column because it was not a very high predictor for the type of tumour. Finally, I removed the 'Single_epithelial_cell_size' column because it was highly correlated with 'Marginal_adhesion'.



In the data, the Class column was distributed as 2 and 4, and I updated it as 1 and 0. I saw that the 0s and 1s were unevenly distributed in the data, so I wanted to equalise them. The problem was that when I multiplied the fewer 1s, my models started to memorise the data and overfitting occurred. Therefore, I tried the method of decreasing the 0's and increasing the 1's. This way I think my models have better generalisation abilities. If I had not equalised the 0's and 1's, my models would have been more biased towards the excess.



3 -) Models I added and Model Interpretation

Model	Accuracy	Precision	Recall	ROC AUC Score
XGBoost Classification	0.967213	0.957894	0.978494	0.989366
Logistic Regression	0.967213	0.967741	0.967741	0.991636
Random Forest Classification	0.972677	0.968085	0.978494	0.992532
Support Vector Classification	0.972677	0.95834	0.989247	0.978494
Neural Network Classification	0.961748	0.957446	0.967741	0.986021

I added classifiers such as kNN and Naive Bayes to the model. Of course they did not perform the best, but I would say that they perform not badly. Now I will quickly summarise all the models and draw some conclusions.

The best performance was given by the XGB and Random Forest models. They basically use decision trees and perform quite well. The difference is that Random Forest uses decision trees in a randomised way, while XGB selects increasingly

relevant features. The reason why decision trees are a good model is that they can separate independent features well and can easily identify malignant tumour cells.

Secondly, Logistic Regression performed well. The reason for this result is that all the features we give to the model can be linearly related to the outcome. For example, malignant tumour cells are usually large and misshapen. Therefore, the larger the values represented by these features, the higher the probability of malignancy.

This is followed by Neural Network. These models generally perform similarly, but better performance can be achieved if certain parameters are more optimised. I think it is a bit complex model for this data set.

Support Vector Classification performed the worst. This is probably due to the lack of outliers in the features.

The kNN model classifies according to the majority of its nearest neighbours. I chose this because malignant tumour cells usually exhibit similar characteristics. For example, malignant tumour cells have a large shape and are thick. For this reason, I found it logical to choose this model. Of course, it did not perform as well as Logistic Regression, but I can say that it performs close.

Finally, I added the Naive Bayes model. Naive Bayes is a classification technique based on Bayes Theorem with the assumption that all features that predict the target value are independent of each other. It calculates the probability of each class and then selects the one with the highest probability. I saw when I researched that it is generally used in NLP. I can say that it performs close to the kNN model.

As a result, the best performing models for this dataset are Random Forest Classification and XGBoost Classification based on decision trees. Decision trees can well distinguish whether a tumour is benign or malignant. Next was Linear Regression, because the independent features in the dataset were found to be linearly effective in determining whether a tumour was benign or malignant.