IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 3, Ser. III (May. –June. 2022), PP 80-89
www.iosrjournals.org

Evaluation of Key Supervised Classifiers with Popular Datasets Using WEKA

Zeeshan Abbas[1], Jianhua Lyu [1], Tahawor Abbas [1]
[1](*School of Computer Science and Engineering, Southeast University, Nanjing, China*)

Abstract:*Now a day's world Machine learning is very popular not only among academic researchers but also among industry researchers as well. In last couple of decades Machine Learning gain tremendous achievements, one of the most famous success was the victory of Deep Blue over then-World Chess Champion Garry Kasparov in 1997, but still there is lot of work which is not explored or need improvements. This technique (Machine learning) is often feasible, cost effective and scalable then manual programing. In this article, we focus on the elementary concept of Machine learning that is classification. Of course, a single article cannot cover all supervised machine learning classification algorithms; however, we hope that this paper gives a short comprehensive overview and the cited references will cover the major theoretical issues, guiding the researcher and practitioners about the selected classifiers and suggest possible combinations for different classification task. The goal of this article is to give a fundamental understanding of supervised classifier algorithms and explain how to implemented it, in a way that is especially useful to new researchers.*
Materials and Methods*: In this paper we discuss five key supervised classification algorithms and evaluate these algorithms using the five popular datasets, which is available online at UCI data repository.*
Key Word*: Machine Learning (ML), Attribute Relation File Format (ARFF), Classifier, WEKA ..*

Date of Submission: 15-06-2022 Date of Acceptance: 30-06-2022

I. Introduction

Machine learning systems automatically learn from data. This is often a very attractive alternate solution as compared with the manual, and in the last couple of decades, the use of machine learning has rapidly spread and grow especially in computer science. Now a day's machine learning is used in almost in everywhere like in web search, recommender system, spam filtering, drug and medicine design, treatment fraud detection and many other applications.

In this paper we discuss five key supervised classification algorithms and evaluate these algorithms using the five popular datasets, which is available online at UCI data repository [1]. The performance of these classifier evaluates according to their result i.e. correctly classified instances, incorrectly classified instances, mean absolute error and Root mean squared error etc. using WEKA and the result will display both statistically and graphically. WEKA is a software developed by The University of Waikato and its stands for Waikato Environment for Knowledge analysis. Five datasets are used from the UCI repository. The dataset format is Attribute Relation File Format (ARFF) that is default file format in WEKA, with Stratified cross-validation 10 folds for each experiment to achieve better result. We evaluate all classifier on each dataset to check which predicts better, the evaluation result shows that, there is no such classifier that perform well on all kind of data. We know that to understand machine learning algorithms requires a subnational amount of hard works, so we tried to ease the way how to learn machine learning algorithms. The main motive of this article is to provide basics idea of the supervised classifier learning and to elaborate how it's being used, that is quite helpful for new researcher.

Although the prior paper[2] done a same type of review, but the experiment only includes one dataset that is not enough, to make decision which classifier algorithm is best it must be test on different data set.

II. Environment Selection

In this article the selected algorithm evaluates using WEKA workbench (Waikato Environment for Knowledge Analysis)[7].

The University of Waikato New Zealand developed WEKA in 2007; the name stands for Waikato environment for knowledge analysis. The system is written in Java and distributed under the terms of the General Public License(GNU). It is platform independent, and can be run on Linux, Macintosh and Windows operating systems.

It offers comprehensive assistance for the entire experimental data categorization process, including data preparation, statistical evaluation of learning schemes, and visualizing the input data and learning results. It provides a large range of preprocessing tools in addition to a wide range of learning algorithms. A common interface is used to access this wide and comprehensive toolbox.

The most common way of using WEKA, is to apply different learning methods to a dataset and analyze the output, to learn more about the data. Another technique is to apply learned models to make predictions for new instances unseen instances [2].

III. Algorithm Selection

The decision, which specified algorithm we used in our experiment is somehow critical. In this section we describe about the selected classification algorithms. The classifiers evaluation is based on mostly on prediction accuracy and time consume. There is more than one way of grouping the classifier algorithms but in this article we group them according to similarity, means in term of function how they work. We pick these classifiers in the way that each of them belong to different function group. To cover as much as possible. The selected classification algorithm is the following:

Logistic regression classifier

Logistic regression is one of the popular technique in for binary classification. It is quite simple classifier algorithm that you can use as a performance baseline, it is easy to implement and it will perform well in many tasks. Therefore, every researcher should be familiar with its concepts. It belongs to Regression model family which include Linear Regression, Stepwise Regression etc., Regression models are workhorse of statistics and have been drafted into statistical machine learning [3]. For beginners, the fact when to use regression and classification is quit confusing, according to problem.

The mathematical modelling approach of logistic regression can be used to represent the relationship between numerous independent factors and a dichotomous dependent variable. Other modelling techniques exist, but logistic regression is by far the most used method for analyzing epidemiologic data to measure is dichotomous during illness. We shall demonstrate why this is correct.[2-4].

To understand logistic regression, we need to explain the logistic function it is sometime also called sigmoid function, which describes the mathematical form on which the logistic model is based.

$$p(x) = \frac{1}{1 + e^{-x}}$$

The p(x) function calculate probability. Where x is the independent variable, e represents the base of natural logarithm (which about 2.718). Figure 1. Illustrate output of the function for x range [-6 to 6].

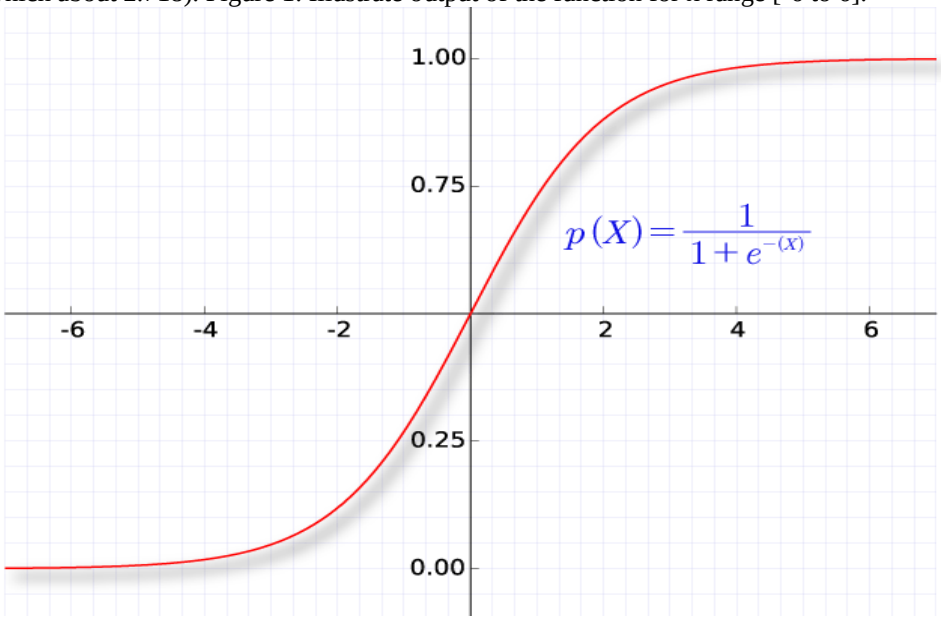


Figure 1: Logistic regression classifier

The fact that the logistic function p(x) has a range of 0 to 1 is the main reason for the logistic model's popularity. The model is intended to describe a probability, which is always a number between 0 and 1 in nature. In epidemiologic terms, such a probability represents the likelihood of a person contracting a disease. As

a result, the logistic model is set up to ensure that whatever risk estimate we get is always a figure between 0 and 1.

As a result, we can never receive a risk estimate higher than 1 or lower than 0. This is not always the case for other models, which is why the logistic model is frequently the first choice for estimating a probability.

When the dependent variable has three or more distinct values, such as Married, Single, Divorced, or Widowed, the method is known as multinomial logistic regression. Although the type of data utilized for the dependent variable differs from that used in multiple regression, the procedure's practical application is the same [4].

Naïve Bayes classifier

Naïve Bayes is one of the simplest and impressive algorithm that belongs to family of simple probabilistic classifiers, based on applying Bayesian theorem. Bayesian networks provide a promising representation for machine learning for the same reasons they are useful to perform different tasks such as diagnosis: they can deal explicitly with issues of noise and uncertainty, which are major problems in any task [5]. Since the 1960s, Naive Bayes has been intensively researched. However, the challenge of evaluating documents as classification to categorize like (such as spam or none spam email classification, sports or politics news classification , etc.) with word frequencies as the features remains a popular way for text categorization [6].

It implements simple probabilistic classifier based on applying Bayes theorem where every feature is assumed to be class-conditionally independent [7].

Naive The underlying premise of Bayes classifiers is that a variable's impact on a particular class is independent of the values of other variables. Class-conditional independence is the term for this presumption. It is designed to make the computation simpler, therefore in this regard, it is viewed as naive. Abstractly, the probability model for a classifier is a conditional model.

$$p(C|\, F1, F2, \ldots Fn)$$

over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F1 through Fn.

When the number of features is large or when it can take on a large number of values.

$$p(C|\, F1, F2, \ldots Fn) = \frac{p(C)\, p(F1, F2 \ldots, Fn|C)}{p(F1, F2 \ldots, Fn)}$$

Support vector Machine

Support vector machine (SVM) is one of the useful and robust supervised machine learning technique for classification as well as for regression but mostly suitable and famous for classification. SVM was initially intended for the binary classification task, in which there are two classes, but the latest SVM can well classify non binary classification. It is based upon the idea of separating hyperplanes.

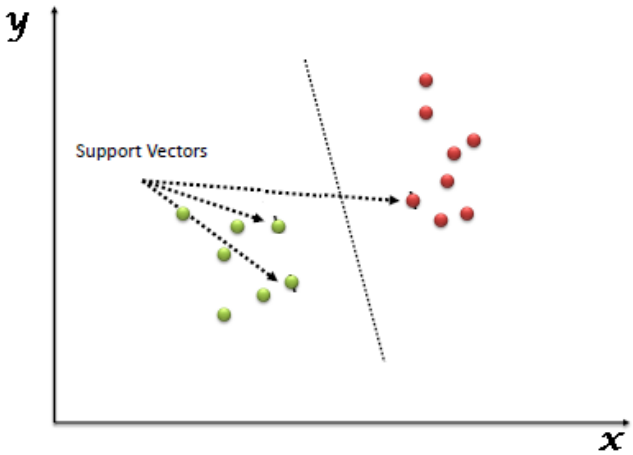


Figure 2: Support vector machine classifier

The support vector classifier, also known as a soft margin classifier, performs better classification for the majority of the training data and has more robustness to individual observations. We allow some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane, rather than

aiming for the biggest margin so that every observation is both on the correct side of the margin and the correct side of the hyperplane [8].
It uses a more natural loss function called hinge loss.

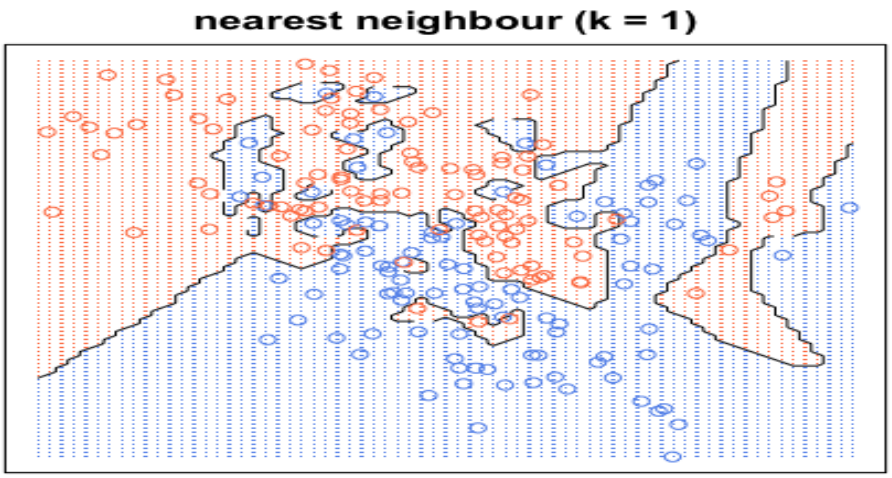
K-nearest neighbor classifier

K-nearest neighbor (KNN) is also sometimes referred as case based learning or instance based learning. Where every training instance is case from the problem domain. Instance-based learning algorithms do not base on abstractions, because they do not maintain abstractions that derived from specific instances. It is well suited for low dimensional data; although it can be used for high dimensional data but aware it performs not well on high dimensional data. It stores entire datasets which it uses as its representation, instead of learning any model it makes prediction on the spot by calculating distance between input and each training instances. [9]

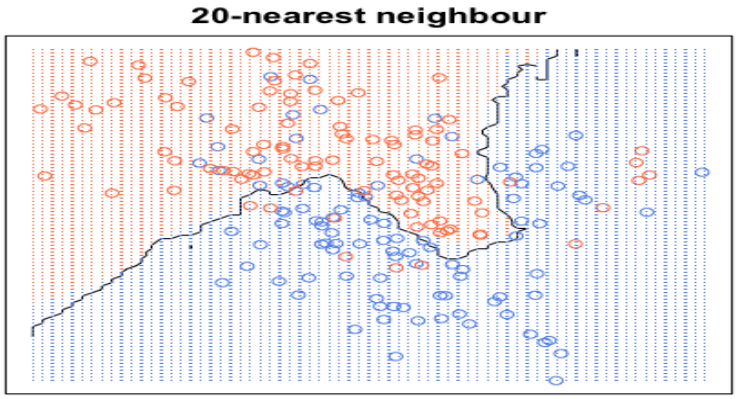
There is no formula for the best K, however it's better to set K odd when you have an even class to forecast and set K even when your data have an odd class. It requires a positive integer K (number of neighbors take part in prediction) and a value for observation x0. The classifier starts by locating the K training data points that are closest to x0, denoted by N0, in the training data. The percentage of points in N0 whose answer values equal j is then used to estimate the conditional probability for class j:

$$Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in x_0}^{\infty} I(y_i = j)$$

The test observation x_0is then classified by KNN using the Bayes rule and assigned to the class with the highest probability [8].
When K = 1, the decision boundary is initially too flexible and uncovers data patterns that don't match the Bayes decision boundary. In line with this, a classifier with low bias but large variance should be used. The approach becomes less adaptable and generates a decision boundary that is almost linear as K increases. This is equivalent to a classifier with strong bias but low variance [8-10].



K-NN with K=1, Images from (G. James et. el.) [10]



K-NN with K=20, Images from (G. James et. el.) [10]

C4.5 classifier

It's a decision tree-based algorithm that's an improved version of descended ID3 developed by (Ross Quinlan). It creates a decision tree for classification, which is why it's also known as a statistical classifier [10]. It can also build classifiers in a more comprehensible ruleset form. [11] Key features of C4.5 algorithms.

- The large decision tree can be viewing as a set of rules which is easy to understand.
- C4.5 algorithm gives the acknowledge on noise and missing data.
- Problem of over fitting and error pruning is solved by the C4.5 algorithms.
- It works very well for continues and discrete data.
- It handles missing data in far better way.

IV. Evaluation Metrics

In this article we pick some of the important and widely used evaluation metrics to evaluate the selected algorithms.

Accuracy

The simple and most widely used metrics to evaluate a classifier and models. Basically the measurement of how often makes the correct prediction.

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction}$$

In detail it's the ratio of True Positive and True Negative Prediction to the total number of predictions.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ number\ of\ prediction}$$

It works well if each classed have number of sample belongs per class is equal.

Relative Absolute Error

Relative absolute error takes total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.

$$RAE = \frac{\sum_{i=1}^{N} |\widehat{\theta_i} - \theta_i|}{\sum_{i=1}^{N} |\bar{\theta} - \theta_i|}$$

Where true value is$\theta_i$and estimated value is $\hat{\theta}$and $\bar{\theta}$ is a mean value of θ.

F-measure

The F-measure is defined as a harmonic mean of precision (P) and recall (R)

$$F - measure = \frac{2PR}{P + R}$$

V. Experimental Work

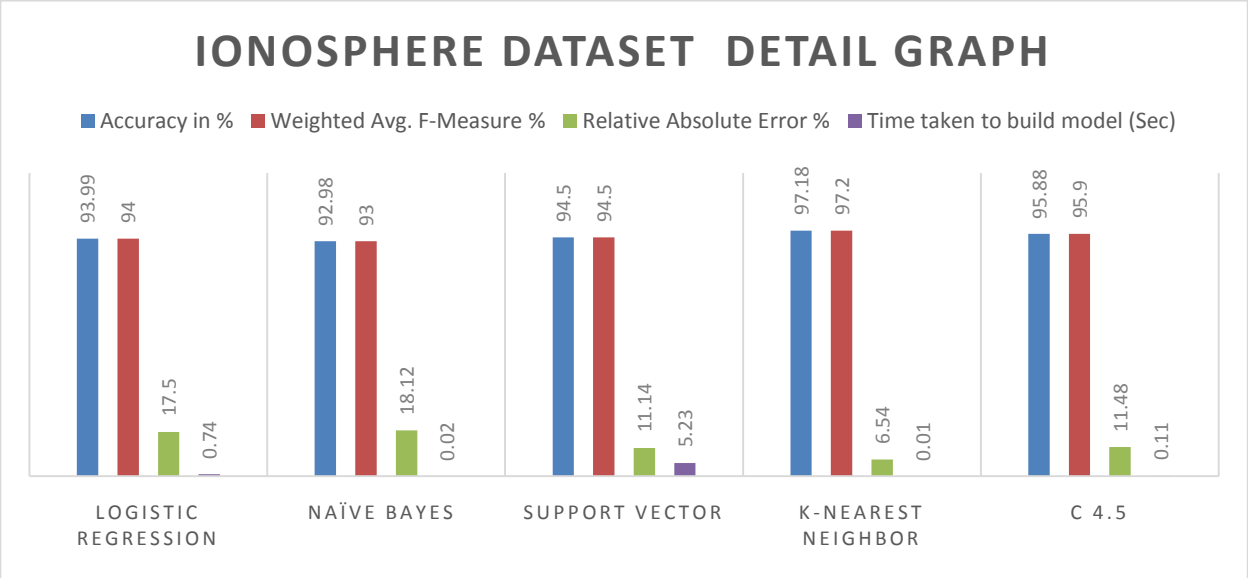
In this experiment we used default parameters and datasets without any extra preprocess from UCI data repository [12]. We use Stratified cross-validation 10 folds for each experiment. The training set is partitioned into mutually exclusive and equal-sized subsets, and the classifier is trained on the union of all the other subsets in cross-validation. As a result, the average of each subset's error rate represents an estimate of the classifier's error rate. Cross validation with one exception is known as leave-one-out validation. There is only one case in each test subgroup. Of course, this method of validation is more computationally expensive, but it is useful when the most precise estimate of a classifier's error rate is necessary.

Ionosphere Dataset

Title: Johns Hopkins University Ionosphere database
Number of Attributes: 35 class attribute
Number of Instances: 351
Source Information: https://archive.ics.uci.edu/ml/datasets/Ionosphere
Relevant Information: A system in Goose Bay, Labrador, collected this radar data. A phased array of 16 high-frequency antennas with a total transmitted power of 6.4 kilowatts makes up this system. Free electrons in the ionosphere were the intended targets. Radar returns that reveal evidence of some form of structure in the ionosphere are considered "good," while those that do not are considered "bad," as their signals flow through the ionosphere.

The time of a pulse and the pulse number were used as arguments in an autocorrelation function that was applied to the received signals. For the Goose Bay system, there were 17 pulse numbers. The complex values given by the function resulting from the complex electromagnetic signal are described by two attributes per pulse number in this Dataset.[13].

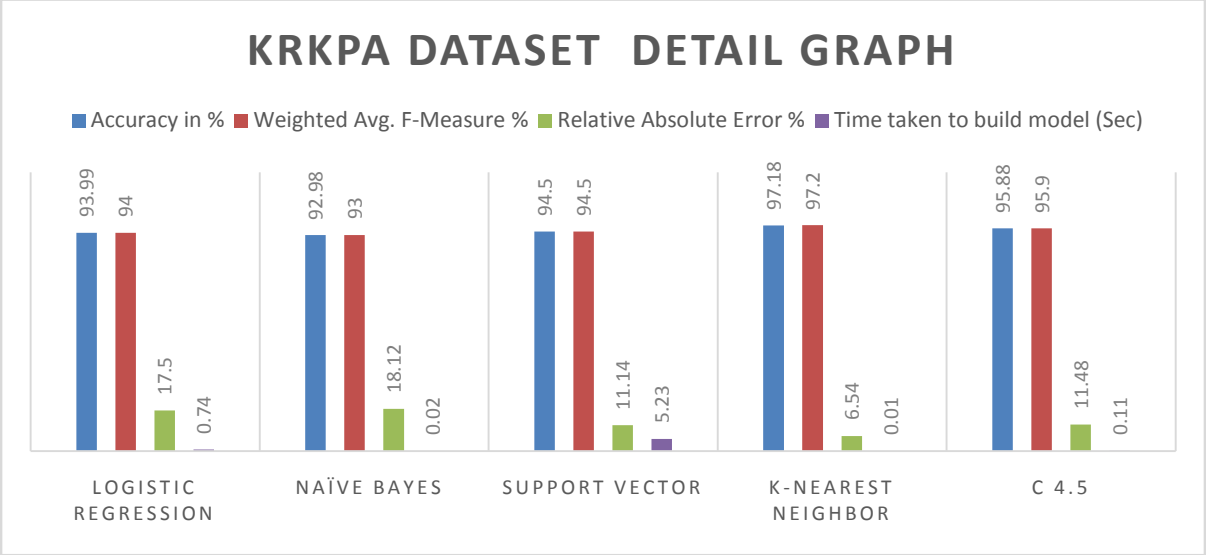
Classifier	Accuracy in %	Weighted Avg. F-Measure %	Relative Absolute Error %	Time taken to build model (Sec)
Logistic Regression	88.89	88.7	27.85	0.02
Naïve Bayes	82.62	82.9	37.70	0.01
Support Vector	93.44	93.3	14.22	0.01
K-nearest neighbor	89.17	88.8	24.30	0.01
C 4.5	91.45	91.3	20.36	1.04



KRKPA7

Title: Chess (King-Rook vs. King-Pawn) Data Set
Number of Attributes: 36
Number of Instances: 3196
Source Information:
https://archive.ics.uci.edu/ml/datasets/ Chess+(King-Rook+vs.+King-Pawn)
Relevant Information: The format of this database was modified on 2/26/90 to conform with the format of all the other databases in the UCI repository of machine learning databases [13].

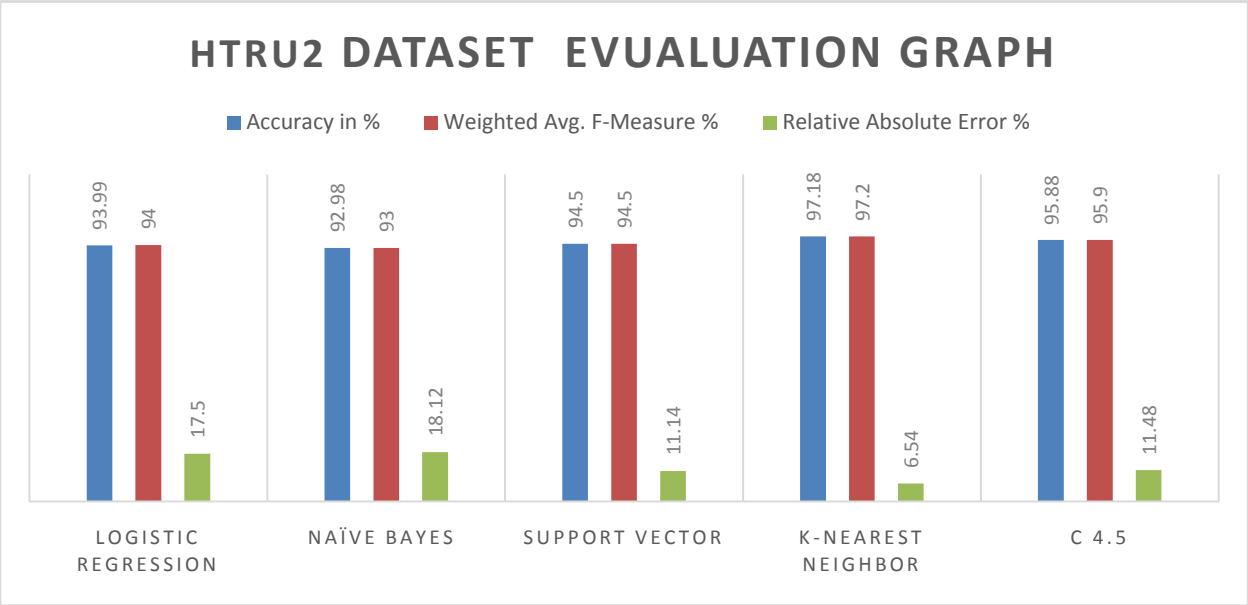
Classifier	Accuracy in %	Weighted Avg. F-Measure %	Relative Absolute Error %	Time taken to build model (Sec)
Logistic Regression	97.56	97.6	8.4	0.02
Naïve Bayes	87.89	87.9	41.82	0.02
Support Vector	93.90	93.9	12.22	0.08
K-nearest neighbor	96.27	96.3	18.92	0.01
C 4.5	99.43	99.4	1.78	0.05



HTRU2 Dataset

Number of Attributes: 9
Number of Instances: 17898
Source Information:
https://archive.ics.uci.edu/ml/datasets/ HTRU2
Relevant Information: HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey.
Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter.[14]

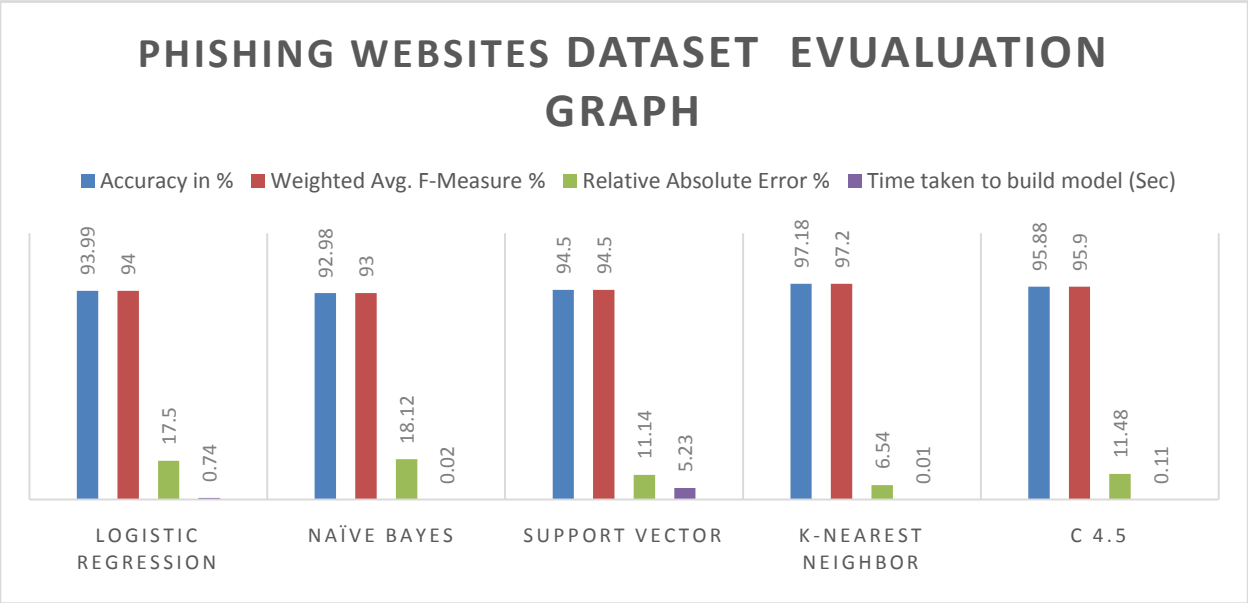
Classifier	Accuracy in %	Weighted Avg. F-Measure %	Relative Absolute Error %
Logistic Regression	97.92	97.9	20.75
Naïve Bayes	94.50	94.8	33.26
Support Vector	91.14	87.2	53.21
K-nearest neighbor	97.79	97.8	18.29
C 4.5	97.84	97.8	20.41



Phishing Websites Data Set

Title: Waveform Database Generator
Number of Attributes: 31
Number of Instances: 11055
Source Information:
https://archive.ics.uci.edu/ml/datasets/ phishing+websites
Relevant Information: One of the difficulties we encountered in our research was the lack of reliable training datasets. In reality, any researcher in the field will encounter this issue. However, despite the proliferation of articles on predicting phishing websites, no reliable training dataset has been publicly released. This may be due to a lack of consensus in the literature on the definitive features that characterise phishing webpages, making it difficult to shape a dataset that covers all possible features. The essential features that have proven to be sound and effective in predicting phishing websites are highlighted in this dataset. We also offer a few new features [13,16].
In this dataset, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we propose some new features [13].

Classifier	Accuracy in %	Weighted Avg. F-Measure %	Relative Absolute Error %	Time taken to build model (Sec)
Logistic Regression	93.99	94.0	17.50	0.74
Naïve Bayes	92.98	93.0	18.12	0.02
Support Vector	94.50	94.5	11.14	5.23
K-nearest neighbor	97.18	97.2	6.54	0.01
C 4.5	95.88	95.9	11.48	0.11

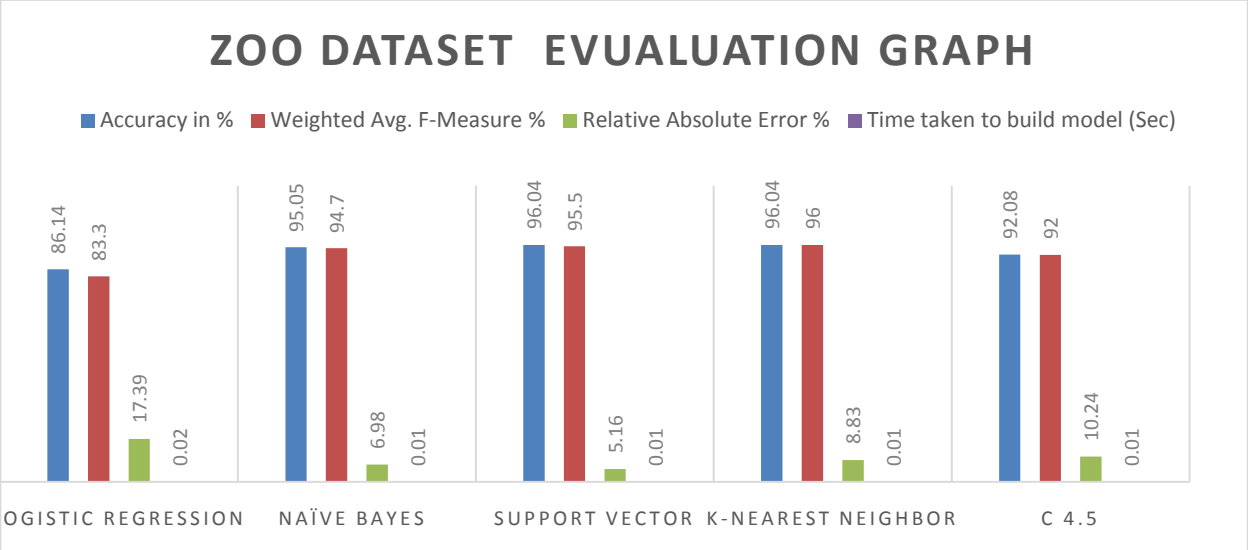


Zoo Dataset

Number of Attributes: 17
Number of Instances: 101
Source Information:
https://archive.ics.uci.edu/ml/datasets/zoo

Relevant Information: A simple database containing 17 Boolean-valued attributes. The "type" attribute appears to be the class attribute. Here is a breakdown of which animals are in which type: (I find it unusual that there are 2 instances of "frog" and one of "girl"!)[13].

Classifier	Accuracy in %	Weighted Avg. F-Measure %	Relative Absolute Error %	Time taken to build model (Sec)
Logistic Regression	86.14	83.3	17.39	0.02
Naïve Bayes	95.05	94.7	6.98	0.01
Support Vector	96.04	95.5	5.16	0.01
K-nearest neighbor	96.04	96.0	8.83	0.01
C 4.5	92.08	92.0	10.24	0.01



VI. Conclusion

Data mining is very popular and vast area that incorporate several techniques from different fields including machine learning, pattern recognition, statistics, and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. There is no single best algorithm that is best for all kind of data, since it highly depends on the data you are working with..

References

[1]. R. kumari Dash, "Selection Of The Best Classifier From Different Datasets Using WEKA," Int. J. Eng., vol. 2, no. 3, pp. 1–7, 2013.
[2]. R. kumari Dash, "Selection Of The Best Classifier From Different Datasets Using WEKA," Int. J. Eng., vol. 2, no. 3, pp. 1–7, 2013.
[3]. E. Frank, M. A. Hall, and I. H. Witten, "The WEKA workbench," Data Min., pp. 553–571, 2016.
[4]. M. Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, Introduction to Logistic Regression, Second Edi. New York, NY: Springer, 2010.
[5]. J. C. le Cessie, S. and van Houwelingen, "Ridge Estimators in Logistic Regression," Appl. Stat., vol. 41, pp. 191–201, 1992.
[6]. Holmes, Geoffrey; Donkin, Andrew; Witten, Ian H. (1994). "Weka: A machine learning workbench" (PDF). Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.
[7]. G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," Elev. Conf. Uncertain. Artif. Intell., pp. 338–345, 1995.
[8]. S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd Ed. Prentice Hall, 2003.
[9]. N. Murty, Pattern recognition : an algorithmic approach, no. 0. 2011.
[10]. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, vol. 103. 2013.
[11]. D. W. AHA, D. KIBLER, and M. K. ALBERT, "Instance-Based Learning Algorithms," Dep. Inf. Comput. Sci. Univ. California, lrvine, CA 92717, vol. 6, pp. 37–66, 1991.
[12]. Q. J. Ross, C4.5: Programs for Machine Learning. San Mateo, CA.: Morgan Kaufmann Publishers, 1993.
[13]. X. Wu et al., Top 10 algorithms in data mining, vol. 14, no. 1. 2008.
[14]. C. Dua, Dheeru and Graff, "UCI Machine Learning Repository." University of California, Irvine, School of Information and Computer Sciences, 2017.
[15]. D. Dua and C. Graff, "{UCI} Machine Learning Repository." 2017.
[16]. J. D. K. R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach," Mon. Not. R. Astron. Soc., vol. 459, no. 1, pp. 1104–1123, 2016.