RevBank Ltd.

Artificial Intelligence (AI) Solution Implementation Report

Table of Contents

[1. Introduction 3](#_Toc172482967)

[2. Business Context 3](#_Toc172482968)

[3. Justification of the Choice of Dataset 3](#_Toc172482969)

[3.1 Data Source 4](#_Toc172482970)

[3.2 Data Preparation 4](#_Toc172482971)

[4. Approach to Developing the Prediction Model 7](#_Toc172482972)

[4.1 Cleaning the Data 7](#_Toc172482973)

[4.2 Data Validation in WEKA 8](#_Toc172482974)

[4.3 Results and Analysis 8](#_Toc172482975)

[6. Machine Learning Algorithms Used 9](#_Toc172482976)

[7. Analysis of Model Outputs and Testing Evidence 9](#_Toc172482977)

[7.1 Decision Tree 10](#_Toc172482978)

[7.1.1 Cross-Validation: 10](#_Toc172482979)

[7.1.2 Percentage Split 10](#_Toc172482980)

[7.2 Logistics Regression 11](#_Toc172482981)

[7.2.1 Cross Validation 11](#_Toc172482982)

[7.2.2 Percentage Split 11](#_Toc172482983)

[7.3 Support Vector Machine (SVM) 12](#_Toc172482984)

[7.3.1 Cross Validation 12](#_Toc172482985)

[7.3.2 Percentage Split 13](#_Toc172482986)

[7.5 Naïve Bayes 14](#_Toc172482987)

[7.3.3 Cross-Validation 14](#_Toc172482988)

[7.3.4 Percentage Split 16](#_Toc172482989)

[8. Application to Identified Business Problem 16](#_Toc172482990)

[9. Conclusion 16](#_Toc172482991)

[10. References 17](#_Toc172482992)

[11. Appendix 18](#_Toc172482993)

# Introduction

RevBank Ltd, a startup in the Financial Services industry, offers a range of services, including Personal and Business Loans, Credit Cards, Investment Advice, and Financial Planning. In the fast-paced, knowledge-intensive financial services market, timely and accurate information is essential for effective decision-making. Embracing AI can enhance our services, streamline operations, and provide a competitive advantage. AI technologies are not just a short-term solution; they are adaptable and can evolve with the market, ensuring RevBank Ltd remains competitive in the long term.

# Business Context

The financial services industry is a vital component of the global economy, encompassing many businesses, including banks, investment firms, insurance companies, and financial technology (FinTech) startups. This sector is responsible for managing the flow of money, facilitating investments, providing loans, and offering insurance and risk management solutions. As a local startup finance company, RevBank Ltd is deeply committed to staying informed about this industry's current landscape and future trends. It is a key to our strategic positioning and growth.

Machine learning revolutionises the banking industry, particularly detecting scams, frauds, or defaulters. With the help of pattern recognition algorithms, this industry makes complex decisions daily. Financial institutions heavily use machine learning's exceptional ability to recognise anomalies and patterns to ensure proper oversight. It provides a solid sense of security and confidence in the industry by significantly reducing the risk of financial fraud and default.

# Justification of the Choice of Dataset

The Credit Risk Dataset is a powerful tool in the financial services industry. It provides a comprehensive collection of information for analysing and predicting individuals' creditworthiness. This dataset, which includes various demographic, financial, and behavioural attributes of borrowers, is a crucial resource for financial institutions in assessing borrowers' likelihood of defaulting on loan obligations. For instance, it includes data on income, employment status, credit history, and other factors that can influence creditworthiness.

The Credit Card eligibility criteria are an integral part of the broader credit risk assessment process, and the credit card eligibility data has been taken into consideration for the purpose of the assessment.

## 3.1 Data Source

This dataset, sourced from Kaggle.com, contains various fields that describe the characteristics of loan applicants. The Dataset Link used is as follows: <https://www.kaggle.com/datasets/sheemazain/credit-card-eligibility-data/data>.

The dataset used in this study contains a substantial 9709 instances. This large size is significant as it allows for a comprehensive analysis and ensures the reliability of the study's findings.

## 3.2 Data Preparation

The dataset includes the following attributes. Short and Long description added for providing explanation of each attribute.

|  |  |  |
| --- | --- | --- |
| Attribute | Short Description | Full Description |
| ID | Unique identifier for each applicant. | Each applicant is assigned a unique number or code to distinguish them from others. |
| Gender | The gender of the applicant | The sex of the applicant is typically categorised as male, female, or other. |
| Own\_Car | Indicates if the applicant owns a car. | A binary field (yes/no or true/false) that specifies whether the applicant owns a car. |
| Own\_Property | Indicates if the applicant owns the property. | A binary field (yes/no or true/false) that specifies whether the applicant owns real estate or property. |
| Work\_Phone | Indicates if the applicant has a work phone | A binary field (yes/no or true/false) that specifies whether the applicant has a phone provided by their workplace. |
| Phone | Indicates if the applicant has a personal phone. | A binary field (yes/no or true/false) that specifies whether the applicant has a personal phone. |
| Email | Indicates if the applicant has an email address. | A binary field (yes/no or true/false) that specifies whether the applicant has an email address. |
| Unemployed | Indicates if the applicant is unemployed. | A binary field (yes/no or true/false) that specifies whether the applicant is currently without employment. |
| Num\_Children | Number of children the applicant has. | An integer value representing the number of children dependent on the applicant. |
| Num\_Family | Total number of family members. | An integer value representing the total number of family members in the applicant's household, including the applicant. |
| Account\_Length | The duration for which the applicant has held their bank account. | The length of time (typically in months or years) the applicant has maintained their current bank account. |
| Total\_Income | The borrower's monthly or annual income. | The applicant's total income is specified on a monthly or annual basis. |
| Age | Age of the applicant | The number of years since the applicant was born. |
| Years\_employment | Number of years the applicant has been employed. | The duration, in years, that the applicant has been continuously employed. |
| Income\_Type | Type of income the applicant receives. | The categories of income sources include salary, business, pension, etc. |
| Education\_Type | The highest level of education achieved by the applicant. | The highest educational qualification attained by the applicant, such as high school, bachelor's degree, master's degree, etc. |
| Family\_status | Marital status of the borrower. | The applicant's marital status, such as single, married, divorced, etc. |
| Housing\_type | Whether the applicant rents or owns their home. | The type of housing arrangement, such as renting, owning, living with parents, etc. |
| Occupation\_type | The occupation or job type of the applicant. | The field or type of job in which the applicant is employed, such as healthcare, education, engineering, etc. |
| Target | Outcome variable indicating the target status (e.g., default or no default). | A field used in modelling to indicate the target outcome, often a binary value representing whether the applicant defaulted on a loan (e.g., 1 for default, 0 for no default). |

The overview of data attributes is shown in WEKA as follows

A screenshot of a computer

Description automatically generated

# Approach to Developing the Prediction Model

The approachtoDeveloping the Prediction Model involves several steps, including data cleaning, feature selection, and model evaluation, to develop an effective predictive model for credit card eligibility.

## 4.1 Cleaning the Data

To construct an effective predictive model, determining which fields provide valuable information and which can be disregarded is crucial. Given the nature of predictive modelling and the provided fields, the following fields can be ignored:

* ID: This is a unique identifier for each applicant and does not provide any predictive value.
* Email: Having an email address does not directly influence credit card eligibility.
* Work\_Phone: Like email, having a work phone does not directly impact credit card eligibility.
* Phone: The presence of a personal phone also does not significantly contribute to determining credit card eligibility.
* Gender: Including gender may introduce gender bias into the model, leading to unfair treatment of applicants based on their gender. At RevBank Ltd, we are committed to fairness and equality. We ensure that our credit card eligibility process does not discriminate based on protected attributes such as gender. This commitment is a reassurance to our customers and a testament to our ethical standards.

These fields can be ignored as they do not contain information that would help in predicting whether an applicant is eligible for a credit card:

* ID: Unique identifier for each applicant.
* Email: Binary indicator if the applicant has an email address (1 = Yes, 0 = No).
* Work\_Phone: Binary indicator if the applicant has a work phone (1 = Yes, 0 = No).
* Phone: Binary indicator if the applicant has a personal phone (1 = Yes, 0 = No)
* Gender: The gender of the applicant (e.g., Male, Female).

## 4.2 Data Validation in WEKA

The remaining fields contain relevant information that could influence the prediction of credit card eligibility and should be retained in the dataset.

1. Load the Dataset: Open WEKA and load the dataset into the Explorer interface.
2. Preprocess the Data: Use WEKA's preprocessing tools to clean and prepare the data.
3. Select Classifier: Choose a classifier (e.g., Naive Bayes or SVM) from the "Classify" tab.
4. Test Options: Select the Test options as “Cross-Validation” and “Percentage Split”
5. Evaluate Classifier: Run the classifier on the selected dataset.

After the classifier finishes running, WEKA summarises the results, including various performance metrics.

## 4.3 Results and Analysis

After the classifier finishes running, WEKA summarises the results, including various performance metrics.

* **Accuracy:** This metric measures the proportion of correct predictions made by the model compared to the total number of predictions. It reflects the overall correctness of the model's predictions. (Brownlee, 2019).
* **Precision:** Precision evaluates the proportion of accurate optimistic predictions out of all positive predictions made by the model. It focuses on the accuracy of the optimistic predictions (Brownlee, 2019).
* **Recall:** Also referred to as sensitivity, recall assesses the model's ability to identify all actual positive instances. It calculates the proportion of true positives among all actual positives (Brownlee, 2019).
* **F1 Score:** The F1 Score combines precision and recall into a single metric by calculating their harmonic mean. It provides a balanced measure of a model's performance by accounting for false positives and negatives (Brownlee, 2019).
* **Refer to the Appendix for more details**

# Machine Learning Algorithms Used

The applicant's financial status and personal history are key determinants in credit card approval decisions. The evaluation process is designed to identify, assess, and mitigate potential financial and other risks that could result in losses for the lending institution. These risks encompass both financial losses from approving high-risk candidates and commercial losses from rejecting acceptable candidates. Credit card applications are approved or rejected based on specific applicant attributes listed on the application form using a credit approval system. AI, particularly machine learning algorithms, plays a significant role in this process by analysing these attributes and predicting the likelihood of an applicant's credit card eligibility.

This thesis aimed to compare the effectiveness of four machine learning algorithms in predicting credit card eligibility.

* Decision Tree
* Logistics Regression
* Support Vector Machine (SVM)
* Naive Bayes

Refer to the appendix for description of the algorithms.

# Analysis of Model Outputs and Testing Evidence

The data following test evidence represent screenshots captured from the WEKA tool using the Test options Use training test and Percentage split.

* Cross-validation is a technique where the dataset is divided into k subsets (folds), and the model is trained and tested k times. Each time, one of the k subsets is used as the test set, and the remaining k-1 subsets are used as the training set. The result is the average of the k results (Bishop, 2006).
* The percentage split involves dividing the dataset into two parts: one part is used for training, and the other part is used for testing. Common splits are 70-30 or 80-20 (Bishop, 2006). In this assessment 70-30 has been used.

## Decision Tree

## **Cross-Validation:**

A screenshot of a computer

Description automatically generated

The test option has Accuracy: 86.7%, Precision: 86.8%, Recall: 100% and F1 Score: 92.9%

## **Percentage Split**

A screenshot of a computer

Description automatically generated

The test option has Accuracy: 86.8%, Precision: 86.8%, Recall: 100% and F1 Score: 92.9%

## Logistics Regression

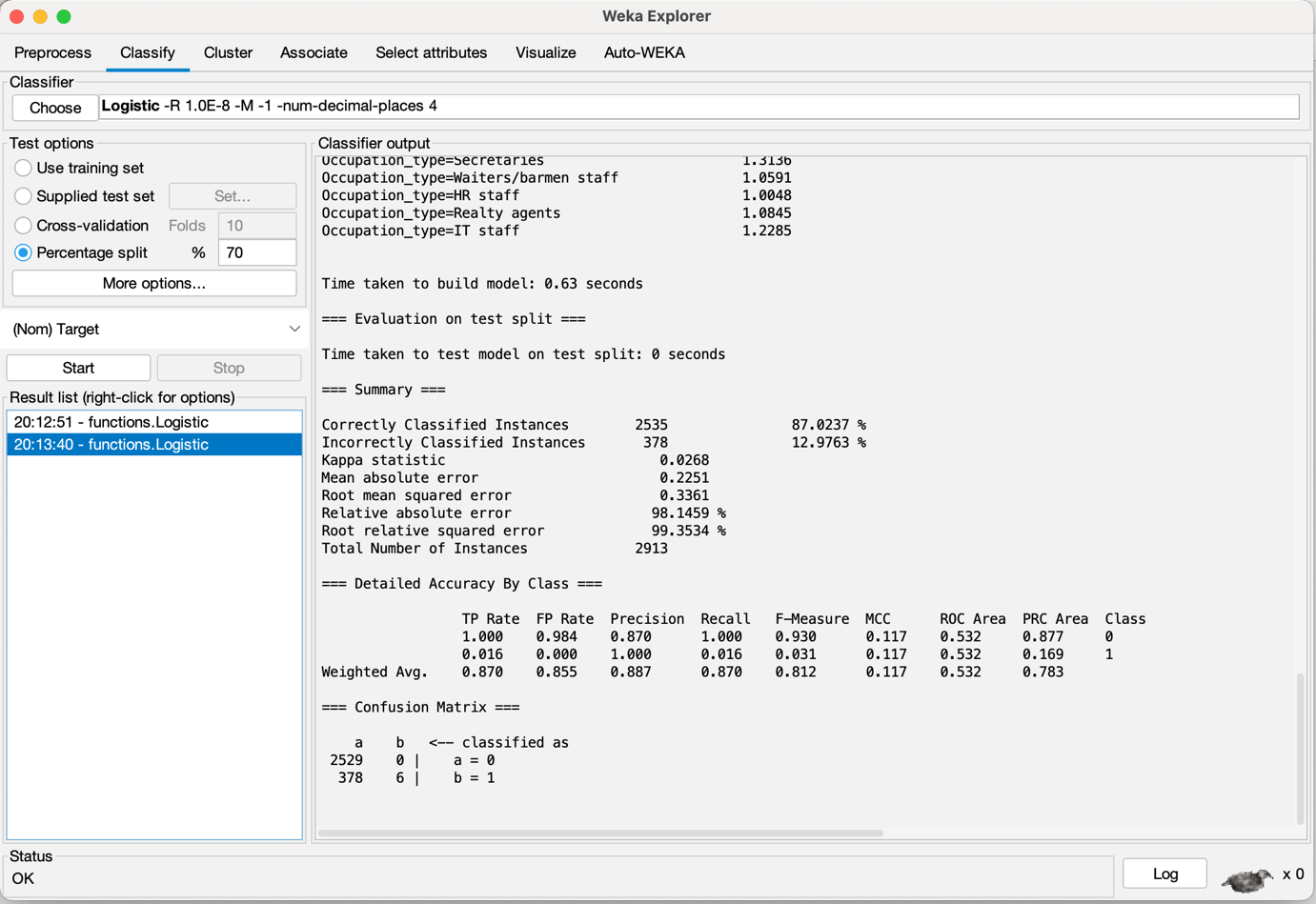
## **Cross Validation**

A screenshot of a computer

Description automatically generated

The test option has Accuracy: 86.9%, Precision: 86.9%, Recall: 100% and F1 Score: 93%

## **Percentage Split**



The test option has Accuracy: 87%, Precision: 87%, Recall: 100% and F1 Score: 93%

## Support Vector Machine (SVM)

## **Cross Validation**

A screenshot of a computer

Description automatically generated

The test option has Accuracy: 86.9%, Precision: 86.9%, Recall: 100% and F1 Score: 93%

## **Percentage Split**

A screenshot of a computer

Description automatically generated

The test option has Accuracy: 87%, Precision: 87%, Recall: 100% and F1 Score: 93%

## Naïve Bayes

## **Cross-Validation**

A screenshot of a computer

Description automatically generated

The test option has Accuracy: 86.7%, Precision: 86.8%, Recall: 100% and F1 Score: 92.9%

## **Percentage Split**

A screenshot of a computer

Description automatically generated

The test option has Accuracy: 86.8%, Precision: 86.8%, Recall: 100% and F1 Score: 92.9%

Here is the comparison of the four models in a tabular format:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Accuracy | Precision | Recall | F1 Score |
| Decision Tree | 86.7% | 86.8% | 100% | 92.9% |
| Logistic Regression | 87% | 87% | 100% | 93% |
| Support Vector Machine (SVM) | 87% | 87% | 100% | 93% |
| Naïve Bayes | 86.8% | 86.8% | 100% | 92.9% |

Here's the bar graph comparing the performance of the four classification models:

# Application to Identified Business Problem

By employing Logistic Regression or SVM, the financial institution can accurately predict credit card eligibility, achieving a high level of accuracy and a balanced performance in terms of precision and recall. This strategy will support well-informed decision-making, enhance customer targeting, and reduce risks.

### Summary of Model Performance:

Logistic Regression and SVM exhibit the highest accuracy, precision, and F1 score, with perfect recall. This optimal balance ensures that the institution can reliably identify eligible customers while minimising the risk of approving those who are ineligible.

### Steps for Implementing Logistic Regression and SVM:

1. **Data Preparation**:
   * Clean and preprocess the data.
   * Handle any missing values.
   * Scale features appropriately.
2. **Model Training:**
   * Train Logistic Regression and SVM models using the training dataset.
3. **Model Evaluation:**
   * Evaluate the models with cross-validation to ensure robust performance metrics.
4. **Model Deployment:**
   * Deploy the model with the best performance in the production environment to make real-time predictions.
5. **Monitoring and Maintenance:**
   * Continuously monitor the model's performance.
   * Retrain the model periodically to maintain its accuracy and relevance.

# 9. Conclusion

In conclusion, AI-driven credit risk analysis tools empower analysts to conduct faster, more comprehensive analyses by leveraging diverse data sources, enhancing scenario analysis, and streamlining communication. A semi-automatic approach balances the benefits of AI with the need for analyst control. As the financial industry evolves, adaptable AI platforms will support well-informed decision-making and effective credit risk management.

The applicant's financial status and personal history are crucial factors in credit card approval. The evaluation process aims to identify, assess, and mitigate potential financial and other risks that could lead to losses for the lending institution. These risks include financial losses from approving a high-risk candidate and commercial losses from not approving a viable candidate.

Credit card application decisions are made using a credit approval system based on specific applicant attributes listed on the application form. This thesis aimed to compare the accuracy of four classification models: Decision Tree, Logistics Regression, Support Vector Machine (SVM), and Naive Bayes in predicting credit card approval. The dataset, sourced from Kaggle, is publicly accessible.

Accuracy was used as the primary performance metric to identify the most accurate algorithm among the selected machine learning models. The study found that Logistic Regression and SVM both appear to be the best methods based on the provided metrics. They both have the highest accuracy (86.9%), precision (86.9%), and F1 score (93%) while maintaining a perfect recall (100%).

# 10. References

* Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
* Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
* Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
* Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).
* Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
* Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
* Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition. Retrieved from <https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf>
* Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.
* Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann.
* Kaggle. (n.d.). Credit Card Eligibility Data. Retrieved from <https://www.kaggle.com/datasets/sheemazain/credit-card-eligibility-data/data>
* Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.
* Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
* Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.
* Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
* Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
* Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). **CRISP-DM 1.0: Step-by-step data mining guide.** SPSS Inc.
* Brownlee, J. (2019) Classification Accuracy is Not Enough: More Performance Measures You Can Use. Available at: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/ (Accessed: 21 July 2024).
* Kuhn, M. and Johnson, K. (2013) Applied Predictive Modeling. New York: Springer.

# 11. Appendix

* **Decision trees** are highly adaptable tools used in both classification and regression tasks. The process begins at the root node and progresses through the tree until reaching the leaf nodes, where a prediction is made. Constructing the tree involves identifying optimal split points to enhance predictions, continuing until the tree reaches a set depth. Decision trees are valuable for elucidating the decision-making process and capturing non-linear relationships. They are easy to understand and interpret, and they handle both numerical and categorical data effectively (Quinlan, 1986).
* **Logistic regression** is a statistical method for analysing datasets with one or more independent variables influencing an outcome, typically binary (e.g., yes/no, true/false). Unlike linear regression, which predicts continuous values, logistic regression estimates probabilities corresponding to binary outcomes (Hosmer et al., 2013).
* **Naive Bayes** is a straightforward, yet powerful probabilistic classifier based on Bayes' theorem. It assumes strong independence between features and is especially effective for large datasets. It works well with both binary and multiclass classification tasks (Rish, 2001).
* **Random Forest** is an ensemble learning technique that generates multiple decision trees during training, producing a class mode (classification) or mean prediction (regression) as output. It is resilient against overfitting and efficiently handles large datasets with high dimensionality (Breiman, 2001).
* **Support Vector Machines (SVM)** are supervised learning models for classification and regression. They operate by finding the optimal hyperplane that separates different classes within a dataset. SVMs perform well in high-dimensional spaces and can handle linear and non-linear classification using kernel tricks (Cortes and Vapnik, 1995).
* **WEKA (Waikato Environment for Knowledge Analysis)** is a widely used open-source machine learning and data mining software suite. It offers a comprehensive array of tools for data preprocessing, classification, regression, clustering, and visualisation (Witten et al., 2011).
* The **CRISP-DM** (Cross-Industry Standard Process for Data Mining) methodology is a widely used framework for managing data mining and analytics projects. It provides a structured approach to planning, organizing, and executing data mining tasks. The methodology consists of six phases, each with specific tasks and deliverables that guide the data mining process from understanding the business problem to deploying the final model (Chapman et al., 2000).
* **Accuracy:** Accuracy represents the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances. It provides an overall measure of the model's performance.

The formula for Accuracy is TP+TN / TP+TN+FP+FN

A higher accuracy indicates better performance. However, accuracy can be misleading if the dataset is imbalanced, as it may not reflect the model's ability to identify all classes effectively (Kuhn and Johnson, 2013).

* **Precision: Precision measures the proportion of true positive predictions among all positive predictions made by the model. It focuses on the quality of the positive predictions.**

**The formula for Precision is TP / TP+FP**

High precision suggests that the model has a low rate of false positives, meaning that when it predicts a positive class, it is likely to be correct (Kuhn and Johnson, 2013).

* **Recall: Recall assesses the model’s ability to correctly identify all actual positive instances. It indicates how well the model captures all positive cases**

The formula for Recall is TP / TP + FN

High recall indicates that the model effectively identifies most of the positive instances, thereby minimizing the number of false negatives (Kuhn and Johnson, 2013).

* **F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that accounts for both false positives and false negatives.**

The formula for F1 Score = 2 X (Precision X Recall / Precision + Recall)

A high F1 Score reflects a good balance between precision and recall, making it useful for evaluating models where both false positives and false negatives are important (Kuhn and Johnson, 2013).