TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DT	Decision Tree
FN	False Negative
FP	False Positive
GBM	Gradient-Boost Machine
KNN	K-Nearest Neighbour
LR	Logistic Regression
MCA	Multiple Correspondence Analysis
MLR	Multinomial Logistic Regression
RF	Random Forest
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

CHAPTER 1

INTRODUCTION

1.1 General Introduction

The lungs are one of the most vital organs in the body as it controls human respiratory system. These respiratory conditions, including tuberculosis, pneumonia, bronchial asthma, and the common cold. Among them, common cold persists as significant public health concern. It is often one of the foremost health problem people consider.

The common cold, often shortened as a cold, is a viral infection affecting the upper respiratory tract, whereas pneumonia is caused by a fungal infection in the lungs. While most colds are minor and can be resolved with rest, their symptoms can overlap with more serious conditions such as pneumonia. Table 1.1 below compares the symptoms of the common cold with those of pneumonia.

Table 1.1: Comparison of symptoms between common cold and pneumonia

Common Cold	**Pneumonia**
Low-grade fever	High-grade fever
Cough (without mucus)	Cough (with or without mucus)
Fatigue	Fatigue
Sneezing	Wheezing
Runny nose	Trouble breathing
Sore throat	Chest pain
	Shaking
	Nausea, vomiting or diarrhoea

(Source: What's the difference between a cold, bronchitis and pneumonia? 2022)

Furthermore, tuberculosis and pneumonia are significant respiratory diseases that shares similar clinical traits which leads to high levels of illness and death in human. They also share similar transmission methods, such as through sneezing, coughing, speaking, or even spitting by infected individuals

(Gweryina et al., 2023). However, tuberculosis is frequently misdiagnosed as pneumonia which is highlighted in a case report by the Department of Pulmonary Diseases in India (Pinto et al., 2011).

On the other hand, studies show that bronchial asthma and tuberculosis are related. Approximately 80% of patients who developed tuberculosis were diagnosed within 3 years after the last inhaled corticosteroids (ICS) prescription (Lee et al., 2019), which is by far the most effective controllers used in the treatment of asthma.

These respiratory diseases pose significantly higher health risks compared to the symptoms of common cold, which emphasizes the urgent need for increased awareness and early detection in order to reduce the negative effects on public health.

1.2 Importance of the Study

The importance of this research arises to address challenges in diagnosing respiratory diseases. In recent years, there has been a big increase in the application of predictive modeling techniques, supported by advancements in data science and artificial intelligence. This increase highlights a growing recognition of predictive modeling's potential to be implemented in various fields, especially healthcare.

In the field of disease diagnosis, predictive modeling offers a promising approach to complement traditional diagnostic methods, which involve in physical examination, medical history, laboratory tests, and other clinical assessments. These traditional diagnostic methods rely on healthcare professionals to interpret symptoms, signs, and test results to arrive at a diagnosis. Predictive modelling significantly enhances this process. By examining large number of datasets and patterns, predictive models can provide valuable insights, potentially improving diagnostic accuracy and efficiency.

Moreover, predictive models excel in handling vast datasets and complex patterns, which can be challenging for healthcare professionals to manage manually due to limitations in time and resources. With manpower often constrained, predictive modeling offers a crucial tool to effectively analyze large amounts of patient profiles, allowing for more comprehensive and efficient diagnosis and treatment.

1.3 Problem Statement

The healthcare industry is facing a critical need for more accurate and efficient disease diagnosis, especially in differentiating between symptoms of common colds and more severe respiratory illnesses such as bronchial asthma, tuberculosis, and pneumonia. This identification of different respiratory diseases is crucial as it influences how patients are treated. A misdiagnosis may worsen a patient's condition and cause delays in getting treatment.

According to information from Ministry of Health Malaysia, it indicates that respiratory diseases rank among the top three causes of death and hospitalization in Malaysia (KKM Health Facts 2017 (final)). Pneumonia is one of the respiratory infections that develops rapidly and has a relatively short duration but could result in long-term conditions, while bronchial asthma is one of the chronic obstructive lung diseases that can be managed at the primary level (Respiratory Disease).

1.4 Aim and Objectives

- To develop predictive models for diagnosing respiratory system disease based on patient symptoms.
- To evaluate the classification performance in developing predictive models.

1.5 Scope and Limitation of the Study

The focus of this study is lies in applying predictive modelling techniques to diagnose respiratory system diseases based on patient symptoms. To facilitate this investigation, the dataset is sourced from Kaggle, which is known as the Disease Prediction Data. Initially, this dataset includes 42 different types of diseases, each associated with 132 symptoms, and resulting in a total of 4920 instances.

In discussing the limitations of the study, the dataset size and its composition should be highlighted. The dataset is relatively small, which may impact the ability to generalize the findings to larger populations or more diverse datasets. Additionally, the presence of many duplicated rows and the limited number of unique rows in the dataset pose challenges for model training and evaluation. These factors could contribute to overfitting, reduced model performance, and potential bias in the models.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This literature review examines various predictive models, including Logistic Regression (LR), K-Nearest Neighbours (KNN), Gradient-Boost Machine (GBM), and Support Vector Machine (SVM), across different applications. It explores the differences between each model and their applications in diverse contexts within the medical field.

2.2 SMOTE Oversampling

In traditional random oversampling, instances of the rare class are duplicated and added to the dataset to increase the representation of the minority class. A more effective oversampling technique, Synthetic Minority Over-sampling Technique (SMOTE), was developed by Chawla, Bowyer, Hall and Kegelmeyer (2002), which synthesizes new instances of the rare class. Specifically, it selects an observation from the rare class and then randomly picks an example from its k nearest neighbours to create synthetic samples. Studies have demonstrated that SMOTE oversampling outperforms random oversampling (Fernandez, Garcia, Herrera and Chawla, 2018) and is widely considered as the standard method for learning from imbalanced datasets.

2.3 Comparison of Different Models for Respiratory Disease Classification

The research focused on classifying respiratory diseases using LR, lasso regularization, and decision tree (DT). Specifically, DT classifications included tree bagging, random forest (RF), and GBM. Following evaluation, GBM demonstrated the best performance and was selected for comparison with other classification models. The aim was to predict occurrences of acute respiratory infection and diarrhoea in both rural and urban settings in Uganda.

After evaluating all DT models and determining the one with the best performance, GBM stood out as the preferred choice (Kananura, 2022). The selection criteria were based on the accuracy and sensitivity of the model in predicting the occurrence of acute respiratory disease and diarrhoea. Ultimately, GBM exhibited superior performance compared to other models, with logistic regression performing the least effectively.

2.4 Comparison of Logistic Regression Across Different Medical Applications

Numerous research attempts have explored into the effectiveness of LR across diverse medical realms, including cardiovascular diseases, kidney injuries, and chronic illnesses. Through careful examination, these studies aim to focus on how logistic regression performs within each unique medical domain, offering insights into its applicability and performance.

2.4.1 Cardiovascular Prediction

In the case study involving the use of LR and other models to predict cardiovascular diseases, LR had the lowest accuracy, the lowest F1-score, and the shortest training time among the various models, including KNN and SVM (Saim and Ammor, 2022). In medical diagnosis tasks, a higher F1-score is preferable, as missing a disease is more critical than falsely diagnosing it. Consequently, LR demonstrated the worst performance when compared to the other models.

2.4.2 Kidney Injury Risk Factor Prediction

In another study investigating the effects of titanium dioxide nanoparticles on kidney injury risk factors, LR was applied along with other models. The findings showed that LR achieved an accuracy rate of 82%, considerably lower than SVM's accuracy of 96% on the training data. Additionally, SVM obtained an F1-score of 96%, while LR scored 81% (Tu et al., 2023). Thus, it was evident that LR was not the optimal model for this case study.

2.4.3 Heart Disease

The research conducted by Binus University from India focused on the classification of heart disease using LR and other models. It aimed to determine the risk of a person acquiring heart disease and explore preventive measures for heart disease.

The results indicated that LR achieved the highest accuracy rate of 92.59 (Boer et al., 2023). In medical applications, high recall is extremely important than precision as it ensures that no patients are missed out. According to Figure 2.2, each model exhibited the same percentage of recall along accuracy. Consequently, it could be concluded that LR stood out as the best model for predicting heart disease.

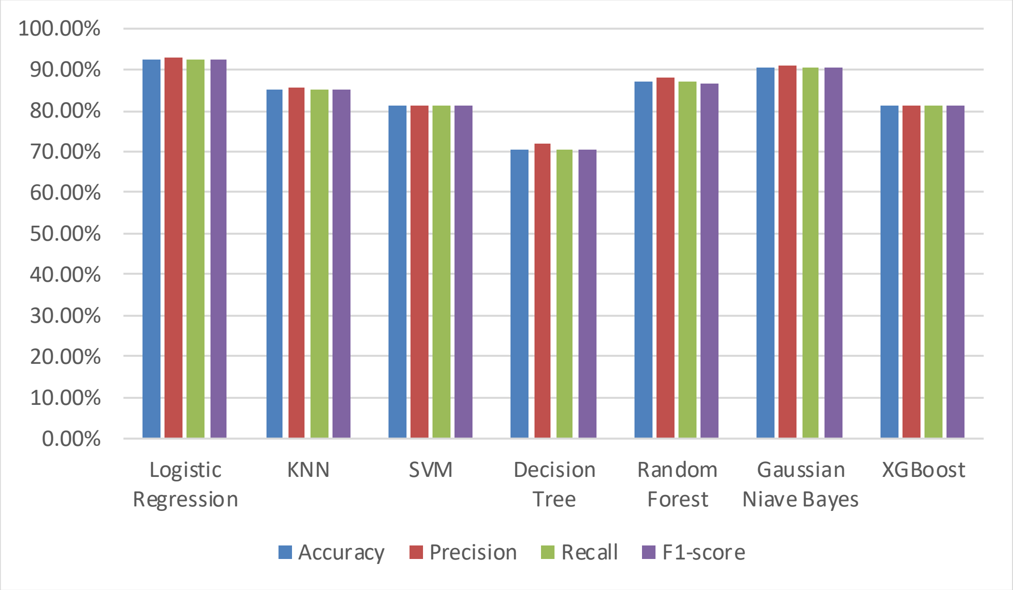


Figure 2.1: Comparison of logistic regression and other models for classifying heart disease (Boer et al., 2023)

2.4.4 Chronic Disease Prediction

On the contrary, a study focused on using LR to predict major chronic diseases demonstrated that LR performed comparably well to other models. In fact, among the various models, LR exhibited the most accurate performance in predicting the risk for chronic diseases. The researchers suggested that LR played a significant role in disease risk prediction (Singh et al., 2024).

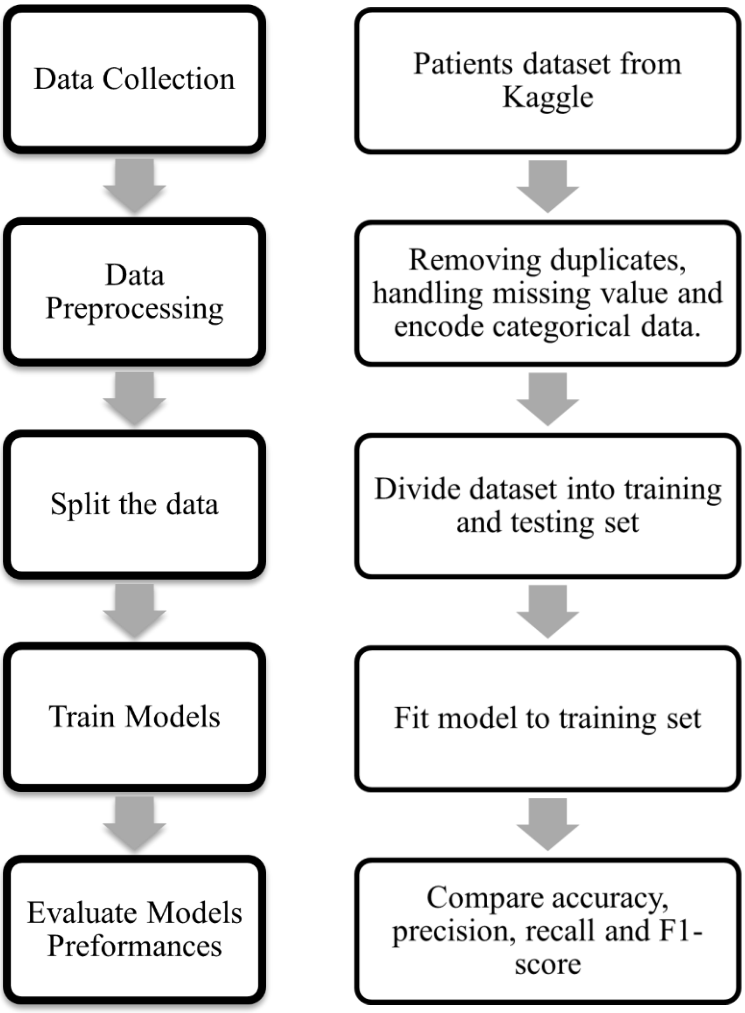
2.5 Summary

In summary, LR demonstrated as a strong model in predicting heart disease, demonstrating high accuracy and recall rates. However, its performance varied across different medical applications, showing mixed results in predicting cardiovascular diseases, respiratory disease, and kidney injury risk factors. Further research is needed to explore its effectiveness across diverse healthcare scenarios.

CHAPTER 3

METHODOLOGY AND WORK PLAN

3.1 Introduction



Data Collection
Patients dataset from Kaggle
Data Preprocessing
Removing duplicates, handling missing value and encode categorical data.
Split the data
Divide dataset into training and testing set
Train Models
Fit model to training set
Evaluate Models Preformances
Compare accuracy, precision, recall and F1-score

Figure 3.1: Flowchart of predictive model

3.2 Data Collection

The dataset used in this project is sourced from Kaggle and was provided by Marslino Edward under the title "Disease Prediction Data." The data encompasses 42 distinct disease types, each associated with 132 patient symptoms, resulting in a total of 4,920 instances.

3.3 Data Preprocessing

The data preprocessing phase involved checking the dataset for missing values and duplicates. There were no missing values, but many duplicated rows were found, resulting in only 304 unique rows. The impact of duplicates on model performance will be assessed through three different training scenarios: using the full dataset with all duplicates, a partial dataset that retains 10 rows from each group of duplicates, and a dataset consisting of only the 304 unique rows. This approach allows for comparisons of model performance across varying levels of data duplication.

Next, data encoding was performed. The categorical target output was encoded into five classes: 'Common Cold' (0), 'Bronchial Asthma' (1), 'Tuberculosis' (2), 'Pneumonia' (3), and all other diseases that are not respiratory disease as '4'. This encoding simplifies the classification process and aids models in learning the relationships between features and target classes more efficiently. Since all features are binary, no additional encoding was needed.

Multiple Correspondence Analysis (MCA) was used for each of the three training scenarios to find the best features for model development since the best features differ for each scenario. By applying MCA separately to each variation of the dataset, it was possible to reduce dimensionality and select the most important features specific to each scenario. This approach allows the models to train on the most relevant and informative features, improving model performance and enabling more meaningful comparisons across the different training scenarios.

Given that the categorical target output was encoded into five classes—'Common Cold' (0), 'Bronchial Asthma' (1), 'Tuberculosis' (2), 'Pneumonia' (3), and all other diseases as '4'—the target variable exhibits class imbalance. The four respiratory diseases (classes 0 through 3) may not be equally represented compared to class 4, which encompasses all other diseases. This imbalance can skew model performance by biasing predictions towards the majority class.

To mitigate this issue and improve model training, oversampling and under sampling techniques were employed. SMOTE (Synthetic Minority Over-sampling Technique) was used to increase the number of instances in the minority classes to a specified count, providing the model with more examples from the less common classes. Additionally, random under sampling was applied to reduce the number of instances in the majority class to a specified amount.

3.4 Model Development

The dataset is divided into training and testing sets for model development, with 70% of the data allocated to the training set and 30% to the testing set, using a test_size of 0.3. Various machine learning models, including Multinomial Logistic Regression (MLR), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Machine (GBM), and Decision Tree (DT), are then developed to compare their efficiency in predicting respiratory diseases.

Based on the literature review conducted, LR was found to be a reliable model with strong performance. However, due to its applicability being limited to dependent variables with only two possible outcomes, the choice was made to employ MLR instead. Both multinomial and binary logistic regression models utilize the logistic function to capture the relationship between independent variables and the outcome. Given that the dataset contains four potential outcomes, MLR was considered the suitable approach.

The decision to utilize KNN is because it is a simple yet effective algorithm for classification and regression tasks. In classification, given a data point, KNN identifies the K-nearest neighbours to that point based on a chosen distance metric in the feature space. KNN then assigns the class label to the data point based on the majority class among its K-neighbours.

SVM is chosen for its performance in some cases conducted in literature review. It works by finding the optimal hyperplane that maximizes the margin between different classes in the feature space (Zafar and Iqbal, 2020). Equation (1) represents the equation of optimal hyperplane,

$$wx + b = 0 \tag{1}$$

where

w = weight vector

b = bias.

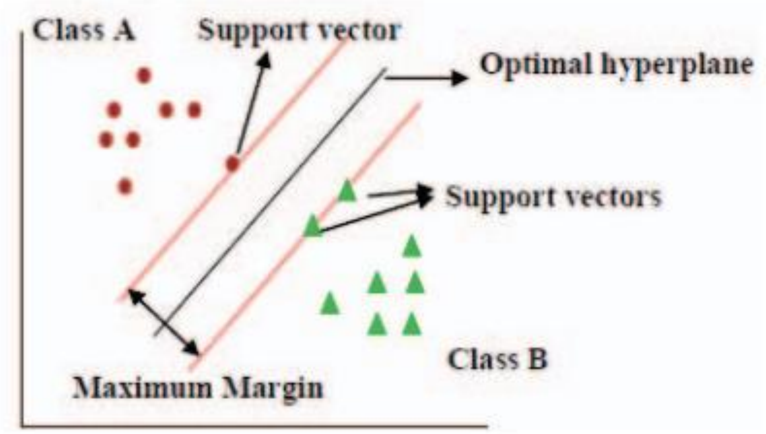


Figure 3.2: SVM (Zafar and Iqbal, 2020)

Additionally, RF is selected for its demonstrated performance, as evidenced by various cases reviewed in the literature. In one study on heart disease prediction, it ranked as the third-best performing model. The functioning of RF involves combining predictions from multiple decision trees. Each tree is trained on a random subset of the data and features. During prediction, each tree offers its own prediction, and the most common prediction among all trees is considered as the final prediction. This approach is simple and can handle a wide range of tasks.

The DT is still chosen despite its poor performance in the heart disease prediction case study mentioned in the literature review. The model operates by dividing the dataset into smaller subsets based on features, starting from the root node, and making decisions at each node to split the data further. This recursive process continues until leaf nodes are reached, providing the final predictions.

Finally, GBM is chosen for its exceptional performance in predicting respiratory diseases, as demonstrated by the literature review. GBM starts with a simple model and progressively improves its accuracy by targeting the errors of previous models in each iteration. It combines predictions from multiple weak models, often decision trees, to form a robust ensemble model.

3.5 Evaluation Metrics

To compare the performance of the models, evaluation metrics such as precision, accuracy, and recall are utilized. A confusion matrix is generated to visualize the number of correct and incorrect predictions made by each model compared to the actual classifications in the test data. High precision and recall indicate accurate diagnosis of respiratory diseases. Additionally, the F1-score, which balances precision and recall, is employed to further evaluate model performance.

Table 3.1: Confusion Matrix

		Actual class	
		Positive	Negative
Predicted class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Precision: $\frac{TP}{(TP+FP)}$

Accuracy: $\frac{TP+TN}{(TP+TN+FP+FN)}$

Recall: $\frac{TP}{(TP+FN)}$

F1-Score: $\frac{2 * Recall * Precision}{Recall + Precision}$

Precision represents the proportion of relevant cases correctly identified, while recall measures the ability of the classifier to capture all relevant cases. Additionally, accuracy measures the overall correctness of the classifier, while the F1-score provides a balanced evaluation by considering both precision and recall in a single metric.

3.6 Decision Making

Following evaluation using the confusion matrix, the model that achieves a balance between precision, recall, accuracy, and F1-score is considered the best choice in predicting respiratory diseases.

CHAPTER 4

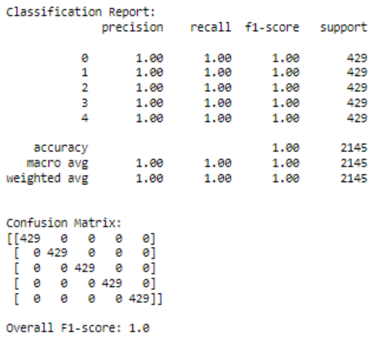
PREMILINARY RESULTS

4.1 Performance of the Models

Table 4.1: Performance of default models

	Full Dataset		**Partial Dataset**		**Unique Dataset**	
	Full Features	**Best Features**	**Full Features**	**Best Features**	**Full Features**	**Best Features**
MLR	1.0	1.0	1.0	0.9944	1.0	0.7477
KNN	1.0	1.0	1.0	0.9944	1.0	0.7477
SVC	1.0	1.0	1.0	0.9944	1.0	0.7477
RF	1.0	1.0	1.0	0.9944	0.7614	0.7477
GBM	1.0	1.0	1.0	0.9944	0.7333	0.7477
DT	1.0	1.0	1.0	0.9944	0.5738	0.6719

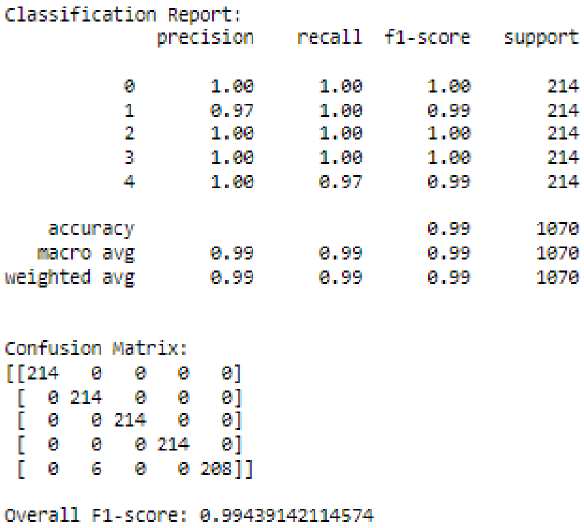
Table 4.1 presents the F1-scores of each model across different scenarios. In every model scenario using the full dataset, an F1-score of 1.0 was observed regardless of whether full or selected best features were utilized. According to Figure 4.1, the confusion matrix reveals that accuracy, precision, recall, and F1-score all reached a perfect score of 1.0. This outcome suggests the presence of overfitting or bias in the models, likely due to the inclusion of duplicated rows that were not removed.



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       429
           1       1.00      1.00      1.00       429
           2       1.00      1.00      1.00       429
           3       1.00      1.00      1.00       429
           4       1.00      1.00      1.00       429

    accuracy                           1.00      2145
   macro avg       1.00      1.00      1.00      2145
weighted avg       1.00      1.00      1.00      2145

Confusion Matrix:
[[429   0   0   0   0]
 [  0 429   0   0   0]
 [  0   0 429   0   0]
 [  0   0   0 429   0]
 [  0   0   0   0 429]]

Overall F1-score: 1.0
```

Figure 4.1: Classification report of every model in full dataset

Given the perfect F1-score of 1.0 in every model trained on the full dataset, the models were retrained using another scenario: a partial dataset that retains 10 rows from each group of duplicates. Table 4.1 shows that while the F1-score for all models remained at 1.0 when using full features, it decreased slightly to 0.9944 when using the best features selected through MCA.

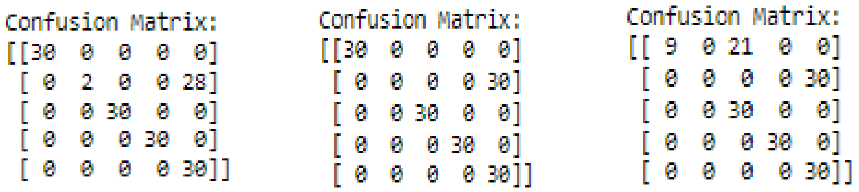


```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       214
           1       0.97      1.00      0.99       214
           2       1.00      1.00      1.00       214
           3       1.00      1.00      1.00       214
           4       1.00      0.97      0.99       214

    accuracy                           0.99      1070
   macro avg       0.99      0.99      0.99      1070
weighted avg       0.99      0.99      0.99      1070


Confusion Matrix:
[[214   0   0   0   0]
 [  0 214   0   0   0]
 [  0   0 214   0   0]
 [  0   0   0 214   0]
 [  0   6   0   0 208]]

Overall F1-score: 0.99439142114574
```

Figure 4.2: Performance of the models using best features in the partial dataset

The classification report and confusion matrix in Figure 4.2 show that the decrease is primarily due to a small number of errors in Class 4, which has a recall of 0.97 (6 instances were misclassified out of 214) compared to the perfect recall scores for the other classes.

When models were trained using the unique dataset with all available features, their performance varied across different models. MLR, KNN, and SVC consistently achieved a perfect F1-score of 1.0, reflecting outstanding classification accuracy. In contrast, the other models displayed notably lower performance. Upon training the models with the unique dataset limited to the best features, the F1-scores for most models are 0.7477, except for the DT. This consistent performance suggests that selecting the best features through MCA helped stabilize model outcomes, resulting in more uniformity across different models. However, DT was an exception, with an F1-score of 0.6719, implying that it may have been particularly affected by the reduction in data size and the selection of optimal features.



```
Confusion Matrix:
[[30  0  0  0  0]
 [ 0  2  0  0 28]
 [ 0  0 30  0  0]
 [ 0  0  0 30  0]
 [ 0  0  0  0 30]]
```

RF

```
Confusion Matrix:
[[30  0  0  0  0]
 [ 0  0  0  0 30]
 [ 0  0 30  0  0]
 [ 0  0  0 30  0]
 [ 0  0  0  0 30]]
```

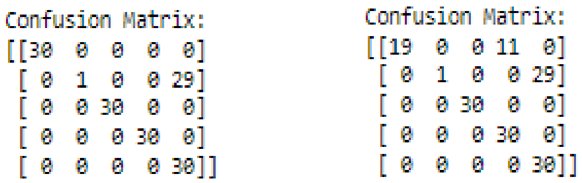
GBM

```
Confusion Matrix:
[[ 9  0 21  0  0]
 [ 0  0  0  0 30]
 [ 0  0 30  0  0]
 [ 0  0  0 30  0]
 [ 0  0  0  0 30]]
```

DT

Figure 4.3: Confusion matrix of RF, GBM, and DT when using full features in unique dataset

Figures 4.3 and 4.4 illustrate that most instances in Class 1 are misclassified as Class 4 across all models and scenarios. In Figure 4.4, the DT model shows confusion between Class 0 and Class 3, while in Figure 4.3, there is a confusion between Class 0 and Class 2. This might be due to a structural similarity between the instances in these classes.



```
Confusion Matrix:
[[30  0  0  0  0]
 [ 0  1  0  0 29]
 [ 0  0 30  0  0]
 [ 0  0  0 30  0]
 [ 0  0  0  0 30]]
```

MLR, KNN, SVC, RF, GBM

```
Confusion Matrix:
[[19  0  0 11  0]
 [ 0  1  0  0 29]
 [ 0  0 30  0  0]
 [ 0  0  0 30  0]
 [ 0  0  0  0 30]]
```

DT

Figure 4.4: Confusion matrix of RF, GBM, and DT when using full features in unique dataset

4.2 Conclusion

Based on this analysis MLR, KNN, and SVC are the top-performing models as they showed consistent and strong performance across all three scenarios. These models are likely the best choices for classifying respiratory diseases and other diseases given their reliable results.

Despite the strong performance of these models, there remains an issue with Class 4 (Other diseases) frequently being misclassified as Class 1

(Bronchial Asthma) in the unique dataset. This suggests that further investigation is needed into the relationship of symptoms between other diseases and bronchial asthma. Understanding the similarities in features or patterns between Class 4 and Class 1 could lead to better model training strategies and improved classification accuracy.

On the other hand, DT consistently underperforms compared to the other models. This poor performance may be attributed to the model's sensitivity to data variations and its propensity to overfit. Consequently, DT might not be the best choice for the classification tasks in this analysis.

4.3 Future Work Plan

The future work plan involves investigating several areas to enhance prediction performance. A detailed approach to model parameter optimization includes exploring different distance metrics for KNN, going beyond the standard Euclidean distance. Analysing various levels of SMOTE oversampling and random under sampling aims to balance classes optimally for improved model training and classification. Moreover, examining the effects of varying MCA thresholds will help identify the most important features and improve model performance. Furthermore, exploring different test sizes in the train-test split process may reveal the ideal data proportions for robust and accurate models. This comprehensive approach seeks to enhance predictive capabilities and overall model quality.

WORK SCHEDULE

Project I

The Gantt chart provides a structured timeline for the Project I over a 12-week period. It begins with foundational tasks such as identifying the project title and analysing the dataset in weeks 1-5. During weeks 4-6, the focus is on writing the proposal draft, researching the literature review, and testing the model. Weeks 7-8 involve completing the proposal. The interim report is drafted in weeks 7-9, and presentation slides are prepared in week 9-10. The final stages include finalizing the interim report in week 9-12, leading up to the project's conclusion.

ACTIVITIES / WEEK	1	2	3	4	5	6	7	8	9	10	11	12
Identifying project title	🟦											
Discussing project details with supervisor		🟦	🟦	🟦	🟦							
Analysing dataset				🟦	🟦							
Writing proposal draft				🟦	🟦	🟦						
Research for literature review				🟦	🟦	🟦						
Testing on model				🟦	🟦	🟦						
Completing proposal							🟦	🟦				
Drafting interim report							🟦	🟦	🟦			
Preparing presentation slide									🟦	🟦		
Finalizing interim report									🟦	🟦	🟦	🟦

Project II

The Gantt chart provides a structured timeline for the Project II over a 12-week period. It begins with discussing project details with the supervisor. The focus then shifts to tuning the models for optimal performance. Following this, a draft of the final report is prepared, summarizing the project's findings and conclusions. The next steps involve finalizing both the report and the poster for submission. The week concludes with preparations for the presentation.

ACTIVITIES / WEEK	1	2	3	4	5	6	7	8	9	10	11	12
Discussing project details with supervisor	🟦	🟦										
Tune the models	🟦	🟦	🟦	🟦	🟦							
Final report draft						🟦	🟦	🟦				
Prepare poster								🟦	🟦			
Finalized report									🟦	🟦		
Finalized poster									🟦	🟦	🟦	
Prepare presentation												🟦

REFERENCES

Boer, Y., Valencia, L., Setiadi, M.R., Setiawan, K.E. and Hasani, M.F. (2023). Classification of Heart Disease: Comparative Analysis using KNN, Random Forest, Gaussian Naive Bayes, XGBoost, SVM, Decision Tree, and Logistic Regression. doi:https://doi.org/10.1109/icoris60118.2023.10352195.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, [online] 16(16), pp.321–357. doi:https://doi.org/10.1613/jair.953.

Fernandez, A., Garcia, S., Herrera, F. and Chawla, N.V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, pp.863–905. doi:https://doi.org/10.1613/jair.1.11192.

Gweryina, R.I., Madubueze, C.E., Bajiya, V.P. and Esla, F.E. (2023). Modeling and analysis of tuberculosis and pneumonia co-infection dynamics with cost-effective strategies. *Results in Control and Optimization*, 10, p.100210. doi:https://doi.org/10.1016/j.rico.2023.100210.

Kananura, R.M. (2022). Machine learning predictive modelling for identification of predictors of acute respiratory infection and diarrhoea in Uganda’s rural and urban settings. *PLOS global public health*, 2(5), pp.e0000430–e0000430. doi:https://doi.org/10.1371/journal.pgph.0000430.

KKM Health Facts 2017 (final) (no date) *Ministry of Health Malaysia. Health Facts Planning Division Health Informatics Centre. MOH/S/RAN/47.17 (AR) 2017*. Available at: https://myhdw.moh.gov.my/public/documents/20186/150084/HEALTH+FACTS+2017/98041185-ce34-4877-9ea1-4d5341e43187?version=1.1&download=true (Accessed: 04 March 2024).

Lee, C.-M., Heo, J., Han, S.-S., Moon, K.W., Lee, S.-H., Kim, Y.-J., Lee, S.-J. and Kwon, J.-W. (2019). Inhaled Corticosteroid-Related Tuberculosis in the Real World Among Patients with Asthma and COPD: A 10-Year Nationwide Population-Based Study. *The Journal of Allergy and Clinical Immunology: In Practice,* 7(4), pp.1197-1206.e3. doi:https://doi.org/10.1016/j.jaip.2018.10.007.

Pinto, L.M., Shah, A.C., Shah, K.D. and Udwadia, Z.F. (2011). Pulmonary tuberculosis masquerading as community acquired pneumonia. *Respiratory Medicine CME,* 4(3), pp.138–140. doi:https://doi.org/10.1016/j.rmedc.2010.11.004.
.908

Respiratory Disease (no date) *Welcome to Department of Primary Care Medicine*. Available at: https://pcm.um.edu.my/respiratory-disease (Accessed: 04 March 2024).

Saim, M.M. and Ammor, H. (2022). Comparative study of machine learning algorithms (SVM, Logistic Regression and KNN) to predict cardiovascular diseases. *E3S Web of Conferences*, [online] 351, p.01037. doi:https://doi.org/10.1051/e3sconf/202235101037.

Singh, P., Kourav, P.S., Mohapatra, S., Kumar, V. and Panda, S.K. (2024). Human heart health prediction using GAIT parameters and machine learning model. *Biomedical Signal Processing and Control*, 88, pp.105696–105696. doi:https://doi.org/10.1016/j.bspc.2023.105696.

Tu, J., Hu, L., Mohammed, K.J., Le, B.N., Chen, P., Ali, E., Elhosiny Ali, H. and Sun, L. (2023). Application of logistic regression, support vector machine and random forest on the effects of titanium dioxide nanoparticles using macroalgae in treatment of certain risk factors associated with kidney injuries. 220, pp.115167–115167. doi:https://doi.org/10.1016/j.envres.2022.115167

What’s the difference between a cold, bronchitis and pneumonia? (2022) *Keck Medicine of USC*. Available at: https://www.keckmedicine.org/blog/what-is-the-difference-between-a-cold-bronchitis-and-pneumonia/

Zafar, A. and Iqbal, A. (2020). Machine Reading of Arabic Manuscripts using KNN and SVM Classifiers. doi:https://doi.org/10.23919/indiacom49435.2020.90836.