Rick van Ravestein (student), Max Verwijmeren (student) and Krishna Viegas-d’Abreu (student)

HAN University of Applied Sciences

29-10-2023



**Project Data Mining  
Asthma Disease Prediction**

**Title:**  Project Data Mining

**Sub-title:** Asthma Disease Prediction

**Students:**  Rick van Ravestein   
 1639258   
 E: [DJ.vanRavestein@student.han.nl](mailto:DJ.vanRavestein@student.han.nl)

Max Verwijmeren

1634976

[MCJ.Verwijmeren@student.han.nl](mailto:MCJ.Verwijmeren@student.han.nl)

Krishna Viegas-d’Abreu

1632691

[KG.Viegasdabreu@student.han.nl](mailto:KG.Viegasdabreu@student.han.nl)

**Study Backgrounds:** Business Administration,   
 Human Resource Management and  
 Applied Psychology

**Class:** MDD-01

**Educational Institution**:  Han University of Applied Sciences

 Ruitenberglaan 26

 6826 CC Arnhem   
 T: 024 353 0500

**Lecturer:**  Witek ten Hove

 E: [Witek.tenHove](mailto:Witek.tenHove)@han.nl

**Date:** 29-10-2023

**Place:** Arnhem

**Version:**  1.0

Table of Contents

[Chapter 1 Introduction 3](#_Toc149337084)

[Chapter 2 Dataset 4](#_Toc149337085)

[2.1 Chosen Dataset 4](#_Toc149337086)

[2.2 Dataset in Detail 4](#_Toc149337087)

[Chapter 3 Predictive models 5](#_Toc149337088)

[3.1 K-Nearest Neighbours model 5](#_Toc149337089)

[3.2 Neural Network Model 5](#_Toc149337090)

[3.3 Bernoulli Naive Bayes 6](#_Toc149337091)

[3.4 Decision Tree 6](#_Toc149337092)

[3.5 Logistic Regression 7](#_Toc149337093)

[Chapter 4 Model outcomes and conclusion 8](#_Toc149337094)

[4.1 KNN Outcome 8](#_Toc149337095)

[4.2 Neural Network Outcome 8](#_Toc149337096)

[4.3 Bernoulli Naive Bayes Outcome 9](#_Toc149337097)

[4.4 Decision Tree Outcome 9](#_Toc149337098)

[4.5 Logistic Regression Outcome 9](#_Toc149337099)

[4.6 Conclusion outcome 10](#_Toc149337100)

[Bibliography 11](#_Toc149337101)

# Chapter 1 Introduction

This report presents an in-depth analysis and evaluation of various machine learning models applied to predict asthma severity based on health-related attributes and demographic information. The used machine learning models applied in this report are the K-Nearest Neighbours model, the Neural Network Model, the Decision Tree and Logistic Regression model.

The dataset utilized for this analysis encompasses a wide range of factors, including symptoms along with demographics such as age groups and gender distinctions. Additionally, the severity of health conditions is classified into three categories: mild, moderate and none. This extensive dataset offers a comprehensive view of health factors and demographics, making it a valuable resource for conducting detailed medical research and predictive analytics within the healthcare domain.

# Chapter 2 Dataset

## 2.1 Chosen Dataset

Dataset: <https://www.kaggle.com/datasets/deepayanthakur/asthma-disease-prediction>

## 2.2 Dataset in Detail

The dataset encompasses a wide array of health-related attributes, including symptoms like tiredness, dry cough, difficulty in breathing, sore throat, pains, nasal congestion, and runny nose, as well as indicators for the absence of symptoms. It also accounts for age groups ranging from 0-9 to 60+ and distinguishes between genders (female and male). Furthermore, it classifies the severity of a health condition into three categories: mild, moderate, and none. This rich dataset offers a holistic view of health factors and demographics, making it a valuable resource for conducting in-depth medical research and predictive analytics in the realm of healthcare (Thakur, 2023).

# Chapter 3 Predictive models

In this chapter, we will explain the models used in the project, their purpose, and how they work.

## 3.1 K-Nearest Neighbours model

**K-Nearest Neighbours (KNN)** is a simple and intuitive machine learning algorithm used for both classification and regression tasks. It's a non-parametric and instance-based learning method, which means that it doesn't make any underlying assumptions about the data distribution and makes predictions based on the data points themselves (Harrison, 2019).

KNN is primarily used for classification tasks, where it predicts the class of a new data point based on the classes of its nearest neighbours. Here's a breakdown of what KNN does:

**Classification**: In classification tasks, KNN categorizes a new data point into one of the predefined classes based on the majority class of its K-nearest neighbours. For example, if you want to classify whether an email is spam or not, KNN looks at the K most similar emails in the training dataset and assigns the class that occurs most frequently among these neighbours to the new email.

**Regression**: KNN can also be used for regression tasks, where it predicts a numerical value for a data point based on the values of its K-nearest neighbours. For instance, it can predict the price of a house based on the prices of nearby houses.

K-Nearest Neighbours (KNN) operates as follows:

* **Select K:** Choose the number of nearest neighbours (K) for prediction.
* **Calculate** **Distances**: Measure the distance between the new data point and all training data.
* **Identify** **Neighbours**: Find the K nearest training data points.
* **Classify** **or** **Predict**: For classification, use majority voting among the K neighbours. For regression, average their values.
* **Make** **Prediction**: Assign the determined class or value as the prediction for the new data point (Harrison, 2019).

## 3.2 Neural Network Model

A neural network, also known as an artificial neural network (ANN), is a computational model inspired by the structure and functioning of the human brain. It's a fundamental component of deep learning, a subfield of machine learning. Neural networks consist of interconnected nodes or artificial neurons organized into layers. These nodes work collectively to process and analyse complex data, such as images, text, or numerical information (Simplilearn, 2023).

Neural networks are versatile and can be used for various tasks, including:

* **Classification**: Neural networks can determine which category or class a given input belongs to.
* **Regression**: Neural networks can predict numerical values based on input data. This is useful for tasks like predicting house prices or stock market trends.
* **Pattern Recognition**: Neural networks excel at recognizing patterns, making them suitable for tasks like speech recognition, handwriting recognition, and even medical diagnosis.
* **Anomaly** **Detection**: They can identify outliers or anomalies in data, which is crucial for fraud detection or quality control.
* **Generative** **Modelling**: Some neural networks, like Generative Adversarial Networks (GANs), can generate new data that is similar to the training data, which has applications in creating art, text, and more.

The fundamental unit of a neural network is the artificial neuron or node. A neural network typically consists of three types of layers:

* **Input** **Layer**: This layer receives the initial data, which could be an image's pixels, words in a sentence, or numerical features.
* **Hidden** **Layers**: These intermediate layers process the input data using weights and activation functions. The number and size of hidden layers can vary, and they are responsible for learning the features and patterns in the data.
* **Output** **Layer**: The final layer provides the network's prediction, whether it's a class label, a numerical value, or a generated image (Simplilearn, 2023).

Neural networks learn through a process called training, which involves the following steps:

* **Forward Propagation**: During training, data is passed forward through the network. Each neuron processes the data it receives from the previous layer and produces an output.-
* **Weight Adjustment**: The network compares its output to the desired output (the ground truth). Any discrepancies result in adjusting the weights of the connections between neurons.
* **Backpropagation**: The error is propagated backward through the network, allowing each neuron to update its weights in a way that minimizes the error.
* **Iterative Learning**: The process of forward propagation, weight adjustment, and backpropagation iterates many times on a training dataset until the network's performance reaches a satisfactory level (Guo et al., 2022).

Once trained, the neural network can be used for making predictions on new, or unseen data.

## 3.3 Bernoulli Naive Bayes

Naive Bayes is among the algorithms which are relatively faster than other classification algorithms. It works on the Bayes theorem of probability to predict the class of unknown data sets. Bayes theorem describes the probability of an event based on the prior knowledge or other certain known probabilities of that event (Sharma, 2021).

The Naive Bayes is a collection of three algorithms: MultinomialNB, BernoulliNB and GaussianNB (Sharma, 2021).

Bernoulli Naive Bayes is a part of the family of Naive Bayes. It only takes binary values. The most general example is where we check if each value will be whether a word that appears in a document (Sharma, 2021).

Bernoulli Naive Bayes works best when applied to real-life scenarios where immediate (fast) results are required. Also, it is used more often to get better results in multi-class problems and independence rules. So, it has a higher success rate than other algorithms. In case of small amount of data or small documents (for example in text classification), Bernoulli Naive Bayes gives more accurate and precise results as compared to other models (Sharma, 2021).

## 3.4 Decision Tree

A decision tree is a non-parametric supervised learning algorithm for classification and regression tasks. It has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. Decision trees are used for classification and regression tasks, providing easy-to-understand models (Saini, 2023).

The purpose of a decision tree is to make decisions or predictions by learning from past data. It helps to understand the relationships between input variables and their outcomes and identify the most significant features that contribute to the final decision (Saini, 2023).

## 3.5 Logistic Regression

Logistic regression is a statistical method used for binary classification problems, where the dependent variable is categorical. It is a type of regression analysis that is used to predict the probability of a binary outcome based on one or more predictor variables. Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using coefficients (b) to predict an output variable (y). The logistic function is a sigmoid function that maps any real-valued number to a value between 0 and 1, which represent the probability of the occurrence of the event. Logistic regression is intended for binary (two class) classification problems. It will predict the probability of an instance belonging to the default class, which can be snapped into a 0 or 1 classification (Brownlee, 2020).

Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output. The main use-case of a logistic model is to be given an observation and estimate the probability that a certain event will occur. Logistic regression can be used for feature selection, as it provides a measure of the importance of each predictor variable in the model. Logistic regression can be used for model interpretation, as it provides estimates of the effect of each predictor variable on the probability of the event occurring (Swaminathan, 2019).

# Chapter 4 Model outcomes and conclusion

In this chapter the outcomes of the models will be addressed and a conclusion based on the outcomes will be presented.

## 4.1 KNN Outcome

The k-nearest neighbours (knn) model that will be discussed uses a K value of 5.

Based on the confusion matrix, the model demonstrates the following performance metrics:  
True Positives (TP): 15,485  
True Negatives (TN): 16,078  
False Positives (FP): 15,398  
False Negatives (FN): 16,399

The overall accuracy of the model is 49.82%. This suggests that the model is slightly better than random guessing for this binary classification task, but it is far from perfect. An accuracy of nearly 50% indicates that there's a substantial amount of misclassification happening.

The model's confusion matrix also reveals a balanced distribution of errors across false positives and false negatives. This means the model is not showing a strong bias towards predicting either class. However, given that both FP and FN values are relatively high, the model seems to be struggling to correctly classify instances from both classes.

## 4.2 Neural Network Outcome

The model was trained for 5 epochs. The training and validation losses were both around 0.6932, indicating little to no improvement in loss across the epochs. The accuracy for both training and validation hovered around 50%, suggesting the model was performing no better than random guessing. The test accuracy was 49.68%, reinforcing the notion that the model did not learn effectively from the training data.

For a better outcome the following can be looked into:

* Data Quality and Preprocessing
* Model Architecture
* Training
* Loss Function
* Evaluate Data Imbalance
* Feature Engineering

## 4.3 Bernoulli Naive Bayes Outcome

The classification report of the Bernoulli Naive Bayes showed the following:

* The accuracy is around 0.50, which means the model is correct about half the time.
* The precision, recall, and F1-score for the class labelled '0' (Severity = None) are relatively better, with a precision of 0.50 and a recall of 1.00, meaning the model correctly identifies this class well.
* However, for the classes labelled '1' and '2' (Severity = Mild or Moderate), the precision, recall, and F1-scores are all very low (close to 0.00). This suggests that the model struggles to accurately classify instances in these classes, as indicated by the low precision and recall values.

In summary, the model performs reasonably well for class '0', but it has significant difficulties in correctly classifying instances in classes '1' and '2'. This is indicated by the low precision and recall values for these classes, which are essential metrics for evaluating a classifier's performance.

## 4.4 Decision Tree Outcome

The classification report of the Decision Tree model showed the same values as the one of the Bernoulli Naive Bayes. So again, the model performs reasonably well for class '0', but it has significant difficulties in correctly classifying instances in classes '1' and '2'.

To gain a better understanding of the importance of each feature in predicting the severity, feature importance was used to check which features have the most influence on the model’s decisions. This showed the following:

* "Age\_Group" is the most important feature, contributing approximately 18.65% to the model's decision-making process.
* "Pains" is the second most important feature, contributing about 13.64% to the model's decision-making process.
* "Runny-Nose," "Nasal-Congestion," "Dry-Cough," and "Gender" also have significant importance in influencing the model's predictions.
* Features like "Difficulty-in-Breathing," "None\_Experiencing," and "None\_Symptom" have relatively lower importance, indicating they have less impact on the model's decisions.

## 4.5 Logistic Regression Outcome

The logistic regression model applied for classification exhibited promising performance in evaluating the severity of instances.

* The evaluation was conducted using 5-fold cross- validation. The cross-validation process involved partitioning the dataset into five subsets, training the logistic regression model on four of these subsets, and then evaluating its performance on the remaining fifth subsets. This was repeated five times, ensuring each subset served as the test set once.
* The average accuracy achieved across all folds was approximately 75%. This indicated the model’s capability to accurately predict the severity level based on the provided symptoms and demographic information.
* The precision, recall, and F1-scores for different severity levels (None, Mild and Moderate) were also examined. The logistic regression model demonstrated commendable precision and recall for classifying instances with ‘None’ severity, achieving scores of around 0.80. However, in distinguishing between ‘Mild’ and ‘Moderate’ severity levels, the model exhibited lower precision and recall, with scores around 0.60.
* The insight drawn from feature importance in logistic regression also shed light on the significant predictors influencing the model’s decision. Features like ‘Age\_Group,’ ‘Pains,’ ‘Runny-Nose,’ ‘Nasal-Congestion,’ ‘Dry-Cough,’ and ‘Gender’ were identified as substantial influencers on the model’s predictions, contributing variably on the severity classification.

## 4.6 Conclusion outcome

This research assessed multiple machine learning models to predict asthma severity based on health attributes and demographics. In conclusion, the logistic regression model, based on the 5-fold cross-validation findings indicate that while each model had strengths and weaknesses, the Logistic Regression model showed promising results in accurately predicting ‘None’ severity instances. However, it faced challenges in distinguishing between ‘Mild’ and ‘Moderate’ severity levels. To enhance predictions, further exploration into model improvements, data quality enhancements, and feature engineering is recommended.

# Bibliography Brownlee, J. (2020, 14 Augustus). *Logistic regression for machine learning*.

MachineLearningMastery.com. <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>

Guo, J., Chen, Z., Liu, Z., Li, X., Xie, Z. Y., Wang, Z., & Wang, Y. (2022). Neural Network Training method for materials science based on multi-source databases. Scientific Reports, 12(1). <https://doi.org/10.1038/s41598-022-19426-8>

Harrison, O. (2019, 14 juli). Machine Learning basics with the K-Nearest Neighbors Algorithm. *Medium*. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

Thakur, D. (2023, September 10). *Asthma disease prediction*. Kaggle. Retrieved September 15, 2023, from <https://www.kaggle.com/datasets/deepayanthakur/asthma-disease-prediction>

Saini, A. (2023, September 13). *Decision Tree Algorithm – A Complete Guide*. Analytics Vidhya. Retrieved October 27, 2023, from <https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>

Sharma, N. (2021, December 14). Bernoulli Naive Bayes and it’s implementation - NANDINI SHARMA - Medium. *Medium*. Retrieved October 27, 2023, from <https://medium.com/@nansha3120/bernoulli-naive-bayes-and-its-implementation-cca33ccb8d2e>

Simplilearn. (2023b, juli 20). An ultimate tutorial to neural networks. Simplilearn.com. <https://www.simplilearn.com/tutorials/deep-learning-tutorial/neural-network>  
  
Swaminathan, S. (2019, 18 januari). Logistic regression — detailed overview - towards data science. *Medium*. <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>