**Table of Contents**

[1. Introduction 1](#_Toc132115417)

[**1.1** **Objectives** 2](#_Toc132115418)

[**1.2** **Methodology** 2](#_Toc132115419)

[**1.3** **Folder Structure of the GitHub Repository** 2](#_Toc132115420)

[2. Business Understanding 2](#_Toc132115421)

[3. Data Understanding 3](#_Toc132115422)

[**3.1** **Data exploration** 3](#_Toc132115423)

[**3.2** **Data quality assessment** 3](#_Toc132115424)

[4 Data Preparation 3](#_Toc132115425)

[**4.1** **Data Cleaning** 3](#_Toc132115426)

[**4.2** **Data Transformation** 4](#_Toc132115427)

[**4.3** **Data Reduction** 4](#_Toc132115428)

[**4.4** **Data Splitting** 5](#_Toc132115429)

[5 Modeling 5](#_Toc132115430)

[**5.1** **Collection of Candidate Models** 6](#_Toc132115431)

[**5.2** **Hyperparameter Tuning** 7](#_Toc132115432)

[**5.3** **Model Comparison** 8](#_Toc132115433)

[6 Evaluation 10](#_Toc132115434)

[**6.1** **Model Performance Comparison** 10](#_Toc132115435)

[**6.2** **Interpretability Analysis** 10](#_Toc132115436)

[7 Deployment 11](#_Toc132115437)

[8 Conclusion and Future Work 12](#_Toc132115438)

[9 References 13](#_Toc132115439)

# **Introduction**

Breast cancer is a type of cancer that develops in the cells of the breast. It is one of the most common types of cancer among women worldwide, but it can also occur in men (DeSantis et al., 2019). Cancer begins when the cells in the breast tissue start growing uncontrollably, forming a tumor. If left untreated, the tumor can spread to other parts of the body, making it a serious and potentially life-threatening disease. However, early detection and treatment can significantly improve the chances of survival (Łukasiewicz et al., 2021).

Traditional methods of breast cancer diagnosis, such as mammography, ultrasound, and biopsy, have limitations. Machine learning techniques have the potential to overcome some of these limitations by building predictive models for breast cancer diagnosis. These models can assist doctors in making informed decisions and improve the accuracy and efficiency of traditional diagnostic methods (Lehman et al., 2022).

The aim of this project is to build a machine learning model for breast cancer prediction that has high accuracy and interpretability. By analyzing large amounts of data, the model can identify patterns and correlations that may not be apparent to humans. This can provide valuable information to healthcare professionals, helping them make informed decisions about diagnosis, treatment, and patient care (Barth, 2022). A highly accurate and interpretable model can reduce the number of false positives and false negatives, leading to improved patient outcomes.

## **Objectives**

The objective of this project is to build a machine learning model for breast cancer prediction that has high accuracy and interpretability. Specifically, I aim to:

* Explore and preprocess the breast cancer dataset.
* Build and compare different machine learning models, including logistic regression, support vector machines (SVM), decision tree, random forest.
* Evaluate the performance of the models using appropriate metrics, including accuracy, precision, recall, and F1-score.
* Select the best model based on performance and interpretability.
* Deploy the selected model in a web application using Flask

## **Methodology**

I will follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to guide this project (Hotz, 2023). This methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In the Business Understanding phase, I will define the problem and the objectives of the project. In the Data Understanding phase, I will explore and analyze the breast cancer dataset to gain insights into its structure and quality. In the Data Preparation phase, I will preprocess the data to prepare it for Machine Learning modeling. In the Modeling phase, I will build and compare different machine learning models, including logistic regression, support vector machines (SVM), decision tree, random forest. In the Evaluation phase, I will evaluate the performance of the models and select the best model based on performance and interpretability. Finally, in the Deployment phase, I will deploy the selected model in a web application using Flask.

## **Folder Structure of the GitHub Repository**

The Folder Structure of the GitHub Repository consists of different directories and files that organize the code and resources for the breast cancer prediction project. The structure includes directories for data, images, models, notebooks, and templates, as well as files for README, and requirements.

# **Business Understanding**

As mentioned above that the breast cancer is one of the most common forms of cancer among women. Early detection and accurate diagnosis are crucial for successful treatment. However, identifying breast cancer requires extensive expertise and can be time-consuming, leading to a significant delay in treatment (Wang, 2017). Therefore, the development of an accurate and interpretable model for breast cancer prediction is of great importance.

The stakeholders of this project include healthcare professionals, medical researchers, and patients. Healthcare professionals and medical researchers can use the model to support their diagnosis and research. Patients can benefit from the model by receiving early and accurate detection of breast cancer, which can lead to successful treatment (Stolovitzky, 2020).

The primary objective of this project is to develop an accurate and interpretable model for breast cancer prediction. This model should have high accuracy and precision and provide meaningful insights into the decision-making process. Additionally, the model should be easy to interpret for healthcare professionals and patients.

# **Data Understanding**

The data used for this project is obtained from the Breast Cancer Wisconsin (Diagnostic) Dataset available in the UCI Machine Learning Repository. This dataset contains 569 samples, each consisting of features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass (*Breast Cancer Wisconsin (Diagnostic) Data Set*, 2016).

## **Data exploration**

Before building the model, I explored the dataset to gain insights and identify potential issues. I found that the dataset contains no missing values, and the features are on different scales. Therefore, I performed feature scaling using the Scikit-learn StandardScaler() function. This helped to normalize the data and ensure that all input features were on the same scale. This is important because many machine learning algorithms assume that all features are on a similar scale, and features with larger ranges can dominate the model training process, leading to biased results. (Chong, 2021). I will use the fit\_transform() method only on the training data to prevent overfitting. On the other hand, I will use the transform() method on the test data to transform it into the same feature space as the training data, without fitting the selector again. This will ensure that the test data is evaluated using the same feature set as the training data, helping to prevent overfitting.

## **Data quality assessment**

I also performed data quality assessment to ensure the accuracy and consistency of the data. This includes checking for outliers, duplicates, and inconsistencies in the data. I did not find any significant issues with the data.

# **Data Preparation**

To prepare the data for the breast cancer prediction model, I performed several steps including data cleaning, transformation, reduction, and splitting. The details of each step are as follows:

## **Data Cleaning**

I checked the data for missing values and duplicates and removed any such records. In order to improve the performance of the model and reduce unnecessary features, the 'id' and 'Unnamed: 32' columns were removed from the dataset as they were unrelated to the diagnosis prediction. These columns contained irrelevant information and did not contribute to the analysis. By removing these columns, the model was able to focus on more relevant features and improve its accuracy in predicting the diagnosis.

## **Data Transformation**

As a part of the data transformation process, the categorical variable "diagnosis" was converted into a numerical variable using the LabelEncoder() function from the scikit-learn library. This conversion was important for the machine learning model to accurately predict the diagnosis of breast cancer. Machine learning algorithms work with numerical data, and converting categorical variables into numerical ones allows the algorithm to better understand the data and identify patterns and relationships. By encoding the diagnosis variable, the machine learning model can more accurately analyze the data and make informed predictions (Ray, 2022). In this case, we can assign the value of 0 to the "B" category (which represents a benign tumor) and the value of 1 to the "M" category (which represents a malignant tumor). This conversion process was a crucial step in building a reliable and accurate machine learning model for breast cancer prediction. Make it concise citation.

Chart, box and whisker chart

Description automatically generated

As seen from the above box plot, it was found that the data was on different scales. To ensure that each feature contributes equally to the model, the StandardScaler was used to normalize the data. This preprocessing step transforms the data so that it has a mean of 0 and a standard deviation of 1. This allows for better comparison between different features and helps the model to learn more effectively (Verma, 2021). Using standardized data can also help to prevent issues such as numerical instability that can arise when working with features that have very different scales. Therefore, it is a common practice to normalize the data before fitting a machine learning model.

## **Data Reduction**

To reduce the risk of overfitting and improve model efficiency, feature selection is an important step in any machine learning project. Feature selection involves reducing the number of features used in the model by identifying and removing those that are irrelevant or redundant. In this project, feature selection was performed using correlation analysis and SelectKBest, which helped to identify the most important features for predicting the target variable (Charonyktakis, 2022). Correlation analysis was used to identify the features that had the strongest relationship with the target variable, while SelectKBest was used to rank the features based on their individual scores. By reducing the number of features, the model becomes more interpretable and easier to understand, while also reducing the risk of overfitting and improving its performance on new, unseen data (Tracyrenee, 2022).

Chart, bar chart, histogram

Description automatically generated

In the above image, the correlation analysis performed on the dataset showed that certain features have a high correlation with the diagnosis. Specifically, the features 'radius\_mean', 'perimeter\_mean', 'concave points\_mean', 'radius\_worst', 'perimeter\_worst', 'area\_worst', and 'concave points\_worst' were found to have a strong relationship with the diagnosis. This indicates that these features could be crucial predictors in determining whether a patient has a malignant or benign tumor. Therefore, these features were given special attention during the feature selection and hyperparameter optimization processes in the development of the machine learning models.

## **Data Splitting**

Finally, the data was split into training and testing datasets, with an 80:20 split. This allowed me to train the machine learning models on the training dataset and evaluate their performance on the testing dataset.

# **Modeling**

In the modeling section, I developed different machine learning models to predict breast cancer using the breast cancer dataset. Since this is a binary classification problem, I selected multiple machine learning models to build and evaluate the predictive models. My chosen models were logistic regression, support vector machines (SVM), decision trees, and random forests. These models were chosen because they are well-suited for binary classification tasks (Gong, 2022).

Gong, D. (2022, July 12). Top 6 Machine Learning Algorithms for Classification. *Medium*. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

## **Collection of Candidate Models**

*Logistic Regression* Model is a binary classification model that predicts the probability of a certain outcome. This model is particularly useful when the dependent variable is binary, as is the case with the breast cancer dataset. Logistic regression is the preferred regression analysis to use where the dependent variable is dichotomous (binary). Logistic regression is called after the logistic function, which is utilized at the heart of the approach. The logistic function, commonly known as the sigmoid, is an S-shaped curve that can take any real-valued integer and translate it to a range between 0 and 1, but never precisely at those boundaries (Pranckevicius & Marcinkevičius, 2017).

*Support Vector Machine (SVM)* is a plane-based classifier that creates a discrete hyperplane in the training data and compounds’ descriptor space. By solving an optimization problem, SVM finds the optimal hyperplane between two groups of data in the training data (Pranckevicius & Marcinkevičius, 2017).

*Decision Tree* Model is a popular classification and regression model that works by recursively partitioning the input space into smaller regions based on the values of the input features. The model selects the best feature and threshold at each node to split the data and create a tree-like structure that represents the decisions and their possible outcomes. Decision trees are particularly useful when there are complex decision rules that can be easily represented in a tree format (Corporate Finance Institute, 2023).

*Random forests* are an extension of decision trees that aim to reduce overfitting by building an ensemble of trees. Each tree is trained on a bootstrap sample of the data, and at each node in the tree, a random subset of the features is considered for the split. The final prediction is obtained by averaging the predictions of all the trees in the ensemble. Random forests can be highly accurate and robust to noise in the data, making them a good choice for our breast cancer prediction project. However, they can be more difficult to interpret than decision trees (Yiu, 2021).

(Pranckevicius & Marcinkevičius, 2017)

Pranckevicius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, *5*(2). <https://doi.org/10.22364/bjmc.2017.5.2.05>

Corporate Finance Institute. (2023). Decision Tree. *Corporate Finance Institute*. <https://corporatefinanceinstitute.com/resources/data-science/decision-tree/>

Yiu, T. (2021, December 10). Understanding Random Forest - Towards Data Science. *Medium*. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

## **Hyperparameter Tuning**

Hyperparameter tuning is the process of selecting the optimal values of hyperparameters for a machine learning algorithm to enhance its performance. Hyperparameters are predetermined parameters that cannot be directly learned from the data. Tuning these hyperparameters can improve the model's accuracy, generalization, and convergence speed. Hyperparameter tuning is a crucial step in machine learning model development, and it involves optimizing the model's hyperparameters using techniques such as GridSearchCV (Jordan, 2018). In this project, GridSearchCV was used to perform hyperparameter tuning for the Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest models, and the best hyperparameters were selected based on their accuracy score to build the final model.

**Table 1:** Hyperparameters optimization grid search results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Featuers Extraction** | **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Best Parameters** |
| All Features | Logistic\_Regression | 97,37% | 97,62% | 95,35% | 96,47% | {'C': 1.623776739188721, 'penalty': 'l2'} |
| SVM | 97,37% | 97,62% | 95,35% | 96,47% | {'C': 0.23357214690901212, 'kernel': 'linear'} |
| Decision\_Tree | 94,74% | 97,44% | 88,37% | 92,68% | {'criterion': 'entropy', 'max\_depth': 4} |
| Random\_Forest | 96,49% | 97,56% | 93,02% | 95,24% | {'max\_depth': None, 'n\_estimators': 200} |
| SelectKBest 7 Features | **Logistic\_Regression** | **98,25%** | 97,67% | 97,67% | 97,67% | **{'C': 11.288378916846883, 'penalty': 'l2'}** |
| **SVM** | **98,25%** | 97,67% | 97,67% | 97,67% | **{'C': 1.623776739188721, 'kernel': 'linear'}** |
| Decision\_Tree | 93,86% | 95,00% | 88,37% | 91,57% | {'criterion': 'gini', 'max\_depth': 4} |
| Random\_Forest | 95,61% | 95,24% | 93,02% | 94,12% | {'max\_depth': 6, 'n\_estimators': 50} |

Table 1 presents the findings of feature extraction and hyperparameter optimization in this study. The first row of the table demonstrates the performance of the machine learning models when all the available features were used. The second row presents the models' performance when only the top 7 features were selected using the SelectKBest method. The best hyperparameters were selected using GridSearchCV, with the best parameters for each model listed in the last column of the table.

Logistic Regression and SVM models were found to perform best for both feature extraction methods. The models achieved an accuracy of 97.37% when using all the features and 98.25% when using SelectKBest. On the other hand, the Decision Tree and Random Forest models had lower accuracy scores compared to the Logistic Regression and SVM models. The results indicate that the feature selection method can have a significant impact on model performance.

When using SelectKBest, the Logistic Regression and SVM models had identical precision, recall, and F1 score values, indicating that both models performed equally well. The Decision Tree model, however, had a lower recall score, implying that it struggled with identifying all the true positive cases. The Random Forest model had a lower precision score, indicating that it predicted more false positives.

## **Model Comparison**

After comparing the results from the trained models. it was observed that logistic regression and SVM performed significantly well in both feature extraction methods. Specifically, logistic regression and SVM achieved an accuracy of 97.37% when using all the features and 98.25% when using SelectKBest. This indicates that both are suitable algorithms for the given dataset and problem.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

To further validate the performance of logistic regression and svm, the following figures were plotted to visualize the model's performance. These figures showed that logistic regression and svm had a high area under the curve (AUC) score, which is a measure of the model's ability to distinguish between positive and negative cases. In addition, the precision-recall curve for logistic regression showed a high precision and recall values, indicating that the model was able to identify most of the true positive cases while minimizing the false positives.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

The confusion matrix shows the number of true positives, true negatives, false positives, and false negatives for each class, which provides a more detailed picture of the model's performance. Both Logistic Regression and SVM models had a high number of true positives and true negatives, indicating that they were able to correctly classify most of the test data.

Chart

Description automatically generated Chart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated

In contrast, the Decision Tree and Random Forest models had a higher number of false positives and false negatives, indicating that they had more difficulty correctly classifying the test data.

Therefore, the confusion matrix supported the findings that the Logistic Regression and SVM had the best performance among all the models tested.

# **Evaluation**

In this project, I will evaluate the best model based on two main parameters: performance measures and interpretability.

## **Model Performance Comparison**

Table 1 provides a comparison of the performance of four different models, namely Logistic Regression, SVM, Decision Tree, and Random Forest, using two feature extraction methods, i.e., all features and SelectKBest top 7 features. Overall, the models performed better when using SelectKBest, with Logistic Regression and SVM achieving the highest accuracy of 98.25% and the best precision, recall, and F1 score of 97.67%. These models also had the best parameters, with Logistic Regression having a C value of 11.288 and an L2 penalty, and SVM having a C value of 1.624 and a linear kernel.

In contrast, the Decision Tree and Random Forest models had lower performance scores, with Decision Tree having the lowest recall score and Random Forest having the lowest precision score. The Decision Tree model had a maximum depth of 4 and used entropy as the criterion for selecting features, while the Random Forest model had a maximum depth of 6 and used 50 estimators. These models did not perform as well as the Logistic Regression and SVM models, which suggests that the latter models are better suited for this classification problem.

## **Interpretability Analysis**

After careful analysis and comparison of the results from various models, logistic regression was selected as the final model for this project. One of the main reasons for selecting logistic regression was its high accuracy and performance, especially when using SelectKBest feature extraction method. The model achieved an accuracy of 98.25%, which is one of the highest among all the models. Additionally, both logistic regression and SVM showed similar performance in terms of precision, recall, and F1-score when using SelectKBest. However, logistic regression had a simpler and more interpretable model structure compared to SVM, which is important in understanding the model's behavior and making it more explainable to stakeholders.

Logistic regression is considered an easy-to-interpret model because it provides a clear understanding of the relationship between the input variables and the output variable. The model estimates the probability of a binary outcome, given the input variables, by fitting a logistic curve to the data. The coefficients of the logistic regression model represent the change in the log-odds of the outcome for a one-unit change in the corresponding input variable. This means that we can interpret the coefficients as the impact of each input variable on the probability of the outcome (Mondal, 2023).

Furthermore, logistic regression models can be easily visualized using graphs, such as bar charts or scatter plots, making it easy for non-technical stakeholders to understand the results. The coefficients can also be converted into odds ratios or probability ratios, which can provide a more intuitive understanding of the impact of each input variable on the outcome (Rashida, 2020). In addition, logistic regression models can be easily updated with new data, making it suitable for real-time applications.

# **Deployment**

In this project, I developed a machine learning model using logistic regression to predict whether a breast tumor is benign or malignant based on several features of the tumor. After analyzing the dataset, it was determined that several features had a strong relationship with the diagnosis, including 'radius\_mean', 'perimeter\_mean', 'concave points\_mean', 'radius\_worst', 'perimeter\_worst', 'area\_worst', and 'concave points\_worst'. Therefore, these features were given special consideration during the feature selection and hyperparameter optimization processes.

After training the logistic regression model using the selected features, I deployed the model as an application using Flask. The web interface of the application allows users to input values for each of the selected features and receive a prediction of whether the tumor is benign or malignant.

To use the application, users simply need to enter values for the seven selected features ('radius\_mean', 'perimeter\_mean', 'concave points\_mean', 'radius\_worst', 'perimeter\_worst', 'area\_worst', and 'concave points\_worst') and click the 'Predict' button. The model will then generate a prediction and display the results on the web interface.

The given image shows the web interface of the application developed using Flask. Users can access this interface to make predictions using the trained logistic regression model. The interface contains a form where users can input the values of the seven selected features: 'radius\_mean', 'perimeter\_mean', 'concave points\_mean', 'radius\_worst', 'perimeter\_worst', 'area\_worst', and 'concave points\_worst'. Once the user submits the form, the model processes the input and generates a prediction for the diagnosis of the tumor as either malignant or benign. The result of the prediction is displayed on the web interface. This web interface provides a user-friendly and easily accessible means for medical professionals and researchers to quickly obtain predictions for tumor diagnoses based on a set of relevant features.

![Graphical user interface, application, table

Description automatically generated with medium confidence]()

# **Conclusion and Future Work**

In conclusion, this project aimed to develop a machine learning model to accurately classify breast cancer tumors as malignant or benign. The dataset was preprocessed by removing irrelevant features and normalizing the data using StandardScaler. Feature selection was performed using correlation analysis and SelectKBest to reduce the number of features and prevent overfitting. The final model was a logistic regression classifier trained on the selected features and achieved an accuracy rate of 98.25%.

One of the strengths of the developed model is its easy interpretability. Logistic regression provides a simple and intuitive way to understand the relationship between the input features and the target variable. This can be especially useful in medical applications where the interpretability of the model is critical for making informed decisions. However, there is still room for improvement in terms of the model's performance. Future work could focus on exploring more complex models or ensemble methods to achieve even higher accuracy rates.

Another area for future work is the collection and analysis of more diverse and representative datasets. The breast cancer dataset used in this project was limited to a specific geographic region and may not generalize well to other populations. In addition, the dataset only included a relatively small number of samples, which could limit the model's ability to capture the full complexity of the problem. By collecting larger and more diverse datasets, it may be possible to develop more robust and accurate models that can be applied to a wider range of patients and populations.

Overall, this project demonstrates the potential of machine learning in medical applications and highlights the importance of careful data preprocessing and feature selection. With further development and refinement, machine learning models could become an invaluable tool for medical professionals in diagnosing and treating a wide range of diseases and conditions.

# **References**

DeSantis, C., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. A., Sauer, A. G., Jemal, A., & Siegel, R. L. (2019, November 1). *Breast cancer statistics, 2019*. CA: A Cancer Journal for Clinicians. https://acsjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.3322/caac.21583

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review. *Cancers*, *13*(17), 4287. <https://doi.org/10.3390/cancers13174287>

Lehman, C. D., Mercaldo, S. F., Lamb, L. R., King, T. A., Ellisen, L. W., Specht, M. C., & Tamimi, R. M. (2022). Deep Learning vs Traditional Breast Cancer Risk Models to Support Risk-Based Mammography Screening. *Journal of the National Cancer Institute*, *114*(10), 1355–1363. https://doi.org/10.1093/jnci/djac142

Barth, S. (2022, November 15). *Machine Learning in Healthcare - Benefits & Use Cases*. ForeSee Medical. https://www.foreseemed.com/blog/machine-learning-in-healthcare#:~:text=Machine%20learning%20in%20healthcare%20can,that%20indicate%20a%20particular%20disease.

Hotz, N. (2023, January 19). *What is CRISP DM? - Data Science Process Alliance*. Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Wang, L. (2017). Early Diagnosis of Breast Cancer. *Sensors*, *17*(7), 1572. https://doi.org/10.3390/s17071572

Stolovitzky, G. (2020, May 1). *The potential benefits of AI for breast cancer detection*. IBM Research Blog. https://www.ibm.com/blogs/research/2020/03/benefits-ai-for-breast-cancer-detection/

*Breast Cancer Wisconsin (Diagnostic) Data Set*. (2016, September 25). Kaggle. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Chong, J. (2021, December 25). *What is Feature Scaling & Why is it Important in Machine Learning?* Medium. https://towardsdatascience.com/what-is-feature-scaling-why-is-it-important-in-machine-learning-2854ae877048

Ray, S. (2022, June 16). *Simple Methods to deal with Categorical Variables in Predictive Modeling*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/

Verma, Y. (2021, October 7). *Why Data Scaling is important in Machine Learning & How to effectively do it*. Analytics India Magazine. https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/#:~:text=So%20if%20the%20data%20in,between%20them%20will%20be%20lower.

Charonyktakis, P. (2022, January 5). *Do We Really Need Feature Selection in a Data Analysis Pipeline?* Medium. <https://towardsdatascience.com/do-we-really-need-feature-selection-in-a-data-analysis-pipeline-dc8401621c6c>

Tracyrenee, T. (2022, January 6). *How to select features using SelectKBest in Python - MLearning.ai - Medium*. Medium. https://medium.com/mlearning-ai/how-to-select-features-using-selectkbest-in-python-c5a5239969f0

Jordan, J. (2018, December 5). *Hyperparameter tuning for machine learning models.* Jeremy Jordan. https://www.jeremyjordan.me/hyperparameter-tuning/

Mondal, S. (2023, February 14). *Regression Analysis | Beginners Comprehensive Guide (Updated 2023)*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/>

Rashida. (2020, August 3). *Logistic Regression Model, Analysis, Visualization, And Prediction – Regenerative*. https://regenerativetoday.com/logistic-regression-model-analysis-visualization-and-prediction/