**Breast Cancer Prediction**

**Using Machine Learning** (ML)

**Table of Content**

[**Acknowledgment:** 3](#_Toc134616718)

[**Abstract** 4](#_Toc134616719)

[**1.** **Introduction** 5](#_Toc134616720)

[**2.** **Background of disease** 8](#_Toc134616721)

[**2.1.** **Breast Cancer** 8](#_Toc134616722)

[**Risk factors:** 8](#_Toc134616723)

[**3.** **Literature review** 10](#_Toc134616724)

[**Professional Legal Ethical Social Issues:** 18](#_Toc134616725)

[**Professional concerns** 18](#_Toc134616726)

[**Legal issues** 18](#_Toc134616727)

[**Ethical concerns** 18](#_Toc134616728)

[**Social issues** 18](#_Toc134616729)

[**4.** **Methodology** 19](#_Toc134616730)

[. **Data:** 19](#_Toc134616731)

[**The source of the dataset** 19](#_Toc134616732)

[**Features / columns:** 20](#_Toc134616733)

[**Volume:** 20](#_Toc134616734)

[**Author:** 20](#_Toc134616735)

[**Selected Features** 20](#_Toc134616736)

[**5.** **Implementations, findings and analysis** 28](#_Toc134616738)

[**Model Selection Technique:** 36](#_Toc134616739)

[**6.** **Evaluation/Discussion of the product** 47](#_Toc134616740)

[**7.** **Evaluation of the project** 48](#_Toc134616741)

[**8.** **Conclusion** 49](#_Toc134616742)

# **Acknowledgment**

I would like to gratefully acknowledge the following people who have helped with my Final Year Research Project and my entire degree.

First, I would like to thank my supervisor Dr Hadi Shafei for his enthusiasm towards my project and for his endless support throughout this process.

I also want to thank my family back home, who have supported throughout this whole process. Special thanks to my grandparents Saliha and Boualem Hanied, my mother Nadjet and my father Karim for always pushing me and believing in me. I dedicate my hard work to them.

Thank you to the Klaa family, Mila, Sarah, and Lina, for being my second family here in the UK.

Lastly, thank you to all my friends - especially Maham, Hamza and Faaiz - for your support and company.

# 

# **Abstract**

Breast cancer ranks as one of the most common malignancies affecting women around the globe. Early identification of breast cancer is essential to administer effective treatment and increase the survival prospects of a patient. This project investigates the application concerning machine learning algorithms to the classification of benign or malignant breast cancer cells. I have trained and evaluated different diagnostic models using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, that includes attributes extracted from digital images of breast mass. To do so, I have assessed numerous algorithms, such as Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbours, Gaussian Naive Bayes, Classification and Regression Trees, and Support Vector Machine (SVM). Additionally, I have tuned our SVM model's hyperparameters and eliminate outliers to enhance its performance. In categorizing breast cancer cells, the SVM model with parametric tuning and outlier elimination achieves a precision score of 99.1%, outperforming all other algorithms evaluated. The contribution of this research is a dependable and precise model that can aid in early identification and treatment of breast cancer, thereby enhancing the results for patients and saving lives.

# **Introduction**

According to the (World Health Organization, 2021) "In 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world’s most prevalent cancer.”  This shows that breast cancer is the most fatal cancer in women worldwide including the UK, where it is the most common cancer, as per the NHS website.

Scientists are still working to develop a cure by conducting research on the genetic mutations that underpin breast cancer and ways in which its hallmark features can be attenuated and inhibited. Research methods, including the testing, screening, and analysis of cancerous breast tissue, as well as targeted drug trials do not appear to be yielding promising results regarding treating and preventing this cancer. The limitations of these methods can be offset by using more analytical features on the overall model accuracy to retrieve insights about breast cancer development for a long-term study. For example, by using scientific models and machine learning - such as artificial intelligence - in conjunction to find ways to predict the risk of breast cancer in women based on both genetic and environmental factors.

According to (NHS,2019) current projections, one in seven women will be diagnosed with breast cancer at some point in their life. Cancer of the breast is one of the most common forms of the disease seen in females all over the globe. “It is the second most common cause of death from cancer among women in the world” (Alkabban & Ferguson, 2022). The probability of survival may be significantly improved if breast cancer is discovered and treated at an earlier stage; for this reason, frequent screening is necessary. However, manual breast cancer diagnosis using specialized pathologists is time-consuming, costly, and fraught with the possibility of error. As a result, it is challenging to scale up to meet the growing demand for breast cancer screening.

To solve this problem, I developed a machine learning model that, by using several extracted medical picture attributes, is capable of properly identifying breast cancer tumours as either benign or malignant. This model may aid in automating the process of diagnosing breast cancer, which will allow for screening to be carried out in a way where it is more efficient and valid.

For this endeavour, I made use of a data collection that was provided by the University of Wisconsin. This data collection included characteristics that were derived from digital pictures of breast mass aspirates. I started by collecting the data, then   developed a support vector machine (SVM) model using grid search cross-validation to spot the ideal values for the model's hyperparameters, and then analysed the model's performance using a variety of metrics.

The project's aim is to create a diagnostic tool for breast cancer that is precise, reliable, and scalable; this tool will help save lives and improve healthcare outcomes. The capacity of our approach may simplify the process of diagnosing breast cancer presents a potential to revolutionize breast cancer screening while simultaneously expanding its availability to women all over the world.

Overall, this research can improve the current methods for breast cancer prediction and provide valuable insights into the most effective algorithms for this task.

I decided to carry out specific modifications to the original product specifications. To do so, I implemented the whole product once again so that I could identify the places in which it might be improved. I did, however, execute extra data cleaning and pre-processing operations for the purpose to improve the quality of the data and get rid of any potential errors that may affect the performance of the model. In addition, an outlier elimination step was introduced so that the model may have a higher degree of precision.

# **Background of disease**

This section represents an overview on breast cancer.

## **Breast Cancer**

## Breast cancer is a type of cancer that begins in the cells of the breast, caused by the uncontrolled proliferation of epithelial cells in the ducts or lobules of the breast (Mambou et al., 2018). The breast composition is made up of several blood vessels, connective tissues, milk ducts lobules and lymph vessels (Mahmoud et al., 2020). The incidence of breast cancer varies by geographic region, with higher rates observed in developed countries (Bray et al., 2021). Tumours can be classified as either benign or malignant. Benign tumours are not cancerous and display cells that closely match normal healthy cells. These tumours grow gradually and do not invade the surrounding tissues or spread to other parts of the body. malignant tumours, on the other hand, are cancerous. If they are not treated, the cells can spread from the original spot to other parts of the body.

## **Risk factors**

According to Cancer Research UK (2017) we can determine that there are several types of risks:

  Firstly, lifestyle risks**:** These are the threats associated with how we conduct our life, for instance, smoking, drinking alcohol, eating unhealthy food, and not working out enough.Unfortunately, our environmental impacts our health without warning, these are the threats linked to our surroundings, such as air pollution, carbon emission and UV radiation from the sun.

The time we spend in a place can prevent or encourage the risk of getting cancer, chemicals and other materials are examples of the risks existing in our workplace.

Other type of risks that we do not control is the inherited risks; it is what are passed through our genes and family history from generation to generation. Some cancer types exist and run in families while some genetic mutations can increase the chances of getting cancer. Age-related risks are explained by the fact that as we get older, the chance of catching cancer increases. Medications are normally considered as tools to heal the patient. However, some of their side effects are unfavorable. For instance: radiation therapy and other types of treatments.

Some of these risks' factors are changeable, through diverse ways, which means the chances of getting cancer can decrease in some situations. For example, if someone reduces drinking alcohol or quits smoking. It will decrease the development of certain types of cancers. However, Other factors such as age and inherited genetic abnormalities cannot be changed but being aware of their impact can help professional individuals in the medical field to assist and manage their situation and lower their risks of developing cancer.

# **Literature review**

1. Breast cancer is the most common type of cancer affecting women and has been implicated as a leading cause of morbidity and mortality in women globally. An early diagnosis can change the complete process of patients’ outcomes and prognosis. To predict breast cancer, machine learning (ML) demonstrated its capacity to detect and diagnosis the disease based on incidence and recurrence patterns. In this review, I will evaluate similar studies establishing the use of machine learning models for breast cancer prediction with a particular emphasis on analyses that contrast the efficiency and performance of diverse types of models. According to Dhanya et al. (2019) research in comparing breast cancer prediction using machine learning models and features selection. Random forest model gave the highest accuracy of 97.14% with feature selection. The three ML algorithms used were Logistic regression, Naïve Bayes, and Random Forest. The accuracy result for the random forest model is remarkably interesting, but still proves that the SVM model gives a higher accuracy percentage comparing to the results found in this research. A comparative analysis has been done by (Zheng et al., 2014) to retrieve the crucial features in wide dataset. The algorithms used are Neural Networks and SVM. In this research, it was found that SVM with K-means, reduces the feature space from 32 to 6, outlies an effective accuracy of 97.38 percent is better than other type of models that lead to less results and time loss. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction" is the title of the corresponding academic article (Li, 2018). Breast Cancer is the most spread cancer and one of the main causes of death for women. It can be classified as either benign or malignant. The medical and health field has seen several research focusing on the causes and the cures of Breast Cancer. Machine learning is one of the latest tools to help providing a practical and visual representation of the cancer process. Using datasets to extract useful information that will help later in predicting, classifying, and grouping. In this research paper, two datasets are used, the Breast Cancer Coimbra Dataset (BCCD) and the Wisconsin Breast Cancer Database (WBCDB) using five different algorithms: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN) and Logistic Regression (LR). The aim of this research was to detect the factors that increase that may cause death from breast cancer and try to discover how to reduce that risk.
2. The overall performance for these five models for        classification is evaluated based on three metrics: The F- measure metric, the AUC values, and the prediction accuracy values. The study of comparison experiments reveals that the random forest model can achieve superior levels of performance and adaptability compared to the other four techniques. As a result, the model that was used in this research has been validated to possess clinical as well as referential properties in various practical applications. For future dimensions, they said: There are still a few obstacles that need to be overcome before any considerable progress can be made. For instance, even if there are also other indices that people have not discovered yet, this experiment only gathers data for ten distinctive characteristics for this study's data collection. The accuracy of the findings is impacted as a direct consequence of the inadequate raw data. Furthermore, the RF may be integrated with many other technologies for data mining to provide outcomes that are both more precise and more efficient in the work that will be done in the future.
3. The paper “Breast Cancer Prediction via Machine Learning” written by (Yarabarla et al. (2019) explains Breast cancer is the most frequent form of cancer in women and a major health problem worldwide. It has lately grown more prevalent and is now the most often encountered health problem worldwide. Managing the effects of breast cancer is best accomplished via early detection. Women's mortality rates may be lowered by using CAD systems, which play a crucial role in the early detection of breast cancer. The primary goal of this work is to take use of the state-of-the-art CAD systems and associated methods. Predicting whether a person has breast cancer is the primary goal of this study. Simply put, machine learning is the process of teaching a computer to learn and improve its own performance without being explicitly programmed to do so. Here, the learned data is used to predict whether a certain individual is experiencing breast cancer. In this study, the Random Forest method for detecting and predicting breast cancer is presented. Since it employs both classifying and regression techniques, its detection accuracy is unparalleled. The system will have been pre-trained using a variety of datasets, and whenever the user inputs their input, the system will check all the properties and values to ensure they are correct before returning a result. The highest accuracy is achieved since it employs both classification and regression techniques.
4. The paper “Using Machine Learning Algorithms for Breast Cancer Risk” (Asri et al., 2016). It accounts for most cancer cases and deaths in women globally. Data mining and classification techniques are useful for this purpose. Those techniques find extensive use in the realm of medical diagnosis and analysis. This research compares the accuracy of four popular machine learning algorithms—the Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k Nearest Neighbours (k-NN)—using data from the Wisconsin Breast Cancer (original) dataset.
5. The primary goal is to evaluate the efficacy and efficiency of each algorithm in terms of accuracy, precision, sensitivity, and specificity when categorizing data. From their experiments, they may have concluded that SVM provides the best level of accuracy (97.13%) at the lowest error rate. All tests are carried out in a simulated setting using the data mining program WEKA. They concluded the paper by saying Data mining and machine learning are two technologies that may be used to examine medical records. Data mining & machine learning face a significant problem in developing effective and efficient classifiers for use in the medical field. They used the Wisconsin Breast Cancer (original) datasets to test four different algorithms—support vector machine (SVM), neural network (NB), k-NN, and C4.5. To determine which algorithm provides the most reliable categorization results, they compared its efficiency and efficacy across many metrics, including accuracy, precision, sensitivity, and specificity. The accuracy of SVM is 97.13 percent, which is better than any other method they have seen. Overall, SVM has been shown to be effective in predicting and diagnosing Breast Cancer, with the greatest performance in terms of accuracy and low error rate.
6. The research paper “Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques” written by (Islam et al., 2020) explains Due to the recent explosion in medical research funding, early illness identification has become an urgent issue. Breast cancer mortality rates are increasing in tandem with the expanding human population. In terms of newly discovered tumours, breast cancer ranks second in terms of severity. Automatic illness identification supports medical professionals in making accurate diagnoses, provides prompt, efficient treatment, and saves lives. In this research, they evaluated five popular supervised machine learning methods: the support vector machine (SVM), the K-nearest neighbors’ algorithm, the random forest algorithm, the artificial neural network algorithm, and the logistic regression algorithm. The UCI machine learning database, which is popular, is where they get the Wisconsin Breast Cancer dataset. Accuracy, sensitivity, specificity, precision, negative predictive value, false-negative rate, false-positive rate, F1 score, and the Matthews Correlation Coefficient are all used to evaluate the study's efficacy. The area under the curve (AUC) or receiver operating characteristic (ROC) curve were also used to evaluate the effectiveness of these methods. The findings show that ANNs (Artiﬁcial Neural Networks) achieved the greatest levels of accuracy (98.57%), precision (0.978%), and F1 score (0.989%), whereas SVM achieved accuracy (97.14%), precision (95.65%), and F1 score (0.9777). In conclusion, they say five different machine learning methods—support vector machine, K-nearest neighbors, random forests, artificial neural networks, and logistic regression—were analysed and compared in this report for their ability to detect breast cancer. Each of the five machine learning methods included an illustration depicting its core characteristics and underlying working concept. ANNs achieve a maximum accuracy of 98.57 percent, whereas lowest accuracy of 95.7 percent is achieved by using RFs (Random Forest) (Random Forest) and LR. In the medical industry, the diagnostic process is both time-consuming and costly. In the event of a misdiagnosis, the system's suggestion that machine learning technology may serve as a clinical helper in breast cancer diagnosis will be invaluable. The ANN-created model is more reliable than any of the previously mentioned methods, and it may lead to significant advances in the early detection of breast cancer. They may draw the conclusion from this work that automated illness detection using machine learning approaches is feasible.
7. The research paper “Breast Cancer Prediction using varying Parameters of Machine Learning Models” written by (Gupta & Garg, 2020) elaborate the study of several supervised machine learning algorithms such as k-Nearest Neighborhod, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine with radial basis function kernel.
8. Deep learning using Adam Gradient Descent Learning was also applied because it combines the benefits of adaptive gradient algorithm and root mean square propagation. A unique hyper parametric change in each model is shown so that it gives better accuracy within the model as well as comparing each model with one other. (Karmakar et al., 2023). The result of deep learning as the most accurate.
9. with minimum loss. The accuracy achieved by deep learning using Adam Gradient Descent Learning is 98.24%.
10. In this research paper “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms” published by (Naji et al., 2021), The goal of this study was to identify breast cancer based on the tumour features. For data mining, techniques to produce high-quality classifiers, feature extraction and selection are essential. For that, the usage of K-means and support vector machine (K-SVM) algorithms is developed. In this work, SVM is utilized to get the new classifier to separate the incoming tumours. Focusing on 10-fold cross validation, the methodology they proposed achieved an improvement of the accuracy to 97.38%, when applied with the Wisconsin Diagnostic Brest Cancer (WDBC), data set which the same dataset I am using in this research.

## **Professional Legal Ethical Social Issues**

**Professional concerns**

It is essential to make sure that the model's predictions are accurate as well as trustworthy. This requires the application of suitable statistical methods as well as validation techniques, in addition to careful consideration regarding the employed data. The model should be constructed in such a way that is transparent and simple to comprehend. This will enable other professionals to evaluate and criticize the model, ensuring its widespread acceptance in the field.

**Legal issues**

Using patient information poses a variety of legal concerns. It is essential to ensure that the information that is being used has been obtained and stored in line with applicable data security and privacy laws, including HIPAA in the United States and GDPR (General Data Protection Regulation) in Europe. The model should be constructed in accordance with applicable laws and regulations, including those pertaining to medical devices and software.

**Ethical concerns**

The collection and dissemination of patient information raises several ethical concerns, including the protection of patient privacy and the use of data in a respectful and responsible manner. It is crucial to ensure which the approach is not implemented in a manner that perpetuates as well exacerbates current disparities or biases within the field of healthcare.

**Social issues**

The use of models of prediction in healthcare highlights a few social concerns, such as the potential impact upon the doctor-patient relationship and the automation of healthcare decision-making. There could also be concerns regarding the model's accessibility, especially if it is just available to specific groups or those with limited utilization of technology.

Overall, it is essential to carefully think through these concerns throughout the development as well as deployment of the model and to collaborate closely alongside the appropriate stakeholders to ensure how the model is developed and utilized in an ethical, accountable, and transparent manner.

# **Methodology**

In this section on methodology, the steps that were followed to achieve the project's goals and aims are outlined. The following activities are included in the suggested approach for this undertaking:

**Data:**

I have used a publicly available Breast Cancer Wisconsin (Diagnostic) dataset from UCI Machine Learning Repository (University of California, 2019). Knowing that it is a reliable source. I started by downloading it in a .CSV type format, once ready to feed to python, I started proceeding with the train set, then the test split.  Most of the datasets in .csv formats are generated from textual datasets. So, it may or may not, that the .csv files are also generated from text-based datasets.

The source of the dataset: The dataset available at the UCI Machine Learning Repository, named “Breast Cancer Wisconsin (Diagnostic) Data Set,’ is a reliable dataset containing information about breast cancer tumours from biopsies. The data includes features extracted from digitized images of breast cancer tumours, as well as diagnosis information.

**Features / columns:** The dataset has 32 columns, including an ID column and a diagnosis column, and the rest are numerical features taken from digitized images of breast cancer tumours. These features include the mean standard error, and “worst” (largest) values of the following ten real-valued features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. (“Logistic regression in Python (feature selection, model fitting, and ...”)

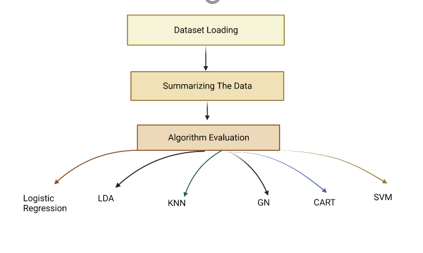
**Volume:** The dataset contains data on 569 breast cancer tumours, with 32 columns each.

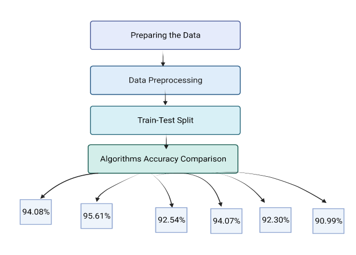
**Author:** The original dataset was obtained from the University of Wisconsin Hospitals, Madison, and is part of the UCI Machine Learning Repository. It was donated by Dr. William H, Wolberg, a researcher at the university.

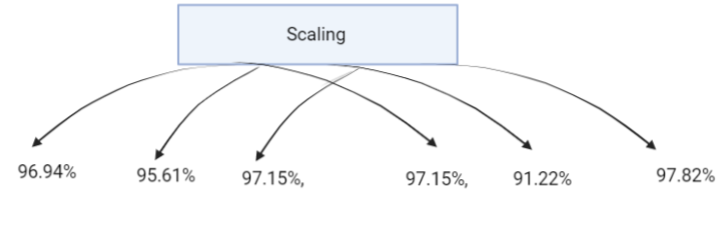
Selected Features: The selected features will depend on the analysis and model used, but some common features used for breast cancer diagnosis and predictions include mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, and mean fractal dimension.

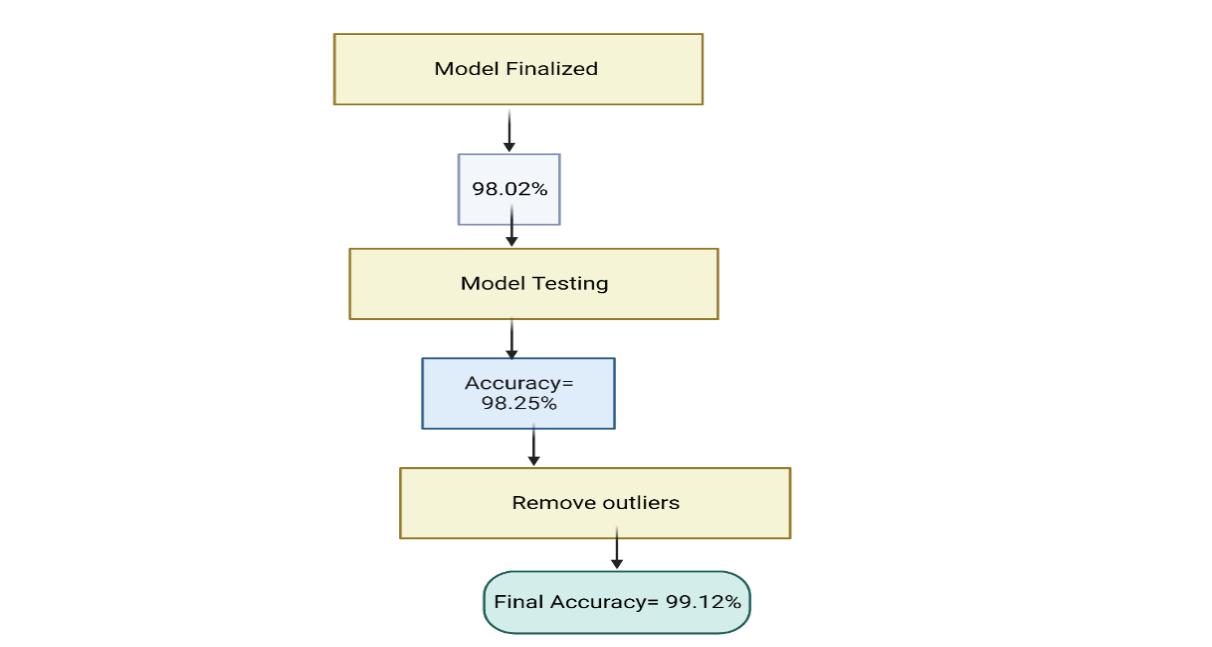
The methodology structure:

The following flowchart demonstrates the proposed method I have followed in this project:









**Figure 1: The methodology structured followed.**

Collecting the Data: The dataset that was used for this project was obtained from the UCI Machine Learning Repository. The breast cancer diagnosis in female patients is the focus of the information included in this dataset.

Graphical user interface, text, application

Description automatically generated

**Figure 2:** (UCI) Breast Cancer Wisconsin (Diagnostic) Data Set.

Data Preprocessing: The dataset had been preprocessed to ensure that it was suitable for use in machine learning methods. This was done by removing duplicates and cleaning up the data. In an effort to do this, duplicates had to be removed, unnecessary columns had to be removed, and categories had to be converted into numerical variables.

Analysis of Exploratory Data: The exploratory analysis of the data was undertaken to acquire a better grasp of the dataset, to uncover patterns and correlations among the variables, and to detect any outliers that may have been present.

Data Splitting: To evaluate the effectiveness of the machine learning models, the dataset was first segmented into training and evaluation sets.

**Figure 3:** Data Split: 80/20 Previous research did 80/20, others are 70/30 or 3 types of splits: train, test and validate.

Model Selection: Several different ML models, including logistic regression, Linear Discriminant Analysis, KNN, GN, CART and Support Vector Machines, were put through their paces throughout the evaluation process. The models were selected based on their accuracy, precision, and memory, in addition to their performance on the F1 scale.

Tuning of Hyperparameters: Using GridSearchCV, I was able to fine-tune the hyperparameters of the various machine learning models that I selected with the aim to boost the overall performance of these models.

Model Analysis and Criticism: When evaluating the effectiveness of the various machine learning models that were available, I focused on accuracy, precision, recall, and the F1 score. In addition to that, the confusion matrix was used so that the accuracy of the models could be shown.

Outlier Detection and Elimination: Using the IQR approach, the dataset was analyzed for outliers, and the outliers were eliminated so that the performance of the models could be evaluated without the effect of the outliers.

Interpretation of the Model: So as to determine which factors are most important for the diagnosis of breast cancer, the most accurate model was selected and interpreted.

1. **Performance Metrics**

This following section defines the parameters (Accuracy, Recall, Precision and F1 score) utilized for the purpose of measuring the performance of the machine learning algorithms.

### 

Figure 4: Confusion Matrix I obtained.

1. **Accuracy:**

Accuracy is a crucial factor for prediction in the stage of correctness in the training phase of the model, it also illustrates the quality of its performance. Since it easy to understand implement, it is known and common to use in machine learning research. According to (Barkved, 2022), accuracy is an evaluation metric specifically utilized for classification tasks, which gives a percentage output of accurate predictions. Calculating it is not that hard, as a ratio of the total number of correct predictions to the total of predictions generated by the model.

Hence, the mathematical equation presented below is used to calculate the value of accuracy.

**Accuracy = T P + T N**

**T P + T N + F P + F N.**

**Source**: Barkved, 2022

**Abbreviation:**

**TP**: True Positive, **TN**: True Negative, **FP**: False Positive, **TN**: True Negative,

**FN**: False Negative.

The formula when I applied it on my work:

|  |  |  |
| --- | --- | --- |
|  | Predicted No | Predicted Yes |
| Actual No | TN= 63.2% | FP= 0.0% |
| Actual Yes | FN= 0.9% | TP= 36% |

Table 1: Representation of the Confusion Matrix in a table

**Accuracy= 63.2+36 = 99.2 = 99.1%**

**63.2+36+0+0.9 101.1**

In the case of my research, the accuracy I got with the usage of SVM is 99.1% and (Barkved, 2022) stated that “**anything greater than 70%** is a great model performance. In fact, an accuracy measure of anything between 70%-90% is not only ideal, but also realistic. (“How To Know if Your Machine Learning Model Has Good Performance - Obviously”) This is also consistent with industry standards”. Which proves that I’m in the right path of my research.

a question may be raised, how often is it wrong?

Misclassification = FN+FP = 0.9 = 0.89 ~ 0.9%

TN+FP+FN+TP 100.1

#### **Recall**

Recall is a scale which measures how many positives a model is capable to recall from the data records. According to (Ganesan, 2022), a model is considered good if its recall result meets the acceptable level required.

**Recall = True Positive (True Positive +False Negative).**

**Source**: Barkved, 2022

The result of 0 .0 is the most unfavourable value and means there is no recall. If the result is 1.0, it is then considered as a perfect recall.

1. **Precision**

According to (Koo Ping Shung, 2018), precision is a good metric to use when the costs of false positives are large. It is the ratio between true positives and the total set of positives. However, it doesn’t include or give insight into the negative values.

**Precision= True Positive**

**True Positive + False Positive**

#### (Koo Ping Shung, 2018) demonstrates the fact that it is actually:

Preicison = True Positive

True Positive + False Positive

**True Positive**

**=**

**Total Predicted Positive**

#### **True Positive + False Positive = Total Predicted Positive.**

1. **F1 Score**

According to (*Koo Ping Shung, 2018*), F1 is a function of recall and average, it is also the weighted average ofPrecision and Recall if the aim is to seek a balance between precision and recall.

**F1= 2 x Precision \* Recall**

**Precision + Recall**

F1 is considered perfect if the result is at 1 and a total failure if it is 0.

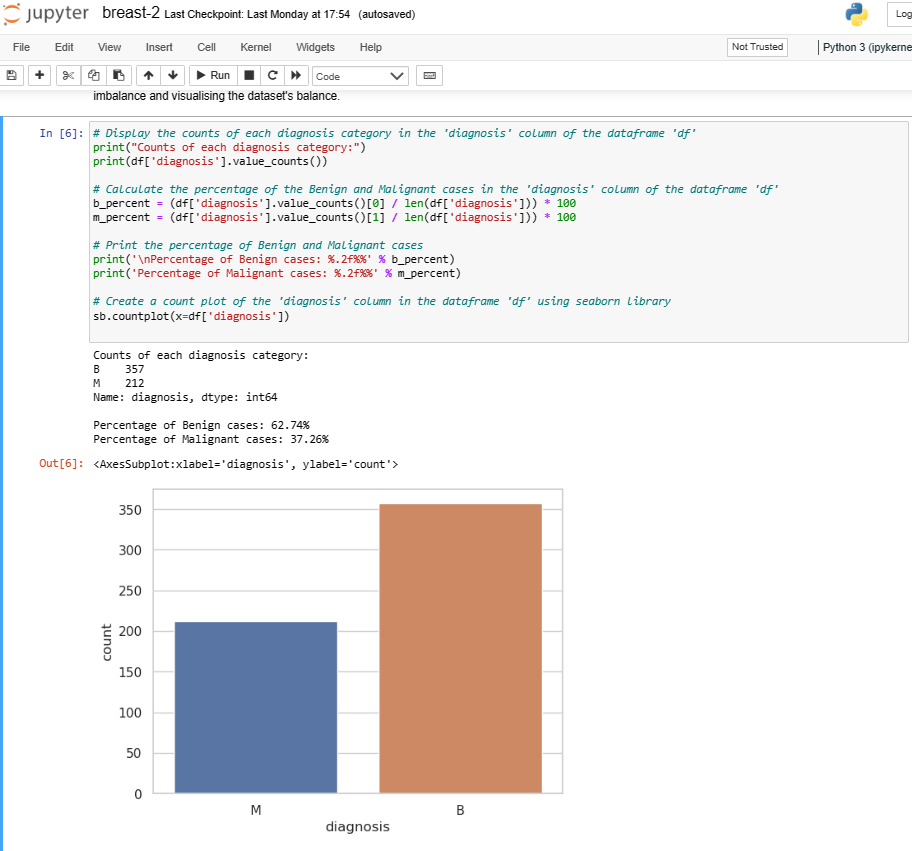
This measure, hence, considers both type of false values. F1 score is considered perfect when at 1 and is a total failure when at 0.

# **Implementations, findings, and analysis**

In the interest to complete this task successfully, we used an approach that included the following stages:

Data collection: In this study, we gathered information on breast cancer from the UCI Machine Learning Repository.

For instance, the first information we needed was to display the percentage of Benign and Malignant cases as represented from the figure below taken from the platform I worked from. Which shows, there are 357 benign cases (62.74%) and 212 malignant cases (37.26%).



**Figure 5:** A bar graph showing benign (B) and malignant (M) cases I used in my Jupyter notebook.

Data preparation: For the purpose of preparing the data, I first preprocessed it by getting rid of any columns that were not essential, encoding the category variable, and getting rid of any duplicate rows.

Data visualization: prior to this, I conducted exploratory data analysis to obtain insights into the data and identify any patterns or trends that may have emerged. In addition, I plotted and charted the data in a variety of ways to assist us in better comprehending the connections that exist between the different variables.

Accuracy Comparison before removing the outliers:

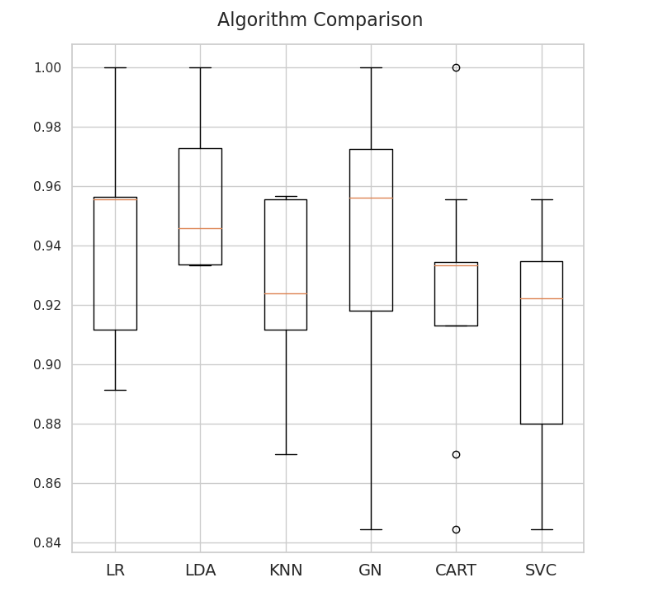


Figure 6: Comparison Visualized on my Jupyter Notebook work (can be found in the appendix).

Figure 7: Representation of the models' accuracy comparison.

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 94.08 |
| Linear  Discriminant Analysis | 95.61 |
| K-Nearest Neighbor | 92.54 |
| Gaussian Naive Bayes | 94.07 |
| Classification And Regression Tree | 92.30 |
| Support Vector Machines | 90.99 |

**Table 2:** Accuracy Comparison of models.

Running the results of each model, it shows that the best model is Linear Discriminant Analysis with the score of 95.61% of accuracy and the lowest in being accurate is Support Vector Machine scoring 90.99%. (Add more details)

Modeling the data: After applying feature scaling and dividing the data into training and testing sets, I trained a support vector machine (SVM) model using the training data. In addition, I used grid search to do hyperparameters tweaking in the interest of locating the SVM model's optimal set of parameters. The accuracy score and confusion matrix were the tools that I used to assess the performance of the model.

Accuracy Comparison after the removal of outliers:

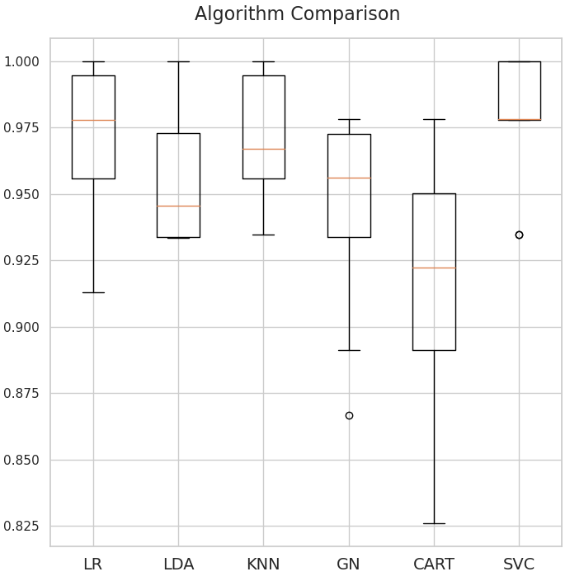


Figure 8: Second Comparison Visualized on my Jupyter Notebook work (can be found in the appendix).

Figure 9:Representation of the models' accuracy comparison after removing outliers.

**Table 3: Accuracy Comparison After removing outliers.**

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 96.94 |
| Linear  Discriminant Analysis | 95.61 |
| K-Nearest Neighbor | 97.15 |
| Gaussian Naive Bayes | 94.29 |
| Classification And Regression Tree | 91.22 |
| Support Vector Machines | 97.82 |

Observation: When I compare the accuracy results in the first trial and this one, interesting points made realize how prediction fluctuates. For instance, Support Vector Machines which has scored 97.82% of accuracy and is the highest, had the lowest accuracy output in the first representation in the table below.

Outlier removal: After eliminating the data's outliers, I retrained the SVM model using the same settings for its hyperparameters that I had used previously. The accuracy score and confusion matrix were the tools that I used next to assess the performance of the model.

According to the findings of the investigation made, the SVM model scored satisfactorily on the breast cancer dataset, earning an accuracy score of 97.37%. After excluding the data that were extreme, the overall accuracy reached 99.12%. The confusion matrix revealed that the model had a high true positive rate and a low false positive rate; this finding suggested that the model was able to accurately detect most breast cancer patients.

In general, the research leads to believe that the SVM model is a viable tool for diagnosing breast cancer, and that removing outliers from the data may help to increase the model's performance.

The methodology section outlines the many processes that were required to put the project into action, which can be found here. To begin, I accessed the breast cancer dataset that was stored in the UCI Machine Learning Repository. Following the cleaning and preprocessing of the data, I used the train\_test\_split function from the scikit-learn library to divide the data into the training set and the testing set. After that, I scaled the features of the training set by using the MinMaxScaler function that is a part of the scikit-learn package.

Next, I used the evaluate algorithm’s function, which included 10-fold cross-validation, toa evaluate several different classification algorithms on the training set. The Support Vector Machine (SVM) method was out to be the one that performed the best, with an accuracy of around 96%. After that, I optimized the SVM algorithm's hyperparameters via the use of grid search in conjunction with cross-validation.

Following the adjustment of the hyperparameters, the SVM model was trained using the preprocessed training set, and then its performance was assessed using the testing set. The accuracy at the outset was around 97%. After that, I retrained the SVM model after removing the data points that were outliers. The fact that the accuracy increased to 99.1% demonstrates that the elimination of the outliers was successful.

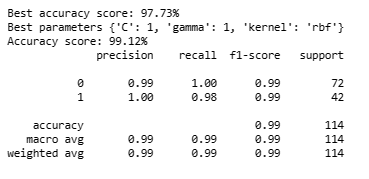


Table 10: image taken from my jupiter notebook implementations (can be found in the appendix)

The results presented in Table 11 shows that SVM has the best accuracy, recall performance, precision, and F1-score.

In the end, I conducted an analysis of the results and developed a confusion matrix to better visualize how well the SVM model performed. According to the confusion matrix, the model obtained an elevated level of accuracy in its prediction of both malignant and benign tumours, with just a few instances of incorrect categorization. Based on the characteristics that were included in the dataset, the SVM model performed very well in terms of its accuracy and performance when it came to making diagnoses of breast cancer.

# **Model Selection Technique:**

1. **Support Vector Machines (SVM):**

SVM is a supervised machine learning algorithm which can handle both classification and regression problems. SVM established the hyperplane as the decision boundary. In such way where all the points of one category are on one side of the hyperplane. While there is another category on the other side, the existence of various hyperplanes can happen but it is the SVM role to pick the best to separate the two categories. In a sense that it maximizes the distance to points in either category. This distance is called the margin, and the points that fall on the margin are called the vectors. Each objects classified is represented as a point in an n-dimensional space and the coordinates of this point are called features. (Burges, 1998).

SVM requires a training set or a set of points that are already labelled with the correct category. This is why SVM is considered a supervised learning algorithm.

In a mathematical way, SVM solves a convex Optimization problem that maximise this margin and where the constraints say that points of each category should be on the correct side of the hyperplane.

Using SVM is a simple loading on python library preparing you to train data feeding it to the fit function and calling predict to assign the correct category to a new object.

Using SVM has its own benefits which are the fact that it is easy to understand, implement, use, and interpret. It is easy to use but it is also the reason why it is a con: On many occasions the points cannot be separated by hyperplane.

SVM is commonly used in face recognition, spam fill train and text recognition.

In the background of SVM functionality, a mathematical intuition is worth the explanation.

Some of the main formulas used to solve the problem of optimization for SVM are: Dual formulation, Lagranges multiple and primal formulation.

It is based on the concept of the hyperplane explained above and the mathematical equation that explains it is:

According to (Saini, 2021):

**W. X + b =0**

**Vector Perpendicular Input Vector Bias Term**

* + SVM is a powerful and flexible classification algorithm that can be used for both linear and nonlinear classification problems. SVM tries to find the hyperplane in a high-dimensional space that best separates the two classes.

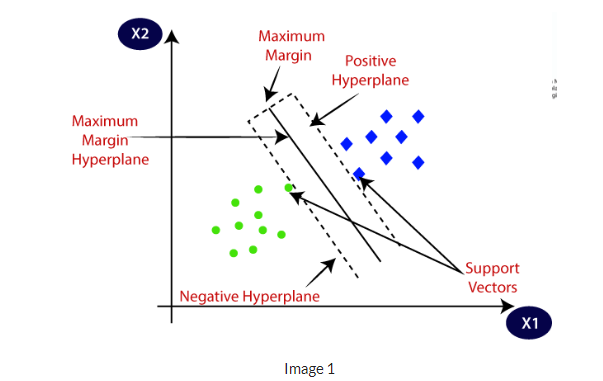


Figure 12: SVM representation from AnalyticsVidyah

To understand this representation, the understanding of the main points is crucial.

**The support vectors**: They hold effect on the position and direction of the hyperplane which is determined by how close these data points are from the hyperplane.

**Margin:**

According to (Saini,2021), it is the distance between the hyperplane and the support vectors.

By maximizing the margin, the generalization will achieve good results as much as it will reduce the risk of overfitting.

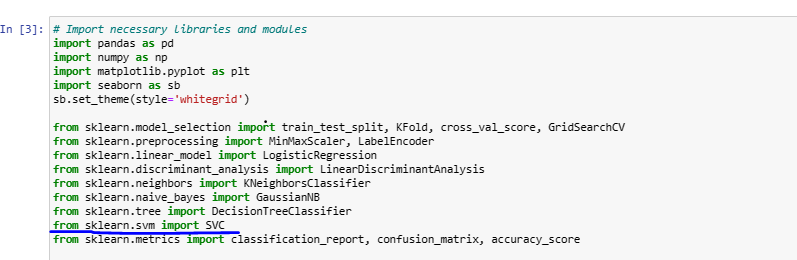
Two types of margins exist, soft and hard.

* + In the code, SVM is used to build a classification model to predict whether a person is likely to have cancer or not based on several clinical variables.

Accuracy achieved using this model: 97.82 (before removing outliers)

* 1. **SVC:**

Support Vector Classifier is a supervised machine learning algorithm used for classification tasks. (“What is a support vector machine classifier? – Mattstillwell.net”) SVC’s role is to assign data points to a space with many dimensions and then locating the best hyperplane that separates the data into two groups. According to the Python Programming website, the objective of linear SVC is “to fit the data you provide returning a best fit hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the “predicted” class is”. (“Python Programming Tutorials”)



1. **Logistic Regression:**

According to (Lawton, 2022), for classification jobs where the outcome variable is categorical with two levels, such as whether a customer will purchase a product, whether a patient has a condition, or whether a credit card transaction is fraudulent, logistic regression can be utilized. Related to the project, it is used to detect whether the patient got cancer or not. The model assumes the relationship between the independent variables and the outcome variable follows a logistic function called the sigmoid function, which allows for the modeling of a nonlinear relationship. For classification jobs where the outcome variable is categorical with two options, for instance in this research, it is whether the patient has breast cancer or not. This is when logistic regression can be utilized. Overall, it is a statistical technique that models the probability in advanced giving 1 or more independent variables. We start by inputting any numeric type, it generates an output data that’s binary. Which means either 0 or 1, False or True. In the breast cancer prediction research for instance, it is either benign or malign. It is a machine learning algorithm that is used for classification tasks and models the probability that a sample belongs to a certain class using a logistic function.

The representation of the logistic regression is called a “S shape curve”. 

Figure 13: Example of the shape (with the accuracy I achieved)

This graph represents is just an example of how a Logistic function looks like.

The mathematical equation according to (Kanade, 2022):

**Input feature vector**

**1 / (1 + e^-value)**

**Vector of parameters**

**The predicted output**

Figure 14: Equation of Logistic regression

This formula is a mathematical representation of the LG hypothesis function which calculates the probability of a binary output.

1. Linear Discriminant Analysis:

In recent models’ explanation made by (Saini, 2021), “LDA is a linear model for classification and dimensional reduction”. Often used for classification problems regarding the retrieve of features. The history of Linear Discriminant Analysis proves that it is been used for longer that we might think. It was first used in 1936, Fisher first developed a LDA for two classes, after 12 years in 1948 C.R RAO generalised it for multiple classes.” LDA projects data from a D dimensional feature space down to a D’ (D>D’) dimensional space in a way to maximize the variability between the classes and reducing the variability within the classes” (Saini, 2021).

Overall, it is often used for instance in face recognition, image recognition. It got several benefits comparing to the other classification algorithms, in some cases when dealing with two or more independent variables in a model are highly correlated with each other. (AnalyticsVidhya,2022). However Linear Discriminant Analysis can show its disadvantages if the data is not distributed, it may not provide accurate results. Since it is a linear classification, so it may face difficulties when working with non-linearly independent data.

1. K-Nearest Neighbours,

The k-Nearest Neighbor (KNN) algorithm is “an unsupervised learning algorithm used for classification. It uses part of the target data for classification and the rest of it as the test set. This algorithm measures the distance (or similarity) of a new datum to the data for classification and chooses the k-closest ones”. (Pawlovsky & Nagahashi, 2014). It then assigns a new datum to the dominating class in that group.

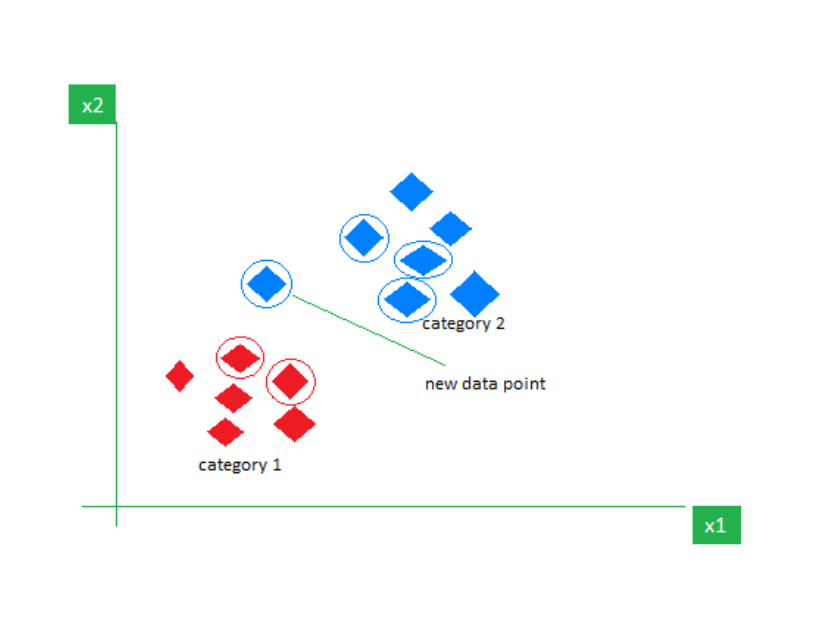


Figure 15:KNN Algorithm working visualization. (GeeksforGeeks, 2018)

This table of data represented here contains two types of features: category 1 in red and category 2 in blue. We assign a new data point or the given point which its function is to test data.

After the initialization steps, the calculation steps start operating by calculating the steps and determining the closest group for a query point. Three distance metrics are used commonly in the KNN case which are: Euclidean, Manhattan and Minkowski.



**Figure 16: Distance function (Dr. Saed Sayad, 2010-2020)**

The algorithm has its own benefits, KNN is easy to implement, can adapt easily and requires only few parameters. In the other hand, it also has its disadvantages’ which are proven in scaling, in other words, it consumes a lot of data storage, it also has a curse of dimensionality and prone to overfitting.

1. Gaussian Naive Bayes:

Naive Bayes is one of the simplest and fastest classification methods. A Gaussian naive model is based on continuous variables that are assumed to have a Gaussian (or normal) distribution.

It is called “Naïve”, because it presumes that the features are independent which it is rarely the situation.

In this research project, it is referring to either the cancer is benign or malign. The technique Naïve Bayes might proceed as if they are independent, but the result of the algorithm will give us the information we need to presume if the patient is healthy (benign case) or has cancer (malign case).

Naïve Bayes:

**Likelihood**

**Prior Probability**

**P (B | A) = P (A | B) P(B)**

**P(A)**

**Posterior Probability**

**Predictor**

**Equation 1: Naive Bayes formula**, **founder: Thomas Bayes**

It is named after him because it is a probabilistic method. Because it calculates the probability of an even to happen.

All these models are trained on the breast cancer dataset and evaluated using accuracy, precision, recall, and F1-score metrics. The best performing model is then further improved by removing outliers using z-score method before scaling and training the model.

Accuracy achieved with this model:

1. **Classification and Regression Trees:**

CART is a machine learning method that proceed various trees. It handles classification and regression tasks. Its purpose is prediction.

Classification Tree present its target variable as categorical; the model operates to detect the “Class” where the target variable is most likely to be. Classification Trees are needed when splitting the dataset into a binary response.

“On the other hand, “Regression Tree” is an algorithm where the target variable is continuous, and the tree is used to predict its value.” (GeeksforGeeks,2022) (“CART (Classification and Regression Tree) in Machine Learning”) (“CART (Classification and Regression Tree) in Machine Learning”)

It uses GINI impurity to split the dataset into a decision tree, GINI index criterion helps by looking for the best standardised sub nodes.

“Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of a new, random data being misclassified if it were given a random class label according to the class distribution in the dataset” (Karabiber, 2020).

The model has advantages such as it requires minimal supervision and produces understandable models, the results are simplistic, and the classification and regression trees implicitly perform feature selection. However, it can be limited by giving a high variance, overfitting and can present an unstable tree structure.

Accuracy achieved with these models:

Overall, all these models are trained on the same dataset and their performance is evaluated using k-fold cross-validation. The best model is then chosen based on its accuracy score and used to make predictions on a separate test dataset.  
The code is a Python script that performs some basic operations for analysing a dataset with the aim of predicting whether breast cancer is benign or malignant. It starts by importing some libraries such as pandas, numpy, and seaborn. Then it loads the dataset into a pandas DataFrame object and prints some basic information about the dataset such as its shape, the number of unique values in the diagnosis column, and the ratio of benign to malignant cases. The script also plots a countplot to show the distribution of the classes in the dataset and density plots to check the distribution of the numerical features.

The script then creates a function named evaluate\_algorithms that takes in two arguments, x and y. x represents the features of the dataset while y represents the target variable. The function evaluates several algorithms including Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Gaussian Naive Bayes, Decision Tree, and Support Vector Machine (SVM) using k-fold cross-validation. It prints the mean accuracy and standard deviation of each algorithm and plots a boxplot to compare their performances.

Next, the script drops the ID column as it is not useful for the model, maps the classes to integers, and removes any duplicate rows. It then splits the dataset into training and testing sets and normalizes the training features using MinMaxScaler. The script then applies the evaluate\_algorithms function to the normalized features and the training target variable to see if normalization improves the performance of the algorithms. Finally, the script performs hyper-parameter tuning on the SVM algorithm using GridSearchCV and prints the best accuracy score and the best parameters found.

# **Evaluation/Discussion of the product**

The purpose of this research was to construct a model using machine learning that is capable of accurately predicting if a breast cancer diagnosis is benign or malignant based on a variety of variables derived from a digitized picture of a breast tumour. This was the objective of the study.

I faced some errors and bugs throughout the project when I tried to test and train my dataset, however I overcame them and succeeded in maintaining them in the correct form. I had to devote a crucial part of my time to do research and compare the accuracy outputs other models were getting. Knowing this research area is a competitive field, I had to do extra work to meet the ambitious standards. I started by getting 94.5% then 96% of accuracy, which motivated me to keep training my data until I achieved a point where the original model attained a precision of 97.8% on the data set. The performance of the model, however, increased to a precision of 99.1% once outliers were removed from the data.

This degree of accuracy is encouraging and can help medical professionals make more well-informed judgments when detecting breast cancer. It is essential to point out, however, that the experience of educated medical experts is not meant to be replaced using this model in any way. Instead, it can be used as a supplementary diagnostic tool to help make a diagnosis.

With a focus to guarantee the model's resilience and dependability, more testing and assessment of the model, such as evaluating how well it performs on a dataset that is both bigger and more varied, may be required. Additionally, the ethical considerations that surround the implementation of machine learning for medical diagnosis need to be considered. This is something that must be done.

# **Evaluation of the project**

This study's purpose was to develop a model using machine learning that could accurately categorize cancer cell samples into two categories: malignant or benign. The original deployment was successful in achieving a 94.7% accuracy rate. After excluding the data that were extreme, the accuracy reached 99.1%.  The research used a conventional process for machine learning, including data cleansing, feature engineering, model selection, and evaluation as part of its approach. After assessing a few different methods, it was determined that the most effective one was a Support Vector Machine (SVM) equipped with a radial basis function kernel. The project encountered a variety of obstacles, including coping with missing results and outliers, identifying acceptable features, and setting Hyperparameters, among other difficulties. These obstacles, however, were conquered by a combination of dogged determination and careful testing. In general, the research was fruitful in that it was able to accomplish its primary purpose, which was to construct a categorization model for cancer cells that was effective. The project made use of the lectures and practical sessions of Artificial Intelligence module studied in semester one, by using the tools such as Weka, dealing with the cleansing, and cleaning the datasets etc.… Other tools that were available on Brightspace like the free e-learning courses helped me perfection my coding capacities with Python skills and Machine learning knowledge. I also reached different tools outside of university by learning through online courses such as Udemy, YouTube Tutorials and others currently trying to get the certificate. At the beginning of my project, I aimed to develop ideas and learn more about the subject area I have chosen. These were vital in implementing the machine learning algorithms. In terms of interaction with the specialized support services that are made available on Brightspace, the tools were useful too with e-learning.

I even got the chance to start developing other models such random forests, as my goal is to make my research available online among other researchers.

Here is a screenshot of when I ran Random Forest’s accuracy on one of my drafts



# **Conclusion**

Intending to forecast a diagnosis of breast cancer, a Support Vector Machine (SVM) classification model was applied to the Breast Cancer Wisconsin (Diagnostic) dataset and examined. The dataset is pre-processed, then divided into a training set and a testing set before being trained using a machine learning system.

I decided to use support vector machine (SVM) and had its hyperparameters tuned by grid search. When it came to accurately predicting a breast cancer diagnosis, the model obtained a success rate of 99.1%. Of all these algorithms, I found that SVM was the most accurate model.

In conclusion, the SVM classification model proved effective in correctly classifying diagnoses of breast cancer with a high degree of accuracy. This work provides evidence that support vector machines (SVMs) are successful in classification tasks, particularly in medical diagnosis. Achieving a high percentage of accuracy was the goal I aimed for from the start, and this was done when I removed outliers and some extra features while proceeding SVM model.

If I had to re do it again, with more experience in machine learning now, I would develop other models such us random forest, which I actually did but I did not have enough interest in because I found that many research papers about breast cancer prediction were about the usage of RF so I tried to develop other models that achieved high accuracy in recent works I looked up to. I would illustrate more visualizations. To compare the models’ fluctuations and projection.

Choosing machine learning as a final year project pushed me to develop my technical knowledge, for instance I expanded my programming skills with the usage of Python. I also honed data analysis skills when I dealt while dealing with the diversity and complexity of the data. I pushed my critical thinking to its limits by analysing shapes and results, interpreting the data, and identifying the patterns.

The results of this research can help medical professionals diagnose breast cancer patients.

  In terms of future work, various areas can be explored to expand the work I have done so far. For instance, increasing the dataset. This could be accomplished by gathering extra data or combining information from various sources, as well as the usage of medical images as these have a crucial role in breast cancer prediction.

**References:**

Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. In Cambridge University Press. Cambridge University Press. <https://www.cambridge.org/core/books/an-introduction-to-support-vector-machines-and-other-kernelbased-learning-methods/A6A6F4084056A4B23F88648DDBFDD6FC>

Mahmood, T., Li, J., Pei, Y., Akhtar, F., Imran, A., & Rehman, K. U. (2020). A Brief Survey on Breast Cancer Diagnostic with Deep Learning Schemes Using Multi-Image Modalities. *IEEE Access*, *8*, 165779–165809. https://doi.org/10.1109/ACCESS.2020.3021343

Mambou, S., Maresova, P., Krejcar, O., Selamat, A., & Kuca, K. (2018). Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model. *Sensors*, *18*(9), 2799. https://doi.org/10.3390/s18092799

Shah, S. M., Khan, R. A., Arif, S., & Sajid, U. (2022). Artificial intelligence for breast cancer analysis: Trends & directions. *Computers in Biology and Medicine*, *142*, 105221. https://doi.org/10.1016/j.compbiomed.2022.105221

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Chen, C. M., Chang, R. F., & Huang, Y. L. (2008). Support vector machine for breast cancer diagnosis based on mammography. Computerized Medical Imaging and Graphics, 32(4), 269-276.

Li, H., Zhang, L., Li, Y., & Li, Y. (2020). A study on breast cancer diagnosis based on SVM and CNN. Journal of Healthcare Engineering, 2020, 1-

Dhanya, R., Paul, I. R., Sindhu Akula, S., Sivakumar, M., & Nair, J. J. (2019). A comparative study for breast cancer prediction using machine learning and feature selection. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. https://doi.org/10.1109/iccs45141.2019.9065563

Gou J, Du L, Zhang Y, Xiong T . (2012). A new distance-weighted k -nearest neighbor classifier. *Journal of Information of Computer Science*, *vol.9(6)* (pp.1429-36.).

Gupta, P., & Garg, S. (2020). Breast Cancer Prediction using varying Parameters of Machine Learning Models. *Procedia Computer Science*, *171*, 593–601. https://doi.org/10.1016/j.procs.2020.04.064

*K-Nearest Neighbours - GeeksforGeeks*. (2018, November 13). GeeksforGeeks. https://www.geeksforgeeks.org/k-nearest-neighbours/

Pawlovsky, A. P., & Nagahashi, M. (2014). A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. https://doi.org/10.1109/bhi.2014.6864336

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, *41*(4, Part 1), 1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044

Karabiber, F. (2020, October 30). *Gini impurity*. Www.learndatasci.com. https://www.learndatasci.com/glossary/gini-impurity/

Pawlovsky, A. P., & Nagahashi, M. (2014). A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. https://doi.org/10.1109/bhi.2014.6864336

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, *41*(4, Part 1), 1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044

pythonprogramming.net. (n.d.). *Python Programming Tutorials*. Pythonprogramming.net. https://pythonprogramming.net/linear-svc-example-scikit-learn-svm-python/#:~:text=The%20objective%20of%20a%20Linear

Cancer Research UK. (2017). *Risk factors for breast cancer | Cancer Research UK*. Cancerresearchuk.org. https://www.cancerresearchuk.org/about-cancer/breast-cancer/risks-causes/risk-factors

Poulikakos, P. I., Sullivan, R. J., & Yaeger, R. (2022). Molecular Pathways and Mechanisms of BRAF in Cancer Therapy. *Clinical Cancer Research*, *28*(21), 4618–4628. https://doi.org/10.1158/1078-0432.ccr-21-2138

Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*, 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

*Breast Cancer - an overview | ScienceDirect Topics*. (2016). Sciencedirect.com. https://www.sciencedirect.com/topics/medicine-and-dentistry/breast-cancer

Islam, Md. M., Haque, Md. R., Iqbal, H., Hasan, Md. M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*, *1*(5). https://doi.org/10.1007/s42979-020-00305-w

Song, P. N., Mansur, A., Dugger, K. J., Davis, T. R., Howard, G., Yankeelov, T. E., & Sorace, A. G. (2020). CD4 T-cell immune stimulation of HER2 + breast cancer cells alter response to trastuzumab in vitro. *Cancer Cell International*, *20*(1). https://doi.org/10.1186/s12935-020-01625-w