Kristu Jayanti Journal of Computational Sciences Volume 1 2021, Pages 01-12

Supervised Learning Algorithms: A Comparison

[1]A.Deva Kumari, [2]Dr.Josephine Prem Kumar, [3] Dr.V.S Prakash, [4]Divya K.S

[1]Asst.Professor, Dept. of Computer Science, Kristu Jayanti College, Autonomous
devakumari@kristujayanti.com
[2]Professor, Dept. of Computer Science, Cambridge Institute of Technology
josephine.cse@cambridge.edu.in
[3]Asst.Professor, Dept. of Computer Science,Kristu Jayanti College,Autonomous
vsprakash@kristujayanti.com
[4]Asst.Professor, Dept. of Computer Science, Kristu Jayanti College, Autonomous
divysks@kristujayanti.com

Abstract

Artificial Intelligence is logical systems where the PCs figures out how to take care of an issue, without unequivocally program them. Machine learning is a subset of AI where machines learn based on the data fed to them. A relative report over various AI managed procedures like Linear Regression, K nearest neighbours, Logistic Regression, Decision Trees, Random Forest, Support Vector Machine and Naive Bayes are made in this paper. The correlation depends on assumptions, influences of co-linearity and exceptions, hyper-parameters, shared examination.

Keywords: supervised learning, co-linearity, exceptions/outliers, hyper parameters

1. Introduction

Machine learning is an implementation of computing (AI) which has control to robotically learn and progress from experience without actually explicitly being programmed. Machine learning concentrates on the computer programs which might access data and make use of it to acquire knowledge for them. The method of learning initiates with observations, comparative examples, unswerving knowledge, or instruction, to identify the patterns in data to draw better decisions within the future supported the examples that are offered. The first goal is to permit the computers to acquire knowledge automatically without human interference or support and adjust activities consequently.

Supervised machine learning procedures can relate to what has been cultured within the past to new data employing labelled examples to foretell future actions. From the analysis of an identified training dataset, the educational algorithm yields an inferred function in order to custom predictions about the output values. The structure is ready to supply objectives for any new input after sufficient training. The procedure also can compare its output with the proper, planned output and catch errors to change the model consequently.

The larger part of viable machine learning employments supervised learning. It may be a method where input features (x) are utilized to urge a yield variable (Y) and calculation learn the mapping work from input to the yield [5].

$$Y = f(X)$$

The objective is to surmise the representing work so fine that when there is current input information (x) that algorithm can anticipate the yield factors (Y) for that data. The

term supervised learning is used as the method of design learning from the preparing dataset can be understood as an instructor overseeing the learning prepare. The calculation iteratively makes expectations on the preparing information, learning halts when the calculation accomplishes an acceptable range of performance [1].

2. Working procedure of comparison based supervised machine learning algorithms:

2.1. Linear Regression

It may be a regression demonstrate, implying, it takes features and antedate a continuous output in other words utilizing independent variables algorithm finds dependent variable which may be a persistent quality, e.g., stock cost, the weight of individual, compensation. Linear regression, the term says, it finds a straight bend arrangement to each problem.

LR apportions weight parameter, beta for every preparing feature. The anticipated yield y will be a linear function of features and β coefficients.

$\beta 0+\beta 1X+\epsilon = Y$ (1)

Y is the target variable to be anticipated, β0 is the y-intercept, β1 is the incline, ϵ is the blunder and X is indicator variable or independent variable. Amid the begin of preparing, each beta is arbitrarily initialized. Gradient incline calculation would be utilized to adjust the values within the precise heading. In the diagram below, each black dots signify the training data then the blue line displays the resulting solution, the pink line indicates the error.

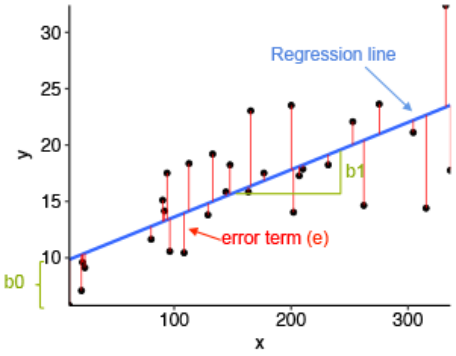

Figure 1. Linear Regression Model

Execution estimation in LR, is done utilizing mean squared error as the measure of loss or mean absolute error. The deviation of anticipated and genuine yields would be squared and summed up. Gradient descend procedure would use the derivative of this loss.

Hyperparameters like Regularization parameter (λ) is utilized to dodge over-fitting of the information. More excellent the λ, greater would be regularization also the arrangement will be exceedingly one-sided. Lesser the λ, the arrangement would be of high fluctuation. Middle esteem is best. Learning rate (α) gauges, by what value the β values ought to be adjusted while relating gradient descends calculation amid preparing. α ought to moreover be direct esteem. Few

presumptions have been made in LR demonstrate, the relationship between the autonomous and dependent factors are accepted to be straight, which is continuously not conceivable. Preparing information is expected to be homoscedastic, meaning the change of the mistakes ought to be, to some degree consistent. Autonomous factors are accepted not to be co-linear.

Co-linearity and outliers are two highlights that must be considered; two features are considered to be collinear if one highlight could be straightly anticipated using the other to absolute degree precision. Co-linearity will magnify the typical mistake also causes a small number of critical highlights to end up inconsequential amid formulating.

2.2. Logistic Regression

Logistic regression, a bit comparable to the above methodology, is the right calculation of classification algorithms, to start with. Although the title 'Regression' appears, it is not a demonstration of relapse, but a model of classification. To outline the dualistic production model, it employs a logistical task. A probability ($0 \leq x \leq 1$) will be the yield of measured regression; it can also be used to estimate binary 0 or 1 as yield (if $x<0.5$, yield= 0, otherwise output=1). Exceptionally similar to linear regression, the existing model behaves to some degree. The linear output, taken after a reservation work for the regression output, is also determined. The Sigmoid function is a logistic function sometimes used. It can be assumed that the z value in (1) is identical to the linear regression output.

$z=\theta_1 x_1+ \theta_2 x_2+ \theta_3 x_3+$.....

$H(\theta)=g(z)$

$G(z)=1/1+e^{-z}$ (2)

The H (θ) value refers to P(y=1), i.e. the yield probability is binary 1, the known input x then P(y=0) is equal to 1-H (θ). If the value of z is 0, then g(z) is 0.5. At any point where z is positive, h(almost) will be additionally noteworthy as binary 1 would also yield 0.5. In addition, the estimation of y will be 0. at any point where z is negative. As we apply a linear condition to decide the classifier, the output model will also be linear, meaning it fragments the input measurement compared to an equivalent mark in two regions for every point in one region. The dispersion of the Sigmoid function (2) appearin the graph be Low.

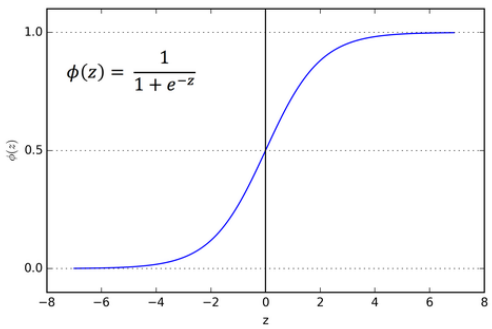

Figure 2. Sigmoid Function

As non-linear sigmoid work is used at the end, the loss function such as mean squared error cannot be used as output estimation in logistic regression (as in linear regression). MSE work can present nearby minimums and affect the procedure of gradient descent. Cross-entropy is also used for functional loss. It will use two conditions, matching y=1 and y=0.0, respectively. If the expectation is purely off-base at some point, the critical logic here is, (e.g., y = 1 & y = 0), the outcome will be -log(0) an infinity cross-entropy loss (3)

$$J(\theta) = \frac{1}{m}\sum cost(y', y)$$

$$cost(y', y) = -log(1 - y') \;\; if \;\; y = 0$$

$$cost(y', y) = -log(y') \;\; if \;\; y = 1 \qquad (3)$$

In the above equation, m represents the magnitude of training data, y' positions expected output, y indicates actual output. In linear regression, hyperparameters are considered, and logistic regression is comparable. The regularisation parameter(λ) also had to be legally altered to achieve good accuracy. Learning rate(a) Logistic regression presumptions have shown to be comparative to linear regression[2].

2.3. K-nearest neighbors

This approach is a non-parametric technique intended for classification and regression. In investigating the neighbourhood, the fundamental logic behind KNN exists to accept the test data value, which must be equivalent to them, even infer the yield. A preponderance vote is related to the KNN classification over the k closest data points, while the mean of k closest data points is known as the yield in KNN regression. As a rule, the calculation selects odd numbers like k. This model can be a sluggish learning model for which runtime calculations occur.

The hyperparameters for KNN mainly involve two attributes, the function of distance and the value of K.K appreciation depends on how many neighbours are interested in the KNN process. Depending upon the confirmation mistake, K should be tuned. Distance is measured using Euclidean distance, which is the primary proximity function used.

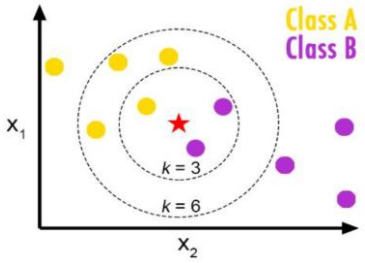

Figure 3.KNN classification for n=3 and n=6

In the graph above, when preparing details, the violet and yellow points are compared to Class B and Class A. The red star focuses on the test details that need to be classified. The algorithm anticipates Class A as output when k = 3, the algorithm forecasts Class B as output even when K = 6. There is no planning included in KNN for output estimation. K neighbours with least distinct, measured using either Euclidean distance or Manhattan distance, will take part in classification / regression in the study.

2.4. Decision Tree

A tree-centred analysis used to unravel classification and regression problems may be a decision tree. An upturned tree is mounted which, for the yield, is divided from a homogeneous probability disseminated root hub to extremely diverse leaf hubs. Regression trees are used when a variable is dependent and has continuous values; when a variable is dependent and has discrete values, classification trees are used. With each node having a condition on a highlight, the decision tree is calculated using the sovereign variables. Depending on the condition, the nodes select the node to explore another. The performance will be expected until the leaf node is reached. The correct grouping of constraints would effectively mark the tree. To pick the conditions for nodes, entropy and information gain are used as the benchmarks. Using greedy and recursive formulas, the structure of the tree is calculated.

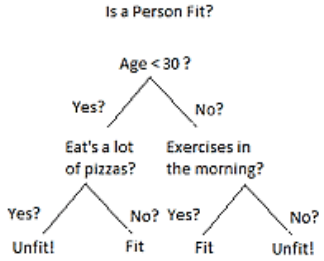

Figure 4. Decision tree to check Fitness of a Person

In the graph above, a tree has a group of internal nodes(constraints) and leaf nodes with names(accept/decline offer).In the case of CART(classification and regression trees), the classification metric used is Gini index[4]. It could be an indicator to find out the blending of data points.

$$giniindex = 1 - \sum P_t^2 \quad (4)$$

The property with a higher Gini index is chosen as the other condition, at each stage of making a decision tree. Gini score will be most extreme when a set is unevenly blended. Entropy, in addition to information gain, is used to choose the next feature. Within the underneath condition, H(s) attitudes entropy while IG (s) stances Information gain. The Information gain computes entropy contrast of internal and child nodes. Class with the most extreme information gain is preferred as another internal node.

$$H(s) = -\sum P_c \cdot log(P_c)$$
$$IG(s) = H(s) - \sum_t P_t \cdot H(t) \quad (5)$$

Hyperparameters in Decision tree includes a condition that would cost the task of selecting the next tree node. Generally used indicators include Gini/entropy, the maximum permissible depth of a tree. A smallest possible sample split is the least nodes necessary for splitting an internal node.

2.5. Support Vector Machine

Support Vector Machine is a tool for both classification and regression that can be used mutually. It primarily has two variants to help linear and non-straight problems. A minimum edge direct arrangement for the problem is found by the Linear SVM. In case of linear or straight SVM, SVM with kernels are used when the structure is not explicitly distinguishable problem space is straightly divided. A hyperplane that capitalises on the classification margin is assumed by the model. The hyperplane would be separated into an N-1 dimensional subspace if there were N highlights. The boundary nodes are termed as support vectors within the highlight space. The most extreme edge is inferred, based on their relative location, and an ideal hyperplane inside the midpoint is drawn.

The margin (m) and||w|| are inversely proportional, w is the collection of weight matrices.||w|| must be minimized in order to maximize the margin.

C is the regularization factor which balances out the miss penalty., w is needed to tune with an outrageous edge among the classes. To put it plainly, C is the degree of numbness over anomalies. Non-straight SVM if the dataset is not directly distinct. A kernel function is utilized to determine another hyperplane for entire training data. The dispersion of labels in newly created hyperplane will be with the end goal that training data would be straightly divisible. Afterwards, a straight bend will arrange the labels in the hyperplane. At the point when classification results are extended back to feature space, an in-direct arrangement is got.

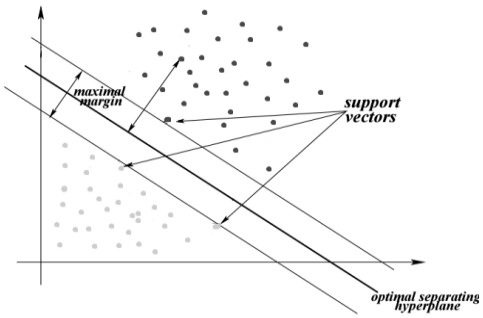

Figure 5. SVM with Hyper planes

Different kernel functions are established. The equation is given by,

$$\text{Minimize } \frac{||\vec{w}||^2}{2} + C\sum_i \zeta_i \quad \text{where } y_i(\vec{w}.\phi(x_i)+b) \geq 1-\zeta_i \text{ for all } 1 \leq i \leq n, \zeta_i \geq 0 \tag{6}$$

The xi would be substituted by ϕ(xi) that would alter the dataset into the newly created hyperplane. The loss function in the above equation can be split as below:

$$1)\ \frac{1}{2}\|w\|^2 \qquad 2)\ C\sum_i \max(0, 1-y_i(w^\intercal x_i + b)) \tag{7}$$

Hyperparameters considered in SVM are, Margin Constant (C), It could be a hyperparameter that chooses the level of penalty above the exceptions. It is straightforwardly converse to regularization parameter. If C is enormous, Exceptions will get high penalty plus the difficult margin is shaped. When C is little, the exceptions are overlooked, and the margin would be varied. Polynomial Kernel (d), if d = 1, it is proportionate to a linear kernel. The Width Parameter (γ), chooses the thickness of the Gaussian curve. With the increment in gamma, thickness also increases [7].

2.6. Random Forest

This method has a group of decision trees ensembled using "bagging method" to get classification and regression outputs. For classification, it estimates the yield utilizing more extensive part voting, while in regression, mean is calculated. Random Forest uses a vigorous, precise model that can lever huge assortments of input data like binary or categorical or continuous features. Loss capacities utilized are entropy/Gini value to determine the loss esteem of the datasets.

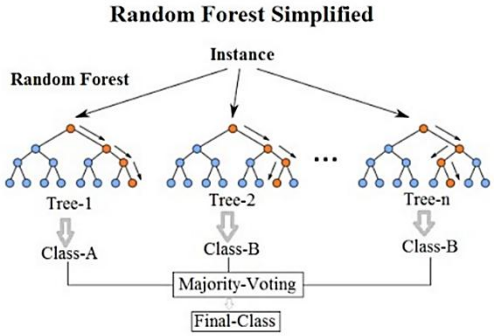

Figure 8.Random Forest

Hyperparameters in the random forest are n_estimators, which is the count of trees within the forest. With a vast number of trees, we get good exactness, but higher computational complication. Most powerful features which are the most incredible number of features permitted in a single tree. Least sample leaf indicates the least number of tests necessary to split an inner node [8].

2.7. Naive Bayes

Naive Bayes could be a propagative probability model utilized for classification issues. It is the critical model utilized for content classifications, where the feature set is exceptionally huge. It is broadly utilized for sentiment investigation, spam sifting etc. The Bayes rule can be expressed as,

$$P(C_i|x_1,x_2,\ldots,x_n) = \frac{P(x_1,x_2,\ldots,x_n|C_i).P(C_i)}{P(x_1,x_2,\ldots,x_n)} \text{ for } 1 \leq i \leq k \quad (8)$$

In the above condition naive Bayes consider all features are free. In the case of numerous class names, P (Ci|X) is intended for each label, the label with the most extreme likelihood is selected as the output. The significant presumption of Bayes theorem is that all features incline to be commonly autonomous. However, in actual situations, this may not be accurate [6].

3. Comparison based Advantages and Disadvantages of Supervised learning algorithms

Table 1: Comparing the advantages and Disadvantages

Algorithm	Advantages	Disadvantages
Linear Regression	• Simple and essential operation. • Space complex result.	•Appropriate if the arrangement is straight. In numerous genuine life

	• Firm training. •Value of β coefficients provides a presumption of highlight significance.	situations, this might not happen. •Algorithm expects the errors to be usual scattered, which is not true always. •Algorithm accepts input features are reciprocally autonomous (no co-linearity).
Logistic Regression	•Simple, firm and direct classification method. •θ parameters clarify the path and intensity of centrality of sovereign factors versus the dependent variable. • Model could be utilized for multi-class classifications. • Loss function is continuously curved.	•Cannot be connected on non-linear cataloguing problems. • Appropriate choice of features is essential. •Decent signal to noise proportion is anticipated. •Colinearity plus exceptions alter the precision of LR model.
K Nearest Neihbors	• Simple and straightforward ML model. • Limited hyperparameters to alter.	• k should be shrewdly selected. • Large computation fetched •Proper scaling must be given for appropriate treatment among features.
Decision Trees	• No pre-processing required on data. • No presumptions on the dispersion of data. • Manages co-linearity capably. •Decision trees can give justifiable clarification over the prediction	• Probabilities of overfitting the model in case tree is repeatedly built to realize high perfection. Decision tree trimming can be utilized to illuminate this issue. • Susceptible to outliers. •Tree might develop to be exceptionally intricate while preparing complicated datasets. •Loses important data while dealing with continuous variables.
Support vector Machine	•To unravel complex solutions, SVM practises kernel trap. •SVM employs a curved optimization feature, which is continuously accomplished by inclusive minima. •Hinge failure provides higher precision. Outliers are addresses which are compatible with soft margin C. •SVM later employs the kernel trick to illuminate non-linear problems, while decision trees infer hyper-rectangles to illuminate the problem in the input space.	• Hinge loss results in sparsity. • Hyperparameters, kernels must be carefully altered for adequate accuracy. • Longer preparing time for bigger datasets.
Random Forests.	• Accurate and capable model. • Manages overfitting competently. •Provisions implicit feature choice and determines feature significance.	• Computationally intricate and slow when timberland gets to be large. • Not a transparent model for the prediction.
Naïve Bayes	• Functions better when preparing data is less. •It is better than other discriminative	• Asserts that the attributes are entirely free to each other, and that in real life circumstances is not acceptable.

	models If conditional independence is fulfilled. • Handles unessential features. •Provisions binary, multi-class classifications.	• While assembling a large population sample, and if P(X = feature) is zero, the posterior frequency will also be null. If the sample does not correctly represent the population, the predicted situation will occur. • To remove discrete values in functions, continuous variables are discarded. In order to maintain a strategic distance from data loss, this job should be sensibly completed.

4. Comparison of various supervised learning algorithms:

4.1. Comparison of Linear Regression with other models :

Linear regression and Decision Tree compared: Decision trees bolsters non-linearity, wherein LR provisions only straight arrangements. When there are vast numbers of features and data set is less, linear regressions might outflank Decision trees. Overall, Decision trees have improved average precision. For categorical free variables, decision trees are the right choice. A decision tree handles co-linearity way better than LR.

LR and SVM compared: SVM uses the kernel trick to support both direct and non-linear arrangements. Exceptions superior to LR are handled by SVM. When the training data is smaller, both strategies perform well, and there are a large range of features.

Comparing LR and KNN: KNN is a non-parametric model, whereas LR could be a parametric model. Later, in real-time, it is moderate because the neighbour node still needs to be discovered to retain all training data, while LR can effectively extricate yield from the tuned β coefficients.

Comparison between LR and Neural Networks: Neural networks need large training data relevant to the LR model, whereas LR actually works better with less training details. Compared with LR, NN will be mild. With neural networks, predicted accuracy can continually be much higher.

4.2. Comparison of Logistic regression and other models :

Logistic regression compared with SVM subtle elements like SVM can manage non-linear arrangements while logistic regression only deals with linear arrangements. Linear SVM addresses exceptions superior because it infers the most outstanding margin solution. Logistic regression compared with Decision Tree shows Decision tree addresses co-linearity way better than LR. Importance of features is determined in LR than Decision trees. For categorical values, Decision trees are superior to LR.

In contrast to the neural network, non-linear configurations can be supported later, while LR cannot. Formerly convex loss job, while NN could suspend, it would not hang in local minima. When training information is less, LR outflanks NN and features are less expansive, whereas NN requires significant training data.

LR when related with Naive Bayes, the later could be a propagative model while LR may be a discriminative demonstrates. Naive Bayes addresses little datasets, while LR with regularization can accomplish comparable performance. Also, LR performs well

compared to naive Bayes for co-linearity; also naive Bayes anticipates all features are autonomous.

LR when equated with KNN later is a non-parametric demonstrates, whereas LR is a parametric model. KNN is comparatively slower than LR. KNN reinforces non-linear arrangements, whereas LR reinforces as it were straightforward solutions. LR can predict to a certain extent, while KNN can only yield the labels.

4.3. Comparison of KNN with other models :

The considerable actual computation time taken by KNN is a typical contrast between KNN and other models. Compared to the naive Bayes, because of KNN's real-time efficiency, NB is much faster than KNN. Whereas KNN is not, Naïve Bayes is parametric. KNN is even easier when the data has a high SNR relative to linear regression.

KNN versus SVM delineates the exceptions superior to KNN are taken care of by SVM. The former is superior to SVM in cases where training data is more prevalent than the number of features (m>>n). SVM is favoured because there are enormous features and less training details.

In contrast, KNN and Neural networks show that NN needs large training data to realise adequate accuracy. Compared to KNN, the neural network needs some hyper parameter modification.

4.4. Comparison of Decision tree and other Models :

RF is an assembly of decision trees compared to Random Forest, and the forest's preponderance survey is selected as the predicted yield. The model of Random Forest is less susceptible to overfitting and offers a more general interpretation. Compared with decision trees, Random Forest is very safe and accurate. KNN and Decision Tree both use non-parametric methods. Due to KNN’s expensive real-time efficiency, the decision tree is speedier. Decision trees are simple and adaptable. Decision tree trimming can ignore a few key values in training data, which can result in toss accuracy. Decision tree versus neural network, both addressing non-linear systems, and dealing with independent variables. Decision trees are selected when a training set of data includes significant categorical values. Compared to NN, decision trees are much easier when a dilemma is based on choice than explanation. If training data is appropriate, NN beats the decision tree.

In contrast to SVM, the decision tree later employs a kernel trick to illuminate non-linear problems while decision trees infer hyper-rectangles in the input space to illuminate the issue. Decision trees are higher than SVM for categorical and collinear details.

4.5. SVM Compared with Other Models:

SVM supports multi-class classification compared to Random Forest portrays RF, while SVM needs multiple models for the same. Although SVM does not deliver, RF can provide an opportunity beyond expectations. In a better way than SVM, Random Forest handles categorical knowledge. With less training data and comprehensive features, both SVM and Naive Bayes demonstrate superior performance. If features are reciprocally dependent, SVM is better. The former is a model of segregation, while the NB is a model of propagation.SVM suggests a curved optimization work for SVM compared to Neural

Networks, while NN might suspend for local minima. For restricted training data and various functions, SVM performs superior to NN. For adequate precision, NN requires massive training data. For SVM, multi-class classification involves different models, whereas with a single model, NN does.

4.6. Random Forest contrast with other models:

Random Forest is somewhat close to Decision Tree Comparisons. Naive Bayes-related Random Forest shows that random forest is a complex and large model, whereas Naive Bayes could be a moderately smaller model. With little training knowledge, NB performs well, while RF requires a more substantial collection of data preparation.
Neural Networks-related Random Forest shows that both are highly competent and high precision systems. Both have to be clever on the inside and are less rational. For random forests, feature scaling is not needed, although NN requires highlights to be scaled. An ensemble model is used for both. In ML, there are no better models that beat all others; performance depends on the form of distribution of training data.

Table 2: Comparison of Supervised earning algorithm

Algorithm	Regression/ Classification	Results interpretability	Ease of understanding	Average predictive accuracy	Training speed	Prediction speed	Amount of parameter tuning needed (excluding feature selection)	Performance with a small number of observations	Handles lots of irrelevant features well (separates signal from noise)?	Automatically learns feature interactions?	Gives calibrated probabilities of class membership?	Parametric?	Features might need scaling?
KNN	Either	Yes	Yes	Lower	Fast	Depends on n	Minimal	Low	No	No	Yes	No	Yes
Linear regression	Regression	Yes	Yes	Lower	Fast	Fast	None (excluding regularization)	Good	No	No	N/A	Yes	No (unless regularized)
Logistic regression	Classification	Partially	Partially	Lower	Fast	Fast	None (excluding regularization)	Good	No	No	Yes	Yes	No (unless regularized)
Naive Bayes	Classification	Somewhat	Somewhat	Lower	Fast (excluding feature extraction)	Fast	Some for feature extraction	Good	Yes	No	No	Yes	No
Decision trees	Either	Somewhat	Somewhat	Lower	Fast	Fast	Some	Low	No	Yes	Possibly	No	No
Random Forests	Either	A little	No	Higher	Slow	Moderate	Some	Low	Yes (unless noise ratio is very high)	Yes	Possibly	No	No
Neural networks	Either	No	No	Higher	Slow	Fast	Lots	Low	Yes	Yes	Possibly	No	Yes

5. Conclusion

The paper incorporates different standard Machine learning Algorithms, their common properties. No algorithm functions best in all scenarios, which implies, dispersion of training data is the significant criteria for choosing an appropriate algorithm. Few common presumptions can be made on the choice of algorithms, depending on training set measure, feature type, number of highlights, computation and space complication etc. By attempting different ML models with diverse hyperparameters on the information, we get a strong opinion of the Calculations. The paper gives a thought over different ML calculations and their comparisons.

References

[1] Alex S. & Vishwanathan, Introduction to Machine Learning. Published by the press syndicate of the University of Cambridge, Cambridge, United Kingdom. Cambridge University Press (2008.

[2] Osisanwo F.Y ,Supervised Machine Learning Algorithms: Classification and Comparison International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 (2017).

[3] Brighton, H. & Mellish, C., Advances in Instance Selection for Instance-Based Learning Algorithms. Data Mining and Knowledge Discovery University of Aberdeen, vol. 6, no 2, (2002), pp 153–172.

[4] Taiwo, O. A Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.), ISBN: 978-953-307-034-6, InTech, University of Portsmouth United Kingdom, (2010) pp 3 – 31.

[5] Jason Brownlee,Supervised and Unsupervised Machine Learning Algorithms, in Machine Learning Algorithms, (2016).

[6] Friedman, N., Geiger, D. & Goldszmidt M., Bayesian network classifiers. Machine Learning vol .29, (1997). pp 131-163.

[7] Keerthi, S. & Gilbert, E., Convergence of a Generalized SMO Algorithm for SVM Classifier Design, Kluwer Academic Publishers. Machine Learning vol. 46, (2002). pp 351–360.

[8] McSherry, D., Strategic induction of decision trees. Knowledge-Based Systems. vol. 12, no 5–6, (1999), pp 269-275.