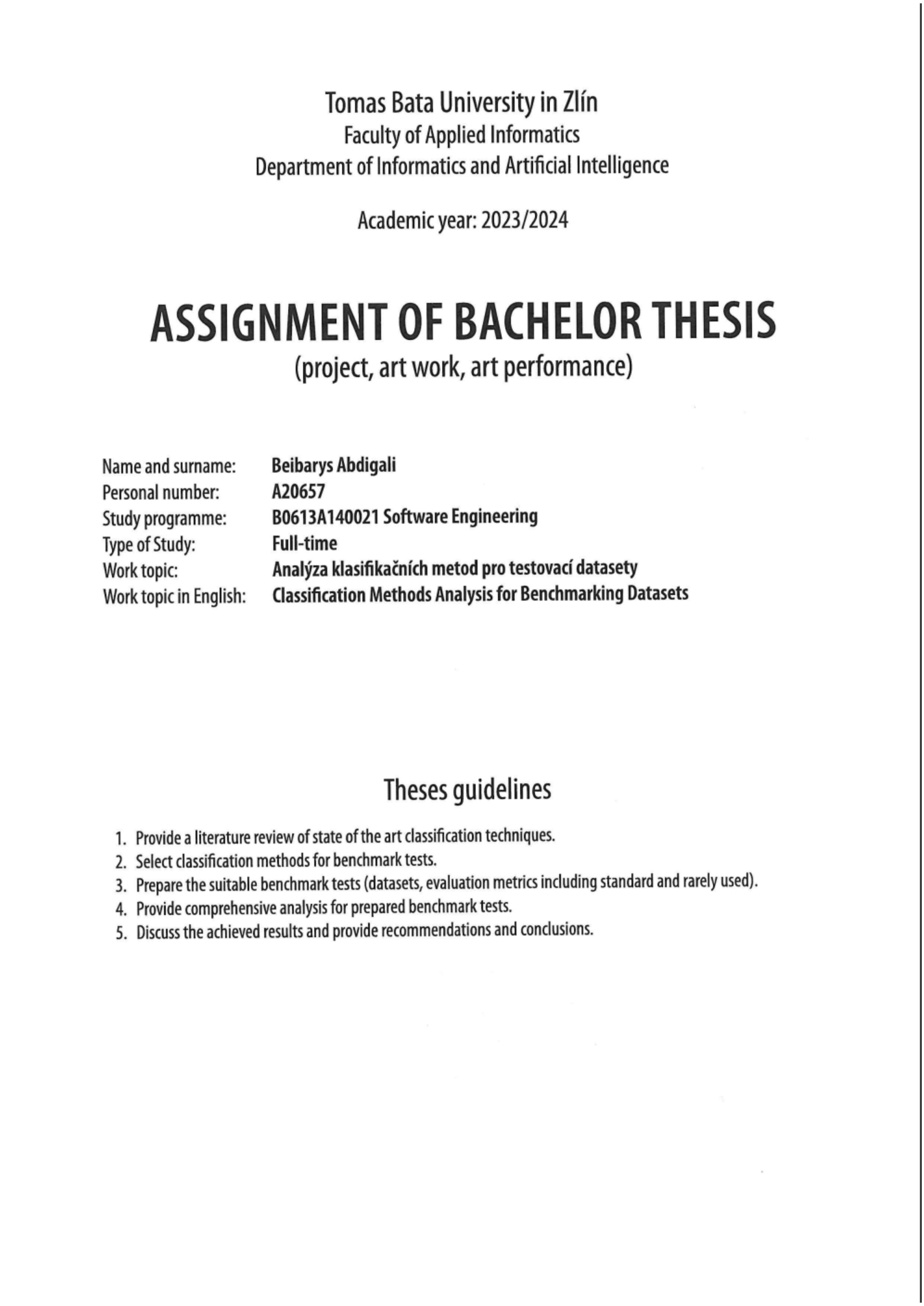
Classification Methods Analysis for Benchmarking Datasets

Bsc. Beibarys Abdigali

Bachelor's thesis

2024

Tomas Bata University in Zlín
Faculty of Applied Informatics



Tomas Bata University in Zlín
Faculty of Applied Informatics
Department of Informatics and Artificial Intelligence

Academic year: 2023/2024

ASSIGNMENT OF BACHELOR THESIS

(project, art work, art performance)

Name and surname:	**Beibarys Abdigali**
Personal number:	**A20657**
Study programme:	**B0613A140021 Software Engineering**
Type of Study:	**Full-time**
Work topic:	**Analýza klasifikačních metod pro testovací datasety**
Work topic in English:	**Classification Methods Analysis for Benchmarking Datasets**

Theses guidelines

1. Provide a literature review of state of the art classification techniques.
2. Select classification methods for benchmark tests.
3. Prepare the suitable benchmark tests (datasets, evaluation metrics including standard and rarely used).
4. Provide comprehensive analysis for prepared benchmark tests.
5. Discuss the achieved results and provide recommendations and conclusions.

Form processing of bachelor thesis: **printed/electronic**

Recommended resources:

1. AN, Shuang, et al. Granularity self-information based uncertainty measure for feature selection and robust classification. *Fuzzy Sets and Systems*, 2023, 470: 108658.
2. SONG, Zihao, et al. Doubly robust logistic regression for image classification. *Applied Mathematical Modelling*, 2023, 123: 430-446.
3. BARANDAS, Marília, et al. Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram. *Information Fusion*, 2024, 101: 101978.
4. VALLE, Denis; IZBICKI, Rafael; LEITE, Rodrigo Vieira. Quantifying uncertainty in land-use land-cover classification using conformal statistics. *Remote Sensing of Environment*, 2023, 295: 113682.
5. JÚNIOR, Joel D. Costa, et al. Novelty detection for multi-label stream classification under extreme verification latency. *Applied Soft Computing*, 2023, 141: 110265.
6. SILVA, Samuel Rocha, et al. On novelty detection for multi-class classification using non-linear metric learning. *Expert Systems with Applications*, 2021, 167: 114193.

Supervisors of bachelor thesis: **prof. Ing. Zuzana Komínková Oplatková, Ph.D.**
Department of Informatics and Artificial Intelligence

Date of assignment of bachelor thesis: **November 5, 2023**
Submission deadline of bachelor thesis: **May 13, 2024**

Tomas Bata University in Zlín
Faculty of Applied Informatics
L.S.

doc. Ing. Jiří Vojtěšek, Ph.D. m.p.
Dean

prof. Mgr. Roman Jašek, Ph.D., DBA m.p.
Head of Department

In Zlín January 5, 2024

Thesis topic:

I hereby declare that:

- I understand that by submitting my Bachelor´s thesis, I agree to the publication of my work according to Law No. 111/1998, Coll., On Universities and on changes and amendments to other acts (e.g. the Universities Act), as amended by subsequent legislation, without regard to the results of the defence of the thesis.
- I understand that my Bachelor´s Thesis will be stored electronically in the university information system and be made available for on-site inspection, and that a copy of the Master´s Thesis will be stored in the Reference Library of the Faculty of Applied Informatics, Tomas Bata University in Zlín, and that a copy shall be deposited with my Supervisor.
- I am aware of the fact that my Bachelor´s Thesis is fully covered by Act No. 121/2000 Coll. On Copyright, and Rights Related to Copyright, as amended by some other laws (e.g. the Copyright Act), as amended by subsequent legislation; and especially, by §35, Para. 3.
- I understand that, according to §60, Para. 1 of the Copyright Act, TBU in Zlín has the right to conclude licensing agreements relating to the use of scholastic work within the full extent of §12, Para. 4, of the Copyright Act.
- I understand that, according to §60, Para. 2, and Para. 3, of the Copyright Act, I may use my work - Bachelor´s Thesis, or grant a license for its use, only if permitted by the licensing agreement concluded between myself and Tomas Bata University in Zlín with a view to the fact that Tomas Bata University in Zlín must be compensated for any reasonable contribution to covering such expenses/costs as invested by them in the creation of the thesis (up until the full actual amount) shall also be a subject of this licensing agreement.
- I understand that, should the elaboration of the Bachelor´s Thesis include the use of software provided by Tomas Bata University in Zlín or other such entities strictly for study and research purposes (i.e. only for non-commercial use), the results of my Master's Thesis cannot be used for commercial purposes.
- I understand that, if the output of my Bachelor´s Thesis is any software product(s), this/these shall equally be considered as part of the thesis, as well as any source codes, or files from which the project is composed. Not submitting any part of this/these component(s) may be a reason for the non-defence of my thesis.

I herewith declare that:

- I have worked on my thesis alone and duly cited any literature I have used. In the case of the publication of the results of my thesis, I shall be listed as co-author.
- That the submitted version of the thesis and its electronic version uploaded to IS/STAG are both identical.

In Zlín; dated: ....

Student´s Signature

ABSTRACT

In this thesis, shown focus on the classification algorithms utilization on the predictive modeling and distinguish them from regression as different breeds of the similar genus with alternate options. Here, the study dwells on important algorithms including K-Nearest Neighbors(KNN), Decision Trees, Random Forests SVM, Logistic Regression and Naive Bayes. The discipline is concerned with the historical aspects, physics, and math formulas of the K-nearest neighbor, decision tree, and random forest models. It continues to explain the avoidance of overfitting in decision trees and the ensemble approach used in random forests. The algorithms are going to be evaluated based on their performance compared to datasets like Digits, Wine, and Diabetes. Besides being in charge here, the job involves tasks like selection of data, processing, splitting and scaling. For this case study a mix of well-known parameters as well as lesser known ones will be used to comprehensively survey the capabilities and limitations of the models studied. This piece stops here to discuss the disparities between classifiers and regressors, with aim of choosing best models along the way by means of characteristics of data and the very detail of research problem. The conclusion has several noteworthy points related to real-life application of algorithms and also points potentially research areas that may require more study. Therefore, the purpose of this thesis is namely to pool together the data for the researchers and professionals in data science and predictive analytics.

Keywords: KNN, SVM, Logistic Regression, Random Forest, Decision Tree, Digits, Wine, Diabetes

Acknowledgements:
I am deeply grateful to Prof. Ing. Zuzana Komínková Oplatková, Ph.D. and Ing. Peng Li, Ph.D. for their valuable guidance and support in my bachelor's thesis. Their expertise and dedication have been instrumental in shaping the outcome of my research, and I am truly thankful for the opportunity to work under her supervision.

I hereby declare that the print version of my bachelor's thesis and the electronic version of my thesis deposited in the IS/STAG system are identical

1 Contents

INTRODUCTION

The field of machine learning is always subject to development and, in the meantime, there is one area that manages to grasp the attention, the technology of classification algorithms. The fact that these algorithms are extremely important in understanding and identifying the consequences of data on the real world's applications is a piece of evidence. A categorization is the root of making decisions in both common and extraordinary events of day-to-day life like the decision making process. This is true whether it is improving the precision of diagnosis or searching trough digital mess to spot only meaningful of them. This thesis aspires to accomplish twofold mission-on the one hand, to unveil the inner workings of these precise algorithms and, on the other hand, to lay a ground for a rigorous benchmarking discipline that documents the performance of these techniques in various circumstances. Classification systems constantly change and get innovative. One should demonstrate the latest and the most innovative classifications systems [62].

Although machine learning categorization is on the right track, the fact is it is full of many diverse approaches and algorithms, all of which have their own set of valuables and demerits. Finally, the biggest classifications algorithms influencing field of data science nowadays will be examined in detail.

K-Nearest Neighbors (KNN): The KNN stands for the simplicity of the presentation in machine learning. Other than that, it distributes data points among neighbors and assigns them a class identifier that is dominant in the neighborhood. Somewhat cliché and paradoxical, the underlying question is whether things close to each other are similar to each other. A key factor to the success of KNN is the careful selection of the distance measure and its neighbor numbers. Such a choice has a great influence on the algorithm's performance this is measurable when it implemented using a large set of different datasets [64].

Support Vector Machines (SVM): SVMs stand out among others for the fact that they can distinguish really complicated class boundaries, representing hyperplanes in a multidimensional space. They are affiliated to the classifier category, which they are believed to be the most developed ones. The fact that SVMs are able to manage non-linearities of any complexity and succeed in out-of-the-way high-dimensional spaces is what makes this feature profound. Also, even though they are very resilience algorithms, SVMs still have hyperparameter optimization and scalability issues and for that reason becoming the subject of current researching and development of them.

Logistic Regression: Two words have the exact opposite meaning from what one would expect of them, but logistic regression is indeed a vital technique for classification where it models the probability that something achieves the class through the aid of logs. One of the simplest approaches that can quickly answer of the clear facts math and computation related is that there is a relation between characteristics and outcomes. Furthermore, the logistic regression stands well against other modernized regularization approaches and still remains one of the best techniques suitable to handle even complicated, high-dimensional data [60].

Naive Bayes: A Naive Bayes classifier calculates the percentages needed for the probability, using Bayes' theorem as a foundation, under the assumption that all features are independent variables. This assumption, rather constantly being contested, still excels the algorithm considerably at such tasks as text classification and some other uses cases with speed and average accuracy year by year [1].

Decision Trees and Random Forests: Decision trees simplify the decision-process by partitioning the dataset in different subsets with overlapping features using stepwise sequential feature splits. The Random forests concept is by such bringing together of plural decision trees, thus improving the prediction correctness and protecting from the overfitting danger. These strategies present a trade-off between interpretability and performance, though from the standpoint of model-building itself they might be somewhat limited by messy or complex data structures [3].

Expanding the Benchmarking Horizon: The aim of this thesis is to meet certain standards in performance evaluation of categorization algorithms in expansion to metrics often used which offers a just preview of their strengths and limitations[61].

Selection of Benchmark Datasets: One of the vital aspects of our data benchmarking strategy is the thoughtful selection of datasets representing real-world-datasets rather than artificially-generated ones from reliable sources like the UCI Machine Learning Repository[72] The dataset generated with its own datasets, which were chosen based on their relationship and representativeness each one of them, was used as a catalyst to uncover whether the algorithms could adapt and function equally well with any data dimensions, distributions, and complexity.

Evaluation Metrics Unveiled: Meanwhile, the metrics which include accuracy, precision, recall and F1-score argue that this paper puts less weight on the measures like Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and κ statistics which were less popular till now as well as the Matthews Correlation Coefficient (MCC). These features, being of finer nature, display the exact elements of classifier performance, which in turn are utilized in the process of producing a comprehensive view of suitability to perform the required tasks. This thesis is aimed at providing a deducted approach using objective information with insights that are further used as a measuring tools. The analytical route presented by this thesis paper is through a systemic measure approach with a primary objective of getting the highest value per category [63].

Not so long ago deep learning methods needed lots of data to work, now they can be used with larger data sets due to transfer learning, fine-tuning, and data extraction. As there are not many labeled samples, the search for high-level feature extraction is incidentally continuing. We should be making uniform standards for observable comparisons. Researchers working in different fields, they have published collections, including nature pictures, medical images, fine categories ads others. With these standards, it is possible to avoid discrimination and assess progress. It is astonishing to see how the market economy may benefit from even the slightest adjustment of the parameters for its success. However, achieving the same level of progress is relatively difficult. undefined Learning with few shots: Drawing wide generalizations from seemingly specific examples. Meta-learning: Training models on how to learn as well. Active learning: deciding on which samples to give the appropriate names. Sample bias and overfitting arising from high uncertainty affect few samples. Lack of diversity results in an incapacity to generalize properly and affects the feature learning [59].

Concluding with Recommendations for Informed Application: The essence of our research will be a set of strategic proposals including those elements that vary under different approaches and a number of common ground ones that remain the same for all of them. The diverse breadth of our findings that range from the rich dimensionality of SVMs to the understandable depth of decision trees serves like a lamp for scholars and practitioners looking for a resort back to learning.

I. THEORY

1. CLASSIFICATION AND REGRESSION

The purpose of the model is to try to make a prediction of the correct label for a specific set of data which has already been provided to another machine learning process (i.e., classification). It is then scored on the training dataset then evaluated with the help of test data which is provided during the performance of the model. It involves the model "learning," which is completed prior to the model being used for prediction on the unknown data that was not happened before.

Lazy Learners Vs. Eager Learners: Within the realm of machine learning, there are two distinct types of learners: those who are eager and those who are you shall see.

Eager Learners is a specific type of machine learning algorithm which this is first build by the trainings dataset to create the model. This checks that the algorithm not only makes predictions for the data it trained on but also on future datasets. Since they spend more time throughout the procedure as the thing is of upmost importance for them and they want to better their generalization, these models take less effort to perform forecasts. Because they are keen on the on – boarding method.

The bulk of machine learning algorithms are rapid learners; some cases are listed below:The bulk of machine learning algorithms are rapid learners; some cases are listed below: Logistic Regression, Support Vector Machine, Decision Trees, Artificial Neural Networks [2]. Invarely, those learners are labeled as lazy ones, also instance-based ones, that being the pace which they create a model using the training set appears to be slow and this results from their lazy nature. Basically speaking, they reckon training data by memory, then whenever they predict the final outcome, they approach the closest neighbor in the data which would lead to extremely slow predictions.[4].

1.1. Difference of regression and classification

1.1.1. Classification Predictive Modelling

Classification predictive modeling is the occurrence of mapping each input variable (X) by a function (F) that takes each input (X) and maps them to a discrete variable (Y), that is mostly called the label or the category. Which is the sole function of the model – it is supposed to classify the observation.[4].Consider it a sample, let us navigate an email message. It may be divided into one of two classes: As Keiser says: “maybe I’m the bad guy, or maybe I’m just ahead of my time… or maybe both.”

Suspicion of discriminator element imposes that every instance be placed into one of two or more classes in some way. It is possible that the classes may use a real or discrete input variables. An example of such a challenge that is usually applied to the two groups of data is called a binary classification that involves two classes. Where a solution more than two classes is involved is referred to as a robust multi-classification. The situation which arises when one instance is a dealing with several classes is referred to as a multi-label classification problems [5].

It is common for binary and multiclass classifications to only train models which utilize a continuous prediction value which is calculated as the probability of an instance belonging to a particular output class. The probabilities could be taken as either as the confidence or probability values for a specific case to be assigned to a particular class. A predicted probability is used to decide the class value as it selects the classes that have a higher likelihood.

Instance of a message, sent in the form of one email could be likened to 0.1 probability of being "spam" and 0.9 probability of being "not spam". Converted chances into a classy imprint through the “Ham” label since it has the highest resulting probability[6]

Some methods are available for accuracy assessment the best one is to find the classification accuracy as people are mostly using it. The classification accuracy measure shows errors in the ratio to all the predictions. For instance, if a classification predictive model made 5 predictions and 3 of them were correct and 2 of them were incorrect, then the classification accuracy of the model based on just these predictions would be the formula of the accuracy is shown in (1) [7]:

$$accuracy = \frac{correct\ predictions}{total\ predictions} \tag{1}$$

1.1.1. Regression Predictive Modelling

Regression predictive softy implies externalizing the function of mapping that resemble taking in the input variables and translates to a continuous output variable [8].

A variable that is always being produced in a continuous way has a real-value such as an integer or floating point number. Similarly is personalization where the most widespread characters are amounts and sizes. [9].

Regression regimen anticipates a certain amount. A regression, for its part, among other possibilities, could be based on real-valued or discrete inputs. When several input variables are put together in one issue, this issue is usually termed as multi-variate regression problem. A regression problem is considered as a time series forecasting when inputs are formed by time in i.e., organized. As the regression predictive model is used to predict a magnitude, the precision of the model is known to be associated with discrepancies in those predictions [10].

A regression prediction model can be assessed in different ways, but standard option is to derive the root mean squared error (RMSE) the formula is shown in (3), where y_i is the actual observed values in your dataset, $\dot{y}_i$ the mean (average) of the observed values y_i, N represents the total number of observations or data points in your dataset, P represents the number of predictors or independent variables in your model. In simpler terms, it's the number of features or factors you're using to make predictions.

$$RMSE = \sqrt{\frac{\sum (y_i - \dot{y}_i)^2}{N-P}} \tag{3}$$

A major advantage of the RMSE is that the units of the value obtained from the error regulatory function are of the same type with the parameters for which the model was designed. A regression algorithm is a program, as most of these algorithms focus entirely on training a model which is based on regression prediction. Some algorithms even use the word regression in their name like linear regression and logistic regression, although, both of these not the regression method, but it is just a misleading name.[11].

1.1.2. Classification vs Regression

The process of predictability modeling consist of two subcategories: classification and regression each with pecularities of its own.

In classification the situation involves suggesting a class value for a given data entity. In the other connection, the regression means when you are predicting a continuous variable. The classification and regression technologies have points in common. For instance:

A classifier works with a numerical value that is interpreted to be a probability of the defined class mark [11]. A regression might suggest a discrete constant, where the constant is also an integer quantity but the input value is discrete [16]. Sometimes, particular techniques can be substituted, used either for classification or regression, requiring only a minor adjustment to the approach, such as decision trees and artificial neural networks [17]. But all the plans are not as adaptable for socio-economic and urban matters. For example, linear regression may be used for a regression predictive modeling but categorical work is often achieved through logistic regression.[12].

It is perhaps the most critical to reflect and understand that the means to assess the calibration and regression predictions are distinct and are not the same at the end of the day [13]. For instance: The accuracy is mostly used on classification but the metric for regression prediction is not efficiently captured by this.[14]. The RMSE is the right approach used while evaluating the predictions in the regression problems, whereas for the classification problems, this measure will not be efficient.[15].

2. CLASSIFICATION ALGORITHMS

2.1. K-Nearest Neighbor (KNN)

2.1.1. History of KNN

KNN algorithm is a non-parametric supervised learning method, in which the reference is an article by Fix and Hodges from 1951. It was after that, which was designed by Thomas Cover, when the printing was enhanced.[18].

The KNN algorithm is an all-purpose algorithm which lets us classify as well as regress. With K-Nearest Neighbors, the subsequent input will be the K closest examples (referred to as the training data set) [19]. For classification, the consequence is a class membership, after units are described to have label by majority of its neighbors. The result proceeds regression for the object property which is the average of the k close neighbor values which is one parameter .[20].

KNN is an ensemble (or cases-based) method of learning, where the function is not approximated, and all the processing is only delayed until the function evaluation takes place. This will influence the performance of the model effectively because this strategy is primarily focused on distance classification so tweaking of the training data could thus make the model more accurate [21].

In time, different upgrades to the KNN algorithm that brought further improvements. Fukunaga and Hostetler [22] obtained bettering rates that were pertinent to the Bayes error rate. In the mid to late 1970s, Dudani and Baily [22] published distance-weighted approaches to this problem. In 1983, an idea of the learning method for a fuzzy KNN rule of Adam Jozwik [22] was developed.

2.1.2. KNN working process

The KNN method is doing so by comparing the new data input with the existing data values (that are differentiated by their classes or categories). Given a certain range of similarities or closeness (K) by this algorithm, the new data is assigned to a class or category as similar to the training data. It takes 4 main steps:

- Assign a value to K.
- Write down the distance between the new data input and other all the existing data entries . Arrange them from the lowest to the highest frequency.

- Locate the K nearest neighbors to the new entry based on the calculated distances.
- Place the new data element of the most frequent class in the neighboring class.

The Euclidean distance between two points p and q in a two-dimensional space is computed as (4), where p_1 and q_1 These are the respective coordinates of the first dimension (often denoted as x axis in a Cartesian coordinate system) of points p and q. So, p_1 is the first coordinate of point p, and q_1 is the first coordinate of point q, similarly, these are the respective coordinates of the second dimension (often denoted as y axis in a Cartesian coordinate system) of points p_2 and q_2. So, p_2 is the second coordinate of point p, and q_2 is the second coordinate of point q:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \tag{4}$$

2.1.3. Advantages and Disadvantages of KNN

Advantages of KNN: This algorithm is not difficult in terms of implementation, which is why it is a favorite among new entrants in machine learning.

No Training Required: The KNN algorithm does not necessitate any training process which implies among others applications that can be used in real-time response where data emerges as such [26].

Accurate and Effective: KNN among other methods is well-known for both its effectiveness and accuracy especially when applied to small datasets of medium sizes. Disadvantages of KNN: Sensitive to Outliers: Outliers give rise to general fragility of the KNN approach, hence the model's performance is likely to be poor.

Computationally costly: A big challenge is the KNN algorithm that may require much computational resources. Therefore, this technique limits the movements to the calculation of the distance between every testing data and any training data, which generally is comprehensive [68].

Requires Good Choice of K: The KNN algorithm necessitates a well-chosen K value, being the number of nearest neighbors being used in classification. If K is insufficient, there is a high chance the algorithm, on one hand, will inundate itself with noise in the data, and on the other hand, it will miss important patterns that appear in the dataset.

Non-parametric: The KNN algorithm does not make assumptions about the distribution type of the data, thus, it tends to be generalized algorithm which can be applied in different settings.

Limited to Euclidean Distance: The choice of the various distance measures sometimes serves as the basis for this scheme measures, though it is the common implementation that still relies on the most popular, which is Euclidean distance, may not necessarily be the best suitable measure [67].

Imbalanced Data: KNN can perform poorly with imbalanced data. If one class has significantly more examples than another, the algorithm will be biased towards the majority class.

2.2. Decision Tree

2.2.1. History of decision Tree

Decision trees have a complex and extensive history, developing from many academic efforts and real-world implementations. Below are significant milestones in their development .[27]:

Historical Background: Ronald Fisher, in 1936, published seminal work [28] in the area of biology. Decision trees are based on the work of a mathematical theory which has proved to have some quite peculiar properties. Ronald Fisher is known in the statistical community for user discriminant analysis, which built the basis for mathematical understanding of the interdependence of input factors and the target (response) variable [28]. AID Project (Adaptive Incentives Decision) tools that were invented by Morgan and Sonquist in the 1960s had greatly increased the application of binary segmentation trees. The principle of the tree was that they will try to figure out the relationship between attributes and result.

Hunt's Publication (1966)[73]: Moreover, Hunt's area of interest also contributed to the emerging decision tree techniques in those respectful decades.

Psychology and Learning Models: Decision tree structures were initially proposed for describing the psychology of learning with the human beings. Surprisingly, the results have shown that in addition to just leaning modeling, it has also found a wider use.

Classification Tree Emergence: The firsts classification tree was developed in the THAID project in 1972 [74] by Messenger and Mandell, which is the project directors. This had a key role in moving the tech closer to the practical deployment.

CART method (1977)[66]: Andrew Ng from the University of California at Berkeley and Charles Joel Stone collaborated with Jerome H. Friedman and Richard Olshen from Stanford University and advanced the Classification and Regression Tree (CART) model. CART created a unified system for dealing with hence foresight would of been a pseudo-utopian state.

Decision Trees Today: Decision trees still rank among the most used about the fact that they are verbally transparent, transparent and efficient. They are mostly used for operations research, decision-making as well as machine learning.

Tin Kam Ho (1995) [1]: With tremendous experience in the Statistics and Learning Research Department, Tin Kam Ho has proposed research algorithm during his work at the Bell Laboratories.

2.2.2. Types of Decison Trees

- CART is a powerful and versatile algorithm that can be used for both classification (predicting discrete categories) and regression (predicting continuous values). The splitting criteria of CART uses a process called binary recursive partitioning. At each node of the tree, it chooses the feature (attribute) that best splits the data into two groups based on a specific impurity measure. Common impurity measures include the Gini index (for classification) and mean squared error (for regression). The process continues recursively until a stopping criterion is met (e.g., reaching a certain depth or minimum data points in a node). CART is known for its simplicity, interpretability, and good performance on various datasets. The use of clear impurity measures allows for easy understanding of how the tree makes decisions. CART can be susceptible to overfitting if not properly tuned. Additionally, it might not always choose the optimal split at each node, leading to potentially less accurate trees compared to some other algorithms. [2]
- ID3 is a foundational decision tree algorithm specifically designed for classification tasks. It's a relatively simple algorithm that forms the basis for more advanced algorithms like C4.5. ID3 uses information gain as the measure to choose the best attribute for splitting at each node. Information gain measures how much a specific attribute reduces the uncertainty (entropy) in the data after the split. The attribute that leads to the highest information gain is chosen for splitting.ID3 is a straightforward algorithm that's easy to understand and implement. It's a good starting point for learning about decision trees. ID3 has some limitations. It can be biased towards attributes with a large number of distinct values (e.g., ID numbers) and doesn't handle continuous attributes well. Additionally, it doesn't consider the possibility of missing values in the data. [3]
- C4.5 is an extension of the ID3 algorithm, addressing some of its shortcomings. It's also designed for classification tasks. C4.5 builds upon ID3 by introducing the concept of gain ratio. Gain ratio addresses a weakness in information gain by considering the number of possible values (splits) an attribute can have. This helps avoid favoring attributes with a high number of values. Additionally, C4.5 can handle continuous attributes by discretizing them using techniques like splitting based on a certain threshold value. Furthermore, C4.5 incorporates pruning techniques to reduce the complexity of the tree and prevent overfitting. C4.5 builds on the strengths of ID3 while addressing some of its limitations. It can handle continuous attributes, considers the number of splits when choosing attributes, and incorporates pruning for better generalization. While C4.5 is an improvement over ID3, it can still be computationally expensive for very large datasets. Additionally, choosing the optimal pruning strategy can be complex. [4]

2.2.3. Decision Tree Working Process

Decision Tree Splits and Accuracy: The strategic choices of development of trees greatly affect its correspondence. The act of node division into sub-nodes is made with the goal of compensating for the low homogeneity in the subsequent subsets. Translated, this statement refers to attempt to apply methods of filtering closer to the parameters which wishes to predict. The method than chooses dimension that provide more pure samples for generating data sets.

Selection Based on Target Variables: The choice of algorithm relies on the kind of target variable: The choice of algorithm relies on the kind of target variable: Categorical Variable Decision Trees: Comment: Used when the target variable is categorical (such as class labels).

This research mainly used ID3 algorithm therfore the steps in ID3 algorithm are detailedly explained as follows: It begins with the original set S as the root node. On each iteration of the method, it iterates over the very underused attribute of the set S and calculates Entropy (H) and Information gain(IG) of this property. It then picks the element which has the least Entropy or Largest Information gain. The set S is then split by the specified attribute to obtain a subset of the data. The method continues to loop on each subset, considering only qualities never picked previously.

Entropy is a measure witnessing the amount of mess or dirtyness that exist in a set of data. In the decision trees, it is the impurity which is level of uncertainty in the target variable (class labels). The purpose is to reduce the degree of disorder by choosing the splits to which an inhomogeneity increases. The formula for entropy is shown as(6), where *c* is defined as the number of various boundaries, p_i is the proportions of cases that is belong to a type of i class.

$$Entropy = \sum_{i=1}^{C} -p_i \cdot \log_2(p_i) \tag{6}$$

Information Gain: Interpretation of information gain refers to the entropy reduction attained by dividing the data into two or more parts with respect to a certain principle. It gives a picture of the disorder before the break and the one after that. The splitting factor which is defined through the highest information gain at that point is chosen as the largest information gain among the properties. Formula for information gain (7), where T and X is the measure of how much information the feature X provides about the class labels (or target variable) T after the dataset is split based on the values of X, Entropy(T) represents the entropy of the original dataset T before the split. Entropy is a measure of impurity or uncertainty in a dataset , represents the entropy of the dataset T after it has been split into subsets based on the values of feature X.

$$Information\ Gain(T, X) = Entropy(T) - Entropy(T, X) \tag{7}$$

Gini Index: The Gini index purity of a node calculates the proportions of misclassified instances and therefore the probability of them to be mistaken for a randomly selected instance. It shifts from 0 (the perfectly balanced nodes) to 0.5 (the entire network is unbalanced). Indexing for gini then is carried out among the splited parts, and the characteristic with the smallest Gini value is selected. Fromula for Gini index show as (8), where *Gini* represents the Gini impurity, which is a measure of the impurity or uncertainty in a set of class labels, *c* represents the total number of classes in the dataset, p_i represents the proportion of instances in the dataset that belong to class i. In other words, it's the fraction of instances in the dataset that are labeled with class *i*

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \tag{8}$$

Gain Ratio: Ratio of gain of information is introduced to take care of the number of the branches originating from a split. However, it hurts men whose qualities have a more universal applicability. The formula for gain ratio shown as(9), *Informaitiom Gain* measure of how much information a feature provides about the class labels after the dataset is split based on that feature, *Splitinfo* measure of how uniformly the dataset is split by the feature, *Entropy(before)* entropy of dataset before the split, *Entropy(after)* entropy of dataset after the split ,*K* presents the number of subsets created by the split, *w* represents the proportion of instances in the jth subset created by the split, “before” is the dataset before the split.

$$Gain\ Ratio = \frac{Information\ Gain}{SplitInfo} = \frac{Entropy(before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j log_2 w_j} \tag{9}$$

Reduction in Variance (for Regression Trees): Here the goal is to lower the spread of the independent variable within the sub-grouping. Reduction in a variance denotes an lowering in variance when splitting on an indicator [35]. In the process of mediation, the most spread-out property which stands for the variance with the lowest value is picked [36].

2.2.4. Ways to Escape Overfitting in Decision Trees

Normally decision trees, even though they may fit into a full table of columns are found to be having difficulty especially during training. At times it looked like the tree just memorized certain samples from training data. A tree that has no limitations on the details results in the only perfectly accurate predictions on the training data set. This is the worst-case scenario because at the end it will result in 1 leaf for each observation. Thus it negatively influences the reliable models for samples that are not part of the learning set. Pruning and Random Forest are the two methods to escape overfitting.

Pruning: During pruning, prune the tree using the strategy that starts from the leaf node where the original tree's accuracy will not be disturbed. This is done by dividing the actual training set into two sets:1)Define the sought-after outcome by creating a training data set. 2) Evaluate the performance of the algorithm using the validation data set. Create a decision tree on the training data set afterwards. Once the validation data set is improved, go on and prune the tree accordingly.[37]. Prining process is showing as Figure 1:

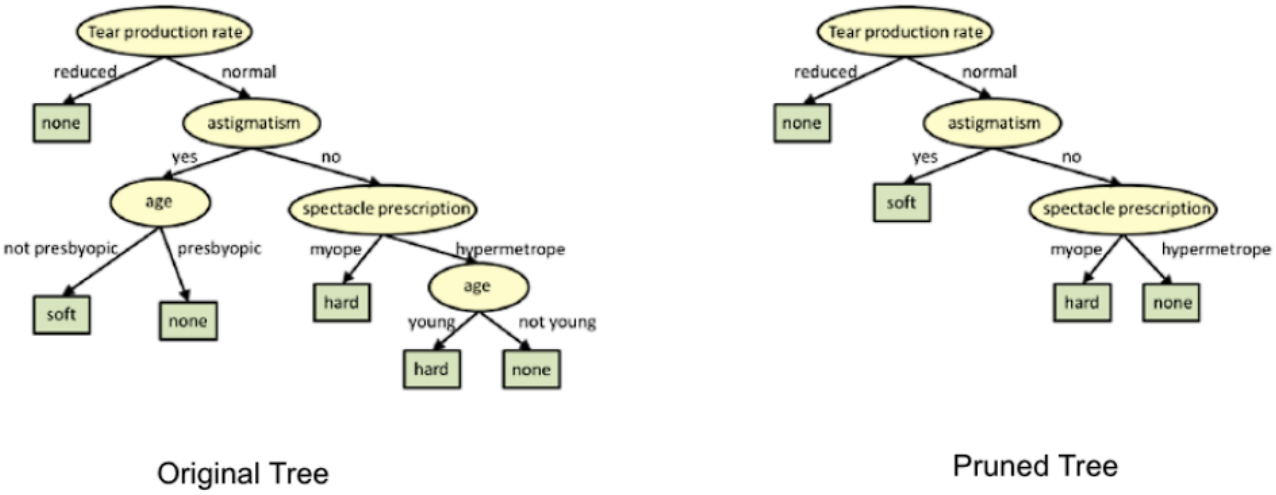


Figure 1 Pruning process [123]

2.2.5. Advantages of Decision Trees

- Simplicity of Interpretation: Decision tree provides the decision-making process in the form of a picture and is therefore easy to understand. A hierarchical nature creates such a convenient framework which compares characteristics with predictable outcomes.
- Robust to Outliers: Decision trees tend to outperform linear methods in the sense that they are less sensitive to outliers. An outlying amount of data would have no impact on the tree formation. They are not putting the missing values at their own against the data.
- Non-Linear: Tables show complex, not linear connection between features and results.
- They possess the ability to acclimate to challenging and often indefinite decision lines.
- Non-Parametric: Decision trees are not based on any particular data distribution as other algorithms. They are the most flexible tools and can trace complex relationships patterns.
- Combining Features to Make Predictions: The decision trees give better results when interactions between features have been considered. They synthesize the data in their own way in the making of the opinions.
- Can Deal with Categorical Values: Categorical features eliminate the need for dimensional transformation called one-hot encoding with decision trees. They jail nodes according to certain categories and attributes.
- Minimal Data Preparation: Decision trees doesn't requires less percentage preparation as compared to their counterparts (e.g., scaling, standardization).They function successfully at the pre-established stage.

2.2.6. Disadvantages of Decision Trees

- Prone to Overfitting: Arrangement trees can end up with too much complexity and contain random noise from data. Frequently regularization operations (e.g., pruning) are used to prevent footing.
- Unstable to Changes in the Data: Keeping tiny variations on the dataset cause the tree pattern to get twisted out. Missed regularity across variations of training data discards decision trees.
- Unstable to Noise: Trees can be divided based on noisy features which give errors in decision trees algorithm. The noise in the form of incorrect splits is most likely to occur.

- Unbalanced Classes: Decision tree models are not good at handling problems where the classes are unbalanced. The best efficiency cannot be reached without the oversampling or undersampling methods.
- Greedy Algorithm: Decision trees begin with a greedy algorithm in their construction process. They can have deficiancies and unable to global optimization of tree an structure.
- Computationally Expensive on huge Datasets: Creating compact trees for large data sets is supercalifragilisticexpialidocious. As capacity increases, the time for training increases.
- Complex Calculations on Large Datasets: Decision tree calculations involve evaluating multiple splits. For large datasets, this can be computationally intensive.

2.3. Random Forest

2.3.1. History of Random Forest

Further detail is needed in order to comprehend the historical evolution of this method. Random Forest is an example of ensemble classification. This application of ensemble learning is utilized to increase the accuracy of classifications [38].

Ensemble learning, which is a machine learning method that differs from the conventional learning where one single model has been applied to the same problem, introduces several models instead of one model to deal with the same problem. In ensembles such as Random Forests, many classifiers are applied alike. This is due to the fact that they are more consistent than any member of the group can be when they are predicting by using different methods.

Then an inner vote is being held by the Random Forest algorithm to determine an accurate class label for unidentifiable instances. Voting is not an easy method of decision-making, but it can still move forward through majority voting, which is majorly a conceptual invention of two academics [5].

In the Method of majority voting every classifier that make up the ensemble is told are to try as much as possible to classify correctly the class label of the instance being probed by each of them [6]. This kind of an ensemble sends out all the classifiers to get asked and then returns the class which gets the biggest score as the final decision. There are different voting systems at work: for example, in the strategic veto voting plan, one classifier vetoes the other. In the concern to arrive the finest result for the classifiers in the ensemble, this classifier should be both reliable and, at the same time, be different from the others.

Nonetheless, the parameters of an ideal classifier includes that the classification error is higher than chance hypothesis. Moreover, two classifiers are inconsistent if there exist errors or misclassifications in case of considering new data points or. When the classifiers are very distinct from each other, get the better learning results. Generally, the ensemble works better than the individual models by nature when there is a difference in the models that make up the ensemble [7].

2.3.2. Random Forest Working Process

The meteorologists miss where the models fail. The particular instance of investments where low correlations are about tries, whose portfolio is of more value than the sum of its parts. Uncorrelated models, also, provide ensemble predictions that are more correct than any of the individual predictions. One of the reasons for the great result is the fact the trees are screening each other from the errors they may commit individually (if all the trees are not erring in the same direction). Some trees may thus be wrong, but by correcting this some trees may also be right, so that the overall group is able to get to the direction they need.

There should be some genuine quality in features so that models that feature these signals build some useful signals and outperform random guessing. The production of different trees ought to have as much difference as possible between the results, leading to the lowest possible correlation.. Though you may already know that having various uncorrelated models is superior among these reasons, still it deserves to be illustrated because it is such a main theme that you have got to be sure that you've really understood it.

The Random Forest algorithm contains various decision trees with the values of the nodes being the same, but across different data sets, each will end up in different leaves. They construct the decision tree ensemble responsible for looking out for each option and finally average it to reach the best solution [42]. By doing category-based executions on Random Forest that use the Gini index, or the principle according to which added nodes on a tree formula (10) :

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \qquad (10)$$

Using the branch and class probabilities the Gini of each node determines the branch that pertains to the highest occurrence in a given node. Here, pi refers to the relative frequency of the class you wish to view in the dataset and c stands for the number of classes. Apply entropy to find the number of branches occurred in decision tree.

Entropy calculates the probability of a specific result and makes decisions regarding the network of the node based on its calculations. Different from the Gini index, it is a more complicated mathematics tool because of the logarithm function that corresponds in computing it. Example of how looks Random Forest working process :

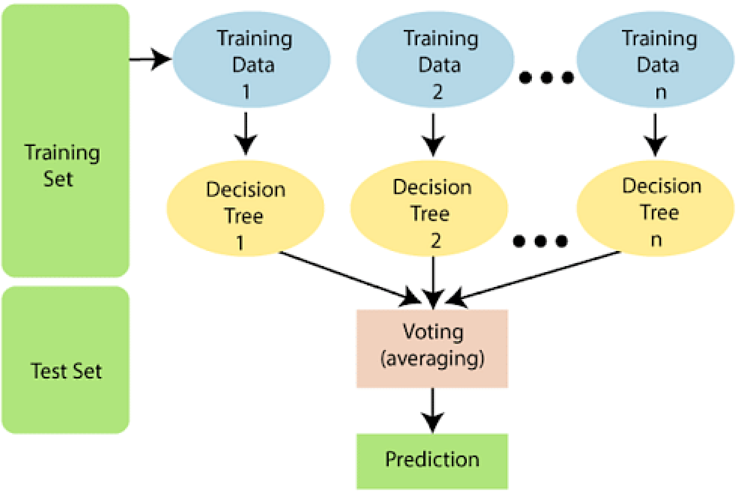


Figure 2 Random forest working process[124]

2.3.3. Advantages of Random Forest

- High Accuracy: Random Forest according to several decision trees. The aggregated forecasts is achieved. Trial in averaging (regression) or voting (classification) lead to the creation of more accurate output. The algorithms on average outperform single decision tree ensemble models mostly.
- Robustness to Noise: There is so much ability in Random Forest to cope with noise. With decision-making that is data-driven, irrelevant data points contribute much less to the projected actions.It works with outliers and has a high tolerance to noise too.
- Non-Parametric Nature: Random Forest does not impose preliminary restrictions on the specific data models as occurs in the case of other methods such as liner regression. It is quickly becoming adoptable among different resource constraints and conflicting vocabulary.It represents data in an intricate manner without having a qualitative limit.

- Feature relevance Estimation: The Random Forest determines the quality of the characteristics. It evaluates per function's impact on variance elimination. Helps at feature selection and doing data analysis.
- Handling Missing Data and Outliers: On top of that, Missing Data is well-managed by this machine learning algorithm. This eliminates the RFI (remotely sensed data), imputation, and outlier elimination. Outliers rate is decreased by the effect of ensemble averaging, this owing.
- Numerical and Categorical Data Handling: Random Forest can manage either an arithmetic or categorical data. At the core, it carries out random selections of the features that are necessary for the training.

2.3.4. Disadvantages of Random Forest

Computational Complexity: On performing this with a large number of trees or training on large datasets, the Random Forest could be computationally intensive.Each separate tree is trained individually taking plenty of work to process these forecasts which involve combining the models. The multi-featured nature of Logistic Regression can contribute to wider training durations and larger memory usage which may be more prominent on systems with limited resources.

Memory Usage: Big data and deep trees forces Random Forest models to use large memory of RAM. Each model tree in the forest has a collection of items: a training data, a feature splits, and a leaf node predictions. The number of trees is directly proportional to how memory is utilized. It follows that the storage issue that may arise from large trees or deep tree structures may become a bigger problem with low hardware specs.

Prediction Time: While the Random Forest model had a really high success rate during training, it takes more time to get the predictions burnt. Through this process of observation, the entire data set will be sent to the trees of the model where they would go through multiple decision trees. This surplus of additional time would damage real-time like applications or those that need quick responses.

Lack of Interpretability: Random forests fall into the category of "black box" algorithms. Making a lot of decision trees too, it becomes hardish to understand the reasons that stand at the back of each prediction.With feature importance metric, features that really matter can be identified, but the complex nature of interactions between feature may still be obscure.

Overfitting: Random Forest can subsequently have the issue of "overfitting". The model might learn noise or some patterns in the training very appropriately and it can under fit to the general situation where the data is new and unknown. Model complexity should to be balanced to avoid the issue of overfitting will allow or maintains the prediction accuracy in the real world.

2.4. SVM

2.4.1. History of SVM

In 1992, a new, exceptional modle was shown - the SVM. The SVM approach–developed by Vladimir N. Vapnik and coworkers at AT&T Bell Laboratories–constituted an abrupt technological leap in machine learning. SVM algorithm is practically designed by Vapnik and Chervonenkis in the early 1970s. It becomes more methodical by using the statistical learning framework, i.e. the VC theory. SVMs ought to be understood as preferable prediction clues for the answer of classification and regression queries whatsoever [43].

Key Concepts of SVMs, binary Classification:Basically SVM Given some labeled training sample set (both of which belong to two different categories) generate a model. New unlabeled cases are assigned to either out of the two depositories [44]. SVMs are non-probabilistic binary linear classifiers that are aimed at recognizing the linear classes to put a margin between them.

Geometric Representation: SVM model includes instances in the form of points of a vector field of large-dimensional space. The idea here is to develop an effective hyperplane (linear classification way) with greater margin distance between the two classes. The newly acquired data points are then arranged in this new space and characterized based on their position in the gap. Kernel approach for Non-Linearity: The SVM kernel feature approach is effective in feature space nonlinearity [45]. With the help of nonlinear association representation by mapping to a high dimensional space, SVMs depict complex relations. Kernel SVMs, on the contrary, increase the concept’s basic meaning, permitting for flexible modeling.

Unsupervised Learning with Support-Vector Clustering: When data retains no labels, then supervised learning isn’t applicable. The vector support clustering method (SV) developed by Hava Siegelmann and Vladimir Vapnik helps to classify unlabeled data. It naturally groups data and is simultaneously capable of assigning new data points correctly.

Practical Applications: SVMs have found uses in several domains [46]:

Biological Sciences: To make animal proteins such as eggs and meat, one needs to accurately classify their species.

Industrial Clustering: Support-vector clustering, a methodology often used in industrial settings.

2.4.2. SVM Working Process

The support-vector machine algorithm specifically selects a hyperplane or a group of hyperplanes in the higher or infinite dimensional space, which are helpful for classification, regression, or other tasks, such as finding outliers. Conceptually a large functional margin, which is the longest distance between the hyperplane and nearest training data point of any class, decreases the average error of the classifier on the general data [47]. Transformation fron Non Linear to Linear svm shown in Figure 3.

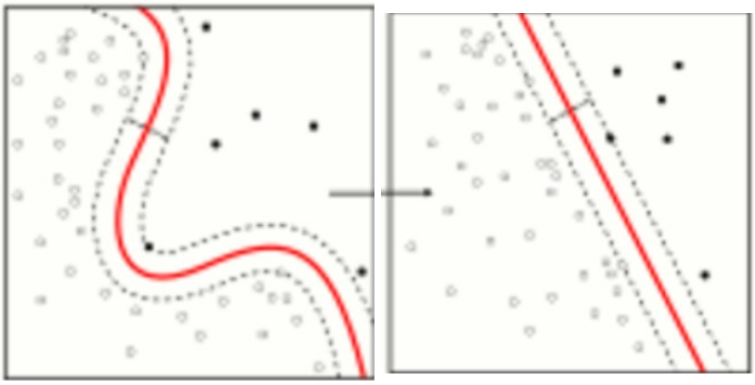


Figure 3 Transforming from Non Linear to Linear[125]

Linear Separation Challenge: Approximately in many cases, the problem is already inside finite-dimensional space. But the kind of discrimination on those sets may not be linear per se in such space. In order to cure this problem researchers proposed that instead of mapping original space on plane space higher dimensional space is mapped. Mapping to Higher Dimensions: The goal is to move data into a higher-dimensional space where it is easier to scrutinize the distances between objects. The computational burden should always remain manageable, and therefore the connecting links are specified with particular care. And use such mappings for the accurate calculation of dot products between input vectors.

Kernel Functions: SVMs are using the kernel functions to define these mappings. A kernel function is used to measure the similarity of the points coming from the original space in the new transformed space. Due to the fact that dot products were mapped out with the help of Kernel function, support vectors machine can be looked at as a nonlinear mapping of data into higher dimensions.

Hyperplanes in Higher-Dimensional Space: In the bigger-dimension space, hyperplanes occur as sets of the points. They are linked with each another through a constant number product with a vector in that frame. These hyperplanes accelerate with the help of vectors which are made up of linear combinations of pictures of feature vectors from their original data.

Relative Nearness Measurement: A possible test point lies at proximity to data points from various closely related data sets. The aggregated computation of the evaluation kernels is calculated to find the relative amount of closeness of the test point to the data points. The result in this is the spread of decision even on complicated non-convex sets when the original domain is in the transformed domain.

Linear SVMs aim to find an uppermost hyperplane in an N-dimensional space which divides the two classes of data. Hyperplane, which is referred to as a hyperplane, is a plane that divides input or output points into two separate classes thus creating a feature space.

Hyperplane and Margin: Consider two independent variables: that such a system should be implemented. Here, added label with blue and red means representing class labels. The purpose of this is to determine the maximum straight line (which may be also a hyperplane) that has data points on both sides. The main idea is to promote the spreading of the subsequent points of various classes into the space among them.

Maximum-Margin Hyperplane (Hard Margin):The right hyperplane in a sense, ensures that the gap (margin) between the classes is the maximum. It just stipulates that classes are separated in the areas they occupy as much as possible. By example, SVMs achieve robustness by ignoring any inner errors and by emphasizing the overall separatrix.

Handling Outliers: SVMs, like the rest of the strategies built on the basis of the statistics methods, are flexible to outliers. It finds the hyperplane that best balances between splitting and adaptive co-evolution. Linear SVM Classifier (12), where *y* is the output of the classifier. It's typically a binary value, meaning it can take only two possible values, often denoted as 0 and 1 , *y* is represents the weight vector, which contains the weights assigned to each feature in the input vector *x, x* represents the input feature vector.

$$\dot{y} =_{0:w^T x+b<0}^{1:w^T x+b\geq 0} \tag{12}$$

Hard margin linear SVM classifier (13),where *w* represents the weight vector, w^t denotes the transpose of the weight vector.

$$\underline{min}\frac{1}{2}w^Tw = \min\frac{1}{2}\left|\left|w\right|\right|^2 \tag{13}$$

Linear SVMs and Their Limitations: A truly first hyperplane method, which produced a linear classifier, known as the maximum- margin method, was introduced by Vapnik in 1963. However, information extraction from real life is oversimplified as it does not abide to the linear separability. Matching with the accuracy of a linear SVM is stated inability to deal with the complication of nonlinear decision boundaries.In the year of 1992, Bernhard Boser, Isabelle Guyon and Vladimir Vapnik introduced a totally novel concept that started the revolution. The authors utilized the “kernel approach” that Aizerman et al. suggested previously and then developed max-margin hyperplanes based on that. The end product, get the identical technique like before, but the vector component is replaced with nonlinear kernel functions.

Kernel Trick and Transformed Feature Space: The kernel procedure SVMs to fit hyperplanes of greatest distance from the sample space. Such way is the most likely to be complex dynamics and even multidimensional. Still the classifier always stays a hyperplane in the maximised space but becomes nonlinear in the primary input space.

Generalization Error and Sample Size: While SVM is operated in higher-dimensional feature space, its generalization error is beginning to increase. But as long as it gets enough data, the computer algorithm still delivers. Non-linear SVMs ranged from complicated to powerful. Kernels:

$$k = \left(x_i, x_j\right) = \left(x_i \cdot x_j\right)^d \tag{14}$$

$$k = \left(x_i, x_j\right) = \left(x_i \cdot x_j + 1\right)^d \tag{15}$$

$$k = \left(x_i, x_j\right) = \exp\left(-y \parallel x_i - x_j \parallel^2\right) \tag{16}$$

$$k = \left(x_i, x_j\right) = \tanh\left(kx_i \cdot x_j + c\right) \tag{17}$$

Polynomial Kernel (14): where: *d* represents the degree of the polynomial (a positive integer).signifies the dot product of the feature vectors.

Shifted Polynomial Kernel (15): Similar to the polynomial kernel, the shifted polynomial kernel also acts on dot products of feature vectors. |Gaussian (RBF) Kernel (16): where: *y* is a scaling factor (typically set to -1 or 1). |x_i - x_j| indicates the Euclidean distance between the feature vectors. Hyperbolic Tangent (Tanh) Kernel (17): where: *k* and *c* are parameters that influence the form of the kernel.

2.4.3. Advantages of Support Vector Machine

- SVM is usually a good tool to when there is a tight grouping of classes or much overlapping between the classes.
- SVM method can operate more efficiently in spaces with large dimensions.
- SVM has been found in the situations where dimensions are greater than the number of samples.
- SVM has some aspects of memory efficiency since it needs to consider only a subset of all the support vectors.

2.4.4. Disadvantages of Support Vector Machine

- SVM method is excellent for moderate sized data; however, it becomes an inefficient method for extremely large data sets.
- SVM badly fails when the data set is flooded by noise with the spillage of target classes.
- Under such conditions when the number of features for each data sample outweighs the number of training data samples, the SVM does not ‘work its magic’.
- Because the Support Vector Machine analyzes samples by which way data points are located, above and below the classifying line, the point of view has no probabilistic argument the classification.

2.5. Logistic Regression

2.5.1. History of Logistic Regression

The logistic regression model or the Logit model that stemmed from the natural situation at the beginning of the twentieth century. The idea of the log-ODDs was introduced in 1944 by Joseph Berkson [8], explaining why the likelihood of a specific event taking place cannot be equal to 1. The econometric model which considered employment as the major economic indicator gained wide dissemination because of its adaptability across different fields:

- Unobserved Latent Variable Approach: The first technique is “logistic regression” which presupposes an unseen or latent variable as a causal factor that is responsible for paying individual’s some observations. Take for example the circumvention of a person as he or she reflects about to take a chance at the workplace. The movement applied here stands on the form of comparison of an unimpaired reservation wage with the present market rate. They will turn up in the market if the pay from the job exceeds their reservation wage. Otherwise, their work force participation will drop [8][9].
- Probability Model: The second method is based on y logistic regression as a probability model for the dependent variable. Essentially, it is a system where logarithms (informally speaking the log) of the odds of an event depending on one or numerous independent factors are used as weights. The transition from the linear combination to probability uses the logistic function that transforms the real number from the linear equation to values between 0 and 1. Thus the terms of regression being "logistic regression" is employed [10].
- In Discrete Choice Model: For the last method to comprehend logistic regression, one has to understand logistic regression through the lens of random utility theory or discrete choice models. In such scenarios, the models have cases when people encounter different choices (for instance, a choice of different option as item, position or means of transportation). With logistic regression, we devise an algorithm for prognosing the chance of opting out of some alternatives upon these explanatory factors. [11]

Applications and Extensions: Binary Logistic Regression [12]: The most typical use case comprises a single binary dependent variable (coded as 0 or 1) and one or more independent variables (either binary or continuous). It has been frequently applied since the 1970s [12] for modeling probabilities connected to binary outcomes, such as forecasting team victories, patient health, and more.

Multinominal Logistic Regression [13]When dealing with categorical variables with more than two potential values (e.g., picture categorization into numerous classes), the binary logistic regression may be expanded to accommodate multiple categories.

Ordinal Logistic Regression: If the categories are ordered (e.g., degrees of satisfaction), ordinal logistic regression offers an appropriate extension [14].

2.5.2. Logistic Regression Working Process

Machine learning involves forecasting both of values types and their qualities. The "backward kind" is commonly referred to as regression, while the latter is a situation where the way to determine the input features is on continuous variables, and the type of prediction is a numerical value. Contrary to a situation face with a quantitative forecasting (routing), have a classification challenge. Classification problems cover such case as an assurance of target user's preference for an item and also a decision on whether the user will get to see the internet advertisement.[48].

While some algorithms deterministically classify the data into categories in this fashion, not all algorithms immediately distinguish themselves to this binary dichotomy. It is a diverse model that should be studied closely. Regression Roots with a Twist: Logistic regression belongs to the regression-based family with the one exception that is its deviation from the linear regression scheme. In contrast with linear regression that only considers the continuous attributes to estimate, logistic regression efficiently works with both continuous and discrete independent features [49]. Along with a numerical form as a result of its operation the provided class is of qualitative character e.g. something like "Yes/No" or "Customer/Non-customer".

Analyzing Relationships and Assigning Probabilities: In the case of practice, logistic regression calculates the correlations between certain variables. For the same, SNN adopts the sigmoid function (aka the logistic function) to check the probability of certain event happens. Through this procedure the analysis of the numerical results helps to strengthen the possibilities of results from less to 1.0. The possibility of an event happening is the basis for the probability," probability" is the term for the probability.For example, have 50% as our cut-off point (which is a threshold). One a threshold is reached (e.g., it is x out of y), then replace the example into one group (e.g., Group A); otherwise, it belongs to the other group (Group B). The figure of Logistic Regression as shown in Figure 4.

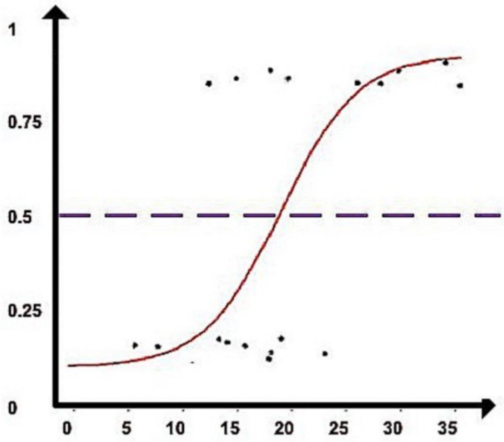


Figure 4 Logistic regression[126]

A hyperplane is used as a decision line to divide two categories (as far as practicable) after data points have been allocated to a class using the Sigmoid function. The class of future data points may then be predicted using the decision boundary.

A logistic function, which goes by the name logistic regression, is used by the Logistic Regression to estimate the value of the observations and their corresponding probabilities. The sigmoid function refers to a curve that plots a real number to the range of 0 and 1, and the shape of the curve is referred to as S [50].

Besides, the estimate of the sigmoid function (obligatory probability) is greater than the threshold of the graph if the model suggests it is of that class. If the probability is less than the threshold point set, the model is predicted the instance is not subscribed to the class.

The sigmoid function is referred to as an activation function for logistic regression and is described as (18):

$$f(x) = \frac{1}{1+e^{-x}} \tag{18}$$

e = base of natural logarithms, value = numerical value one wishes to transform

Logistic Regression sigmoid function (19):

$$y = \frac{e^{(b_o+b_i X)}}{1+e^{(b_o+b_i X)}} \tag{19}$$

Where, x = input value, y = predicted output, b_0 = bias or intercept term, b_i = coefficient for input (x)

2.5.3. Advantages of Logistic Regression

- Linear Separability and Performance: In the situation when the information is separable on a line, logistic regression does a good job. The point of dividing two classes can be roughly presented by a line whose slope is a linear function. Thus logistic regression is appropriate. Since it works simply and good, it can utilize in particular cases where the connection between character and achievements require a straight line to conclude.
- Resistance to Overfitting: The logistic has less possibility of a fallacy like "the more connected model" (such as decision trees and neural networks) becoming even better than what it was intended to do. Even though overfitting becomes much less of a problem in high-dimensional samples, it still can take place.
- Interpretable Coefficients: The main views of logistic regression is to explain how much contribution has each trend for accountability to the final decision making. The bivariable correlation between each trait on the one hand and the result on the other hand is depicted by the graphs Putting of the factor with bigger magnitude the influence on the prediction will be bigger. In general, small factors mean small error.
- Convenience of Use and Understanding: Logistic regression is easy to use, especially in the light of the models that have a lot of differences. The categories define all the areas that can be affected by certain decision. Interpretability is a key advantage: this allows you to notice easily how the effect of all traits is put all together so the expected chance can be calculated. Furthermore, the model's result (probabilities) bees parsed into classifications that can be use for decision-making.

2.5.4. Disadvantages of Logistic Regression

- Assumption of Linearity: Apart from the straight-line assumption about the dependent variable and the independent factors, logistic regression is one of the critical constraints. The line between real life data and ideal sentences is rather blurred. Furthermore, majority of the datasets present intricate activities which cannot be simply fitted with linear modeling. In the event that the underlying bond is intricate, logistic regression could possibly give the wrong judgment as a consequence.

- Observations and Features: A further aspect that take attention is the ratio between the number of variables (samples) and number of traits (predictors).In case the data is really a small fraction of the number of features, logistic regression may not come up with a very accurate answer. Gen these instances, the model will quite possible overfit the model and will fit the noise in the data instead of the underlying trend.
- Discrete Predictions Only: Logistic Regression is geared to classifying incoming data into either an event or non-event. The consequence of this is an estimate which has a certain level of chance, which after thresholding becomes the basis of binary decisions (e.g. “Yes/No” or “Customer/Non-customer”). As a result of that, logistic regression is totally consumed by a specific number set (0 or 1 most likely). Brief characteristic of this limitation causes shortcomings when using data which is constantly updated. As an example, the precision of the estimated numeric values such as (e.g., regressing of house prices) is much passed the limit of logistic regression.

2.6. Naive Bayes

2.6.1. History of Naïve Bayes

Bayes’ Theorem [15]: Naive Bayes’ depend on the foundation of Bayes theorem which is named after Thomas Bayes (1702–1761) who was a reverend by religion. Bayes’ formula is the type of tool to revise our standpoint about an event after a stage of gaining new information. Thomas Bayes: The English mathematician and Presbyterian minister, Thomas Bayes, is the next prestigious mathematician to be studied in this article. He was into learning how to find a distribution for a chance distribution. He worked with the binomial distribution.

Posthumous Publication [16]: During Bessesses’ death, his colleague Richard Price have edited and distance his research at the beginning of the 1763 year. Rather than unveiling the paradigms of Statistical Inference, the featured article introduced derivations of the famous rule of probability named after Thomas Bayes.

The Naive Bayes Classifier: Naive Independence Assumptions: The name “naive” in naive Bayes testifies to the fact that making strong assumption of independency in the classifier. Certain of the assumptions are meant to simplify the model for the purpose of representation but may not be realistic in the real-world situations.

Classification: Naive Bayes is a statistical classifier classification. It implies the case is more likely out of a collection comprising of a single class because of its distinctive features.

Sigmoid Function [17]: The meaning of naive Bayes is a probabilistic output. Applying the Sigmoid function, there is a transition from 0 to 1.0 of the model input ordinal values. This step of the probability is an opportunity to evaluate the chances that it will happen.

Historical Debates: Stephen Stigler's Suggestion in 1983 [18], Stephen M. Stigler indicated that the method of disclosing probability, known nowadays as Bayes' theorem, was invented by Nicholas Saunderson about 60 years before Bayes. Saunderson was an unseeing whom worked on probability theory, not accounting for the misbehavior of the dice.

Edwards' Dispute [18]: Although Stigler did not quote it, A.W.F. Edwards re-read Hartley's Works (1749) and came upon a clear reference to Bayesian argument [51].

2.6.2. Naive Bayes working process

The Naive Bayes classifier is a classification method which is directly derived from Bayes' Theroem. This misconception can lead to the wrong assumption that underlying all the traits linked to the goal value there is total independence.

Probabilistic Decision Making: Bayes theorem calculates probabilities of each class based on quantitative characteristics of given data. Following that, the expected number of days in class to complete the course will be selected among the classes with the highest probability [52].

Applications and Strengths: Naive Bayes has been widely applied for multifarious objectives and it is the very reason for this classification algorithm to stand out in the issues of natural language processing (NLP). Its simplicity, quickness and competency to deal with the vast set of data make it a solution for some problems.

Bayes' Theorem: Bayes Theorem considers probability related to the establishment of chances of occurrence of a given event based on the past experience of factors related to the same event in the past. It handles the pre-existing views (prior probability) along and with new data (likelihood) in order to make us able to understand the case more clearly (posterior probability) [53].

Naive Bayes classifier assumes that the traits uses to predict the target are separate and do not affect each other. While in real-life data, traits rely on each other in identifying the target, but this is ignored by the Naive Bayes algorithm. Though the independence claim is never right in real-world statistics, but often works well in reality. so that it is called "Naive" [54]. Formula of Naïve Bayes step by step (19), Where: y represents the class labe,X represents the feature vector (e.g., (x_1, x_2, x_n)),(*P(X|y))* is the likelihood of the features given the class,*(P(y))* is the prior probability of the class,*(P(X))* is the evidence (total probability of observing the feature vector) and acts as a normalization factor.

$$P(y\mathrm{I}X) = \frac{P(X\mathrm{I}y)\cdot P(y)}{P(X)} \tag{19}$$

Evidence(20):

$$P(y\mathrm{I}X) = P(x_1, x_2, \dots, x_n \mathrm{I}\, y) = P(x_1 \,\mathrm{I}\, x_2, \dots, x_n,\ y) \cdot P(x_2 \mathrm{I} x_3, \dots, x_n,\ y) \dots P(x_n \,\mathrm{I}\, y) \tag{20}$$

Likelhood break up (21):

$$P(XIy) = P(x_1Iy) \cdot P(x_2Iy) \dots P(x_2 \mathrm{I}\, y) \tag{21}$$

Posterior Probability(22):

$$P(y\,\mathrm{I}\,X) = \frac{P(x1Iy)\cdot P(x2Iy)\dots P(x_n \mathrm{I} y)\cdot P(y)}{P(x_1)\cdot P(x_2)\dots P(x_n)} \tag{22}$$

Result(23):

$$y\ =\ \arg\frac{max}{y}\mathrm{x}P(y)\prod_{i=1}^{n} P(x_i \,\mathrm{I}\, y) \tag{23}$$

2.6.3. Advantages of Naive Bayes

- Efficiency and Speed: Naive Bayes is quick and accurate, thus, its use of parallel processing as well as hardware-based technology for it to be appropriate for real-time applications and huge data sets. The simplicity of its learning and prediction has a convenience advantage of ease at training.
- Multi-Class Prediction: With NAIVE-Bayes, problems of multiclass classification become easily manageable. It is capable, when there are more probable classes than when you want to predict, to handle situations.
- Less Data Dependency: If the assumption (indicates a correlation) of independence that features being unrelated still holds true, the implementation of Naive Bayes method may be successful by virtue of its being data-based.It can make recommendations faster taking fewer steps, because compared to certain other models, it is more reliable.

- Categorical Input Variables: Naive Bayes face no problem when working with categorical input variables thus the attainment of high accuracy is possible. It has no difficulty dealing with discrete features (indicators: word categories, frequency of words).

2.6.4. Disadvantages of Naive Bayes

- Strong Independence Assumption:One of the most typical fallacies is a thought that certain characteristics are independent of each other in real world data. In reality, personality traits often demonstrate complicated relationships which are not that simple and refreshes this notion.
- Zero-Frequency issue: Bayesian Naives construct zero-frequency problem. If the category of a categorical variable was not detected at training time and was unseen, the model will assign this category zero probability. Once encounter this issue, use of smoothing techniques (like Laplace smoothing) prevents zero probability.
- Estimation Errors: Naive Bayes classifier can generate probabilities at any point, but they can be wrong sometimes. It is possible that this weakening of the method may lead to inaccurate predictions based on its narrow approach.

3. BENCHMARKING DATASETS

3.1. Datasets

In the field machine learning as well as Artificial intelligence (AI), the dataset means a collection of data (multiple sources) that serves as a training and evaluating tool for algorithms and models. This data serves as a basis of successful development of machine learning systems since it provides the needed input and output formats that create a learning landscape for algorithms to acquire information from [19].

Essential aspects about datasets: Structured Data [20]: By " structured data", means information that is presented is a specified one, such as that of a spreadsheet or a database table. Working the software is user friendly it comes as preformatted.

Unstructured data, similarly, consists of the data that do not have a specific structure or layout. Accordingly, examples in question will cover typographic elements, graphic representations as well as audio and video recordings. Machine learning and AI may begin with deriving unstructured data and then require further processing to arrive at the desired conclusions.

Researchers and developers can easily use the public datasets for evaluating their algorithm's performance and analyzing the AI systems. The general public could make these databases to their own use. When data is held by companies it is determined a private dataset. It remains undisseminated in the public record. Through these sources, data access is limited to certain people or groupings.

Generated Datasets: Differ from these data sets that is created for machine learning and artificial intelligence algorithms only. Their role is instrumental in the process of getting new regimes alive and thriving [71].

3.2. Key difference of classificationn and regression datasets

Classification

Objective: The classification predicts the group name or class that the data point belongs to.Target Variable: The final variable in the model of classification possesses a categorical nature; its task is to show various classes that are defined as separate labels. Let's say that the process consists of checking whether an email is spam or not.

Examples: Predicting a customer leaving (could be yes or no). Pinpointing the species classification of the plant is given as an example (iris setosa, versicolor or virginica). One critical application is identifying fraudulent transactions (identifying/fraud or not fraud) [21].

Regression

Objective: Unlike in Regression where the target is a continuous value rather that classes.

Target Variable [22]: The continuous variable (where the regression relationship is wide-ranging) in the regression function is the output variable, since it connects to real-valued quantities. In such cases as predicting house prices, temperature or rainfall as examples.

3.3. Benchmarking

Baselines datasets are indeed the foundations for what is the proper performance assessment and comparison of the classifier algorithms.

Steps for effective Benchmarking:

1) Select Diverse Datasets: The selection of varied data sets is key since it enables us to test that how leveling algorithms will work in a different context. By compilation of datasets endevouring a variety of problem settings, aim to make the models more generalized. Model overfits with particular traits of that given dataset. The variability of datasets help us assess whether our algorithms are working properly when dealing with data sets distributed differently, feature space and class imbalance [56].

2) **Preprocess Data** [23]**:** Dataset will first undertake a data preparation process, including cleaning and transforming the raw data for the training of machine learning models, which will yield the desired results. Natural language processing consists of three basic stages pre-processing in this case guarantees that the data is consistent, free from noise and ready to receive analysis. Missing data handling, feature scaling, and local encoding are the most frequent preprocessing steps at the initial stage.

3) **Split Data** [24]**:** It is of extreme necessity to split the dataset into different folders “training, validation, and test sets”. Separate subgroups to: Train the model: The model therefore take into account all this data. Use our artificial intelligence writing assistant to unlock your full potential as a writer.

4) **Tune hyperparameters:** Parameter adjustments are attempted in aiming for best performance.

5) **Assess generalization:** The output of the model is tested to check how accurately the data it did not come across before is guessed [57].

6) **Scaling Process** [25]**:** Scaling adjusts the range of features to a common scale. It helps algorithms work better by preventing features with larger scales from dominating those with smaller scales.

7) Evaluate Algorithms [26]: Do multiple training classifications (like Support Vector Machines, K-Nearest Neighbors) on the training data. Comparing the different algorithms allows us to choose the best-preforming one algorithm for our particular settings. Some algorithms may be better at one task than the other but some other algorithms could be very good at different things. Finally that is how to make a decision.

8) **Metrics:** One of the things you need to do when comparing different algorithms' performance is to use accurate evaluation metrics. Typically people work with, for example, accuracy, precision, recall and F1-score [58]: 1**)**Accuracy evaluates total correctness, 2)Precision focuses on true positive predictions, 3)Recall puts more emphasis on getting all rational affirmances, 4)F1-score balances precision and recall.

Except typical metrics, there are also rare metrics which are aslo used in practical part of this thesis:

Out-of-Distribution Detector [27]: An OOD detector system makes the decision of whether a given input belongs to an already known distribution or falls into a new, unseen one. Purpose: OOD detection has to do with the reliance, safety, and generalization of the system since whatever is not recognized is most likely an outlier. Through identification of samples that have significant dissimilarity from the training data, our algeria paper checking system wouldn't be prone to errors on unknown samples.

Anomaly Detection Score [28]: Anomaly detection scores assess how unlikely, out-of-place, or uncommon a data point is within a data set. Purpose: Classifying irregularities, or outliers, is vital in such applications like scam identification, fault detection, and quality checking.

Transferability [29]: Transferability means to what extent model is able to transfer information it learned on one task to another semblable task. Purpose: Knowing the transferability, in turn, makes more effective use of pre-trained models and fine-tuning techniques for faster work with the better speed and quality.

Prediction Interval [30]: Average prediction interval is an area showing the range of expected future observations. Purpose: In contrast with point estimates (for instance, mean), range estimates take into account aleatoric and epistemic factors and are useful in making decisions (adapted).

Average Brier Score Loss [31]: The Brier score measures the statistics of calibration Purpose: The smaller Brier score implies that the model is effectively calibrated that assist to identify the probability of correct prediction for uncertainty calculation.

Adversial Distance: Adversarial distance measures the degree to which the model's forecast amplifies after exposure to negative shocks. Purpose: Being immune to attacks by unwelcomed threatens plays an imperative role in security and dependability.

Noise Evaluation [32]: Measurements in noise evaluation is used to examine how the evaluation of the noise (random errors) affect models' performance. Purpose: In addition to noise resistance, strong models are also built to function even in the most difficult working conditions.

Mutual Information [33]: The factual relationship between two random variables is captured by mutual information a measure. Purpose: That embraces feature selection, knowledge of relationships and knowledge gain

II. ANALYSIS

4 ANALYZES FOR PREPARED BENCHMARK TEST

In experiment used 100,200 and 300 amount of data for train and test 3 datastets with classification and regression by appearing on experiment of 3 different sources.

4.1. Select Diverse Datasets

Digits Dataset [34]**:** The digit dataset comprises of labelled 1797 grayscale 8×8 images of handwritten digits. **Data Source:** The dataset contains digit images that have been written by ground:1) Using it in our analyzes: The load_digits function from the scikit-learn module is used to load a dataset of handwritten digits into the variable data,2) Choosing amount (300) [35] of data which will train and text: Create a pandas DataFrame named "digits" that contains the pixel data of handwritten digit pictures from the data dictionaries. Then, slice the top 300 rows of "digits" to create a new DataFrame named "digits_300" that only includes these entries.

Wine Dataset [36]**:** As a representative of multi-class classification datasets, Wine dataset is a classic. Data Source: This tabulation will contain the outcome of the wine chemical analysis with a specific place of the production:1) Using it in our analyzes: The load_wine function from the scikit-learn module is used to load a dataset of handwritten digits into the variable data, 2) Choosing amount (100) [37] of data which will train and text : Create a pandas DataFrame named "digits" that contains the pixel data of handwritten digit pictures from the data dictionaries. Then, slice the top 100 rows of "digits" to create a new DataFrame named "wine_100" that only includes these *entries.*

Diabetes Dataset [38]**:** The Diabetes dataset, known as a regression dataset, is famously used in diabetes model determination and validation. Here are the key details:Here are the key details: Data Source: The database is NIH-maintained and can be found at the National Institute of Diabetes & Digestive and Kidney Diseases website: 1) Using it in our analyzes: The load_diabetes function from the scikit-learn module is used to load a dataset of handwritten digits into the variable data, 2) Choosing amount (200) [39] of data which we will train and text : Create a pandas DataFrame named "digits" that contains the pixel data of handwritten digit pictures from the data dictionaries. Then, slice the top 200 rows of "digits" to create a new DataFrame named "diabetes_200" that only includes these entries.

4.2. Preprocess Data

Preprocess for digits, wine and diabetes datasets: **Step 1:** x is allocated the values of all rows and all columns save the final column, thus extracting the features (pixel values of the pictures), **Step 2:** y is assigned the values of the last column of all rows, typically containing the target labels or outputs (though in the context provided, this might be an error unless the last column indeed represents labels, as the digits DataFrame originally contained only pixel data from data['data'] and no target data). **Step 3:** dataset is allocated the values of x, which are the characteristics without the target. **Step 4:** x and y are now numpy arrays with the relevant feature and (hopefully) goal values from the digits DataFrame. In both classification dataset have function dataset = x, due to the fact for function in rare metrics which for ***noise evalution.***

4.3. Split Data

For all three datasets are used same way for splitting: divides the arrays x and y into training and testing sections using the train_test_split function from the scikit-learn library: 1) xtrain and ytrain is the training set where feature and targets are subsets of them, 2) The training sets associated with the features are sometimes called xtest and the targets are named ytest. The test_size=0. The second statement is that 20% of the data should be absconded for testing set, 3) The function is called with a seed value 1 which ensures that the split argument is repeatable; every-time the code is run the same random split will occur.

4.4. Scaling Process

For all three datasets i used same way for scaling MinMaxScaler: 1) scales the training features using the fitted scaler object, 2) apply the same scaling to the test features (xtest) without refitting the scaler.

This ensures that both the training and test data are scaled in the same way, maintaining consistency in the feature range across both sets. MinMaxScaler : Scales features to a specified range (e.g., [0, 1])

4.5. Evaluate Algorithms for Classification datasets

- Naive Bayes: trains a Gaussian Naive Bayes classifier on the training data, predicts labels for the test features, then outputs both the predicted labels and the real labels for assessment.
- KNN: trains a KNN classifier on the training data, predicts labels for the test features, then outputs both the predicted labels and the real labels for assessment.
- DecisionTree: trains a Decision Tree classifier on the training data, predicts labels for the test features , then outputs both the predicted labels and the real labels for assessment.
- RandomForest: trains a RandomForest classifier on the training data, predicts labels for the test features, then outputs both the predicted labels and the real labels for assessment.
- SVM: trains a SVC classifier on the training data, predicts labels for the test features (xtest), then outputs both the predicted labels and the real labels for assessment.
- LogisticRegression: trains a LogisiticRegression classifier on the training data, predicts labels for the test features, then outputs both the predicted labels and the real labels for assessment.

4.6. Standard Metrics and Analyzes

For checking accuracy, precision, recall, f1-score same way for both classification datasets: calculates the confusion matrix that is a table that contains the counts of true positive, true negative, false positive, and false negative predictions. Sums up a textual report of the precision, recall, F1-score, and support that is calculated for each class and gives the average values across the classes.

For checking Mean Squared Error(MSE),Mean Absolute Error(MAE) and R2 Squared(R2) used that functions for diabetes regression dataset: 1) Compute its MSE between the actual target values and the prediction results , 2) The MAE was computed for the true target values and the fitted values 3) The R-squared (R-squared), computed as the bivariate correlation coefficient between the true target values and the predicted values and which indicates, in proportion, the variance of the dependent variable that is predictable from the independent variables is elaborated.

4.7. Digits dataset benchmarking by metrics

Table 1 Digits datasets benchmarking by standart metrics

Classification Methods	f-1 score	accuracy	precison	recall
KNN	0.07	0.93	0.06	0.07
Decision Tree	0.10	0.93	0.09	0.10
Random forest	0.11	0.93	0.10	0.13
SVM	0.09	0.93	0.08	0.09
Logistic Regression	0.12	0.94	0.11	0.15
Naive Bayes	0.08	0.41	0.13	0.10

As shown in Table 1: 1)The KNN has the lowest scores under all above measures. A low score of F1 comes across as a statistical representation of the high amount of error rates which arise from the imbalance between precision and recall. 2)Decision Tree proves to display a smaller enhancement compared to KNN. 3)Random Forest works in the same way as Decision Tree; it only considers the learning process is at a rather simplistic level. Recall is slightly better. 4)SVM is closely comparable to Decision Tree and Random Forest models in terms of accuracy. 5)Logistic regression approach model performs the best. F1-score scores higher and recall remembers better. 6)Naive Bayes is great in terms of precision but tends to have looked lower on the F1-Score. The cooled F1-score indicates the problem expands the precision and recall balance.

By appearing on that table benchmarking of digits dataset is;higher F1 score indicates better overall performance and Logistic Regression has highest and Naive Bayes has lowest score . And all methods have similar accuracy (93%), expect Naive Bayes (41%) and Logistic Regression (94%). By appearing of all used standard metrics, Logistic Regression most suitable for digits dataset and Naive Bayes is most not suitable for that dataset.

4.8. Wine dataset benchmarking by metrics

Table 2 Wine datasets benchmarking by standard metrics

Classification Methods	f-1 score	Accuracy	Precison	recall
KNN	0.98	0.97	0.98	0.97
Decision Tree	0.89	0.89	0.91	0.88

Random forest	0.98	0.97	0.98	0.97
SVM	0.97	0.97	0.98	0.96
Logistic Regression	0.98	0.97	0.98	0.97
Naive Bayes	1.00	1.00	1.00	1.00

As shown in Table 2: 1) KNN has strong performances on all scores of assessment model. A high F1 score implies that the classifier exhibits good behavior of precision and recall in the same time. 2) Decision tree algorithm performs better when compared to K-nearest neighbors algorithm. However, KNN performs better when it comes to F1-score and accuracy. The accuracy is rather good, but the recall rate is sadly poor. 3) Random Forest appears to be adequate, maintaining a score similar to the KNN algorithm and with a good balance of precision and recall. 4) SVM everything is done on high precision, and such system is characterized by the good general disposition.The recall value had a relatively inferior result to the random forest and KNN algorithm. 5) Logistic Regression also performs well compared to both KNN and Random Forest models. Succeeding to get the course of F1-score not alongside the no balanced trade-off of precision-recall. 6) Naive Bayes is an ideal method because of all standard metrics, it reports zero errors. In the meantime, the accuracy of a model should not be overestimated unless the regularization is incorporated to avoid overfitting and the unique features of the data are taken into consideration.

By appearing on that Table 2, Naive Bayes is absolutely winner and perfect performance with testing of all 4 metrics. Other algorithms have a bit same scores around 0.97-0.98 on the middle expect Decision Tree which has around 0.9 only. Naive Bayes is most suitable for wine dataset , cause it uses probability function and predict something is easier with probability when you have only 3 types of wine. At the same time Decision did not made high performance like other algorithms.

4.9. Diabetes dataset benchmarking by metrics

Table 3 Diabetes datasets benchmarking by standart metrics

Regression Methods	MAE	MSE	R2 squared
KNN	47.04	3934.55	0.26
Decision Tree	63.20	6476.64	-0.22
Random forest	46.93	3791.25	0.29
SVR	55.64	4517	0.15
Logistic Regression	41.97	2992.58	0.44

As shown in Table 3 :1) The KNN model has quality ratings such as MAE and MSE. The high R2 score of 0.26 implies that KNN gives only 26% . 2) Decision Tree performs better than KNN from the aspect of MAE and MSE, with MAE and MSE of 0.465 and 0.501. The R-squared coefficient (R2) is negative (–0.22) indicating that, using this model, performance is even lower than just a mean prediction. 3) Random Forest was better than KNN and Decision Tree based on MAE and MSE for each target variable. So, the fact that R2 score, which equals 0.29, is higher compared to KNN and Decision Tree, shows a better action at a certain level, but still it is not purely predictive. 4) Compared to SVR, Random Forest has greater MAE and MSE values. The R-Squared of 0.15 credits the model with limited explanatory power. 5) The Logistic Regression is the best solution of the methods executed. MAE and MSE of R2 are the lowest, whereas R2 (0.44) is indicating more fit compare to others.

From the provided indicators, conclude that in the cases Logistic Regression shows better performance than the others. Because of its lower error level and better fit, it has wide application.

4.10. Rare metrics and analyzes

For all methods , same functions to get our rare metrics for analyzes .

Below shown results of them with Naive Bayes with digits dataset ,as an example ,to be able to imagine how our table ,will look like

- **Out-of-Distribution Detector:** 1)IsolationForest Generates a Contamination Forest Isolating Model with an alternative value set to zero. This first assumption referred to the anticipated proportion of outliers in the data and thus a way of preventing false conclusions to be drawn from the data, 2)Run the Isolation Forest algorithm on the input data based on the input features *x*, 3)Calculation of the anomaly score of each sample of the test set xtest, then, 4) OOD fraction Determines the proportion of samples being the outlier (having the score lower than 0).

- **Anomaly detection score:** IsolationForest Entrust the model building of the Isolation Forest along with the contamination parameter set to 0 this states the expected rate of anomalies in the data, and a random seed for the reproducibility of the whole result. Fits the Isolation Forest classifier into this set of x. It predicts the labels of each sample as normal or abnormal in total the number of the anomalies which is elaborated by the count of the -1 labels, representing the outliers.
- **Tranferability:** 1)Stores the cosine similarity between the logits and the correct labels by computing the cosine similarity heuristic, because each output vector from the model serves as a single sample, 2) counts the mean Cosine distance between all the occurrences to determine the shiftability across the classifier in this task.
- **Prediction Interval:** 1)Sets the number of times global and local models should be generated in order to apply the Bootstrap method, 2) Prepare a space for the collection and analysis of the bootstrap predictions, 3) The all process of loop is carried out by bootstrap sampling, which is done by randomly drawing indices with replacement and then aggregating the predictions made on those samples, 4) lower_bound and upper_bound are computed using np. **main**. tal value up to the 2.5th and 97.5th percentile of the bootstrapped prediction predictions, respectively, lower and upper bands of the prediction interval.

Lower Bound: If the bottom bound of the confidence interval is small, often it relates to the fact that the performance or the metric that is being benchmarked is at the lower end of the spectrum. This may indicate that the system or process is behind the expectations or standards set. This could be due to factors like environmental spectacles, charm of the rural areas, or the stimulation of a new experience into her consciousness.Alternatively, if the lower limit is higher, it indicates that the safety is really good or even above the average.

Upper Bound: It implies that the upper level limit of the prediction interval is very low that makes the performance or metric being measured at the higher end of the scaled range. It means that the outcome is better than expected or competitive similar indexes. The situation of upper bound can be taken as indicative for lower of the performance or later achievement of the expected results.

- **Average Brier Score Loss:** 1) determine the number of existing classes in the one-hot encoded vector representation of both ypred and ytest, 2) ypred_prob does it by converting ypred into probability distributions as one-hot encoded vectors, each element representing the prediction, 3) ytest_binary is the function responsible for converting ytest to a binary type using the format of one-hot encoding, 4) Brierscore loss is named when a loss of node across all classes is computed. They are measured using Brier_score_loss which is the mean squared difference between the predicted probabilities and the outcomes, 5) mean_brier_score empirically analyzes the average number of classes, which can be considered as Brier losses.
- **Adversarial Distance:** 1) found Mean Squred Error(MSE) by using the same which was used in Standard Metric and Analyzes, 2) $\sqrt{mse} = adversial\ distance$
- **Noise evaluation:** No function for noise evaluation in library and after checking formula ,it was written handly : The purpose is to carry out the data arrangement and instead of the mean, the first quartile (q1) and the third quartile (q3) are calculated with np. percentile. And finally, for the calculation of the : IQR $= Q_3 - Q_1$ Ireturns the IQR (noise evaluation) of the given dataset.
- **Mutual Information:**which is the mutual information between actual labels and predicted labels.

4.11. Digits dataset benchmarking by rare metrics

Table 4 Digits datasets benchmarking by rare metrics

Classification Methods	ODD	Anomaly detection	Transferability	Prediction interval 95% (lower bound/upper bound)	AVG Brier score	Adversial Distance	Noise evaluation	Mutual Information
KNN	1.000	180	0.0033	0/5.6	0.0088	1.3365	50.0	0.2133
Decision Tree	1.000	180	0.0046	0/8	0.0084	0.9874	50.0	0.2696
Random Forest	1.000	180	0.0022	0/7.8	0.0081	1.1353	50.0	0.1565
SVM	1.000	180	0.0000	0/0	0.0078	2.1409	50.0	0.0000
Logistic Regression	1.000	180	0.0031	0/6.2	0.0075	1.1055	50.0	0.2208
Naive Bayes	1.000	180	0.0398	0/8.9	0.0696	2.1036	50.0	0.2198

As shown in Table 4: KNN achieves good one-day dynamics and low anomaly detection, but it is being carried out poorly / not well. The niche limit the dependability of the forecast deteriorates. Over all, KNN is a decent algorithm which is half-decent. Decision Tree is comparable to KNN in its so-called features. The larger the variance, the more uncertain it is estimated. Decision tree performs moderatly. Random Forest has many similar features with both K-NN and decision-tree estimation methods. The prediction margin indicates the probability of error. To sum up, Random Forest reasons mediumly.

SVM features outstanding ODD but still with an extremely small amount of transferable learning capabilities.The point estimate which is the prediction interval makes no sense. In short, SVM are rather average for reasons of being vulnerable in close combats. Logistic regression behaves the same way as KNN and a decision tree does. Larger forecasting interval hints of some level of the uncertainty. Overall Logistic Regression seems performs modestly but good enough to use for this task. Naive Bayes has a big discriminative ability but still has a limited generalization. The latter depicts this by wider confidence limits. High brier score signifies the fractional estimation capability which implies low probable matrix. The potential for worsening adverse conditions and increased noises cause worry. Naive Bayes gets scored bad mostly because of extreme adversarial set distance and high noise interference.

Logistic Regression ranks among the most powerful competitors on the market in accordance with various criteria including ODD, transferability and overall stability. It might be a sign that it is not only Logistic Regression gives the best approximation, but also it is often accurate and stable in many contexts.

Random forest and KNN (K-Nearest Neighbors) have a moderate performance among the other models such as logistic regression. They demonstrate the efficacy as a positive thing to say but they can not prevail over Logistic Regression in the matter of predictive accuracy and stability.

From another angle this Naive Bayes method appears to be deficient. The fact that it can only travel for rather long distances if the light intensity is alarmingly high and susceptible to noise. Therefore, it is not suitable for the task in hand. The adversarial distance signifies how sensitive the model is to small tweaks in the inputs leading to attacks and which is a broad issue of complex tasks training.

Consequently, from the evaluation Logistic Regression could be said to be superior model owing to the fact that it is robust to complexities, whereas Naive Bayes depicts a weakness to errors in training data.

4.12. Wine dataset benchmarking by rare metrics

Table 5 Wine datasets benchmarking by rare metrics

Classification Methods	ODD	Anomaly detection	Transferability	Prediction interval 95% (lower bound/upper bound)	AVG Brier score	Adversial Distance	Noise evaluation	Mutual Information
KNN	1.000	18	0.3565	0/2	0.0185	0.1667	504.5775	0.9796
Decision Tree	1.000	18	0.3735	0/2	0.0740	0.3333	504.5775	0.7201
Random Forest	1.000	18	0.3565	0/2	0.0185	0.1667	504.5775	0.9796
SVM	1.000	18	0.3735	0/2	0.0185	0.1667	504.5775	0.9816
Logistic Regression	1.000	18	0.3565	0/2	0.0185	0.1667	504.5775	0.9796
Naive Bayes	1.000	18	0.3735	0/2	0.0	0.000	504.5775	1.0817

As shown in Table 5 KNN has a good origin dependence and weak transferability, its anomaly detection ability is very poor. The short dispersion of the estimated range of prediction proves their dependence upon predictions. This means the subjects have a strong hearing capacity that the inner ear can sense a high level of sound. Overall, KNN performs moderately. Decision Tree is also pertinent with K nearest neighbors. Interval of prediction suggests obvious uncertainty. Testing the extent of faking and the extent of noising. To summarize, the graft of Decision Tree obtains mediocre marks.

The Random Forest feature is also similar or comparable to that of KNN. It is due to this that forecasting is of a high degree of accuracy.The ability of noise evaluation to reflect its sensitivity brings noise to its light.While Random Forest has medium quality of performance, the Ensemble method with gradient boosting techniques outperformed others such as Support Vector Machines, Logistic Regression, and Neural Networks.SVM posses similar attribute as KNN and

Random Forest in many respects.The reliability of these forecast systems is largely dependent on tight prediction intervals. The Dosimeter in fact registers the sound level and illustrates the human sensitivity to sound. Overall, SVM performs moderately.

Logistic Regression as quickly as KNN and Random Forest and gives similar results as they do. Reliability of forecasts increases by a mechanism of close prediction intervals. Subjective noice-evaluation studies, which constitute the basis for assessing people's quality of life, are highly sensitive to noise.

Summing up, Logistic Regression shows good results. Naive Bayes has good ODD, however lack of transferability can be referred to it as well. The small margin of error in the forecasting statement means that it is easy to rely or trust on the prediction. Probability of perfect prediction converts into 0 in the Brier score case. A low threshold of adversariality and many connections with other countries. While the high noise evaluation pleads for sensitivity to noise, it does not imply the exact numbers. Logistic Regression appear to have a good performance on a lot of different evaluation metric, which always outperform the other models by having better overall diagnostically accuracy, adaptability and stability. Its regular performance consistency enables it to be a perennially practical solution to many issues. Random Forest, SVM are slower like Logistic Regression between all the techniques used to classify. Although Lagging Indicators and Leading Indicators do their jobs well, they still do not surpass the Logistic Regression in terms of the prediction power and stability. Naive Bayes method is shows potential but quite a few kinks still to be revealed of its hypersensitivity to noise. It helps to deliver results under certain aspects, however, it may be noisy in the real world and therefore the results are unreliable. Therefore in order to eliminate any ifs and buts as far is safety and effectiveness this measure should be examined and possible improvement suggested.

CONCLUSION

In summary, classification schemes are one of the pillars of machine learning with no doubt about it, and they incredibly help in data analysis in addition to decision making and solving of very complex problems on the technological and social levels. Their contributions do not end with the plausible practice but also involve a crucial part for the technological development, engagement, and scientists' responsible achievement in AI within a society. While steady get a grasp of the capability of this classifying method, taking note of its ethical, social, and mental implications is also super important as this raises up the value of such tool towards empowerment and the forward movement of the society.

An overall conclusion from our attempt to benchmark the discriminator using conventional measures, Logistic Regression comes out victorious and shows better efficiency than the others on several evaluation criteria at the same time in another experiment Logistic Regression shows bettter and stable results and Naive Bayes loses too. It is a complete glaring clear picture for the Naïve Bayes Model that the model performs well on the wine dataset with a perfect fit to the unique peculiarities of the dataset. Besides that, Logistic Regression is more than just a categorization value and it is the smart choice for the diabetes dataset, thus exhibiting multifacetedness and power to operate in varied data environments.

Since after the fact that Logistic Regression revealed itself as the defining measure upon the in-depth investigation, the latter displays its dominance as the class type even in the subtle kind of evaluation frames, showing the latter's superiority even in the complicated conditions of evaluation. Though, it is hard-to-be-ignored for the fact that Naive Bayes can outperform logistic regression in certain occasions specifically when an evaluating measure is used such as the Brier score. Though the generated marginal difference of reliability among varied methods manifests their similar capacity, their performance is not as satisfactory as Logistic Regression.

In conclusion, our comprehensive assessment justifies the superiority of Logistic Regression not only for classification assignments but in regression complexity overcoming also for all the datasets and evaluation protocols. In this kind of intelligence field, Logistic Regression is a must requirement for successful intelligence which does not change. This undeniable endorsement proves its individual in machine learning and data analysis itself as the most important technique among the many others.

BIBLIOGRAPHY

[1] A novel selective naïve Bayes algorithm☆ Author links open overlay panelShenglei Chen a, Geoffrey I. Webb b, Linyuan Liu a, Xin Ma c 2020 https://www.sciencedirect.com/science/article/pii/S0950705119306185

[2] Zhu, Y.-Z., Zhang, J. et al. Comparison among four deep learning image classification algorithms in AI-based diatom test. 2022, Available:https://www.semanticscholar.org/paper/Comparison-among-Four-Deep-Learning-Image-in-Diatom-Zhu-Zhang/d5110ec322e2d73113d7c6dea9f3c6d2e21aed92

[3] Ensemble of optimal trees, random forest and random projection ensemble classification Zardad Khan1,2 · Asma Gul2,3 · Aris Perperoglou2 · Miftahuddin Miftahuddin2,4 · Osama Mahmoud2,5,6 · Werner Adler7 · Berthold Lausen2 2020 https://link.springer.com/content/pdf/10.1007/s11634-019-00364-9.pdf

[4] Wang J. & Zucker D. Solving the multiple-instance problem: A lazy learning approach. [online]. 2020, Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=08c63a3e1c883d68a3556f12ec49cafd9d98e02d

[5] Lai K. et al. Application of data mining on partial discharge part I: Predictive modeling classification. [online]. 2023, Available: https://ieeexplore.ieee.org/abstract/document/5492258

[6] Gulzat T. et al. Research on predictive model based on classification with parameters of optimization. [online]. 2021, Available: https://www.researchgate.net/profile/Naizabayeva-Lyazat/publication/348490516_Research_on_predictive_model_based_on_classification_with_parameters_of_optimization/links/603ccf6e4585158939d9ea57/Research-on-predictive-model-based-on-classification-with-parameters-of-optimization.pdf

[7] Terrin N. et al. External validity of predictive models: A comparison of logistic regression, classification trees, and neural networks. [online]. 2020, Available: https://www.sciencedirect.com/science/article/abs/pii/S0895435603001203

[8] Prairie Y. T. Evaluating the predictive power of regression models. [online]. 2019, Available: https://www.researchgate.net/profile/Yves-Prairie/publication/235554912_Evaluating_the_predictive_power_of_regression_models/links/0c960520e3db97e44a000000/Evaluating-the-predictive-power-of-regression-models.pdf

[9] Breiman L. & Friedman J. H. Predicting multivariate responses in multiple linear regression. [online]. 2020, Available: https://www.stat.berkeley.edu/users/breiman/curds-whey-all.pdf

[10] Krackhardt D. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. [online]. 2020, Available: https://www.sciencedirect.com/science/article/pii/0378873388900044/pdf?md5=6494818d44bf3cfa2b016f1a535a1c60&pid=1-s2.0-0378873388900044-main.pdf

[11] Pearce J. & Ferrier S. Evaluating the predictive performance of habitat models developed using logistic regression. [online]. 2021, Available: https://www.whoi.edu/cms/files/Ecological_Modelling_2000_Pearce_53557.pdf

[12] Liaw A. & Wiener M. Classification and regression by randomForest. [online]. 2022, Available: https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

[13] Loh W. Classification and regression trees. [online]. 2023, Available: https://imsarchives.nus.edu.sg/oldwww/Programs/014swclass/files/loh.pdf

[14] Torgo L. & Gama J. Regression by classification. [online]. 2020, Available:

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ad0353a75df9e81420a2a77b849a95db31b23b63

[15] Lou Y. et al. Intelligible models for classification and regression. [online]. 2021, Available: https://www.cs.cornell.edu/~yinlou/papers/lou-kdd12.pdf

[16] Muthukumar, V. et al. Classification vs regression in overparameterized regimes: Does the loss function matter? [online]. 2022, Available: https://www.jmlr.org/papers/volume22/20-603/20-603.pdf

[17] Pintea S. L. et al. A step towards understanding why classification helps regression. [online]. 2021, Available: https://openaccess.thecvf.com/content/ICCV2023/papers/Pintea_A_step_towards_understanding_why_classification_helps_regression_ICCV_2023_paper.pdf

[18] Zhang R. et al. Nearest neighbors meet deep neural networks for point cloud analysis. [online]. 2021, Available: https://openaccess.thecvf.com/content/WACV2023/papers/Zhang_Nearest_Neighbors_Meet_Deep_Neural_Networks_for_Point_Cloud_Analysis_WACV_2023_paper.pdf

[19] Shi X. et al. Towards faster k-nearest-neighbor machine translation. [online]. 2021, Available: https://arxiv.org/pdf/2312.07419

[20] Zardini E. et al. A quantum k-nearest neighbors algorithm based on the Euclidean distance estimation. [online]. 2022, Available: https://link.springer.com/content/pdf/10.1007/s42484-024-00155-2.pdf

[21] Xu F. et al. Why do nearest neighbor language models work? [online]. 2023, Available: https://proceedings.mlr.press/v202/xu23a/xu23a.pdf

[22] Ritzkal S. et al. K-nearest neighbor algorithm analysis for path determination in network simulation using software-defined network. [online]. 2023, Available: https://beei.org/index.php/EEI/article/download/4868/3332

[23] Suwanda R. et al. Analysis of Euclidean distance and Manhattan distance in the K-means algorithm for variations number of centroid K. [online]. 2020, Available: https://iopscience.iop.org/article/10.1088/1742-6596/1566/1/012058/pdf

[24] Hidayati N. & Hermawan A. K-nearest neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation. [online]. 2020, Available: https://www.academia.edu/download/97662081/pdf.pdf

[25] Rosa T. et al. Comparison of distance measurement methods on K-nearest neighbor algorithm for classification. [online]. 2022, Available: https://www.academia.edu/download/99960392/125939888.pdf

[26] Pathak A., & Pathak S. Study on Decision Tree and KNN Algorithm for Intrusion Detection System. [online]. 2021, Available: https://www.academia.edu/download/91101318/study-on-decision-tree-and-knn-algorithm-for-intrusion-detection-system-IJERTV9IS050303.pdf

[27] Azam Z. et al. Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree. [online]. 2022, Available: https://ieeexplore.ieee.org/iel7/6287639/6514899/10185955.pdf

[28] Custode Leonardo L. et al. Evolutionary Learning of Interpretable Decision Trees. [online]. 2023, Available: https://ieeexplore.ieee.org/iel7/6287639/6514899/10015004.pdf

[29] Xu Q. et al. Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review. [online]. 2020, Available: https://downloads.hindawi.com/archive/2023/9919269.pdf

[30] Klusowski J. M. & Tian P. M. Large Scale Prediction with Decision Trees. [online]. 2023, Available: https://www.researchgate.net/profile/Jason-Klusowski/publication/375626829_Large_Scale_Prediction_with_Decision_Trees/links/65676dd7b1398a779dc6f801/Large-Scale-Prediction-with-Decision-Trees.pdf

[31] Zarzoor A. R. et al.Intrusion detection method for Internet of Things based on the spiking neural network and decision tree method. [online]. 2022, Available: https://www.academia.edu/download/97071498/105_28257_EMr_16oct22_25apr22_10_K.pdf

[32] Labera E. et al. Shallow decision trees for explainable k-means clustering. [online]. Available: https://www.sciencedirect.com/science/article/am/pii/S003132032200718X

[33] Ma R. et al. Developing Machine Learning Algorithm Literacy with Novel Plugged and Unplugged Approaches. [online]. 2021, Available: https://www.academia.edu/download/101306612/3545945.pdf

[34] Mankar A. D. & Bhoite S. D. A Comparative Study of Recursive Partitioning Algorithms (ID3, CART, C5.0) for Classification. [online]. 2020, Available: https://irjhis.com/paper/IRJHISIC2302052.pdf

[35] Shahhosseini M. et al. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. [online]. 2018, Available: https://www.sciencedirect.com/science/article/am/pii/S2666827022000020

[36] Reynara F. J. et al. The Comparison of C4.5 and CART (Classification and Regression Tree) Algorithm in Classification of Occupation for Fresh Graduate. [online]. 2018, Available: https://www.academia.edu/download/90025682/15680.pdf

[37] Abdul Salam M. et al. The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem. [online]. 2020, Available: https://www.researchgate.net/profile/Mustafa-Abdul-Salam/publication/351312502_The_Effect_of_Different_Dimensionality_Reduction_Techniques_on_Machine_Learning_Overfitting_Problem/links/609da16792851cfdf32f3728/The-Effect-of-Different-Dimensionality-Reduction-Techniques-on-Machine-Learning-Overfitting-Problem.pdf

[38] Hu J. & Szymczak S. A review on longitudinal data analysis with random forest. [online]. 2023, Available: https://academic.oup.com/bib/article-pdf/24/2/bbad002/49559948/bbad002.pdf

[39] Avci C. et al. Comparison between random forest and support vector machine algorithms for LULC classification. [online]. 2020, Available: https://dergipark.org.tr/en/download/article-file/1943956

[40] Jin K. et al. Sustainable Digital Marketing under Big Data: An AI Random Forest Model Approach. [online]. 2022, Available: https://eprints.lse.ac.uk/121402/1/Sustainable_Digital_Marketing_under_Big_Data_An_AI_Random_Forest_Model_Approach.pdf

[41] Wali S. & Khan I. Explainable AI and Random Forest Based Reliable Intrusion Detection System. [online]. 2021, Available: https://www.techrxiv.org/doi/pdf/10.36227/techrxiv.17169080.v1

[42] Savargiv M. et al. A New Random Forest Algorithm Based on Learning Automata. [online]. 2021, Available: https://downloads.hindawi.com/archive/2021/5572781.pdf

[43] Lin Y. On the Support Vector Machine. [online]. 2023, Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=40f4e66f0f6f26ca452a160a9ff5bc3eee02f9a5

[44] Jun Z. The Development and Application of Support Vector Machine. [online]. 2020, Available: https://iopscience.iop.org/article/10.1088/1742-6596/1748/5/052006/pdf

[45] Tanveer M. et al. Comprehensive Review On Twin Support Vector Machines. [online]. 2023, Available: https://arxiv.org/pdf/2105.00336

[46] Andersen S. Forecasting Economic Downturns in The Scandinavian Countries using The Yield Curve. [online]. 2020, Available: https://oda.oslomet.no/oda-xmlui/bitstream/handle/11250/2823981/Andersen_Sondre.pdf?sequence=3

[47] Pavithra C. & Saradha M. Classification And Analysis Of Clustered Non-Linear Separable Data Set Using Support Vector Machines. [online]. 2021, Available: https://migrationletters.com/index.php/ml/article/download/7365/4796

[48] Dumitrescu E. et al. Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision-Tree Effects. [online]. 2022, Available: https://www.sciencedirect.com/science/article/am/pii/S0377221721005695

[49] Cartus A. R. et al. The Impact of Undersampling on the Predictive Performance of Logistic Regression and Machine Learning Algorithms: A Simulation Study. [online]. 2022, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7871213/

[50] Andi H. K. An Accurate Bitcoin Price Prediction using logistic regression with LSTM Machine Learning model. [online]. 2023, Available: https://www.researchgate.net/profile/Hari-Andi-2/publication/354710727_An_Accurate_Bitcoin_Price_Prediction_using_logistic_regression_with_LSTM_Machine_Learning_model/links/6187a5fe07be5f31b753b6fd/An-Accurate-Bitcoin-Price-Prediction-using-logistic-regression-with-LSTM-Machine-Learning-model.pdf?origin=journalDetail&_tp=eyJwYWdlIjoiam91cm5hbERldGFpbCJ9

[51] Vishwakarma M. & Kesswani N. A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection. [online]. 2020, Available: https://www.sciencedirect.com/science/article/pii/S2772662223000735

[52] Arumugam K. et al. Multiple disease prediction using Machine learning algorithms. [online]. 2021, Available: https://www.researchgate.net/profile/Mohd-Naved/publication/353651472_Multiple_disease_prediction_using_Machine_learning_algorithms/links/61279a342b40ec7d8bc8275c/Multiple-disease-prediction-using-Machine-learning_algorithms.pdf

[53] Vangara R. et al. Opinion Mining Classification using Naive Bayes Algorithm. [online]. 2021, Available: https://www.academia.edu/download/85635035/E2402039520.pdf

[54] Tabash M. et al. Intrusion Detection Model Using Naive Bayes and Deep Learning Technique. [online]. 2020, Available: https://iajit.org/PDF/Vol%2017,%20No.%202/17046.pdf

[55] Subarkah P. et al. Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification. [online]. 2023, Available: https://link.springer.com/content/pdf/10.1007/s11227-020-03481-x.pdf

[56] Raji I. D. et al. AI and the Everything in the Whole Wide World Benchmark. [online]. 2021, Available: https://arxiv.org/pdf/2111.15366

[57] Li Z. et al. Searching for an Effective Defender: Benchmarking Defense against Adversarial Word Substitution. [online]. 2020, Available: https://arxiv.org/pdf/2108.12777

[58] Bednárová L. et al.A Model for Streamlining Benchmarking in Sustainable Development of Industries. [online]. 2021, Available: https://www.mdpi.com/2071-1050/16/6/2587/pdf

[59] Brigato, L. et al. Image Classification with Small Datasets: Overview and Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). [offline], 2020.

[60] Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models Author links open overlay panelAlexandre Bailly a b, Corentin Blanc a b, Élie Francis a, Thierry Guillotin a, Fadi Jamal c, Béchara Wakim d, Pascal Roy b 2022 https://www.sciencedirect.com/science/article/am/pii/S0169260721005782

[61] DEWP: Deep Expansion Learning for Wind Power Forecasting WEI FAN, University of Oxford, UK YANJIE FU∗ , Arizona State University, USA SHUN ZHENG and JIANG BIAN, Microsoft Research, China YUANCHUN ZHOU, Computer Network Information Center, Chinese Academy of Sciences, China HUI XIONG∗ , Hong Kong University of Science and Technology, China 2024 https://arxiv.org/pdf/2401.00644

[62] A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets Md Tahmid Rahman Laskar 2023 https://arxiv.org/pdf/2305.18486

[63] Development of an Artificial Intelligence Model for the Classification of Gastric Carcinoma Stages Using Pathology Slides Shreya Reddy , Avneet Shaheed , Yui Seo , Rakesh Patel 2024 https://www.cureus.com/articles/238679-development-of-an-artificial-intelligence-model-for-the-classification-of-gastric-carcinoma-stages-using-pathology-slides.pdf

[64] Volume 146, October 2023, 110631 Applied Soft Computing Trust your neighbours: Handling noise in multi-objective optimisation using kNN-averaging Author links open overlay panelStefan Klikovits a 1, Cédric Ho Thanh c 1, Ahmet Cetinkaya d 1, Paolo Arcaini b

https://www.sciencedirect.com/science/article/pii/S156849462300649X

[65] Volume 32, April 2024, 100979 Measurement: Sensors A decision tree approach for enhancing real-time response in exigent healthcare unit using edge computing Author links open overlay panelEram Fatima Siddiqui a, Tasneem Ahmed b, Sandeep Kumar Nayak c

https://www.sciencedirect.com/science/article/pii/S266591742300315X

[66] Volume 182, November 2022, 107911 Microchemical Journal Evaluation of ensemble data preprocessing strategy on forensic gasoline classification using untargeted GC–MS data and classification and regression tree (CART) algorithm Author links open overlay panelMd Gezani Bin Md Ghazi a b, Loong Chuen Lee a c, Aznor Sheda Binti Samsudin d, Hukil Sino a

https://www.sciencedirect.com/science/article/abs/pii/S0026265X22007391

[67] Volume 202, September 2021, 108026 Building and Environment

Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm Author links open overlay panelLei Xiong, Ye Yao

https://www.sciencedirect.com/science/article/abs/pii/S0360132321004285

[68] Volume 61, Issue 5, May 2022, Pages 3645-3655 Alexandria Engineering Journal

Fuzzy rank cluster top k Euclidean distance and triangle based algorithm for magnetic field indoor positioning system Author links open overlay panelCaceja Elyca Anak Bundak, Mohd Amiruddin Abd Rahman, Muhammad Khalis Abdul Karim, Nurul Huda Osman

https://www.sciencedirect.com/science/article/pii/S1110016821005883

[69] 21 Volume 176, 2020, Pages 156-165 Procedia Computer Science Multidimensional Decision Tree Splits to Improve Interpretability Author links open overlay panelFrank Höppner a

https://www.sciencedirect.com/science/article/pii/S187705092031841X

[70] 19 Volume 149, May 2024, 110220 Pattern Recognition A tree-based model with branch parallel decoding for handwritten mathematical expression recognition Author links open overlay panelZhe Li a 1, Wentao Yang a 1, Hengnian Qi b, Lianwen Jin a, Yichao Huang c, Kai Ding c

https://www.sciencedirect.com/science/article/abs/pii/S0031320323009172

[71] Volume 137, January 2023, 104274Journal of Biomedical Informatics

Methodological Review Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals Author links open overlay panelKathrin Blagec a 1, Jakob Kraiger a 1, Wolfgang Frühwirt b, Matthias Samwald a

https://www.sciencedirect.com/science/article/pii/S1532046422002799

[72] Markelle Kelly, Rachel Longjohn, Kolby Nottingham,

The UCI Machine Learning Repository, https://archive.ics.uci.edu

[73] Hunt, W.E., Meagher, J.N. and Hess, R.M. (1966) Intracranial Aneurysm. A Nine-Year Study. The Ohio State Medical Journal, 62, 1168-1171.

[74] Macià N, Bernadó-Mansilla E, Orriols-Puig A, Kam Ho T. Learner excellence biased by data set selection: A case for data characterisation and artificial data sets. Pattern Recognition 2013; 46: 1054-1066.

[75] Ghiasi MM, Zendehboudi S, Mohsenipour AA. Decision tree-based diagnosis of coronary artery disease: CART model. Computer methods and programs in biomedicine 2020; 192: 105400.

[76] Nagasaka M, Miyajima C, Inoue Y, Hashiguchi S, Suzuki Y, Morishita D, Aoki H, Toriuchi K, Katayama R, Aoyama M, Hayashi H. ID3 is a novel target gene of p53 and modulates lung cancer cell metastasis. Biochemical and biophysical research communications 2024; 708: 149789.

[77] Shanthi J, Rani DGN, Rajaram S. A C4.5 decision tree classifier based floorplanning algorithm for System-on-Chip design. Microelectronics Journal 2022; 121: 105361.

[78] Suen CY, Xu Q, Lam L. Automatic recognition of handwritten data on cheques – Fact or fiction?. Pattern Recognition Letters 1999; 20: 1287-1295.

[79] Suen CY, Xu Q, Lam L. Automatic recognition of handwritten data on cheques – Fact or fiction?. Pattern Recognition Letters 1999; 20: 1287-1295.

[80] KP V, AB R, HL G, Ravi V, Krichen M. A tweet sentiment classification approach using an ensemble classifier. International Journal of Cognitive Computing in Engineering 2024; 5: 170-177.

[81] Martín-Baos JÁ, García-Ródenas R, García MLL, Rodriguez-Benitez L. PyKernelLogit: Penalised maximum likelihood estimation of Kernel Logistic Regression in Python. Software Impacts 2024; 19: 100608.

[82] Yu H, Li Z, Zhang G, Liu P. A latent class approach for driver injury severity analysis in highway single vehicle crash considering unobserved heterogeneity and temporal influence. Analytic Methods in Accident Research 2019; 24: 100110.

[83] Bera B, Saha S, Bhattacharjee S. Forest cover dynamics (1998 to 2019) and prediction of deforestation probability using binary logistic regression (BLR) model of Silabati watershed, India. Trees, Forests and People 2020; 2: 100034.

[84] Zhu W, Si W. Predicting choices of street-view images: A comparison between discrete choice models and machine learning models. Journal of Choice Modelling 2024; 50: 100470.

[85] Bacquaert G, Raude S, Alves-Fernandes V, Voldoire F, Kondo D. A standard thermodynamic-based extension of the Modified Cam-Clay soil model and its applications. European Journal of Mechanics - A/Solids 2024; 103: 105122.

[86] Jayaprakash D, Kanimozhiselvi CS. Multinomial logistic regression method for early detection of autism spectrum disorders. Measurement: Sensors 2024; 33: 101125.

[87] Jayaprakash D, Kanimozhiselvi CS. Multinomial logistic regression method for early detection of autism spectrum disorders. Measurement: Sensors 2024; 33: 101125.

[88] Jetti HV, Ferrero A, Salicone S. A modified Bayes' theorem for reliable conformity assessment in industrial metrology. Measurement 2021; 184: 109967.

[89] Hans JD. Posthumous gamete retrieval and reproduction: Would the deceased spouse consent?. Social science & medicine 2014; 119: 10-17.

[90] Hou Y, Li G, Zhang H, Wang G, Zhang H, Chen J. Affine projection algorithms based on sigmoid cost function. Signal Processing 2024; 219: 109397.

[100] LAI LWC, DAVIES SNG, CHAU KW, CHOY LHT, CHUA MH, LAM TKW. A centennial literature review (1919–2019) of research publications on land readjustment from a neo-institutional economic perspective. Land Use Policy 2022; 120: 106236.

[101] Divasón J, Pernia-Espinoza A, Martinez-de-Pison FJ. HYB-PARSIMONY: A hybrid approach combining Particle Swarm Optimization and Genetic Algorithms to find parsimonious models in high-dimensional datasets. Neurocomputing 2023; 560: 126840.

[102] Ebbers T, Takes RP, Smeele LE, Kool RB, van den Broek GB, Dirven R. The implementation of a multidisciplinary, electronic health record embedded care pathway to improve structured data recording and decrease electronic health record burden. Internati onal journal of medical informatics 2024; 184: 105344.

[103] Omelina L, Goga J, Pavlovicova J, Oravec M, Jansen B. A survey of iris datasets. Image and Vision Computing 2021; 108: 104109.

[104] Esmaeiloghli S, Lima A, Sadeghi B. Lithium exploration targeting through robust variable selection and deep anomaly detection: An integrated application of sparse principal component analysis and stacked autoencoders. Geochemistry 2024;: 126111.

[105] Tawakuli A, Havers B, Gulisano V, Kaiser D, Engel T. Survey:Time-series data preprocessing: A survey and an empirical analysis. Journal of Engineering Research 2024;.

[106] Zheng J, Chen Y, Lai Q. PPSFL: Privacy-Preserving Split Federated Learning for heterogeneous data in edge-based Internet of Things. Future Generation Computer Systems 2024; 156: 231-241.

[107] Mula C, Zybura N, Hipp T. From digitalized start-up to scale-up: Opening the black box of scaling in digitalized firms towards a scaling process framework. Technological Forecasting and Social Change 2024; 202: 123275.

[108] Xu J, Zhang C, Xie M, Zhan X, Yan L, Tao Y, Pan Z. IMVis: Visual analytics for influence maximization algorithm evaluation in hypergraphs. Visual Informatics 2024;.

[109] Koo J, Choi S, Hwang S. Generalized Outlier Exposure: Towards a trustworthy out-of-distribution detector without sacrificing accuracy. Neurocomputing 2024; 577: 127371.

[110] Hong J, Kang S. Score distillation for anomaly detection. Knowledge-Based Systems 2024; 295: 111842.

[111] Shi X, She Q, Fang F, Meng M, Tan T, Zhang Y. Enhancing cross-subject EEG emotion recognition through multi-source manifold metric transfer learning. Computers in biology and medicine 2024; 174: 108445.

[112] Lin F, Zhang P, Chen Y, Liu Y, Li D, Tan L, Wang Y, Wang DW, Yang X, Ma F, Li Q. Artificial-intelligence-based risk prediction and mechanism discovery for atrial fibrillation using heart beat-to-beat intervals. Med 2024; 5: 414-431.e5.

[113] Wang J, Liu X, Pan H, Xu Y, Wu M, Li X, Gao Y, Wang M, Yan M. Construction and validation of a risk-prediction model for anastomotic leakage after radical gastrectomy: A cohort study in China. Laparoscopic, Endoscopic and Robotic Surgery 2024; 7: 34-43.

[114] Wang C, Li YG, Li GM, Li HL. Equipment noise evaluation based on auditory saliency map. Applied Acoustics 2022; 201: 109125.

[115] Lu J, Wang R, Zuo G, Zhang W, Jin X, Rao Y. Enhancing CNN efficiency through mutual information-based filter pruning. Digital Signal Processing 2024; 151: 104547.

[116] Rayeed SM, Tuba ST, Mahmud H, Mazumder MHU, Mukta SH, Hasan K. BdSL47: A complete depth-based Bangla sign alphabet and digit dataset. Data in Brief 2023; 51: 109799.

[117] Kumar A, Gorai AK. Design of an optimized deep learning algorithm for automatic classification of high-resolution satellite dataset (LISS IV) for studying land-use patterns in a mining region. Computers & Geosciences 2023; 170: 105251.

[118] Visalli M, Dubois M, Schlich P, Ric F, Cardebat J, Georgantzis N. A dataset on the sensory and affective perception of Bordeaux and Rioja red wines collected from French and Spanish consumers at home and international wine students in the lab. Data in Brief 2023; 46: 108873.

[119] Ge Y, Song S, Yu S, Zhang X, Li X. Rice seed classification by hyperspectral imaging system: A real-world dataset and a credible algorithm. Computers and Electronics in Agriculture 2024; 219: 108776.

[120] Chowdhury MM, Ayon RS, Hossain MS. An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. Healthcare Analytics 2024; 5: 100297.

[121] Miguel JPM, Neves LA, Martins AS, do Nascimento MZ, Tosta TAA. Analysis of neural networks trained with evolutionary algorithms for the classification of breast cancer histological images. Expert Systems with Applications 2023; 231: 120609.

[122] Zelenkov Y, Fedorova E, Chekrizov D. Two-step classification method based on genetic algorithm for bankruptcy forecasting. Expert Systems with Applications 2017; 88: 393-401.

[123] https://www.cs.cmu.edu/~bhiksha/courses/10-601/decisiontrees/DTprune.png

[124] https://blogs.sas.com/content/subconsciousmusings/files/2017/04/kernalSVM.png

[125] https://www.kdnuggets.com/wp-content/uploads/jessica_logistic_regression_work_1.jpg

[126] https://www.kdnuggets.com/wp-content/uploads/jessica_logistic_regression_work_2.jpg

LIST OF ABBREVIATIONS

R2	Coefficient of determination (R-squared)
MSE	Mean Squared Error
F1 Score	Harmonic mean of precision and recall
ML	Machine Learning
NLP	Natural Language Processing
SVM	Support Vector Machine
KNN	k-Nearest Neighbors
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
OOD	Out-of-Distribution Detector
Avg	Average
MAE	Mean Absolute Error
CVDT	Continuous Variable Decision Trees
AI	Artificial Intelligence
RAM	Random Access Memory
RMSE	Root Mean Squared Error

LIST OF TABLES

LIST OF FIGURES