8/31/2023

Machine Learning

# Introduction

Machine learning is the fusion of statistical principles and computer science insights. Machine learning is the process of training a model to make useful predictions using a dataset. Coined by Arthur Samuel in 1959, it now resides as a component within the Artificial Intelligence (AI), encompassing algorithms empowering computers to autonomously process and categorize novel data using historical information (Bansal, 2022). This computational approach enables systems to automatically learn from a data, identify patterns, and make decisions with minimal human intervention.

Machine learning is essential in various applications, from voice recognition in smartphones to fraud detection in banking. It has also transformed numerous industries by enabling automation and providing insights and personalization at an unprecedented scale. The core idea of machine learning is to create and use models that are trained on preexisting data to predict future data or other outcomes of interest. It’s an iterative process where the model makes predictions, receives feedback, and then adjust itself in response to the feedback, continually improving its performance.

Machine learning can be categorized into three main types namely: Supervised learning, a model trained on a label dataset where each training example is paired with an output label, Unsupervised learning, a model tries to learn the patterns and structure from the data without any guidance or hints, and Reinforcement learning, where an agent learns to make decisions by interacting with the environment, the agent receives feedback in the form of rewards or penalties and learns to navigate the environment accordingly.

# Literature Review

In the field of machine learning, supervised learning involves the task of acquiring knowledge from a training dataset that contains labelled examples. These examples consist of paired input instances and their corresponding desired outputs. The is to enable machines to make accurate predictions on new, unseen datasets. Supervised learning encompasses both classifications, where the goal is to assign labels to inputs, and regression, which focuses on predicting numerical outcomes. Among the frequently employed supervised learning techniques are neural networks, k-nearest neighbours, support vector machines, naïve Bayes, large margin nearest neighbours, and extended nearest neighbours. The upcoming subsections will delve into K-nearest neighbours, Naïve Bayes, and Support Vector Machine (SVM).

## K nearest neighbour (KNN)

KNN stands out as a fundamental and highly effective algorithm for classifying data, making it a top contender for implementation, particularly in situations where the data is intricate and lacks clarity. KNN was developed in 1951 by Evelyn Fix and Joseph Hodges. It was initially designed handle discriminant examination challenges related to probabilistic density estimation using parametric methods. This algorithm falls within the category of supervised learning techniques and is renowned for being one of the most user-friendly algorithms in the field of machine learning. While it can be applied to both classification and regression tasks, its primary use is in the classification of objects (Bansal, 2022).

KNN classification assigns a label to an unlabelled object based on the majority labels of its nearest neighbour from the training set. The value of ‘k’, representing the number of neighbours considered, can reduce the impact of the noise but may blur the group boundaries. For binary classification, the optimal ‘k’ value is often chosen through empirical analysis using bootstrap method (Khan *et al.*, 2018). Selecting the correct value of ‘k’ is crucial for improved accuracy, this task is known as parameter tuning (Bansal, 2022). Using a very small ‘k’ value like 1 or 2 in KNN can lead to results that are affected by noise (Bansal, 2022). Conversely, using a very high ‘k’ value can sometimes result in confusion, particularly depending on the dataset being used (Bansal, 2022). While there isn’t a fixed value for ‘k’, a common choice is 5 (Bansal, 2022). This means that for the process of majority voting, the algorithm considers the nearest 5 neighbours to the new data point (Bansal, 2022). It’s generally recommended to use an odd value for ‘k’ to avoid ambiguity between 2 classes of datasets (Bansal, 2022). An alternative to determine ‘k’ is through a formula-based calculation which is as follows:

Where n is overall count of the data points (Bansal, 2022).



Figure : Classification of the new data point based on neighbours (Bansal, 2022).

## Naïve Bayes

The Naïve Bayes method is a straightforward probabilistic classifier that determines probabilities by analysing the frequency and combinations of values within a dataset. It employs Baye’s theorem and operates under the assumption that all features are mutually independent given the class variables. This assumption of conditional independence is often overly simplistic for real-world scenarios, hence the term ‘Naïve’. However, the algorithm exhibits rapid learning in a range of controlled classification tasks. Bayes’ theorem, named after the 18th century British mathematician Thomas Bayes, is a mathematical equation used to calculate conditional probability (Saritas & Yasar, 2019). According to (Dimitoglou *et al.*, 2012), Naïve Bayes uses the following equation:

Where: P(A|B) is the probability of the occurrence of event A when event B occurs.

P(A) is the probability of the occurrence of A.

P(B|A) in the probability of the event B when event A occurs.

P(B) is the probability of the occurrence of B.

Naïve Bayes is particularly effective for text analysis, especially in areas like Natural Language Processing. One of its unique characteristics is the use of mixture models for classification. These models are adept at determining the probability of their individual components. Furthermore, Naïve Bayes is sometimes referred to as ‘simple Bayes’ or ‘independence Bayes’ (Rahat *et al.*, 2019).

## Support Vector Machine (SVM)

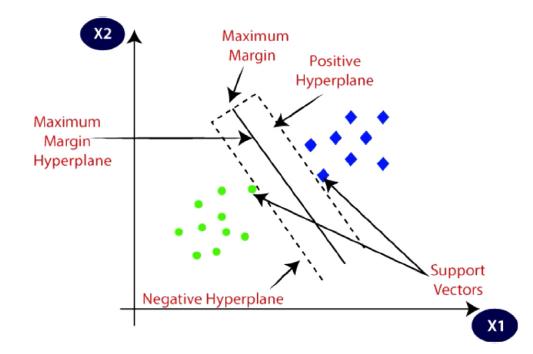


Figure : Illustration of a decision boundary in SVM (Bansal, 2022).

SVM is a powerful algorithm developed by Alexey Chervonenkis and Vladimir Vapnik in 1963. It’s widely used in various fields, from image and text classification to biology for protein sorting. Essentially, SVM works by drawing a boundary, known as a hyperplane, to best separate data into classes. The unique feature of SVM is its use of specific data points called “Support Vectors” to define this boundary. Over the years, SVM has been applied in many areas, including face detection, image classification, text categorization, and even in advanced technologies like self-driving cars and chatbots. Its main goal is to classify or predict data accurately. An example is illustrated in Figure 2. There are two types of SVM namely Linear and Non-linear SVM’s. Linear type SVM is used when data can be split into two groups with a straight line while a Non-linear type SVM is used when data can’t be separated by a straight line (Bansal, 2022).

SVM can create intricate models and uses advanced algorithms, which include a wide variety of neural networks and other classifiers. However, despite its complexity, SVM is easier to study mathematically because it transforms input into a higher dimensional space using non-linear process. SVM can process high-dimensional data without doing computations in that complex space, thanks to kernel tricks. While SVM is primarily for binary classification, it can be used in power systems to identify loads. For even better image identification, there’s the Hybrid Multiclass SVM (Khan *et al.*, 2018).

### Linear SVM

To illustrate linear SVM imagine you have red and yellow dots on a paper, representing two types of objects, and you want to draw a line that separates them. There can be many lines that do this, but SVM helps find the best one. This “best line” is called a hyperplane. It’s chosen based on its distance from the closest dots of both colours, ensuring it’s in the middle and equally distant from both. These closest dots are known as support vectors. In short, SVM finds the most balanced line to separate two categories (Bansal, 2022).

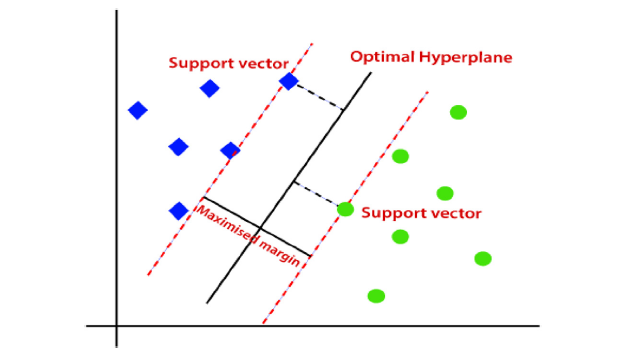


Figure : Implementation of SVM (Bansal, 2022).

### Non-linear SVM

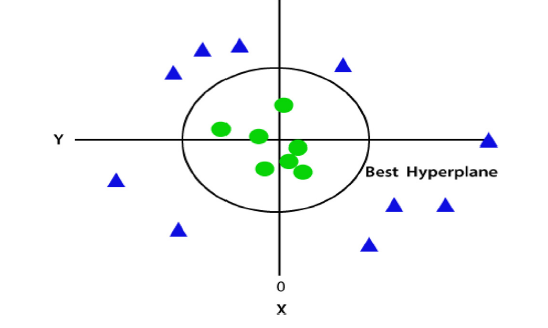


Figure : 3D Arrangements (Bansal, 2022).

When data isn’t arranged in a straight line, we can’t use just a single line to separate it. While we only needed two dimensions (x and y) for linear data, non-linear data requires a third dimension, z. This z can be found using the formula Think of this as moving from a flat plane to a 3D space, where the new plane is parallel to the x-axis. When we simplify this back to 2D with we get a circle with a 1-unit radius as illustrated in figures 3 and 4.

# Material and Methods

## Material

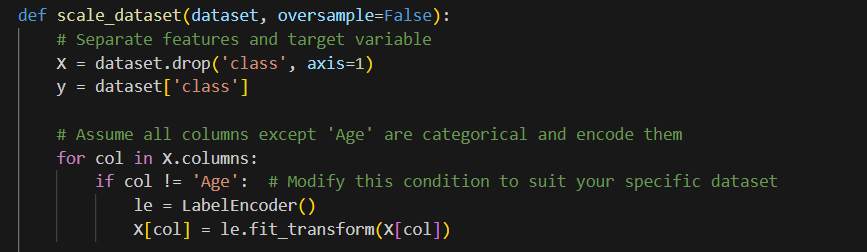
The dataset comprises 520 records, each representing individual data entries, and spans across 17 distinct variables. This dataset appears to be a medical dataset, related to diabetes diagnosis, given the nature of the variables. The variables encompass both categorical and numerical data:

1. Age (Integer): This variable captures the age of the individuals. Being numerical, it can range from infants to the elderly, providing insights into age-related trends or susceptibilities.
2. Gender (Categorical): This delineates the biological sex of the individuals, typically classified as male or female. It's essential in understanding gender-specific patterns or predispositions.
3. Polyuria (Categorical): Polyuria is a condition characterized by excessive or abnormally large production or passage of urine. This variable indicates its presence or absence.
4. Polydipsia (Categorical): Polydipsia is excessive thirst or excess drinking. This variable highlight whether individuals exhibit this symptom.
5. Sudden Weight Loss (Categorical): This indicates if the individual has experienced any unexpected reduction in weight.
6. Weakness (Categorical): This variable captures instances of generalized weakness or fatigue.
7. Polyphagia (Categorical): This denotes excessive hunger or increased appetite, another symptom that might be associated with various medical conditions.
8. Genital Thrush (Categorical): Genital thrush is a common yeast infection. Its presence or absence is captured here.
9. Visual Blurring (Categorical): This variable notes instances of blurred vision, a symptom that can be associated with several medical conditions.
10. Itching (Categorical): This represents the presence or absence of itching, which can be a symptom of various conditions, from skin disorders to internal ailments.
11. Irritability (Categorical): Captures mood fluctuations or instances of individuals being easily agitated.
12. Delayed Healing (Categorical): Indicates if wounds or sores take an unusually long time to heal, a potential symptom of underlying conditions.
13. Partial Paresis (Categorical): Paresis refers to weakened muscle movement, and this variable captures its occurrence.
14. Muscle Stiffness (Categorical): This highlights instances where individuals experience rigidity or lack of flexibility in their muscles.
15. Alopecia (Categorical): Alopecia represents hair loss. The variable indicates its presence or absence.
16. Obesity (Categorical): These variable captures whether the individual is medically classified as obese, a condition characterized by excessive body fat.
17. Lastly, the Class (Categorical) variable serves as the target for this dataset. It indicates the diagnosis outcome, determining if an individual is positive or negative for diabetes.

In essence, this dataset is a comprehensive medical dataset that encapsulates a wide range of symptoms and conditions, potentially geared towards diagnosing diabetes.

## Method

The dataset contains 520 entries, each representing an individual in a diabetes study. It comprises one numerical column and sixteen categorical columns. The "Age" column, which is numerical, has values ranging from 16 to 90, with a mean age of approximately 48 years. The other columns are categorical, capturing various symptoms and characteristics such as "Polyuria," "Polydipsia," and "Obesity," among others, with "Yes" or "No" responses. The "Gender" column predominantly consists of male participants, and the "Class" column, which is the target variable, has more "Positive" responses than "Negative." The categorical features exhibit varying distributions, with some having a balanced number of "Yes" and "No" responses, while others are skewed towards one of the responses. The dataset has no missing values, indicating a complete dataset.



Because the “Age” column is the only numerical attribute in the dataset and machines understand numerical data better than categorical data, the code above was used to convert all the categorical attributes in the dataset to numerical values.

The dataset underwent a series of preprocessing steps to ensure it was primed for the modelling process. Initially, the data was standardized using the StandardScaler, which adjusted the features to have zero mean and unit variance. This standardization is crucial for algorithms that are sensitive to the scale of input features. To address class imbalance in the dataset, two techniques were employed: RandomOverSampler and SMOTE (Synthetic Minority Over-sampling Technique). The former technique duplicates instances from the minority class, while the latter generates synthetic samples.

Three models were trained on this pre-processed dataset:

1. KNN: This non-parametric algorithm was trained with the parameter n\_neighbors set to 5. Essentially, it classifies a data point based on the majority class of its 5 nearest neighbours in the training data.
2. Gaussian Naive Bayes: This probabilistic classifier was employed without any specific parameter modifications, meaning it utilized its default settings. The algorithm is based on the Bayes theorem and assumes that features follow a Gaussian distribution.
3. Support Vector Classifier (SVC): A Support Vector Machine was also trained on the dataset. The exact parameters used were not specified in the extracted portion of the notebook, suggesting it might have been trained with default settings. SVC works by finding a hyperplane that best separates the classes in the input feature space.

Although the train/test split ratio wasn't directly extracted from the provided segment, based on the prior examination, it's inferred that the data was divided with 60% allocated to training, 20% to validation, and 20% to testing. This division ensures that the model can be trained, fine-tuned, and then evaluated on unseen data.

### Flowcharts

Figure : KNN Flowchart

Figure : Naive Bayes Flowchart

Figure : SVM Flowchart

# Results and Discussion

Figure : KNN vs Naive Bayes vs SVM

In the machine learning experiments conducted, three classification algorithms were compared: KNN, Naive Bayes, and the SVM. The results of the experiments showcased distinct performance characteristics for each classifier:

1. SVM emerged as the top performer, excelling in all metrics. It achieved an impressive accuracy of 97% and consistently maintained high precision, recall, and F1-scores for both classes. This underscores SVM's robustness and capability to draw clear decision boundaries, especially for this dataset.

A graph of a function

Description automatically generated with medium confidence

Figure : SVM ROC curve

1. KNN displayed commendable results with an overall accuracy of 92%. It stood out particularly for its high recall for the first class (class 0) and precision for the second class (class 1), indicating its effectiveness in identifying true positives.

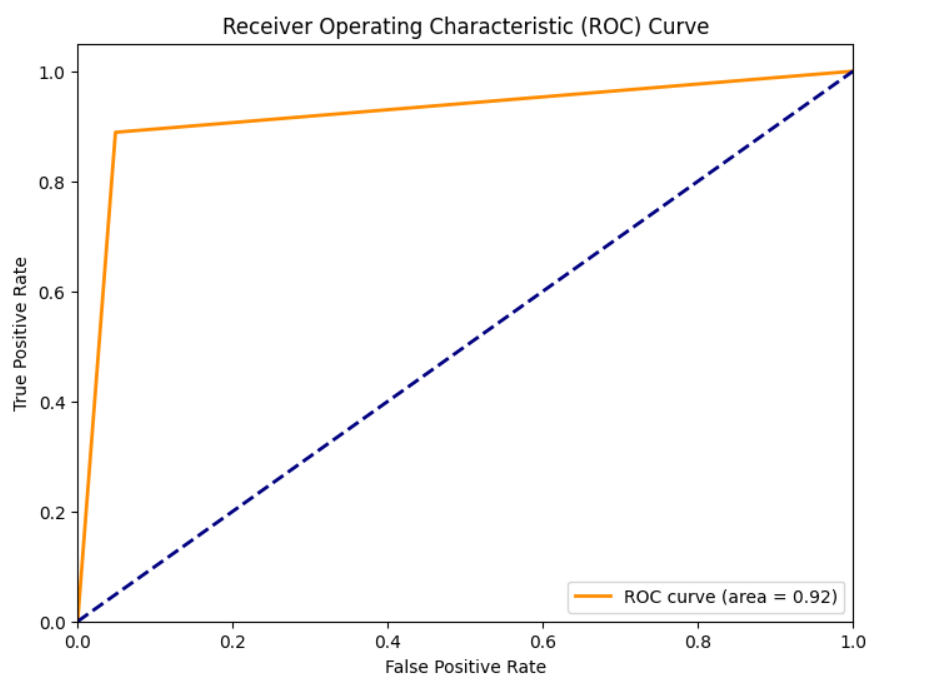


Figure : KNN ROC curve

1. Naive Bayes, while slightly lagging the other two in terms of accuracy at 87%, demonstrated a strong recall for the second class (class 1). This suggests that while it might misclassify some instances of the first class, it's quite adept at correctly identifying instances of the second class.

A graph of a line

Description automatically generated

Figure : Naïve Bayes ROC curve

In essence, while all three classifiers demonstrated their strengths, SVM clearly outperformed the others in this experimental setup, making it the most suitable choice for this dataset.

# Conclusion

In conclusion, the field of machine learning encompasses a range of powerful algorithms that enable computers to autonomously process and categorize data, thereby making predictions and decisions with minimal human intervention. This fusion of statistical principles and computer science insights has revolutionized various industries, automated processes and offering unprecedented insights and personalization. Among the notable machine learning techniques are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning, where models are trained on labelled datasets, is crucial for making accurate predictions. This category includes classification and regression tasks, with techniques like KNN, Naïve Bayes, and SVM being widely used.

KNN is a user-friendly algorithm for classification tasks, determining labels based on the majority labels of its nearest neighbours in the training set. The choice of 'k', the number of neighbours, impacts its performance, with the value often chosen empirically or through formula-based calculations. Naïve Bayes, a probabilistic classifier, assumes conditional independence of features given class variables. Despite its "naïve" assumption, it excels in various controlled classification tasks, particularly in text analysis and Natural Language Processing. SVM are powerful algorithms that draw decision boundaries to separate data into classes. SVM's unique use of support vectors to define these boundaries, and its ability to handle both linear and non-linear data through kernel tricks, make it versatile for tasks like image and text classification.

A medical dataset with 17 variables related to diabetes diagnosis was processed and used to train KNN, Naïve Bayes, and SVM models. The results showed that SVM emerged as the top performer with an accuracy of 97%, maintaining high precision, recall, and F1-scores for both classes. KNN achieved an accuracy of 91% with strengths in recall and precision for different classes, while Naïve Bayes exhibited an accuracy of 89% with strong recall for one class. Overall, while all three classifiers demonstrated their strengths, SVM proved to be the most suitable choice for this dataset due to its robust performance. This experiment underscores the significance of choosing the right machine learning algorithm based on the specific characteristics and goals of the dataset at hand.

# References

Bansal, M. 2022. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory Algorithms in machine learning. *Decision Analytics Journal*, 3, <https://www.sciencedirect.com/science/article/pii/S2772662222000261#sec4>

Dimitoglou, G., Adams, J.A. & Jim, C.M. 2012. Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability.,

Khan, M.M.R., Arif, R.B., Siddique, M.A.B. & Oishe, M.R. 2018. *Study and Observation of the Variation of Accuracies of KNN, SVM, LMNN, ENN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository*. Paper presented at the 4th Internation Conference on Electrical Engeneering and Information & Communication Technology.

Rahat, A.M., Kahir, A. & Masum, A.K.M. 2019. Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review

Dataset*.* In. 8th International Conference on System Modeling & Advancement in Research Trends, College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India.

Saritas, M.M. & Yasar, A. 2019. Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal ofIntelligent Systems and Applications in Engineering*, 7(2):88-91.