Predicting Bacterial Harmfulness: An Evaluation of Machine Learning Models for Classification Based on Family and Environment

by

Ellyanna Hyman

In partial fulfilment of the requirements for the
degree of
MSc in
Artificial Intelligence and Applications

Department of
Computer and Information Sciences

August, 2024

Abstract

This study explores the use of several supervised machine learning techniques in classification task, to predict the harmfulness of bacterial species based on their family and environment of origin. Due to the increasing complexity of biological data, scientists aim to further integrate machine learning techniques in their identification methods to enhance efficiency, reduce costs, and increase accessibility.

The dataset used comprises of 200 samples, with which we employed five machine learning models, namely: Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbours, and Multinomial Naïve Bayes. These models were trained to classify bacteria as either harmful or non-harmful to humans. The objectives of this study were to develop and train the predictive models, evaluate their performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC scores, and to optimise their performance through hyperparameter tuning using GridSearchCV. The key findings revealed that the KNN emerged as the top-performing model, showcasing the highest accuracy and stable AUC scores post-tuning. Random forest and logistic regression showed slight decreases in AUC, however still remained reliable based on other metrics. The support vector machine model experienced the largest decline in performance, highlighting the need for further investigation into an optimisation strategy. Multinomial naïve bayes maintained its performance throughout the study.

The study faced limitations, including a relatively small dataset size and limited diversity, which constrained the generalisability of the models. Additionally, the optimisation process did not explore a full range of possible parameters. Future research should focus on expansion of the dataset, incorporating diverse bacterial environments, and employing other machine learning techniques such as deep learning and unsupervised methods. These improvements may enhance model accuracy and provide a deeper understanding into bacterial harmfulness, ultimately benefitting public health and our environment.

Declaration

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc in Software Development of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself.

Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

(please tick) Yes [X] No [ ]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices is 10077.

I confirm that I wish this to be assessed as a Type 5 Dissertation.

Signature: Ellyanna Hyman

Date: 08/08/2024

Contents

1. INTRODUCTION 1
1.1 Background and Context 1
1.2 Problem Statement 2
1.3 Significance of the Study 2
1.4 Scope of the Study 3
1.4.1 Research Questions and Objectives 3
1.4.2 Project Structure 3
1.4.3 Key findings 5
2. LITERATURE REVIEW 6
2.1 Traditional Methods for Bacterial Identification 6
2.2 Integration of Machine Learning in Microbiology 7
2.3 Machine Learning Techniques in a Biological Context 8
3.METHODOLOGY 10
3.1 Data Source 10
3.2 Data Preprocessing 13
3.2.1 Count Vectorisation 14
3.2.2 Term Frequency-Inverse Document Frequency 14
3.2.3 Label Encoding 15
3.2.4 Data Splitting 15
3.3 Model Selection 16
3.4 Training and Evaluation 18
3.5 Optimisation 18
4. ANALYSIS 21
4.1 Overview 21
4.2 Initial Model Performance 22
4.3 Random Forest 22
4.4 Logistic Regression 23
4.5 Support Vector Machine 24
4.6 K-Nearest Neighbours 26
4.7 Multinomial Naïve Bayes 27
4.8 Comparison 29
4.9 Hyper-parameter tuning 29
5. CONCLUSION 34
5.1 Key Insights 34
5.2 Conclusion of Objectives 35
5.3 Limitations 36

5.4 Recommendations for further research 37
REFERENCES 38

1. INTRODUCTION

1.1 Background and Context

Microbiology, the study of microorganisms, confronts significant challenges in identifying and classifying bacteria with high accuracy and efficiency, due to their vast diversity and complex interactions with the environment. These challenges are especially crucial to address, as they are essential to our understanding of diverse aspects of life. They particularly play a crucial role in public health, agriculture, and environmental management (Xie et al., 2018).

In recent years, computational advancements have greatly enhanced our ability to understand complex biological data (Peiffer-Smadja et al., 2020; Amgarten et al., 2018). Machine learning, in particular, has emerged as a powerful tool in this field of study, enabling development and assessment of algorithms to identify, classify and discern patterns in large datasets. Examples of machine applications in biology include medical image analysis, healthcare informatics, and cancer genomics (Zou et al., 2017). Through leveraging machine learning techniques, researchers can analyse vast amounts of data are a more efficient degree, leading to faster and more accurate predictions of microbial classification such as bacteria harmfulness.

Bacterial cells, known for abundancy and uniqueness, exhibit a vast array of effects on their hosts and environment, spanning from beneficial, pathogenic, or even non-effective. Accurate identification of harmful bacterial strains is vital for preventing infections and managing disease, as well as protecting the ecosystem. Traditional methods used in the identification of bacteria, such as Polymerase Chain Reaction (PCR) and culture-based techniques, may be time-consuming and costly (Tabit 2016). These procedures require expert knowledge and substantial resources, which may result in delaying the development of crucial information needed for timely decision-making in a clinical and environmental context. Consequently, the application of machine learning techniques may provide a more efficient and scalable approach to predict bacterial harm. By analysing large datasets containing information about bacterial strains and their environment, machine learning models may be able to identify patterns more rapidly and make predictions with high accuracy. This will not only reduce costs associated with more traditional methods, but it will also yield faster results for potential bacterial threats. In a world characterised with constantly evolving environmental conditions, it is imperative to develop techniques that enable us to comprehensively understand the dynamic impacts on bacteria that pose risks to human health.

Machine learning can analyse data related to bacterial strains and their environments, identifying patterns and making predictions to high accuracy. These models draw from per-existing data to train and refine their algorithms, improving their predictive power over time. The integration of machine learning techniques into various scientific fields has significantly enhanced our understanding and capabilities, ultimately safeguarding public health and the environment.

1.2 Problem Statement

The task of the identifying harmful bacteria presents ongoing challenges in the field of microbiology. Current techniques, while effective, often rely on expert analysis, time-consuming laboratory work and expensive equipment, making them less accessible or impractical in scenarios where analyses demand fast responses or operate under resource constraints. Traditional methods may not be able to keep up with fast, ever-changing data and knowledge. The fast advancement of high-throughput technology is reportedly too rapid for cultivation methods (Qu et al., 2019). As the healthcare landscape continually evolves, the need for alternative approaches that can effectively manage and interpret this information grows (Wu and Gadsden 2023).

This project addresses the critical challenge of bacterial harmfulness prediction using machine learning algorithms. Specifically, it seeks to determine the potential risk if bacterial strains to human health based on bacterial family and the environmental context in which they were found. The application of machine learning machine learning offers the potential to streamline and enhance the process of prediction, providing a faster, more scalable solution for real-world applications.

1.3 Significance of the Study

The complexity of identifying the harmfulness of bacteria is related to several factors. Bacteria are intricate organisms with the remarkable ability to adapt and survive when resources are scarce. For example, there are many bacteria that use oxalate to produce carbon and electron sources, thus expanding their niche (Herve et al. 2016). Furthermore, bacteria have many different morphological features, many of which can look deceptively similar – therefore relying solely on physical traits for identification is insufficient (Kysela et al. 2016). Micro-organisms exist in complex communities with multiple species interacting, requiring more sophisticated methods of identification (Nagarajan and Loh 2014; Xie et al., 2019).

Given these complex factors, the ability to predict bacterial harm becomes increasingly challenging, however the development of machine learning models could potentially revolutionise how we approach microbial threat detection, offering more precise methods for identifying bacteria that pose a risk to human health. For example, in healthcare settings, a predictive model may accelerate the identification of harmful bacteria, leading to speedier treatments.

1.4 Scope of the Study

This study focuses on a dataset of bacteria with attributes related to their family and location. It employs various supervised machine learning algorithms to develop predictive models and evaluates their performance, as well as implementing a grid search for optimisation. The research is limited to the dataset provided and does not cover all possible factors influencing bacterial harmfulness.

The overall scope of the study is to determine what supervised machine learning methods may be best for the simple classification task of harmful and non-harmful bacterial species.

1.4.1 Research Questions and Objectives

Given the critical role of accurate bacterial identification in public health, this project is guided by the following research questions:

1. Can machine learning algorithms be effectively developed for the task of predicting bacteria as harmful or not harmful based on their family and environment of origin?

2. How do different machine learning algorithms perform in terms of accuracy and reliability when classifying harmful and not harmful bacteria?

3. What is the effect of hyperparameter tuning on model performance, and which model ultimately offers the best balance of accuracy and generalisability?

The primary objectives derived from the research questions are:

1. **To Develop a Predictive Model:** Create and train machine learning models which predict whether bacteria are harmful to humans or not based on their family and location of origin.
2. **To Evaluate Model Performance:** Compare the performance of different machine learning algorithms in accurately classifying bacteria as harmful or non-harmful to humans. This study will use accuracy, classification reports, confusion matrixes, and cross-validation scores for evaluation.
3. **To Optimise Model Performance:** Enhance the performance of the predictive models through hyperparameter tuning, ultimately selecting the model that achieves the highest accuracy.

1.4.2 Project Structure

Literature Review

This section will investigate the existing research on machine learning applications in the biological field. It will begin by exploring the traditional methods used for identifying bacteria, such as PCR and

culturing, and will highlight the challenges associated with them. The review will then examine the integration of machine learning into microbiology, discussing how algorithms are increasingly being utilised to conquer the current challenges of traditional methods. This part will also include the various machine learning algorithms in use at this current time, what their roles entail, and their applications in microbial research, providing an overview of the evolving field of identification, classification, and prediction of bacterial behaviour, paved by machine learning.

Methodology

This section will provide a detailed explanation of the machine learning models employed, including Random Forest, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Multinomial Naïve Bayes (MNB). It will cover the preprocessing methods applied to the dataset, such as encoding variables and feature scaling, ensuring the data is suitable for model training. Additionally, this section will include the optimisation process, including hyperparameter tuning through GridSearchCV to enhance model performance. The aim is to present a thorough understanding of the steps taken to develop, train, and optimise the predictive models used for bacteria harmfulness classification.

Analysis

In this section, a discussion and presentation of results obtained from the machine learning models will be exhibited. There will be a focus on the initial models and the following fine-tuned models. A detailed comparison will be made to evaluate the performance improvements achieved from hyperparameter tuning. An analysis of performance metrics including accuracy, precision, recall, F1-score, and AUC-ROC will be used to highlight model strength and weaknesses. Here we will provide tables and visual representations to provide a clear understanding of model performance. This section will also delve into the trade-offs observed, such as AUC score changes, and discuss these implications in the context of bacterial classification.

Conclusion

The conclusion will provide a full summary of the study, highlighting key insights drawn from machine learning model analysis in prediction of bacterial harmfulness. It will reflect on the effectiveness of model selection and hyperparameter tuning. Limitations encountered during the research will be discussed. Finally, the conclusion will offer recommendations for further research, addressing the limitations including the exploration of alternative or enhanced machine learning techniques, the use of larger datasets, and potential areas where this research may be expanded to improve bacterial classification in different contexts.

1.4.3 Key findings

- This project demonstrates significant variations in model performance post-tuning. This underscores the need for careful model selection and hyperparameter optimisation. This also highlights the importance of evaluating the models on multiple metrics to gain a thorough assessment of model performance.
- While some models showed improved accuracy post-tuning, this was sometimes accompanied by a reduction in AUC scores, suggesting potential overfitting. This trade-off highlights the importance of a balanced approach when hyperparameter tuning to ensure model performance is robust.
- The ROC-AUC curves showed that all models performed well, with the KNN model achieving the highest score after hyperparameter tuning.

2. LITERATURE REVIEW

The integration of machine learning applications in microbiology, particularly in prediction or classification tasks used for unknown microbial species identification, signifies an intersection of computer science and biological research where there have been significant advancements in the modern age. This literature review explores existing research within three main areas: traditional methods for bacterial identification, the integration of machine learning in microbiology, and specific machine learning techniques used for identification of bacteria and similar projects.

2.1 Traditional Methods for Bacterial Identification

In this field of study there have been many developments of identifying bacteria methods over the years. Traditional methods typically involve culture-based techniques and other molecular biology-based approaches. Before delving into machine learning, it is important to understand the traditional methods, their limitations, and how machine learning may be beneficial as an alternative.

Culture-based techniques are inexpensive and simple, yet they are extremely time consuming and labour-intensive (Law et al., 2015; Qu et al., 2019). Early pioneers, including Robert Koch and Louis Pasteur, developed these techniques, which have been refined over time (National Research Council 2004). Culturing requires the growth of bacteria to confirm the species; however, this technique is considered limited due to low sensitivity (Lee et al., 2014). In scenarios where prompt results are required, sensitivity is extremely important – for example, a single pathogen present in food can cause infection in humans (Law et al., 2015). Despite these drawbacks, culture-based methods remain a cornerstone in microbiology due to their simplicity and cost-effectiveness. Advances in other traditional methods have attempted to streamline these processes, yet they still require significant manual intervention and expertise (Tabit, 2016).

There are, however, other methods such as PCR, which are more time-efficient and reduce the risk of human error (Mandal et al., 2011). Molecular-based approaches have been widely used in pathogen detection and diagnoses using DNA-based techniques commonly involving PCR, which has proven to be sensitive in identifying certain bacterial strains. Several PCR methods are popular, notably used as a diagnostic tool for the detection of plant, airborne and foodborne pathogens (Ali et al., 2022). Quantitative PCR (qPCR) is known for its rapid results and has become increasingly popular as an alternative to culturing, as some bacteria are unculturable (Hospodsky et al., 2010). In aerosol science, qPCR is largely popular as humans are constantly exposed to allergenic, toxic, or pathogenic particles in air. However, PCR methods are not without its limitations. They may encounter false-negatives and false-positives. The PCR may be less sensitive to certain bacterial strains, causing inaccurate results (Law et al., 2015). Additionally, if processing bacterial pathogens in food, natural inhibitors present in certain foods may interfere with the findings (Wilson 1997; Maurer 2011). Despite these challenges,

PCR and its variants remain vital tools in modern microbiology, providing a balance between speed, sensitivity, and specificity.

Despite significant advancements in traditional methods, they often remain time-consuming, expensive, and require expert knowledge. Furthermore, there is a shortage of pharmacists trained in infectious disease in the US, and with increasing demand, these limitations present substantial challenges in efficiently identifying harmful bacteria (Guiliano et al. 2019). As these traditional methods continue to evolve, the integration of newer technologies aims to address their inherent limitations. The need for rapid, accurate, and scalable identification techniques has led to researchers exploring machine learning as a potential solution.

Ultimately, traditional methods for bacterial identification, while foundational and still widely used across the field, face challenges related to time consumption, labour intensity, required expertise, and potential inaccuracies. Advances in PCR and new-generation sequencing are addressing some of these issues, however each have their own set of limitations which must be considered. The ongoing development and refinement of these technologies are crucial for improving efficiency and accuracy of bacterial identification, ultimately enhancing our ability to respond to microbial threats in various contexts.

2.2 Integration of Machine Learning in Microbiology

The relationship between machine learning and biology has been a long and intricate one, marked by significant milestones that have progressively intertwined these two fields. An early and notable intersection was the development of the perceptron to model neuronal behaviour, which laid the foundation for the development of the artificial neural network (Rosenblatt 1968; Tarca et al., 2007).

By the mid-20th century, the potential for computer-assisted analysis in biological sciences began to be recognised more broadly. In 1966, a computer-assisted service for the identification of gram-negative bacteria was introduced to aid public health and hospital laboratories (Willcox et al., 1980). This service focused on mainly rod-shaped bacteria which, even after culturing on agar, hospitals struggled to identify (Lapage et al., 1973). The system, perhaps rudimentary for todays standards, set a significant precedent for future developments in microbiological diagnostics, illustrating the profound impact that computational methods could have in the field.

Machine learning is growing rapidly; however, many microbiologists lack the adequate training in quantitative methods and often have insufficient statistical backgrounds (Asnicar et al., 2024). Therefore, there is a significant focus on making machine learning accessible and intuitive for professionals in the field. Efforts are being made so software tools are becoming increasingly more convenient and easier to adopt.

These include the need for more comprehensive training for microbiologists, the development of robust and user-friendly tools, and addressing the inherent complexities of biological data (Zou et al., 2017). As these challenges are met, the integration of machine learning in microbiology will continue to grow, promising new insights and more efficient solutions in the study and management of microorganisms.

2.3 Machine Learning Techniques in a Biological Context

There have been various machine learning algorithms employed in the biological field. These techniques are generally categorised into supervised learning, unsupervised learning, and deep learning, each offering unique advantages and disadvantages in a biological context.

Supervised learning techniques are when a model is trained on data where the outcomes are known and labelled, so that predictions can be made on new data. These methods include decision trees, SVM, logistic regression, naïve bayes algorithm, and KNN. A typical example includes taxonomic classification. For instance, the Ribosomal Database Project uses a naïve bayes model to associate rRNA sequences to their respective taxonomic labels, facilitating classification of new rRNA sequences (Asnicar et al., 2024).

Decision trees are intuitive in nature and easy to interpret, making them a popular choice for various biological applications, such as identifying gene expression patterns associated with specific diseases. MARVEL is a proposed tool used to identify and predict bacteriophages, utilising the random forest decision tree algorithm (Amgarten et al., 2018; Grazziotin et al., 2017). Another example included use of random forest for image analysis for the identification of bacterial colonies, resulting in a high accuracy around 98% (Croxatto et al., 2017). The random forest algorithm can handle high-dimensional data and has high resilience to noise, making it well suited for tasks where biological data can be intricate and variable (Wu and Gadsden 2023).

Several studies report random forest and SVM as their highest performing classification approaches (Asgari et al., 2018). Support vector machines are also effective in high-dimensional spaces. and have been successfully used for tasks like protein structure prediction and gene classification. Logistic regression, a fundamental method in statistics, is frequently employed for binary classification problems, such as predicting the presence or absence of a disease based on a set of biomarkers. However, this method is easily affected by correlations between features and may be prone to underfitting (Cox,1958; Jiang et al., 2022). K-nearest-neighbours has also been utilised in many studies, showing promising results. For example, it has been used to predict the time interval after death, by using data from nose and ear samples (Johnson et al., 2016).

Unsupervised learning are methods where hidden patterns in data are identified without any prior knowledge of relationships between the samples. In microbiology, clustering algorithms like k-means and hierarchical clustering are particularly useful for understanding microbial community structures

and their dynamics. These methods organise samples into groups based on similarities, revealing relationships between different microbial species and their environments. There have been multiple studies using k-means clustering for identifying bacteria (Nurlaila et al., 2021; Satyanarayana et al., 2022; Nakao and Magariyama 2021) A number of these studies are still undergoing developments, highlighting the dynamic and evolving nature of this field. A relatively earlier example involved the hierarchical clustering analysis of breast tissue which aided in the prognosis of breast cancer (Makestrov et al., 2004). This study demonstrated potential of clustering to group tissue samples based on characteristics, thereby providing insights into disease prognosis. A slightly more recent study utilised hierarchal clustering to survey antibiotic resistance, illustrating the method's application in understanding the spread and dynamics of resistant bacterial strains (Berrazeg et al., 2013). Through clustering of bacterial samples, scientists were able to identify the patterns and trends in antibiotic resistance, contributing valuable information in the ongoing fight against resistant pathogens.

These examples of unsupervised learning underscore versatility and power of machine learning in microbiology, as they have helped provide valuable insights into complex biological data. Continuous advancement of these methods promises to further enhance our understanding of microbial communities and interactions in the environment.

Deep learning, another subset of machine learning, involves the use of neural networks with many layers to discern complex patterns in data. These models have shown great promise in various fields as they are able to handle large volumes of data and extract meaningful insights without extensive feature engineering. Convolutional neural networks (CNNs) are particularly well-suited for handling high-dimensional data. In an early development, scientists implemented a neural network to predict the length of stay (LOS) of patients following cardiac surgery, which produced an AUC-ROC score of 0.7094, indicating a reasonable performance (Tu and Guerriere 1993). Predicting LOS is difficult due to variability in patient conditions and limited resources; however, the neural network approach demonstrated the potential of machine learning in improving hospital resource management and patient care. This early example underscores the promise of machine learning techniques in tackling complex biomedical problems, paving the way for more sophisticated models and applications in various medical domains. A more recent example involved combining Raman spectrometry with a CNN, achieving more accurate bacterial identification (Ho et al., 2019). Although CNNs often demonstrate extremely accurate results in comparison to other conventional methods, they come at a cost of high computational time (Wu and Gadsden 2023).

This project primarily utilises supervised machine learning techniques, however it is important to acknowledge other approaches within the spectrum of machine learning methodologies.

3.METHODOLOGY

In this section, the systemic approach used to develop, evaluate, and optimise the machine learning models for predicting harmful bacteria will be outlined. It will describe the experimental components including data collection, preprocessing, and the development, evaluation, and optimisation of the models.

Material

The programming language used in this project is Python 3.11.4, packaged by Anaconda inc. Python was chosen due to its extensive libraries and frameworks which are commonly used for data analysis and machine learning problems. Below is a brief description of the primary libraries and tools involved:

4. Pandas : Used for data manipulation and analysis. Offers data structure and operations for manipulating tables and time series.
5. NumPy : Supports large, multi-dimensional arrays and matrices. Also provides mathematical functions to operate on arrays.
6. Matplotlib and Seaborn : Used for data visualisation, allows understanding of data distributions and model performance.

The machine learning components of the project were supported by the following libraries:

Scikit-learn: Provided tools for implementing and evaluating the machine learning algorithms, including:

- Preprocessing: LabelEncoder, CountVectorizer, TfidfTransformer
- Model Selection: train_test_split, cross_val_score, StratifiedKFold
- Metrics: roc_cuve, auc, confusion_matric, classification_report, accuracy_score, precision_recall_curve, average_precision_score
- Classifiers: RandomForestClassifier, AdaBoostClassifier, ExtraTreesClassifier, GradientBoosting Classifier, LogisticRegression, MultinomialNB, KNeighboursClassifier, SVC

As for the environment, Jupyter Notebooks was used for writing and running the code, as this was easier for visualisation.

3.1 Data Source

The dataset used in this study was sourced from Kaggle, which is a popular online website for open source datasets usually used for data science purposes. The dataset contains information on 200 unique

bacterial species and their respective scientific classification, habitat, such as soil, human tissue, or water, and an indication of whether it is harmful or not to humans. The data was chosen due to its comprehensive representation of different bacterial strains and ease of access.

The information provided in the dataset was sourced from the Centers for Disease Control and Prevention and the World Health Organization along with published microbiology journals, indicating the information was reliable and accurate.

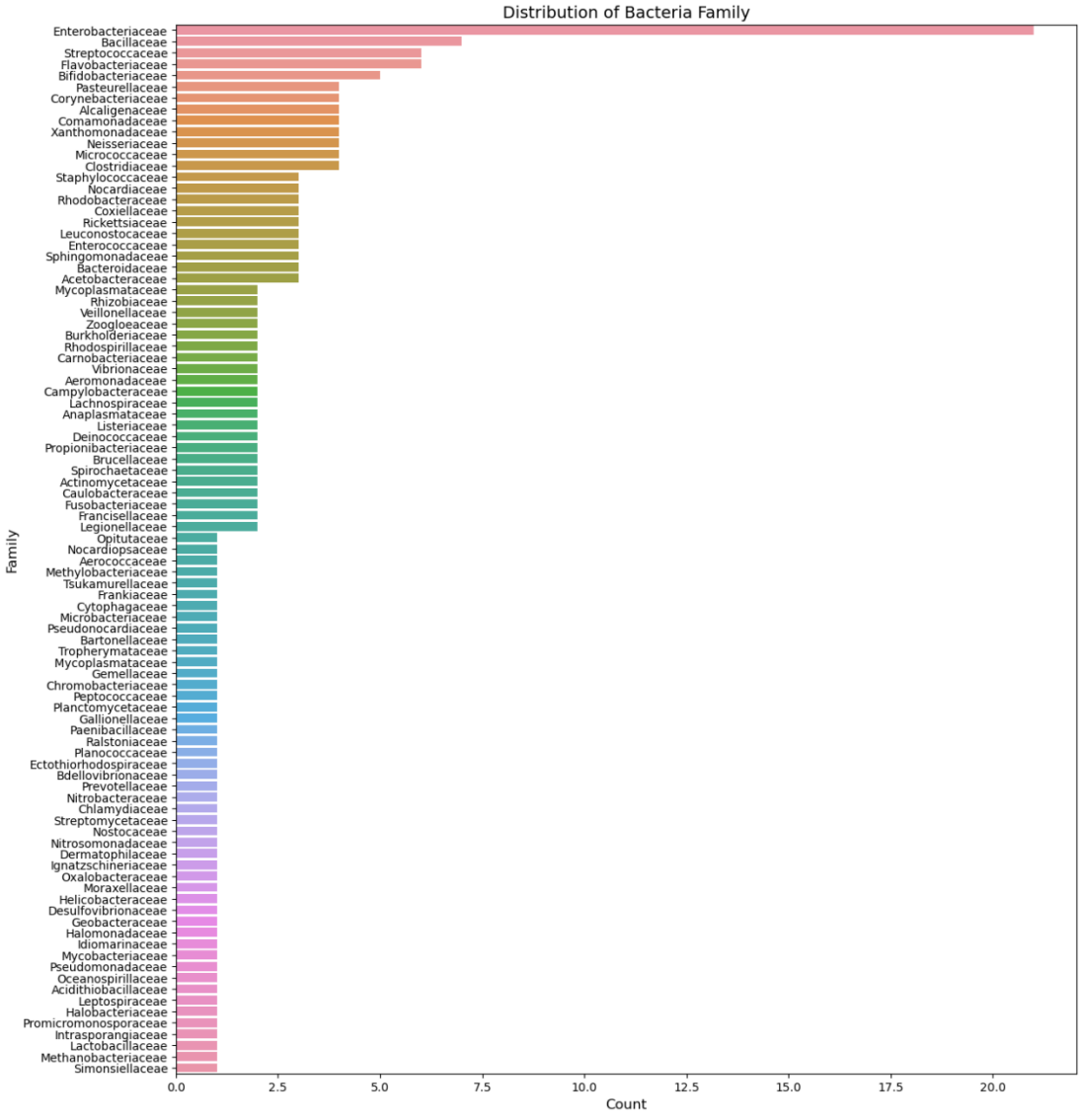

Figure 1: Bar chart depicting the distribution of different bacteria families. Each bar represents a specific family of bacteria, and the length of the bar corresponds to the count or frequency of occurrences of that family within the dataset.

Figure 1 provides a summary of the bacterial families in the dataset. We can see the most prevalent families, the most being Enterobacteriaceae, and the less common families towards the bottom of the graph. It shows a clear imbalance, however, there is no need for re-sampling due to the model's ability to handle the data distribution effectively. The algorithms used, especially those like Random Forest

and SVM, are robust to class imbalance and can provide reliable performance without the need for artificial rebalancing.

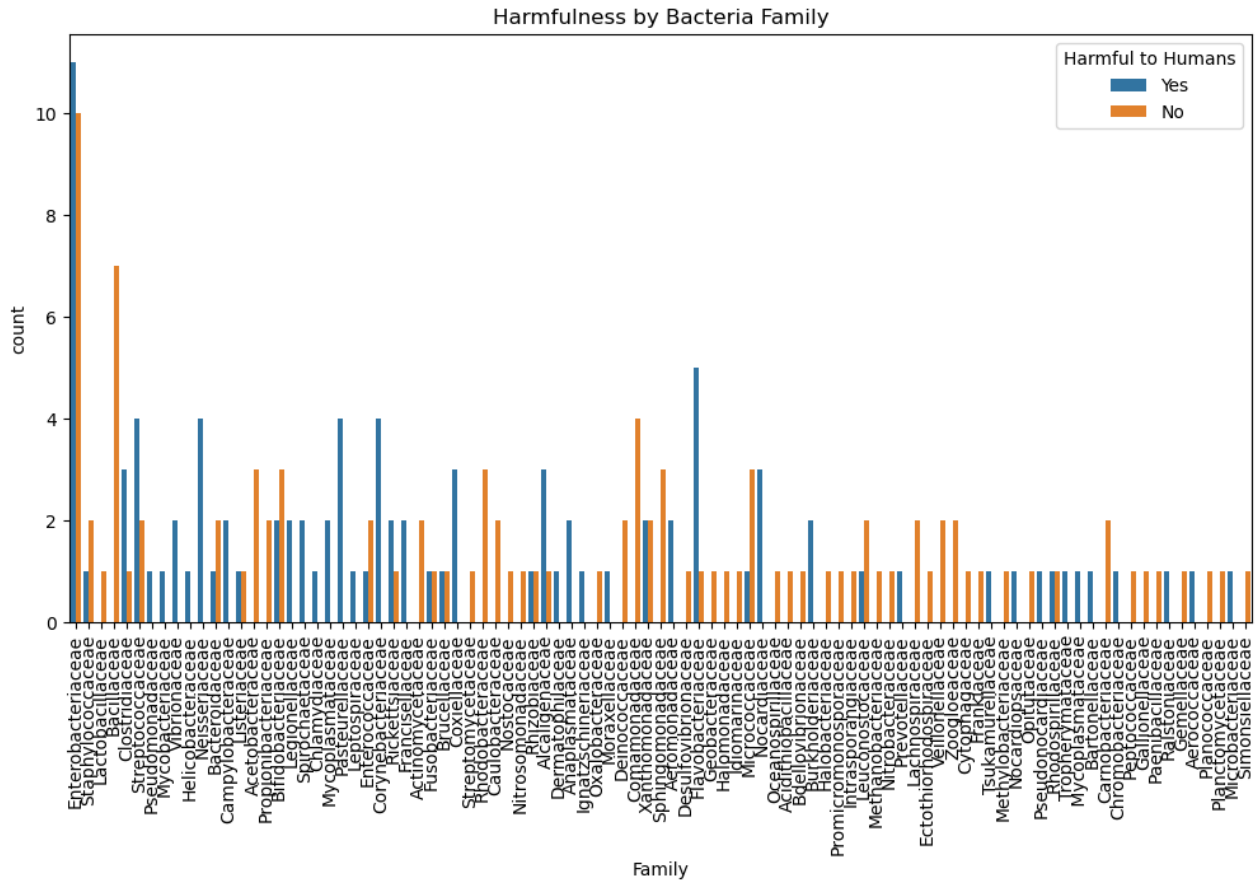

Figure 2: Bar chart showing distributions of the bacteria families categorised by their harmfulness to humans. On the x-axis are different bacteria families, while the y-axis is the count of occurrences. The bars are colour-coded to indicate whether the bacteria family is harmful to humans (blue) or not (orange).

The above chart shows insights into the distribution of harmful and non-harmful bacteria families. From this we can discern the balance of harmfulness in each family. For example, Enterobacteriaceae is predominantly harmful, as the blue bar is larger than the orange. The strain with the second largest harmful count is Flavobacteriaceae. We can see that many families have a relatively balanced distribution, with most counts at 1. This shows that bacteria family may be a significant predictor of harmfulness – for example, if the bacteria is in the Bacillaceae family, it is most likely not harmful to humans.

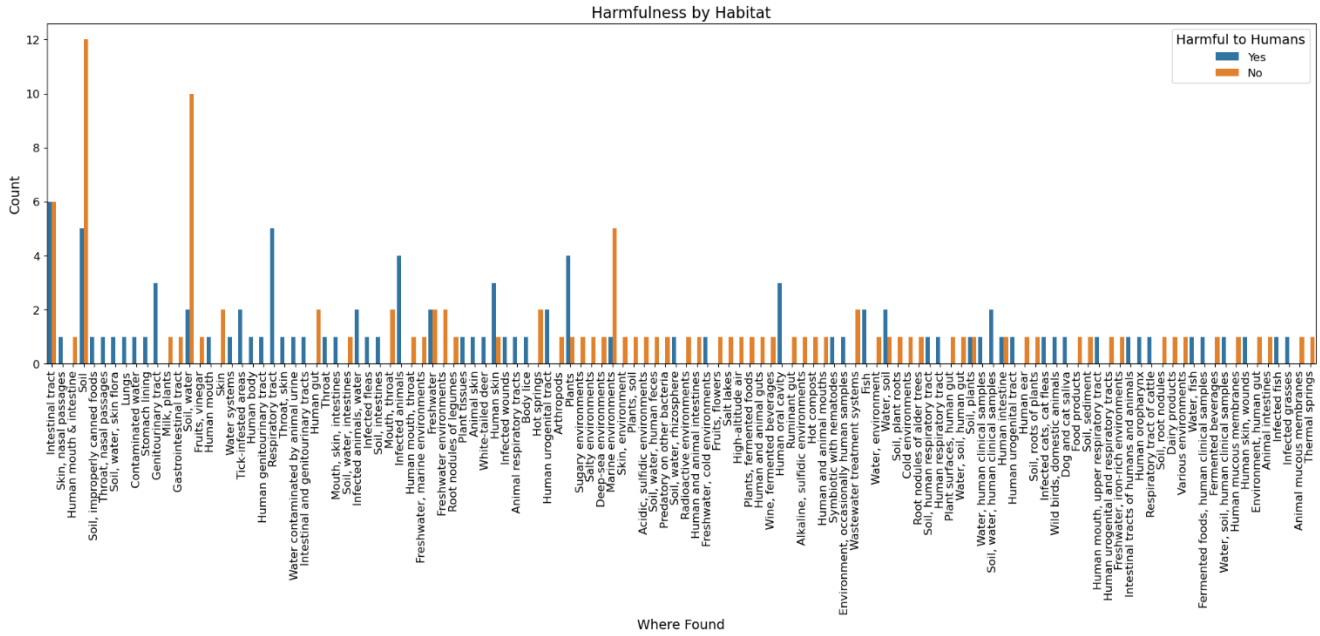

Figure 3: Bar chart showing distributions of the bacteria locations categorised by their harmfulness to humans. On the x-axis are different locations, while the y-axis is the count of occurrences. The bars are colour-coded to indicate whether the bacteria family is harmful to humans (blue) or not (orange).

Figure 3 provides insights into the distribution of harmful and non-harmful bacteria across different habitats. This helps in the understanding of the relationship between habitat and harmfulness, which can be valuable for subsequent predictive tasks. This chart reveals that habitat may be a significant predictor of bacteria harmfulness. For example, bacteria found in the intestinal or respiratory tract may have a higher probability of being harmful compared to those found in soil or water.

3.2 Data Preprocessing

In this phase, the dataset underwent various essential steps to ensure the suitability for machine learning algorithms. First, the general information of the data was printed to understand its structure, including the number of samples, data types, and any missing values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199 entries, 0 to 198
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Name              199 non-null    object
 1   Family            199 non-null    object
 2   Where Found       199 non-null    object
 3   Harmful to Humans  199 non-null    object
dtypes: object(4)
```

Figure 4: General information about the dataset, output of print(data.info())

The initial exploration of the dataset (Fig. 4) laid the groundwork for understanding its structure and identifying potential issues. Initially, the dataset appeared relatively straightforward, featuring a manageable number of attributes at four columns, each representing a distinct feature related to bacterial species. After inspection, it was revealed that the dataset was complete with no missing values detected across the columns. Furthermore, all columns were consistent in data type.

While the preliminary findings suggested a relatively clean and well-organised structure, it is imperative to analyse further aspects. A discrepancy was discovered after investigating the count of categories in 'Harmful to Humans' that required correction. The expectation was that this column would have just two categories: 'Yes' and 'No'. However, further investigation revealed an unexpected third category ' Yes', which included an extra space before the Y. This anomaly would cause inconsistencies in further analyses. The problem was solved using the str.replace() method, thus ensuring that the data was uniform.

To enhance the richness of the dataset and make it more suitable for the machine learning algorithms, the two categorical columns 'Where Found' and 'Family' were combined into a single text feature. This step aimed to capture the context in which bacteria are found, helping models understand certain bacteria are more likely to be found in specific locations, thus enriching the dataset with meaningful relationships.

Next, the features and labels were defined. The new combined text feature served as the input, while the 'Harmful to Humans' column was the target label.

3.2.1 Count Vectorisation

The CountVectorizer was used to convert the combined text feature into a matrix of token counts. This method creates a vocabulary of all the unique words found in the text feature and transforms the text data into a sparse matrix representation.

3.2.2 Term Frequency-Inverse Document Frequency

Following count vectorisation, TF-IDF (Term Frequency-Inverse Document Frequency) was applied to the combined feature. It is a statistical measure which is used to evaluate the importance of a word in a collection of documents.

Term frequency alludes to how often a term appears in a document, relative to the total number of terms in that document (Qaiser and Ali 2018). This allows the user to highlight the terms more relevant within the individual document. For example, if a particular bacteria name or location occurs frequently, term frequency would be high. On the other hand, Inverse Document Frequency measures the importance of the term across the entire set of documents. Therefore, if a bacteria name or location appears frequently, the IDF score would be low, suggesting the term is common and less distinctive. Conversely, terms

found in less records will have a high IDF score and suggests they are potentially more important. These values help reduce the impact of common terms and emphasize terms that are more specific to the context.

The TF-IDF score is calculated by multiplying the term frequency by the inverse frequency. This process ensures terms with high relevance to specific documents are highlighted. For example, if a certain bacteria name is rare but highly relevant in certain locations, the TF-IDF score will be high.

The TD-IDF transformed the bacteria names and locations into numerical vectors, which made them suitable for machine learning algorithms. This transformation allowed for a more nuanced understanding of the data, as it emphasised the important terms while downplaying less informative terms. Ultimately, it helped differentiate between bacteria names that are unique to specific locations, versus those that are more commonly found across multiple locations.

In summary, TF-IDF was applied to facilitate a more accurate analysis and modelling. This approach ensured the most relevant terms were properly represented in the machine learning models, leading to a better classification and insights.

3.2.3 Label Encoding

In the ‘Harmful to Humans’ column, the data were transformed from categorical values to numerical values using LabelEncoder from the sklearn.preprocessing library. In this case, the 'Harmful to Humans' column contained three categories ('Yes', 'No', ' Yes'). After correcting the third category to 'Yes', the LabelEncoder mapped the categories to integers.

3.2.4 Data Splitting

The data was then split into a train and test set using train_test_split from sklearn.modelselection, in anticipation of the models to follow. The dataset was divided such that 80% of the data was allocated to the training set, and the remaining 20% data allocated to the testing set. The testing subset of data was in turn kept separate and was used to evaluate the model’s performance. An 80-20 split is a common practice in machine learning and appropriate for the size of the dataset, striking a balance between having enough data to train the model effectively and ensuring there is a sufficient amount of data for rigorous evaluation. The purpose of the spit was to simulate how the model would perform on new, unseen data, to provide an unbiased assessment on predictive accuracy.

A random state parameter was set to 42, to ensure reproducibility of the results. This ensures the data split is consistent across different runs.

3.3 Model Selection

In this classification project, five different supervised machine learning models were employed to predict whether a bacterial species is harmful to humans. These models were chosen based on their diverse approaches to classification problems, providing a comprehensive evaluation of the dataset.

Random Forest

Random Forest is an ensemble learning method. Figure 5 shows the process. Once started, a multitude of decision trees are built during the training process, where each tree is trained is on a random subset of the data and features (Parmar et al., 2018). The final prediction is then made by taking a majority vote for classifying tasks (aggregating predictions). This method belongs to integrated learning, often with high accuracy. It has the capacity to run on larger datasets, as well as be resistant to noise (Breiman 2001; Jiang et al., 2022).

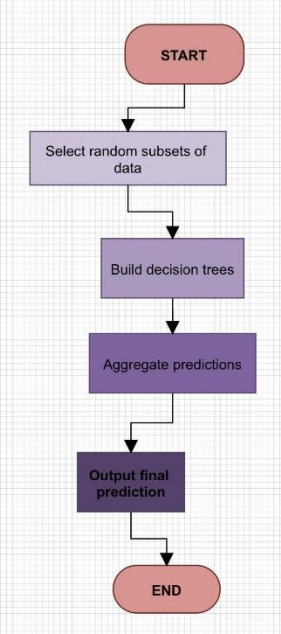

Figure 5: Random Forest algorithm flowchart. Own illustration.

This particular method was selected for its superior performance in many studies, often surpassing SVM methods in some experiments, in regard to sensitivity, specificity, and accuracy metrics (Wu and Gadsden 2023). This model captures complex interactions between features, making it suitable for diverse datasets. Its robustness and ability to reduce overfitting through averaging of multiple trees contribute to its reliability and effectiveness, which is crucial in a biological context where data can be noisy.

Logistic Regression

Logistic regression is a linear model used for binary classification tasks. In this project, it was employed to calculate the probability of a binary outcome, which in this case was harmful or not harmful, by applying a logistic function to a linear combination of input features, this is formulised as:

$$p = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

where p is the probability of the binary outcome, w represents the weights for the features, x represents the input features, and b is the bias term (LaValley, 2008).

This model was chosen due to simplicity and ease of interpretation. In the context of predicting bacterial harmfulness, it serves as an excellent baseline model, providing a straightforward comparison for more complex models. Its ability to produce interpretable results allows for an understanding of how individual features, such as habitat and family, influence the harmfulness prediction.

K-Nearest Neighbours (KNN)

KNN is a non-parametric learning algorithm. It a simple, but effective method that classifies a sample based on the majority class among its k-nearest neighbours in the training set. The KNN method is essentially biased by the chosen k-value, which determines the number of neighbours to consider (Guo et al., 2003). In this project, it will help identify bacterial species' harmfulness based on similarities to known harmful or non-harmful species.

KNN was selected for its simplicity and effectiveness in capturing local data structures. It is particularly useful in cases where the decision boundary maybe irregular (Hamilton et al., 2020). The KNN has various methods to calculate distance between points. Below is the formula for Manhattan distance:

$$d(x_i, x_j) = | x - y | 1 = \sum_{i=1}^{n} | x_i - y_i |$$

where x and y represent two vectors in the feature space and x_i and y_i are their coordinates (Gao and Li 2020).

Support Vector Machines (SVM)

SVM is a powerful classifier that, like logistic regression, can perform binary classification of data. SVMs employ their decision boundary based on the maximum-margin hyperplane of the learning sample (Qu et al., 2019). It uses kernel functions to handle non-linear relationships.

As a simple example, there are two classes, A and B, that need to be classified. The SVM will first find a line or 'hyperplane' that best separates the two classes. The line position will then be adjusted to maximise the margin between the two classes.

Selected for its robustness and effectiveness in high-dimensional spaces, the SVM ensures high accuracy and generalisability, making it a strong candidate for complex biological datasets. The SVM has demonstrated high accuracy and speed in some image classification tasks compared to the KNN (Hamilton et al., 2020).

Multinomial Naïve Bayes (MNB)

This method is particularly efficient and effective in handling discrete data, making it suitable for text classification tasks. MNB is a probabilistic learning algorithm, applying Bayes' theorem with the assumption that features are conditionally independent given the class (Kibriya et al., 2004). Given that the features in this study are represented as text, this algorithm is appropriate for TF-IDF conversion.

For text classification examples, the MNB would first count the occurrences of each word in documents belonging to different categories. The counts are then used to estimate probabilities of words occurring in each category. A new document will be classified by calculating the posterior probability for each

category based on the word counts in the document, and then choosing the category with the highest probability.

Although some studies found that SVM often achieves a higher accuracy, MNB is still computationally efficient and provides fast training and prediction times, making it a practical choice for this project (Wu and Gadsden 2023).

3.4 Training and Evaluation

Each of the chosen models were trained on the transformed dataset and evaluated using various metrics, including accuracy, classification reports, confusion matrices, and cross-validation scores. This thorough evaluation process ensured that each of the model's performance was assessed to high accuracy, allowing for the selection of the most effective model for predicting the harmfulness of bacterial species.

In this step, the initial models were run with default parameters to provide a baseline for performance assessment. As an example, the Random Forest classifier was initialised with 100 decision trees, and the Logistic Regression model was set with a maximum of 1000 iterations to ensure it converged. The Support Vector Machine (SVM) used a linear kernel and probability estimates, while the K-Nearest Neighbours (KNN) and Multinomial Naive Bayes (MNB) models were employed with their default settings.

3.5 Optimisation

The next step involved more detailed hyperparameter tuning to optimise model performance. This process included the use of a Grid Search through gridsearchcv, where various combinations of parameters were systematically tested to identify the best set of parameters for each of the models. GridSearchcv is an exhaustive search method which compares model performances using model metrics such as accuracy, precision, and recall (Veeralagan and Priya 2022). This refinement aimed to enhance model accuracy and reliability for predicting the harmfulness of bacterial species, ensuring the selection of the best-performing model for the specific application. The accuracy, cross validation and mean cross validation scores were printed for all fine tuned models.

The parameter grids used for each model, along with justifications for the chosen hyperparameters, are as follows:

Random Forest

- n_estimators: Number of trees in the forest. Chosen values to test were 50, going up in increments of 50 until 250. A higher value would improve model performance by reducing

variance, but this also increases computational time. Therefore, a balance between performance and computational efficiency was considered.

- max_depth: Maximum depth of the trees. The values tested were None (no limit), 10, 20, and 30, 40, and 50. By limiting the depth, overfitting may be prevented therefore making the model more generalisable.
- min_samples_split: The minimum number of samples required to split an internal node. Tested numbers were 2, 5, and 10. Lower values allow the model to capture more complex patterns but may increase risk of overfitting.
- min_samples_leaf: Determines number of samples required to be at a leaf node. Tested values were 1, 2, and 4. Higher values smoothen out the model and helps prevent overfitting.
- bootstrap: This can be true or false. Determines whether random samples with replacement are used. If used, each tree is built on a random subset of the data.

Logistic Regression

- C: Inverse of regularisation strength. Smaller values specify stronger regularisation. Tested values were 0.01, 0.1, 1, 10, and 100. With increased regularisation strength, more large coefficients will be penalised.
- solver: Algorithm to use in the optimization problem. Tested values were 'liblinear' and 'saga'. Different solvers have different strengths, and testing both can help find the optimal one for the dataset.
- penalty: Determines the type of regularisation for the model to use. The tested penalty types are l1, l2, elasticnet, and none. Regularisation helps prevent overfitting, helping the model generalise over the unseen data.
- max_iter: This value determines the maximum number of iterations for the model to converge. The values chosen to test were 100, 200, and 300. Fewer iterations may lead to incomplete convergence, however more iterations will increase computational time.

Support Vector Machine (SVM)

- C: Regularization parameter. Tested values were 0.01, 0.1, 1, 10, and 100. This value controls trade-off between achieving a low training error and low testing error.
- kernel: Specifies the kernel type to be used in the algorithm. Tested values were 'linear', 'rbf'., ‘poly’, and ‘sigmoid’. The chosen kernels will transform input data into higher dimensions, to make it more linearly separable.
- gamma: Kernel coefficient for 'rbf', 'poly', and 'sigmoid'. Tested values were 'scale' and 'auto'. This value will determine the influence of a single training sample.

- degree: This is only applicable if a polynomial kernel is chosen. This value defines the degree of the polynomial. Values tested included 2, 3, and 4. At lower degrees, complexity is less and it may underfit, however higher degrees may capture more complexity but at increased risk of overfitting.

K-Nearest Neighbors (KNN)

- n_neighbors: These are the number of neighbours to use. The tested values were 3, 5, 7, 9, and 11. These values determine the balance between bias and variance.
- weights: Weight function used in prediction. Tested values were 'uniform' (all points in each neighborhood are weighted equally) and 'distance' (weight points by the inverse of their distance).
- algorithm: determines the type of algorithm used to compute the nearest neighbour. Tested types were auto, ball_tree, kd_tree, and brute. Some are more suitable for higher dimensional data, whereas others for lower dimensional data.
- p: Determines the distance metric used for finding neighbours. p=1 is for Manhattan distance, p=2 is for Euclidean distance.

Multinomial Naive Bayes (MNB)

- alpha: Smoothing parameter which is used to handle zero probabilities. Tested values were 0.01, 0.1, 1, and 10. Lower values mean lower smoothing, useful when zero probabilities are common, whereas higher value means larger smoothing which is useful if there are many zero probabilities.
- fit_prior: This determines whether the model learns class prior probabilities from training data, or uses a uniform prior. The options are true or false.

For all models, a Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) curve was printed to evaluate their performance. The ROC-AUC curve provides a graphical representation of a model's ability to discriminate between positive and negative classes across various threshold values. The AUC value indicates the overall performance in distinguishing between positive and negative classes – the higher, the better.

4. ANALYSIS

4.1 Overview

In this section, results and evaluation of the models implemented will be presented. Model performance was assessed using several key metrics: accuracy, classification reports, and cross-validation scores. The analysis will demonstrate how each model performed in predicting harmfulness of the bacterial species and will identify the best-performing model based on these metrics.

Before delving into the model results, we first need to understand the metrics chosen for evaluation. This understanding provides context for interpreting the performance of each model and ensures that comparisons are meaningful.

Accuracy

- The ratio of correctly predicted observations to the total observations. It is the most intuitive performance measure.
- Accuracy provides a straightforward measure of model performance, particularly when the classes are balanced. It gives an overall sense of how well the model is performing. As aforementioned, one of the main objectives in this project is to select the model that achieves the highest accuracy, ensuring that the model can correctly identify the majority of the bacterial species as harmful or not harmful to humans.

Classification Report

- This report includes precision, recall, F1-score, and support for each class.
    - **Precision**: This is the ratio of correctly predicted positive observations to the total predicted positives. A higher precision value would indicate a low false positive rate.
    - **Recall**: Ratio of the correctly predicted positive observations to all the observations in the actual class. A higher recall indicates low false negative rate.
    - **F1-Score**: This value represents the weighted average of precision and recall, useful for when the distribution of classes in the dataset is imbalanced.

Cross-Validation Scores

- Cross-validation was chosen as an output as it provides a more reliable estimate of the model's performance by mitigating overfitting. It involves dividing the dataset into multiple 'folds' and training on some folds while validating on others. This process is repeated to ensure the model is consistent.

4.2 Initial Model Performance

Each model was evaluated on the transformed dataset, and the results were compared to determine the most effective model for our classification task. The primary values of each models' results are provided below.

4.3 Random Forest

Accuracy: 0.7750

Random Forest achieved an accuracy of 0.7750, which suggests the model was impressively accurate in correctly predicting 77.5% instances. It effectively balances bias and variance through the ensemble of decision trees, each trained on different subsets of the data. It suggests a solid performance, that the model effectively differentiates between harmful and not harmful bacterial species. However, as accuracy is a relative measure, it is important that we consider other metrics to understand model performance fully.

Table 1.1: Classification report for initial random forest model. Class 0 is not harmful bacteria, while class 1 is harmful bacteria.

Metric	**Class 0**	**Class 1**	**Weighted Avg**
Precision	0.75	0.80	0.78
Recall	0.79	0.76	0.78
F1-Score	0.77	0.78	0.78

Table 1.1 above exhibits the classification report, with values that overall suggest a good performance. Precision is seen higher for class 1 than class 0, indicating that the model is correctly predicting a bacterium as harmful 80% of the time. For class 0 the precision is at 75%, meaning 25% of the predictions for non-harmful bacteria were incorrect. This suggests that the model is less likely to make false positive errors, beneficial in contexts where it is crucial to avoid them such as medical diagnoses where predictions of harm may lead to unnecessary treatment. Conversely, for recall, class 0 is slightly higher than class 1, meaning the model is somewhat more effective at identifying non-harmful bacteria than harmful ones, reducing false negatives. The relatively high recall for both classes indicates that the model is effective in detecting both harmful and non-harmful bacteria, although there is a small trade-off between precision and recall. The close f1-scores, 0.77 for class 0 and 0.78 for class 1, highlights this and indicates the model maintains a good precision-recall balance.

Table 1.2: Confusion matrix for initial random forest model

	Predicted 0	**Predicted 1**
Actual 0	15	4
Actual 1	5	16

Table 1.3: Cross Validation scores for initial random forest model

Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean CV Score**
0.625	0.750	0.725	0.700	0.821	0.7241

Table 1.2 shows the confusion matrix, where values indicate a higher number of true positives and negatives which is desirable. The trade-off between false positives and false negatives is important, and depending on the application, one might be more critical than the other. Cross validation scores shown in table 1.3 show variability from 0.625 up to 0.821. The mean cross-validation score, at 0.7241, is fairly close to the initial accuracy score, suggesting consistent performance.

Overall, the random forest model performed well. However, there is always room for improvement, especially in reducing the number of false positives and false negatives.

4.4 Logistic Regression

Accuracy: 0.800

The Logistic Regression model achieved indicates that 80% of the predictions made by the model were correct. This high accuracy reflects the model's ability to differentiate between harmful and non-harmful bacterial species effectively.

Table 2.1: Classification report for initial logistic regression model. Class 0 is not harmful bacteria, while class 1 is harmful bacteria.

Metric	**Class 0**	**Class 1**	**Weighted Avg**
Precision	0.76	0.84	0.80
Recall	0.84	0.76	0.80
F1-Score	0.80	0.80	0.80

The precision is higher for class 1 than class 0, meaning when the model predicts a bacterium is harmful, it is more likely to be correct compared to when it predicts a bacterium as non-harmful. This indicates that the model has a stronger ability to correctly identify harmful bacteria, reducing false positives in this category. Recall is higher for class 0, suggesting the model is better at correctly identifying non-harmful bacteria more, reducing false negatives in this category. This model is more sensitive to non-harmful bacteria. The balanced F1-scores demonstrate that the model performs equally well in balancing precision and recall for both categories.

Table 2.2: Confusion matrix for initial logistic regression model

	Predicted 0	**Predicted 1**
Actual 0	16	3
Actual 1	5	16

Table 2.3: Cross-validation scores for initial logistic regression model

Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean CV Score**
0.625	0.725	0.65	0.825	0.821	0.7291

Table 2.2 shows a confusion matrix of a well-performing model. The true values are high, while the false values are low in comparison. Additionally, the cross-validation fold scores are relatively balanced, meaning the mean cross-validation score indicates the model's performance is consistent across different subsets of the data, indicating that the model generalizes well to unseen data.

4.5 Support Vector Machine

Accuracy: 0.8250

An accuracy score of 0.825 shows that this model correctly classifies 82.5% of the instances. This high accuracy reflects the model's overall robustness in classification tasks within this dataset.

Table 3.1: Classification report for initial support vector machine model. Class 0 is not harmful bacteria, while class 1 is harmful bacteria.

Metric	Class 0	Class 1	Weighted Avg
Precision	0.80	0.85	0.83
Recall	0.84	0.81	0.82
F1-Score	0.82	0.83	0.83

Precision is higher for class 1 than class 0, therefore this model is more likely to correctly identify harmful bacteria. However, recall is slightly less for class 1, indicating the models sensitivity to non-harmful bacteria. Again, this balance suggests the model minimises both false positives and false negatives.

Table 3.2: Confusion matrix for initial support vector machine model

	Predicted 0	Predicted 1
Actual 0	16	3
Actual 1	4	17

Table 3.3: Cross-validation scores for initial Support Vector Machine model

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean CV Score
0.7	0.7	0.725	0.825	0.795	0.7490

Cross-validation scores are also well-balanced, indicating model consistency. In summary, the SVM model exhibits a strong and balanced performance, making it a reliable choice for predicting the harmfulness of bacterial species in this dataset. The model's ability to balance precision and recall while maintaining high accuracy ensures that it effectively identifies harmful bacteria without excessively misclassifying non-harmful ones. This is also reflected in table 3.2, where false positives and false negatives are relatively low.

4.6 K-Nearest Neighbours

Accuracy: 0.8250

Accuracy for the KNN was the same as SVM, indicating it had the same high level of reliability in its predictions.

Table 4.1: Classification report for initial K-nearest neighbours model. Class 0 is not harmful bacteria, while class 1 is harmful bacteria.

Metric	Class 0	Class 1	Weighted Avg
Precision	0.80	0.85	0.83
Recall	0.84	0.81	0.82
F1-Score	0.82	0.83	0.83

The classification report follows the same pattern as SVM – precision for harmful bacteria is higher, while it is lower in recall. This model also has a sensitivity for non-harmful bacteria.

Table 4.2: Confusion matrix for initial K-nearest neighbours model

	Predicted 0	Predicted 1
Actual 0	16	3
Actual 1	4	17

The confusion matrix also exhibits the same values as the SVM. Misclassifications were rare.

Table 4.3: Cross-validation scores for initial K-nearest neighbours model

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean CV Score
0.525	0.7	0.725	0.75	0.795	0.6990

The mean cross-validation score is slightly lower for this model, however, still indicates a reasonable consistency of performance.

4.7 Multinomial Naïve Bayes

Accuracy: 0.8250

Accuracy remained consistent compared to KNN and SVM. This high accuracy is particularly notable given the probabilistic nature of the Naive Bayes algorithm and its assumption of feature independence.

Table 5.1: Classification report for initial multinomial naïve bayes model. Class 0 is not harmful bacteria, while class 1 is harmful bacteria

Metric	Class 0	Class 1	Macro Avg	Weighted Avg
Precision	0.80	0.85	0.82	0.80
Recall	0.85	0.81	0.83	0.80
F1-Score	0.82	0.83	0.82	0.80

Again, the classification report exhibited very similar values to SVM and KNN, suggesting they all have a sensitivity to non-harmful bacteria. The f1-score also indicates consistent performance.

Table 5.2: Confusion matrix for initial multinomial naive bayes model

	Predicted 0	Predicted 1
Actual 0	16	3
Actual 1	4	17

Again, the confusion matrix stayed the same, suggesting SVM, KNN and MNB are performing similarly, rarely misclassifying.

Table 5.3: Cross-validation scores for initial multinomial naïve bayes model

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean CV Score
0.65	0.725	0.775	0.825	0.821	0.7591

Cross-validation scores across all folds remain relatively consistent, indicating a solid model performance.

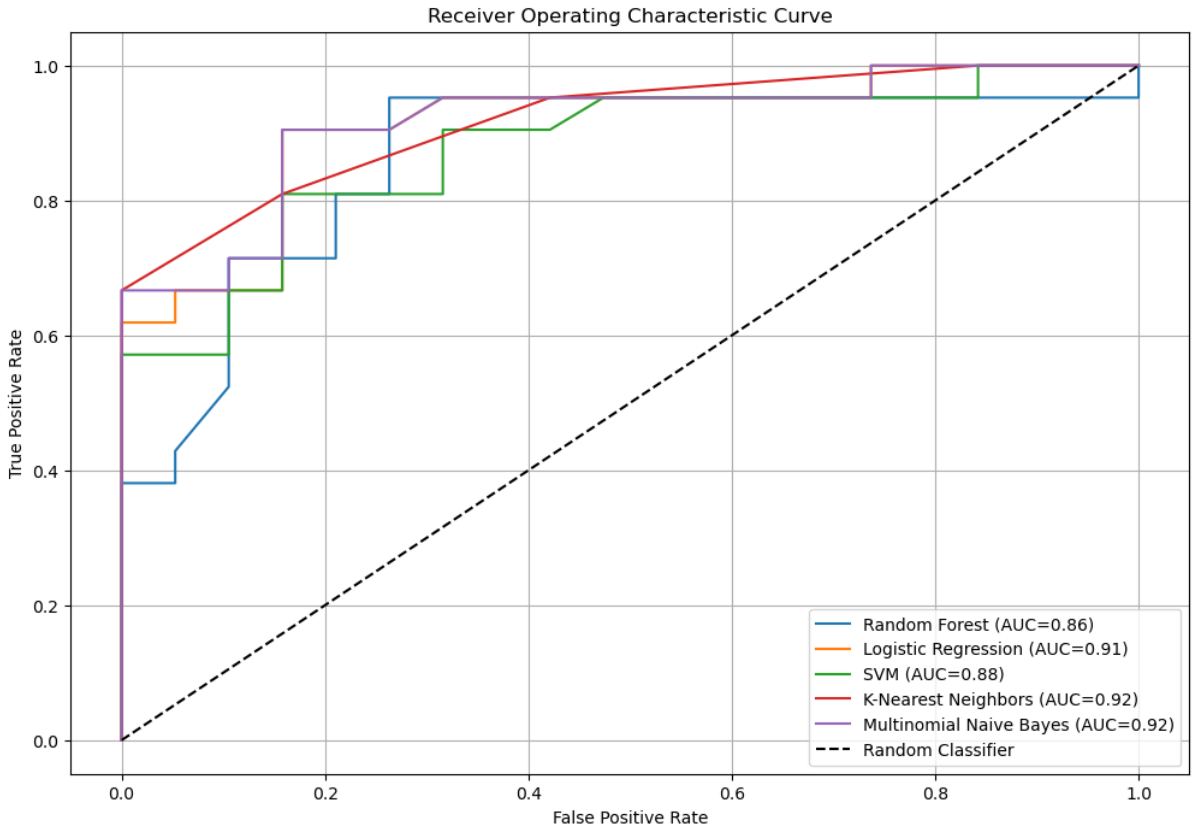

Figure 6: ROC-AUC curve comparing the initial models performance. The ROC curve plots the True Positive Rate (y-axis) against the False Positive Rate (x-axis), and the Area Under the Curve (AUC) is provided for each model as a measure of performance.

Figure 6 is a Receiver Operating Characteristic (ROC) curve comparing the performance of the five different machine learning models. Generally, all models seem to have performed well, as their AUC scores are high, and the curve stays high in true positives. According to AUC scores, KNN and MNB seem to be the strongest contenders at 0.92.

For random forest, the curve shows a moderate performance and the AUC of 0.86 demonstrates good discriminative ability, however it is lower than some of the other models. Logistic regression maintains a high true positive rate across a range of false positive rates and has a high AUC of 0.91 showing its effectiveness in distinguishing between classes. SVM also shows a good true positive rate but shows some fluctuations as the false positive rate increases. Its AUC score is slightly lower, coming in second to last.

4.8 Comparison

Table 6: The initial models primary performance metrics

Model	Accuracy	Precision	Recall	F1-Score	Mean Cross-Validation Score
Random Forest	0.7750	0.78	0.78	0.78	0.7241
Logistic Regression	0.8000	0.80	0.80	0.80	0.7291
SVM	0.8250	0.83	0.82	0.83	0.7490
KNN	0.8250	0.83	0.83	0.83	0.6990
MNB	0.8250	0.83	0.83	0.83	0.7591

The above table showcases all model with their primary performance metrics. From this, we can conclude that SVM, KNN, and MNB achieved the highest accuracies, however, MNB achieved the best mean cross-validation score as well. KNN, although one of the highest in accuracy, achieved the lowest cross-validation score, suggesting that this was the least consistently performing model in comparison to all others. Although all models had a reasonably good performance, random forest performed the weakest.

4.9 Hyper-parameter tuning

After using a grid search for optimisation, the parameters found were as follows:

Random Forest

- bootstrap: True – Each tree in the forest was trained on a random subset of the data, helping reduce overfitting and increasing model robustness.
- max_depth: 20 – This limits the growth of the tree. While balancing the prevention of overfitting and allowing the model to capture enough complexity of the data, the search identified this was the best fit for depth.
- min_samples_leaf: 1 – Having a minimum of 1 sample per leaf allows the model to fully utilise the data, although this may slightly increase the risk of overfitting.
- min_samples_split: 5 – With at least 5 samples required to split a node, the creation of nodes with very few samples is prevented which reduces overfitting.

- n_estimators: 250 – The search identified using 250 trees is optimal, as it reduces variance and helps achieve better generalisation. However, this may also increase computational time.

Logistic Regression

- C: 0.01 – By having the regularisation parameter at 0.01, overfitting is prevented by penalising any large coefficients.
- max_iter: 100 – A maximum 100 iterations balances ensuring convergence while considering computational time.
- penalty: none - This means there is no regularisation applied, therefore the model can fit the training data more closely.
- solver: saga - This choice of solver is efficient for large datasets, also supporting regularisation and therefore making logistic regression more robust.

Support Vector Machine

- C: 1 – This chosen value represents a balanced approach between bias and variance, which provides a good trade-off.
- degree: 3 – This value enables the model to capture non-linear relationships in the data.
- gamma: scale - Means the parameter is adjusted based on the number of features. This ensures the kernel can effectively handle the feature space.
- kernel: poly – By having the kernel polynomial, the SVM will be able to capture complex relationships in the data that a linear kernel might miss.

K-Nearest Neighbours

- algorithm: auto – This chooses the most appropriate algorithm for the dataset, optimising model performance.
- n_neighbors: 5 – This number of neighbours balances between capturing local patterns and reducing noise from too many neighbours.
- p=1 – Corresponds to Manhattan distance. May be more robust to outliers in certain scenarios, compared to Euclidean distance.
- weights=distance – ensures closer neighbours will have influence the prediction more, improving accuracy in heterogenous regions.

Multinomial Naïve Bayes

- alpha:1 – this applies Laplace smoothing which in turn will handle zero frequencies in data and prevents overfitting
- fit_prior=true – allows model to adjust based on class distribution in training data, improving performance for imbalanced datasets

Below, we will compare the accuracies and mean cross-validation scores for the initial models and the fine-tuned models.

Table 7: Table comparing the accuracy and mean cross-validation scores of the initial vs fine-tuned models

Model	**Initial Accuracy**	**Fine-tuned accuracy**	**Initial mean CV score**	**Fine-tuned mean CV score**
Random Forest	0.7750	0.8000	0.7241	0.7673
Logistic Regression	0.8000	0.8000	0.7291	0.7548
SVM	0.8250	0.7500	0.7490	0.7861
KNN	0.8250	0.8500	0.6990	0.7417
MNB	0.8250	0.8250	0.7591	0.7486

Generally, the accuracies have either increased or stayed constant after hyperparameter tuning. Interestingly, we see that accuracy has decreased from 0.8250 to 0.7500 for the SVM model, however, mean cross-validation increased. The unexpected change of accuracy may be attributed to several factors. One is that the fine-tuned parameters may have caused the SVM to overfit the training data, and even though the cross-validation improved, the model may have captured noise or irrelevant patterns resulting in decreased generalisation. Overfitting may have occurred due to overcomplexity of the kernel, which was changed to 3. Furthermore, as C was set to 1, even though bias and variance was balanced, this may not have been optimal paired with the polynomial kernel.

Nevertheless, all other models show more promising results. The increased accuracies indicated the fine-tuned models are better are classifying results correctly. We can see all cross-validation values have increased, showing an increase in consistency across the models. Despite the accuracies staying the same for logistic regression and MNB, we still see an improvement in cross-validation scores. It suggests that consistency in performance suggests robustness in predictions, but also indicates the model was already operating near its optimal performance for this dataset.

Notably, the KNN had the highest accuracy out of all. Furthermore, we also see a greatly improved mean cross-validation score. Initially, the KNN accuracy was comparable to the SVM and MNB, however its lower mean cross-validation was lower, placing it in a less favourable position. After fine-tuning, the mean cross-validation score increased to 0.7417 from 0.6990, not only enhancing credibility but also it demonstrates that KNN's performance is more stable and reliable across different data splits when tuned appropriately.

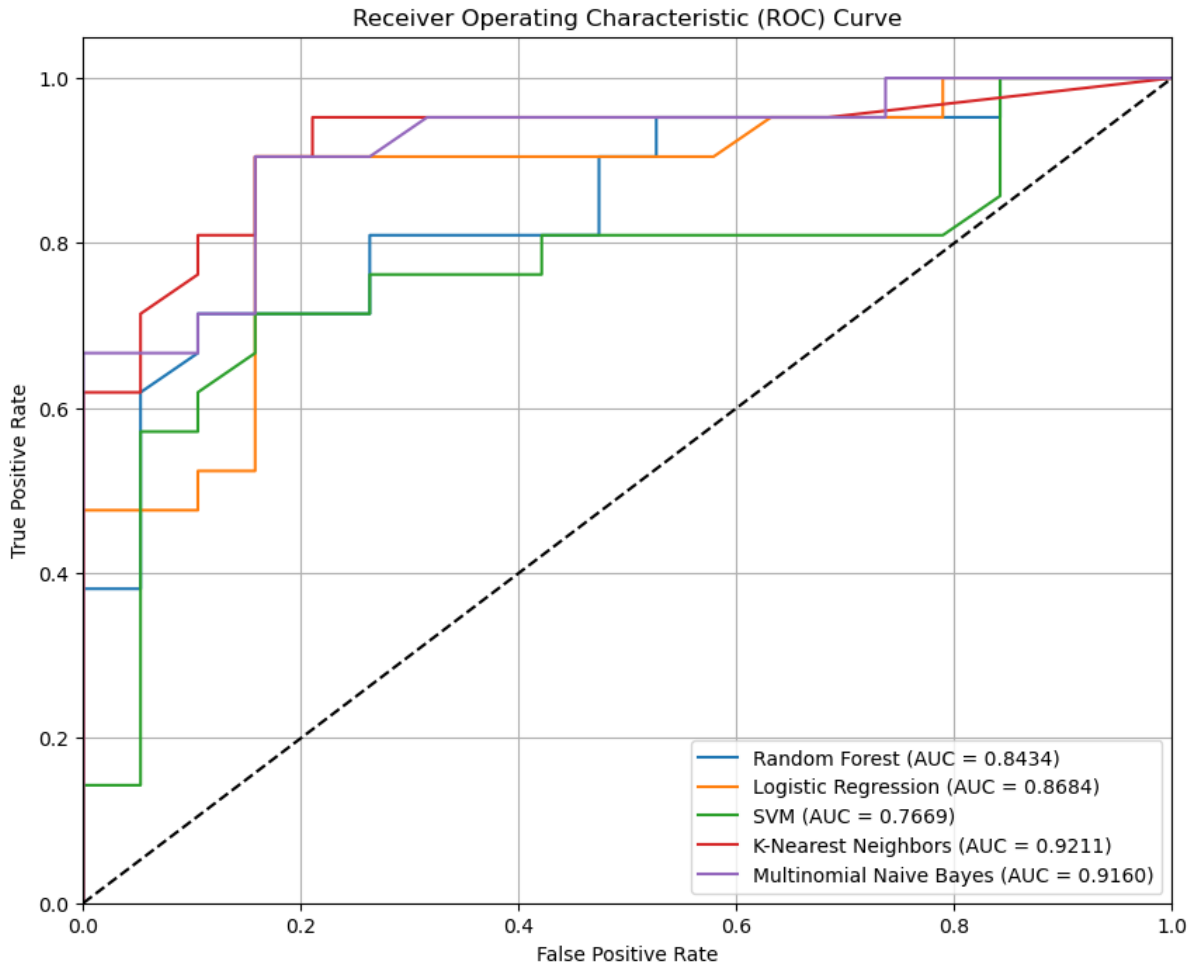

Figure 7: ROC-AUC curve comparing the fine-tuned models performance. The ROC curve plots the True Positive Rate (y-axis) against the False Positive Rate (x-axis), and the Area Under the Curve (AUC) is provided for each model as a measure of performance.

The ROC curve exhibits impressive performance from all models, demonstrating their ability to effectively distinguish between harmful and non-harmful bacteria. Each model's ROC curve indicates a strong ability to correctly classify instances across different decision thresholds, underscoring a general effectiveness.

Among all models evaluated, KNN stands out with the highest AUC score of 0.92. This value indicates that KNN has the most robust performance in distinguishing between the two classes. It reflects a superior model performance in capturing the underlying patterns in the data.

Table 8:Comparison of initial vs fine-tuned models AUC scores

Model	**Initial AUC**	**Fine-tuned AUC**
Random Forest	0.8600	0.8434
Logistic Regression	0.9100	0.8684
SVM	0.8800	0.7669
KNN	0.9200	0.9211
MNB	0.9200	0.9160

Above we can view the comparisons of the initial vs fine-tuned AUC scores. For random forest and logistic regression there is a slight decrease in AUC, indicating while the overall accuracy improved, the model's ability to discriminate between classes slightly diminished. The SVM showcased the most significant drop in AUC after fine-tuning, suggesting the hyperparameters chosen during fine-tuning did not generalise well to the validation data. The MNB did not drop significantly, indicating that performance was maintained. The KNN demonstrated a slight increase in AUC, suggesting that fine-tuning did not negatively affect the model's performance.

5. CONCLUSION

5.1 Key Insights

This project aimed to predict the harmfulness of bacterial species from the dataset, using several machine learning models. Initially, five models were chosen to experiment: Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbours, and Multinomial Naive Bayes. These models were then evaluated on metrics including accuracy, precision, recall, F1-score, and ROC-AUC curves. Following the initial evaluations, hyperparameter tuning was conducted using GridSearchCV to optimize model performance.

After analysis it was revealed that:

- KNN emerged as the top-performing model. After hyperparameter tuning, it showcased the highest accuracy and cross-validation score, along with a stable AUC.
- Random forest and logistic regression displayed slight decreases in AUC post-tuning, however they remained reliable in other metrics.
- The SVM experience a significant drop in AUC, highlighting the need for further model optimisation.
- MNB maintained its performance.

The analysis showcased the importance of model selection and hyperparameter optimisation in predictive modelling. The study demonstrated there was a significant variation in performance post-tuning, emphasizing the need of careful consideration in both model selection and parameter settings. By employing a diverse array of supervised learning algorithms, the research was able to perform a comprehensive evaluation of different models. This approach highlights both the strengths and weaknesses of each, thus facilitating the selection of the most appropriate model for predicting the harmfulness of bacterial species.

The fluctuations in performance metrics, including accuracy and AUC scores, distinctly illustrate the inherent trade-offs involved in hyperparameter tuning. While some models experienced an increase in accuracy post-tuning, this was sometimes accompanied by a reduced AUC score. Such discrepancies suggest potential overfitting, where the model may perform well on the training data but may fail to effectively generalise on unseen data, which is less than ideal in a real-world scenario. This underscores the necessity of balancing model complexity with generalisation ability to achieve an optimal performance.

In particular, the KNN and MNB demonstrated a notable stability after fine-tuning. This consistent performance suggests that these models are robust and suitable for similar classification tasks. Their

performance improvements were sustained across various metrics, indicating their effectiveness in handling the dataset without overfitting. Conversely, however, the SVM exhibited a decline in performance after fine-tuning, showing a significant drop in AUC score. This decrease highlights the challenges associated with hyperparameter optimisations for SVMs particularly and points the need for further investigation into parameter settings. The decline in performance suggests that the selected hyperparameters may not have been optimal, resulting in overfitting or there may possibly be other performance issues. This emphasises the importance of rigorous and iterative approaches to hyperparameter tuning to avoid such pitfalls.

Overall, the use of multiple models in this study proved invaluable. Not only did it allow for a thorough evaluation of different algorithms, but also provided insights into their relative performance and suitability for the classification task at hand. The findings reinforce the need for a nuanced approach to model selection and fine-tuning, highlighting that no single model may be universally superior across all metrics. Instead, a combination of models and careful parameter optimisation is essential for achieving the best results in a predictive analysis.

5.2 Conclusion of Objectives

In conclusion of this study, it is also essential to reflect on the effectiveness of how the objectives were met and how the insights contributed to our understanding of predictive modelling for bacterial harmfulness.

1. **To Develop a Predictive Model**: The primary goal of this project was to create and train machine learning models with the task of bacteria harmfulness prediction based on family and location of origin. This objective was achieved – a range of algorithms were successfully implemented, including random forest, logistic regression, support vector machines, k-nearest neighbours, and multinomial naïve bayes. Each model was trained on the dataset, enabling the models to learn patterns and make predictions about bacterial harmfulness.

2. **To Evaluate Model Performance**: The next objective involved comparing model performance of these different algorithms to assess accuracy in classifying bacteria as harmful or non-harmful. Models were evaluated on various metrics including accuracy, precision, recall, F1-score, AUC, and cross-validation. Results highlighted strengths and weaknesses of each model. For example, KNN and MNB emerged as best performing, while SVM demonstrated a decline after fine-tuning, illustrating the complexity of tuning and model stability.

3. **To Optimise Model Performance**: The last objective focused on enhancing the performance of the predictive models through hyperparameter tuning. Grid Search was implemented to identify the optimal parameters for each model. This process led to some improvements in the accuracy and cross-validation scores of several models. However, it also revealed the effect of

trade-offs, shown by a decline in AUC for some models, indicating potential overfitting. The results show that KNN and MNB have the most consistent performance improvements, while SVM required further investigation to refine hyperparameters to avoid a decline in performance.

To summarise, this study effectively met its objectives, by implementing various supervised machine learning models, evaluating their performance using classification reports, confusion matrixes, and cross-validation, and optimising them through using a grid search.

5.3 Limitations

While the study successfully achieved its objectives, limitations should be acknowledged:

Dataset Size and Diversity

The size of the dataset was relatively small, at 200 samples. Although this may be reasonably sized, a larger dataset would provide a more robust evaluation of model performance. Additionally, the study did not address class imbalance explicitly, it can be a common issue in many biological datasets. Imbalanced datasets may lead to biased models that favour the majority class. Techniques such as resampling or synthetic data generation e.g. SMOTE could be employed to address this issue. In this project, resampling was not performed to maintain the original distribution of the dataset and to avoid synthetic data which may not accurately represent a real-world scenario. Furthermore, the dataset's diversity was limited in terms of bacterial families and locations represented. A more diverse dataset would help in creating a more comprehensive model that would be able to generalise well across different bacterial species and environments.

Hyperparameter Tuning:

As aforementioned, the SVMs decline in performance after fine-tuning demonstrated the need for further investigation into its chosen hyperparameters. The process, while extensive, may not have explored the full range of possible parameters. It is possible to expand the range of hyperparameters tested in the grid search, however this would come at a cost of more computational time. Furthermore, other optimisation methods could be tested, including genetic algorithms or Bayesian optimisation. The computational resources available for this study limits the extent of the hyperparameter tuning, therefore experiments with larger parameter grids and longer training times may provide a more comprehensive evaluation.

Addressing these limitations in future work could lead to more accurate, reliable, and interpretable models for predicting bacterial harmfulness, ultimately contributing to better public health strategies and interventions.

5.4 Recommendations for further research

Upon reflections on the findings and limitations identified in this study, several possibilities for future research arise, recommended to enhance the understanding and predictive capabilities regarding the harmfulness of bacterial species.

Given the small nature of the dataset used, replicating this project with an expanded dataset would yield more robust results and provide further insights into possible predictors of bacterial harmfulness. For example, combining bacterial samples from a wider variety of environments would enhance the model's predictive power. Alternatively, the study could focus on a narrower selection of bacterial families, such as those most commonly found in hospital settings, while integrating more diverse factors, such as optimal temperature, genetic markers, and resistance profiles. This way, results may offer more targeted insights and improve our understanding of specific conditions and characteristics that contribute to bacteria harmfulness.

Further study could also explore advanced techniques utilising deep learning approaches, such as CNNs or recurrent neural networks (RNNs), which may capture complex patterns in the data that traditional machine learning models might not. Furthermore, unsupervised machine learning algorithms, such as clustering and dimensionality reduction methods, could be implemented to uncover hidden patterns in the data. These methods could identify natural groupings and associations in the bacterial samples, offering insights that are not constrained by predefined labels seen in supervised methods. For instance, clustering bacteria based on genetic markers may reveal additional subcategories of harmful bacteria that previously went unrecognised.

REFERENCES

A KNN Model Based on Manhattan Distance to Identify the SNARE Proteins | IEEE Journals & Magazine | IEEE Xplore [WWW Document], n.d. URL https://ieeexplore.ieee.org/abstract/document/9119343

(accessed 8.8.24). Ali, Q., Zheng, H., Rao, M.J., Ali, M., Hussain, A., Saleem, M.H., Nehela, Y., Sohail, M.A., Ahmed, A.M., Kubar, K.A., Ali, S., Usman, K., Manghwar, H., Zhou, L., 2022. Advances, limitations, and prospects of biosensing technology for detecting phytopathogenic bacteria. Chemosphere 296, 133773. https://doi.org/10.1016/j.chemosphere.2022.133773

Asgari, E., Garakani, K., McHardy, A.C., Mofrad, M.R.K., 2018. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. Bioinformatics 34, i32–i42. https://doi.org/10.1093/bioinformatics/bty296

Asnicar, F., Thomas, A.M., Passerini, A., Waldron, L., Segata, N., 2024. Machine learning for microbiologists. Nat Rev Microbiol 22, 191–205. https://doi.org/10.1038/s41579-023-00984-1

Berrazeg, M., Drissi, M., Medjahed, L., Rolain, J.M., 2013. Hierarchical clustering as a rapid tool for surveillance of emerging antibiotic-resistance phenotypes in Klebsiella pneumoniae strains. Journal of Medical Microbiology 62, 864–874. https://doi.org/10.1099/jmm.0.049437-0

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Centers for Disease Control (CDC), 1990. Increase in National Hospital Discharge Survey rates for septicemia--United States, 1979-1987. MMWR Morb Mortal Wkly Rep 39, 31–34. Croxatto, A., Marcelpoil, R., Orny, C., Morel, D., Prod'hom, G., Greub, G., 2017. Towards automated detection, semi-quantification and identification of microbial growth in clinical bacteriology: A proof of concept. Biomed J 40, 317–328. https://doi.org/10.1016/j.bj.2017.09.001

Evason, D.J., Claydon, M.A., Gordon, D.B., 2001. Exploring the limits of bacterial identification by intact cell-mass spectrometry. J. Am. Soc. Mass Spectrom. 12, 49–54. https://doi.org/10.1016/S1044-0305(00)00192-6

Giuliano, C., Patel, C.R., Kale-Pradhan, P.B., 2019. A Guide to Bacterial Culture Identification And Results Interpretation. P T 44, 192–200. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. KNN Model-Based Approach in Classification, in: Meersman, R., Tari, Z., Schmidt, D.C. (Eds.), On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

Hamilton, D., Pacheco, R., Myers, B., Peltzer, B., 2020. kNN vs. SVM: A comparison of algorithms. In: Hood, Sharon M.; Drury, Stacy; Steelman, Toddi; Steffens, Ron, [eds.]. Proceedings of the Fire Continuum-Preparing for the future of wildland fire; 2018 May 21-24; Missoula, MT. Proceedings RMRS-P-78. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 95-109. 78, 95–109. Hervé, V., Junier, T., Bindschedler, S., Verrecchia, E., Junier, P., 2016. Diversity and ecology of oxalotrophic bacteria. World J Microbiol Biotechnol 32, 28. https://doi.org/10.1007/s11274-015-1982-3

Ho, C.-S., Jean, N., Hogan, C.A., Blackmon, L., Jeffrey, S.S., Holodniy, M., Banaei, N., Saleh, A.A.E., Ermon, S., Dionne, J., 2019. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. Nat Commun 10, 4927. https://doi.org/10.1038/s41467-019-12898-9

Hospodsky, D., Yamamoto, N., Peccia, J., 2010. Accuracy, Precision, and Method Detection Limits of Quantitative PCR for Airborne Bacteria and Fungi. Appl Environ Microbiol 76, 7004–7012. https://doi.org/10.1128/AEM.01240-10

J, Veeralagan., Priya, S.M., 2022. Hyper Tuning Using Gridsearchcv on Machine Learning Models for Prognosticating Dementia. https://doi.org/10.21203/rs.3.rs-2316713/v1

Johnson, H.R., Trinidad, D.D., Guzman, S., Khan, Z., Parziale, J.V., DeBruyn, J.M., Lents, N.H., 2016. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. PLoS ONE 11, e0167370. https://doi.org/10.1371/journal.pone.0167370

Khajehali, N., Alizadeh, S., 2017. Extract critical factors affecting the length of hospital stay of pneumonia patient by data mining (case study: an Iranian hospital). Artificial Intelligence in Medicine 83, 2–13. https://doi.org/10.1016/j.artmed.2017.06.010

Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G., 2005. Multinomial Naive Bayes for Text Categorization Revisited, in: Webb, G.I., Yu, X. (Eds.), . Springer, Berlin, Heidelberg, pp. 488–499. https://doi.org/10.1007/978-3-540-30549-1_43

Kysela, D.T., Randich, A.M., Caccamo, P.D., Brun, Y.V., 2016. Diversity Takes Shape: Understanding the Mechanistic and Adaptive Basis of Bacterial Morphology. PLoS Biol 14, e1002565. https://doi.org/10.1371/journal.pbio.1002565

LaValley, M.P., 2008. Logistic Regression. Circulation 117, 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

Law, J.W.-F., Ab Mutalib, N.-S., Chan, K.-G., Lee, L.-H., 2015. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Front. Microbiol. 5. https://doi.org/10.3389/fmicb.2014.00770

Makretsov, N.A., Huntsman, D.G., Nielsen, T.O., Yorida, E., Peacock, M., Cheang, M.C.U., Dunn, S.E., Hayes, M., Van De Rijn, M., Bajdik, C., Gilks, C.B., 2004. Hierarchical Clustering Analysis of Tissue Microarray Immunostaining Data Identifies Prognostically Significant Groups of Breast Carcinoma. Clinical Cancer Research 10, 6143–6151. https://doi.org/10.1158/1078-0432.CCR-04-0429

Mandal, P.K., Biswas, A.K., Choi, K., Pal, U.K., 2011. Methods for Rapid Detection of Foodborne Pathogens: An Overview. American J. of Food Technology 6, 87–102. https://doi.org/10.3923/ajft.2011.87.102

Maurer, J.J., 2011. Rapid detection and limitations of molecular techniques. Annu Rev Food Sci Technol 2, 259–279. https://doi.org/10.1146/annurev.food.080708.100730

Nagarajan, K., Loh, K.-C., 2014. Molecular biology-based methods for quantification of bacteria in mixed culture: perspectives and limitations. Appl Microbiol Biotechnol 98, 6907–6919. https://doi.org/10.1007/s00253-014-5870-9

Nakao, H., Magariyama, Y., 2022. Simple and rapid method for selective enumeration of lactic acid bacteria in commercially prepared yogurt by image analysis and K-means clustering. ANAL. SCI. 38, 191–197. https://doi.org/10.2116/analsci.21P273

National Research Council (U.S.) (Ed.), 2004. Science, medicine, and animals: a circle of discovery. National Research Council, National Academies Press, Washington, DC. Nurlaila, I., Irawati, W., Purwandari, K., Pardamean, B., 2021. K-Means Clustering Model to Discriminate Copper-Resistant Bacteria as Bioremediation Agents. Procedia Computer Science 179, 804–812. https://doi.org/10.1016/j.procs.2021.01.068

Parmar, A., Katariya, R., Patel, V., 2019. A Review on Random Forest: An Ensemble Classifier, in: Hemanth, J., Fernando, X., Lafata, P., Baig, Z. (Eds.), International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Springer International Publishing, Cham, pp. 758–763. https://doi.org/10.1007/978-3-030-03146-6_86

Peiffer-Smadja, N., Dellière, S., Rodriguez, C., Birgand, G., Lescure, F.-X., Fourati, S., Ruppé, E., 2020. Machine learning in the clinical microbiology laboratory: has the time come for routine practice? Clinical Microbiology and Infection 26, 1300–1309. https://doi.org/10.1016/j.cmi.2020.02.006

Qaiser, S., Ali, R., 2018. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. IJCA 181, 25–29. https://doi.org/10.5120/ijca2018917395

Qu, K., Guo, F., Liu, X., Lin, Y., Zou, Q., 2019. Application of Machine Learning in Microbiology. Front. Microbiol. 10, 827. https://doi.org/10.3389/fmicb.2019.00827

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408. https://doi.org/10.1037/h0042519

Satyanarayana, K.V., Rao, N.T., Bhattacharyya, D., Hu, Y.-C., 2022. Identifying the presence of bacteria on digital images by using asymmetric distribution with k-means clustering algorithm. Multidim Syst Sign Process 33, 301–326. https://doi.org/10.1007/s11045-021-00800-0

Tabit, F.T., 2016. Advantages and limitations of potential methods for the analysis of bacteria in milk: a review. J Food Sci Technol 53, 42–49. https://doi.org/10.1007/s13197-015-1993-y

Tarca, A.L., Carey, V.J., Chen, X., Romero, R., Drăghici, S., 2007. Machine Learning and Its Applications to Biology. PLoS Comput Biol 3, e116. https://doi.org/10.1371/journal.pcbi.0030116

Tu, J.V., Guerriere, M.R.J., 1993. Use of a Neural Network as a Predictive Instrument for Length of Stay in the Intensive Care Unit Following Cardiac Surgery. Computers and Biomedical Research 26, 220–229. https://doi.org/10.1006/cbmr.1993.1015

Wu, Y., Gadsden, S.A., 2023. Machine learning algorithms in microbial classification: a comparative analysis. Front Artif Intell 6, 1200994. https://doi.org/10.3389/frai.2023.1200994

Xie, K., Guo, L., Bai, Y., Liu, W., Yan, J., Bucher, M., 2019. Microbiomics and Plant Health: An Interdisciplinary and International Workshop on the Plant Microbiome. Molecular Plant 12, 1–3. https://doi.org/10.1016/j.molp.2018.11.004

Zou, Q., Chen, L., Huang, T., Zhang, Z., Xu, Y., 2017. Machine learning and graph analytics in computational biomedicine. Artif Intell Med 83, 1. https://doi.org/10.1016/j.artmed.2017.09.003