International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03731
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

A Review Paper on Classification Models

Ms. Shuchi Sharma
Department of AI-ML
ADGIPS - Delhi
New Delhi, INDIA
303shuchi@gmail.com

Abhishek
Department of AI-ML
ADGIPS - Delhi
New Delhi, INDIA
akp83218@gmail.com

Shahastransh Tiwari
Department of AI-ML
ADGIPS - Delhi
New Delhi, INDIA
shahastranshtiwari@gmail.com

Akash Kumar
Department of AI-ML
ADGIPS - Delhi
New Delhi, INDIA
akashkrsahu6207@gmail.com

Abstract* — *Classification models play a crucial role in various machine learning applications. This paper provides a comparative analysis of some of the widely-used classification algorithms (also referred to as models or machines). It includes definitions, examples, use cases and comparisons among the different models in order to differentiate among their pros and cons and determine the best one depending on the requirement.

The comparative analysis includes interpretability, accuracy and computational efficiency along with some other factors that can help differ one model from another.
Expectations generally foretell that a complex model will be more accurate but computation (time, memory and energy) intensive than a simpler model and thus, their use cases will be dependent on not just requirements but also resources and feasability. Results also indicate a similar result. This paper also involves a survey of Machine Learning students that tells us whether classification models are replaceable and how much are these models used in our daily needs and/or other sectors.
Keywords-Classification Models,KNN(K-Nearest Neighbours),SVM(Support Vector Machines), Random Forest, Accuracy, Precision, Recall, F1-Score

I. Introduction.

Classification models are fundamental in the field of machine learning. They are used to categorize data into distinct classes or groups. Here's a brief introduction to get you started:

A. *What Are Classification Models?*

Classification models are algorithms/models / machines that are a type of supervised learning algorithms. These models learn from labeled training data to make predictions about unseen data. They are used when the output is a categorical variable. For instance, a classification model could be used to determine whether an email is spam or not spam, or whether a tumor is malignant or benign.

B. *Common Classification Algorithms*

- Logistic Regression: Despite its name, logistic regression is used for binary classification problems. It models the probability of a binary outcome using a logistic function.
- Decision Trees: These models use a tree-like graph of decisions to predict the outcome. They split the data into subsets based on the value of input features.
- Random Forest: An ensemble method that uses multiple decision trees to improve the accuracy and control over-fitting.
- Support Vector Machines (SVM): These models find the optimal boundary (or hyperplane) that separates the data into classes.
- k-Nearest Neighbours (k-NN): This algorithm classifies data based on the majority class among the k nearest neighbours of a given data point.

C. *How They Work*

Classification models work in two main phases:

- Training Phase: The model is trained, that is, learns from a set of labeled data. The algorithm identifies relationships between the data and their labels, noting patterns.
- Prediction Phase: Once trained, the algorithm can be used to predict the labels/outputs for unseen data or datasets.

D. *Evaluation Metric*

In the field of machine learning, especially in the problem of classification models, a **c**onfusion matrix, also called error matrix, is used to visualise and help determine the performance of an algorithm, typically a supervised learning one. Confusion matrix (a.k.a. matching matrix) is a specific table layout that maps the true positives, false positives, true negatives and false negatives.

Each row of the matrix can either represent the instances in an actual class or instances in a predicted class while the columns contain the class not represented in rows - both variants are found in literature. In both variants, one diagonal (starting from top left to bottom right) represents all instances that are correctly predicted while the other diagonal represents the incorrectly predicted instances. The name stems from the fact that it makes it easy to see whether the system is confusing two classes that is mislabeling one as another commonly.

E. Metrics

It is important to evaluate performance of models using various metrics such as:

a. Accuracy: The ratio of correctly predicted instances over all instances.

b. Precision: The ratio of true positive instances to the instances that were predicted positive (including both the correctly and incorrectly predicted instances).

c. Recall: The ratio of true positive predictions to the actual positives.

d. F1 Score: Harmonic mean of precision an recall.

F. *Applications:*

Classification models have a wide range of applications and use cases:

a. Healthcare**:** Diagnosing diseases based on medical data.

b. Marketing: Grouping customers based on preferences (customer segmentation), ranking potential customers based on their conversion likelihood (lead scoring), etc.

c. Finance: Assessing credit score of loan applicants, detecting fraud, stock movement prediction, etc.

II. K-Nearest Neighbours (KNN) Model

A. Definition of KNN

K-Nearest Neighbours (KNN) is a classification model where data points are categorised based on the majority of its neighbouring classes among their k-nearest neighbours in the feature space. This works on the principle that data points that are close to one another tend to be in the same category.

The value 'k' is an important and vital parameter, representing the number of neighbours that are to be considered from the feature space for the voting, and the model's performance is highly sensitive to its selection.

KNN is considered a non-parametric model as it doesn't assume any specific data distribution and relies directly on the training dataset for classification, storing it entirely and using it for classification rather than learning or memorising something like a fixed set of coefficients like in logistic regression.

KNN is also considered an instance-based learning approach (a.k.a. "lazy learner") as it uses the entire dataset leading to the complexity being in the *data* rather than in models that are fixed with a set number of parameters.

KNN can create complex, non- l inear boundaries (decision boundaries) depending on the data.

The model's working is straightforward: it calculates the distance between a new data point and all existing data points, making it a brute-force approach. The model, then, identifies 'k' closest neighbours and assigns the new point to the class through the process of voting and selecting the most prevalent among these neighbours.

The choice of distance metric like Euclidean, Manhattan or Minkowski, significantly impacts the performance of the model.

B. Importance of KNN

KNN's simplicity and versatility makes it valuable for various classification tasks, especially when the boundary is irregular or non-linear.

It is easy to understand and implement. Its instance-based approach enables it to adapt quickly to new data, making it suitable for cases in which data distribution changes over time.

It is also useful in cases where interpretability is not a big concern. Understanding the relationships between features is not its strength.

Though its non-parametric nature providing flexibility help deal with datasets that don't conform to standard distributions, its effectiveness still hinges on selecting an appropriate 'k' value and distance metric, thus requiring careful tuning to obtain optimal performance.

C. Use Cases of KNN

KNN has found applications in various fields ranging from image recognition, recommender systems to medical diagnosis and financial market prediction.

Some recent use cases:

- T. Purboyo demonstrated its use case in social media filtering regarding disaster data using KNN. [1]
- R. Thangamani employed KNN to analyse handwriting dynamics for prediction of Alzheimer's disease. [4]

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03731
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

III. Support Vector Machine

A. *Definition of SVM*

This model tries to find an optimal hyperplane that effectively separates data into distinct classes. This hyperplane is chosen to maximise the margin that is the distance between the hyperplane and the nearest data points of each class, known as support vectors.

The primary goal of SVM is to achieve robustness and better generalisation by creating separation among classes.

SVM is specially effective for high-dimensional spaces and handles non-linear data by the help of kernel functions.

These functions map the non-linearly separable data into a higher dimension space where the data might be linearly separable, allowing SVM to classify non-linear data effectively. These include : linear, polynomial, radial basis function (RBF) and sigmoid.

The model's objective is to maximise the margin, which in turn reduces the risk of overfitting and improves model's ability, thus increasing accuracy and allowing for better performance.

This makes SVM a robust and reliable choice for various classification tasks.

B. *Importance of SVM*

SVM's strength lies in its ability to handle high dimensionality complex datasets and non-linearly separable data. The use of kernel functions allows it to intricately decide decision boundaries for non-linearly separable data.

SVM is considered to be relatively more robust than others as it is less sensitive to noisy data compared to other algorithms because the decision boundary or the hyperplane is primarily determined by the support vectors. However, this also makes SVM's performance sensitive to the choice of kernel function and hyperparameters, requiring careful optimisation.

It is also very valuable due to its ability to handle high-dimensional data effectively. It is used in various applications with numerous features as image classification and text mining. The use of regularisation techniques further enhances its robustness.

C. *Use Cases of SVM*

SVM has been applied in various domains, showcasing its versatility and effectiveness, ranging from face detection to bioinformatics.

- **Vinita Sangwan** applied SVM in predicting water quality classification. [3]
- **R. Thangamani** used it for human activity recognition. [2]

IV. Random Forest Model

A. *Definition of Random Forest*

Random Forest is an ensemble learning method, that is, where multiple models (often called "weak learners") are combined to produce a stronger, more accurate model.

In the context of Random Forest, it is applied through a process known as bagging (bootstrap aggregation) in which multiple base models (Decision trees in this case) are given samples of data either with repetition across data samples or no repetition across data samples but with no repetition inside a sample in and of itself, and the outputs generated by every weak learner is aggregated to a single output.

The algorithm creates a forest of decision trees, which employ a tree-like structure to make decisions based on data features, each trained on a random subset of of the data and features.

Random Forest is robust and can handle large number of features and complex interactions, making it a popular choice for various classification tasks. The combination of multiple trees reduces the risk of overfitting, making it more reliable than a single decision tree.

B. *Importance of Random Forest*

Its strength lies in its ability to reduce overfitting and improve performance compared to decision tree. The ensemble approach reduces the impact of individual tree errors and improving the overall accuracy.

It also provides a measure of feature importance, identifying the most relevant features for classification.

It is relatively easy to use and requires minimal parameter tuning. This makes it accessible to a wide range of users.

It can handle both numerical and categorical data and is relatively robust to outliers and missing values.

Random Forest models can be more complex and have lower interpretability than single decision tree.

Its ability to handle complex interactions and high-dimensional data makes it helpful in numerous use cases.

These models can be computationally intensive especially with higher number of trees.

C. *Use Cases of Random Forest*

Random Forest has been applied in various domains, ranging from health care to agriculture.

Some recent use cases include:

- T. Purboyo also used Random Forest in classifying disaster data from social media. Random Forest can categorise social media posts related to disasters, helping to identify

and respond to emergencies more effectively. [1]

- Susandri Susandri used Random Forest in sentiment labelling and text classification for WhatsApp groups. [5]

Comparison

For simplicity sake, the data chosen to compare the aforementioned three classification models is the same and consists of custom hand-drawn images with basic features like width and height or point count and path count selected. The algorithms' accuracy, computational overhead, processing time, complexity regarding time and space, interpretability alongside precision, recall, F1-score are compared.

Point of comparison	K-Nearest Neighbours	Support Vector Machine	Random Forest
Accuracy	Low(~38% accuracy)	Medium-High	Highest in this case among these options
Precision and Recall	Low	Medium-High	Usually high
F1-scor	Low	Medium	High
Training time	Fast	Slow	Moderate
Prediction Time	Slow	Fast	Moderate
Complexity	Low due to its simple logic	Medium-High depending on kernel	Medium (many trees)
Interpretability	Highly interpretable	Somewhat interpretable (Medium)	Generally lower than KNN/SVM but **feature importance** can give some insight

References

1. Purboyo, T. W. ., Wijaya, R. ., Latuconsina, R. ., Setianingsih, C. ., & Ruriawan, F. . (2024). Behavioural Indonesian disaster data classification in social media using KNN, random forest, and RNN in machine learning. Edelweiss Applied Science and Technology, 8(6), 169–183. https://doi.org/10.55214/25768484.v8i6.2033
2. R. Thangamani, D. Sathya, G. Sreeka, V. Kavila, K. Nithish and S. Santhosh, "Unveiling Novel Techniques for Human Activity Recognition with Smartphone Sensors," 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2023, pp. 1714-1722, doi: 10.1109/ICECA58529.2023.10394941. https://doi.org/10.1109/ICECA58529.2023.10394941
3. Vinita Sangwan, Rashmi Bhardwaj; Machine learning framework for predicting water quality classification. Water Practice and Technology 1 November 2024; 19 (11): 4499–4521. doi: https://doi.org/10.2166/wpt.2024.259
4. M. Vimaladevi, R. Thangamani, P. S, V. B and T. A, "Prediction of Alzheimer's Disease by Analyzing Handwriting Dynamics Using Machine Learning Algorithms," 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2024, pp. 1298-1304, doi: 10.1109/ICESC60852.2024.10690124. https://ieeexplore.ieee.org/document/10690124
5. Susandri, S., Defit, S., & Tajuddin, M. (2023). SENTIMENT LABELING AND TEXT CLASSIFICATION MACHINE LEARNING FOR WHATSAPP GROUP. JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer), 9(1), 119–125. https://doi.org/10.33480/jitk.v9i1.4201
6. Img 1 from https://www.researchgate.net/figure/Evaluation-metrics-accuracy-precision-recall-F-score-and-Intersection-over-Union_fig2_358029719 captioned Accuracy, precision, recall, f1-score