

CHAPTER 2 - LITERATURE REVIEW

2.1 Machine Learning

Machine learning is one of the branches of the field of Artificial Intelligence that focuses on developing statistical models and algorithms that let computers carry out tasks without the need for explicit programming (Koza *et al.*, 1996). Providing systems with the ability to learn from data and gradually enhance their performance is the main objective of machine learning.

Traditionally, machine learning systems have been classified into three broad groups based on the type of "signal" or "feedback" that the learning system can use. These categories correlate to different learning paradigms: reinforcement learning, supervised learning, and unsupervised learning (Jordan and Mitchell, 2019).

Supervised learning systems typically use a learned mapping of labelled data to generate an output (or a probability distribution of all possible outputs) for each input in order to make predictions. In unsupervised learning, unlabelled data is typically analyzed while making assumptions about the data's structural characteristics. Reinforcement learning combines supervised and supervised learning techniques. Rather than providing training examples that show the correct output for a given input, reinforcement learning relies on training data that only indicates whether an action is correct or not. If an action is incorrect, the challenge of determining the correct action persists (Jordan and Mitchell, 2019).

The learning system that is used in this research is supervised learning. In this approach, the model is trained on a labelled dataset, where each example in the training set has both input features and corresponding output labels. The dataset is classified and split into training and test dataset. Training dataset is used to train the model while testing dataset functions as a benchmark to get the score of the model performance.

2.1.1 Multi-class Classification

In machine learning, multi-class classification refers to classification tasks that involve more than two classes (Grandini, Bagli, and Visani, 2020). Each instance is assigned to one class without any overlap with other classes. By implementing multi-class classification when training a model, the output given is the class with the highest probability among all the classes. This means that the output is mutually exclusive to one class only.

2.1.2 Multi-label Classification

Multi-label classification is the opposite of multi-class classification. It is possible to assign many non-exclusive labels to each instance in this classification approach. There is no restriction on how many classes an instance can be assigned to in the multi-label problem, and the labels are non-exclusive (Read *et al.*, 2011).

The problem transformation method used for this study is transformation into binary classifications. This requires an independent training of one binary classifier for each label in the dataset. The combined model predicts all labels for this instance whenever the classifiers signal a positive result.

2.2 Algorithms

There are five supervised machine learning algorithms that are used for this research: Linear Regression, Naïve Bayes, K-Nearest Neighbor, Random Forest Algorithm, and Support Vector Machine. These algorithms are chosen because they are commonly used supervised machine learning algorithms (Islam *et al.*, 2020; Shah, Patel and Bharti, 2020; Uddin *et al.*, 2019).

2.2.1 Logistic Regression

Logistic regression is a powerful and widely used technique in supervised learning for classification. It extends from standard regression and can identify which factors have the most correlations with an outcome and gives an estimate of the magnitude of their influence (Tolles and Meurer, 2016).

According to Tolles and Meurer, logistic regression has three limitations. First, the quantity and appropriateness of the measured independent predictor variables determine a regression model's validity. The second limitation is that the variables' magnitude of association must be consistent over the variable's value range. Lastly, a lot of logistic regression analyses make the assumption that the value of one predictor has no bearing on the effect of another. In actuality, the value of one predictor affects the influence of another, and in order to guarantee the validity of the predicted associations, this "interaction" must be specifically incorporated in the analysis.

2.2.2 Naïve Bayes

Naïve Bayes is a basic learning method that employs Bayes' theorem and assumes that attributes are independent of the class under specified conditions. Even though the independence assumption is frequently not met in real-world scenarios, Naïve Bayes often achieves competitive classification accuracy. Its efficiency in computation and other attractive features make Naïve Bayes widely utilized in real-world applications (Geoffrey, 2017).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2.1)$$

As shown in *Equation 2.1*, the Bayes Theorem is used to calculate the likelihood of an event occurring based on the likelihood of an earlier event.

- The probability of A for occurring while B is true is denoted by $P(A|B)$).
- The probability B occurring while A is true, is denoted as $P(B|A)$).
- The probability of detecting occurrence A and B separately are denoted by $P(A)$) and $P(B)$).

The variant of Naïve Bayes that is used in this research is Gaussian Naïve Bayes. The probability distribution of the characteristics inside each class is represented by Gaussian (normal) distributions in Gaussian Naive Bayes. For a dataset with m features and n classes, part of the representation includes estimating the mean (μ) and variance (σ^2) for each feature in each class. The formula for Gaussian Naïve Bayes is shown in *Equation 2.2*.

$$P(X_i|C_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}} \quad (2.2)$$

2.2.3 K- Nearest Neighbor

The KNN algorithm, or K-Nearest Neighbor, is a non-parametric method that organizes data points according to their proximity and relationship to other relevant data. The algorithm assumes that comparable data points are often located close to each other. Therefore, its goal is to establish the distance between data points, commonly using Euclidean distance, and then assign a category based on the most frequent category or mean.

While most data scientists usually use this method for its simplicity and fast processing time, its processing time increases with the size of the test dataset, making it less useful for classification tasks.

2.2.4 Random Forest Algorithm

A random forest is formed by aggregating various tree predictors, where each tree's decision-making process is influenced by values randomly drawn from a vector that shares the same distribution across all trees (Breiman, 2001). When the forests are constrained to prioritize specific feature dimensions, these groupings of trees splitting with oblique hyperplanes can achieve better accuracy as they expand without facing overtraining (Ho, 1995)

2.2.5 Support Vector Machine

Support-vector machines, or support-vector networks, are a family of similar supervised learning techniques used in regression and classification. An SVM training algorithm develops a model that predicts if a new instance will belong to one of two categories, using a training dataset where each example is labeled as belonging to one of these categories (Cortes and Vapnik, 1995). Then, it identifies the hyperplane that maximizes the margin between each class and minimizes errors in classification to separate data items into two distinct classes as illustrated by Cortes and Vapnik shown in *Figure 2.1*.

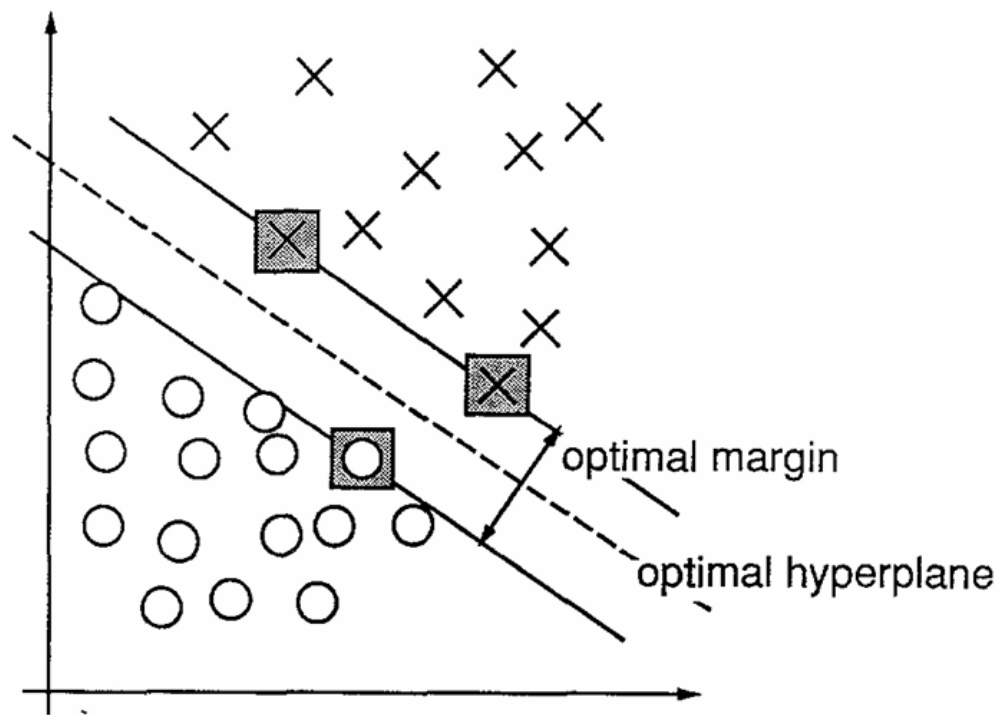


Figure 2.1:Support Vector Machine (Cortes and Vapnik, 1995)

2.3 Hyperparameter Tuning

Hyperparameter tuning involves selecting the best set of hyperparameters for a learning algorithm. A hyperparameter serves as a control parameter during the learning process.

According to Feurer and Hutter (2019), hyperparameter tuning has a number of significant applications, including lowering the amount of human labor required to apply machine learning, enhancing the efficiency of machine learning algorithms, and enhancing the validity and impartiality of scientific research

2.4 Expert System

An expert system is a program that mimics the decision-making and actions of a person or a group of people with expertise and experience in a certain area by leveraging artificial intelligence (AI) technologies. Expert systems are designed to solve complex problems using knowledge bases from experts, often expressed as if-then rules, unlike traditional procedural code.

In an expert system, there are two subsystems: the knowledge base and the inference engine. The knowledge base contains rules and information, and the inference engine applies these rules to derive new facts based on the system's accumulated knowledge.

2.5 Related Works

2.5.1 Optimization of Naïve Bayes Method using Genetic Algorithm to Diagnose Cattle Disease

The authors from State University of Semarang developed a model to diagnose diseases on cows based on given symptoms. The dataset consisted of 7 diseases, each consisted of a combination of different symptoms from a list of 23 symptoms. The authors did a comparison of diagnoses between using Naïve Bayes Algorithm and a combination of Naïve Bayes Algorithm and Genetic Algorithm. From testing 20 test data, the model produced 18 accurate test data using the Naïve Bayes Algorithm and 19 accurate test data using the combination of Naïve Bayes Algorithm and Genetic Algorithm (Firmansyah, Sugiharti and Arifudin, 2020).

2.5.2 Diagnosis of Fever Symptoms Using Naive Bayes Algorithm

The authors from State University of Malang trained a model using Naïve Bayes Algorithm to predict the category of fever on human patients based on its symptoms. The model is trained to detect fever in the category of typhoid fever (TF) or Dengue Hemorrhagic Fever (DHF). The dataset which the model was trained on consisted of 11 attributes (10 predictors and 1 class) and contained 53 instances of TF and 47 instances of DHF with a total of 100 instances. The model produced an average accuracy score of 94.00%, a precision score of 95.36%, and a recall score of 84.75% based on a k-fold cross-validation ($k = 10$) testing method (Widiyaningtyas, Zaeni and Jamilah, 2020).

2.5.3 Heart Disease Prediction using Machine Learning Techniques

Based on a number of influencing factors, the authors trained a model to predict whether or not a human patient will acquire heart disease. The dataset that was used comprises 300 examples of data with 14 attributes (13 predictors and 1 class). The

authors trained the model using several supervised machine learning algorithms: Naïve Bayes, Decision Tree, K-Nearest Neighbor, and Random Forest Algorithm. After comparing the results, the authors found that the K-Nearest Neighbor algorithm ($k = 7$) produced the best result with an accuracy of 90.789% followed by Naïve Bayes with an accuracy of 88.157% (Shah, Patel and Bharti, 2020).

2.5.4 Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques

The authors made a comparative study of breast cancer prediction using 5 different machine learning algorithms: Support Vector Machine, K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Logistic Regression. From the authors' research, the algorithm with highest accuracy obtained is Artificial Neural Network with 98.57% whereas the lowest accuracy is from Random Forest and Logistic Regression with 95.7% (Islam *et al.*, 2020).

2.5.5 Comparing Different Supervised Machine Learning Algorithms for Disease Prediction

The authors made a comparative study regarding the accuracy of seven supervised machine learning algorithms (ANN, DT, KNN, LR, NB, RF, SVM) that are implemented on four diseases that are frequently modelled (heart disease, diabetes, breast cancer, and Parkinson's disease). Additionally, the authors enumerated the benefits and drawbacks of every supervised machine learning technique they examined as shown in *Table 2.1* (Uddin *et al.*, 2019).

Table 2.1: Advantages and Limitations of Each Algorithm the Authors Tested

Supervised algorithm	Advantages	Limitations
Artificial neural network (ANN)	- Can detect complex nonlinear relationships between dependent and independent variables. - Requires less formal statistical training. - Availability of multiple training algorithms. - Can be applied to both classification and regression problems.	- Have characteristics of 'black box' - user can not have access to the exact decision-making process and therefore, - Computationally expensive to train the network for a complex classification problem. - Predictor or Independent variables require pre-processing.
Decision tree (DT)	- Resultant classification tree is easier to understand and interpret. - Data preparation is easier. - Multiple data types such as numeric, nominal, categorical are supported. - Can generate robust classifiers and can be validated using statistical tests.	- Require classes to be mutually exclusive. - Algorithm cannot branch if any attribute or variable value for a non-leaf node is missing. - Algorithm depends on the order of the attributes or variables. - Do not perform as well as some other classifier (e.g., Artificial Neural Network) [80]
K-nearest neighbour (KNN)	- Simple algorithm and can classify instances quickly. - Can handle noisy instances or instances with missing attribute values. - Can be used for classification and regression.	- Computationally expensive as the number of attributes increases. - Attributes are given equal importance, which can lead to poor classification performance. - Provide no information on which attributes are most effective in making a good classification.
Logistic regression (LR)	- Easy to implement and straightforward. - LR-based models can be updated easily. - Does not make any assumptions regarding the distribution of independent variable (s). - It has a nice probabilistic interpretation of model parameters.	- Does not have good accuracy when input variables have complex relationships. - Does not consider the linear relationship between variables. - Key components of LR - logic models, are vulnerable to overconfidence. - May overstate the prediction accuracy due to sampling bias. - Unless multinomial, generic LR can only classify variables that have two states (i.e., dichotomous).
Naive Bayes (NB)	- Simple and very useful for large datasets. - Can be used for both binary and multi-class classification problems. - It requires less amount of training data. - It can make probabilistic predictions and can handle both continuous and discrete data.	- Classes must be mutually exclusive. - Presence of dependency between attributes negatively affects the classification performance. - It assumes the normal distribution of numeric attributes.
Random forest (RF)	- Lower chance of variance and overfitting of training data compared to DT, since RF takes the average value from the outcomes of its constituent decision trees. - Empirically, this ensemble-based classifier performs better than its individual base classifiers, i.e., DTs. - Scales well for large datasets. - It can provide estimates of what variables or attributes are important in the classification.	- More complex and computationally expensive. - Number of base classifiers needs to be defined. - It favours those variables or attributes that can take high number of different values in estimating variable importance. - Overfitting can occur easily.
Support vector machine (SVM)	- More robust compared to LR - Can handle multiple feature spaces. - Less risk of overfitting. - Performs well in classifying semi-structured or unstructured data, such as texts, images etc.	- Computationally expensive for large and complex datasets. - Does not perform well if the data have noise. - The resultant model, weight and impact of variables are often difficult to understand. - Generic SVM cannot classify more than two classes unless extended.

2.5.6 A Neural Network Based Expert System for Diagnosis of Diabetes

To assist in diagnosing pregnant women with diabetes or not, the authors created a neural network model. The system is an expert system since the authors gathered data from experts while adhering to knowledge engineer-established guidelines. For user convenience, the developed model has also been integrated into a web application (Alade *et al.*, 2018).

2.6 Comparison of Related Works

Based on all related works mentioned before, a comparison of the strength and weakness of each research is shown in Table 2.2.

Table 2.2: Comparison of Related Works

Title	Strength	Weakness
Optimization of Naïve Bayes Method using Genetic Algorithm to Diagnose Cattle Disease	None	The dataset used for training was too small.
Diagnosis of Fever Symptoms Using Naive Bayes Algorithm	Good amount of dataset and pre-processing techniques. The authors also did a validation for the model's performance.	None
Heart Disease Prediction using Machine Learning Techniques	The authors compared several supervised machine learning methods for predicting cardiac disease in an eloquent manner.	The authors stated that in order to get better accuracy for early cardiac disease prediction, more sophisticated models must be used in combination with one another.
Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques	The authors compared a number of supervised machine learning methods for predicting breast cancer in an excellent way.	None

Comparing Different Supervised Machine Learning Algorithms for Disease Prediction	The authors compared seven supervised machine learning methods for four commonly modelled diseases in an eloquent manner. The authors carefully filtered researches that fulfilled their requirement.	None
A Neural Network Based Expert System for the Diagnosis of Diabetes Mellitus	The authors explained about expert system clearly and the integration of it with artificial intelligence.	The authors didn't explain further regarding the neural network system, instead they focus more on the expert system.