COMBINED ALGORITHM BASED ON PREDICTION OF DIABETICS

Dr. G. Satyavathy[1], Mrs. K. Jayanthi Vagini[2], Dr. T. Deepa[3],
[1]Professor & Head, Department of Computer Science with Data Analytics, KPR College of Arts Science and Research, Coimbatore – 07. satyavathy.g@kprcas.ac.in
[2]Assistant Professor, Department of Computer Science with Data Analytics, KPR College of Arts Science and Research, Coimbatore – 07. jayanthivagini.k@kprcas.ac.in
[3]Professor & Head, Department of Computer Science, KPR College of Arts Science and Research, Coimbatore – 07. deepa.t@kprcas.ac.in

Abstract:

When computers are not explicitly instructed to complete a task, they resort to statistical models and machine learning (ML) methods. Learning algorithms are used in a variety of popular programs. Internet search engines like Google owe much of their popularity to an algorithm that has gained how to properly order search results. In this article, we will examine the greater context of artificial intelligence and how supervised learning, unsupervised learning, and recurrent neural networks function within it. It provides a thorough examination of several well-known ML methods, such as the DT, RFA, SVM & BP algorithm. Machine learning is a subfield of artificial intelligence and computer science that aims to simulate human learning using data and algorithms to produce increasingly competent outcomes. The Support Vector Machine (SVM) approach seeks the optimum dividing line (decision boundary) across classes in an n-dimensional space, allowing fresh data to be readily categorized. This optimal decision-making frontier is defined by a hyperplane. Predictions of diabetes are a popular application of machine learning techniques due to their high degree of precision. The decision tree, which excels at classification, is one of the most popular machine-learning methods in medicine. The random forest produces numerous decision trees. The objectives are to expand people's understanding of machine learning and accelerate its spread. In this paper must be research a prediction of diabetics using machine leaning algorithms.

Keywords: SVM, DT, RFA (random forest algorithm), ANN

Introduction:

Since the dawn of time, people have used a variety of tools to make life simpler. Human ingenuity has permitted the development of a huge array of tools. These devices helped human existence by addressing fundamental requirements such as transportation, industry, and data processing. Learning by machines is one of them. In the field of artificial intelligence, the rapid development of science and technology has created whole new fields of study and businesses. Theoretical findings from several fields, such as statistics and algorithm complexity, are applied to computer-based machine technology, therefore boosting the practical capabilities of artificial intelligence. By analyzing various machine learning algorithms, it may be possible to

prepare the way for future machine learning advancement, so enhancing the algorithms' value and making them more conducive to the industry's overall growth and prosperity. When we examine data, we do not always know how to use it. In this instance, machine learning is implemented. The availability of enormous data sets for training have expanded the popularity of machine learning. Various sectors use machine learning to get insights from data. Machine learning's objective is to learn from data. Long ago, scientists began investigating how robots may acquire knowledge without being explicitly programmed. Numerous mathematicians and computer scientists are trying to solve this challenge, which involves incredibly large datasets, by using various approaches.

Literature Review

Pisapia et al. predicted hydrocephalus using image analysis and machine learning. They were able to extract 77 distinct imaging characteristics of the brain using ventriculomegaly. The ventricular characteristics of 25 newborns were examined using the automated learning approach support vector machines. Problems developed when determining which patients really need shunts. The data was gathered and examined. Three out of every four children will need shunts, according to a 75% positive predictive value and a 95% negative predictive value. In this paper, we provide a novel fuzzy rule-based categorization system. An expectation-maximization and fuzzy-rule-based classifier is used to build analytics and cluster creation approaches.

Precision, response time, false positive rate, and computation cost are evaluated between the suggested technique and regularly utilized alternatives. The results indicate promise for the suggested approach, since they outperform Bayes networks, multi-layer models, and decision tables.

Diabetes, its complications, and the spectrum of therapies available were all expected. Map-reduce technology from Hadoop was used to generate informed estimates about which patient subgroups might react better to certain therapies. Hadoop is used to process a large dataset gathered from healthcare institutions, clinics, EHRs, and PHRs, and the resulting data is distributed across many computers throughout the world. Jiang Zheng and Aldo Dagnino conducted a comprehensive literature review on big data and analytics within. This study's primary objective was to apply machine learning techniques to industrial power systems and related applications in order to predict failures and energy consumption.

The scientists proposed a novel approach based on convolutional neural networks for forecasting the risk of many illnesses at the same time. We tested our proposed method using real-world data from hospitals in central China from 2013 to 2015. The major emphasis of the investigation was chronic disorders induced by brain transgressions. According to the trials, the Naive Bayes technique works better on structured data, whereas the recommended approach performs better on data that mixes structured and textual information.

As illustrated by Muhammad Azeem Sarwar et al., predictive analytics in healthcare has the potential to drastically modify how healthcare professionals and researchers utilize data

in their everyday work. In this article, statistical analyses are carried out using six widely-used machine learning techniques. Techniques such as a support vector machine, k-nearest neighbor, logistic regression, deep learning, radio frequency, and neural network (NB) are all at your disposal. Diabetes predictions were provided using data from all 768 publications in the PIMA Indian dataset. There are eight characteristics in each of the prediction model's training and testing sets. Experiments reveal that SVM and KNN are the most efficient diabetes prediction algorithms. Their total accuracy of 77% is much greater than the average of the other four algorithms utilized in this study. As a result, SVM and KNN may be used to predict diabetes.

Machine Learning Algorithms:

There are types of algorithms that is 1. Supervised 2. Unsupervised 3. Reinforcement algorithms

Supervised Algorithm
- ✓ Regressions
- ✓ DT
- ✓ RFA
- ✓ Classify
- ✓ Decision Tree

Unsupervised Algorithm
- ✓ Clustering
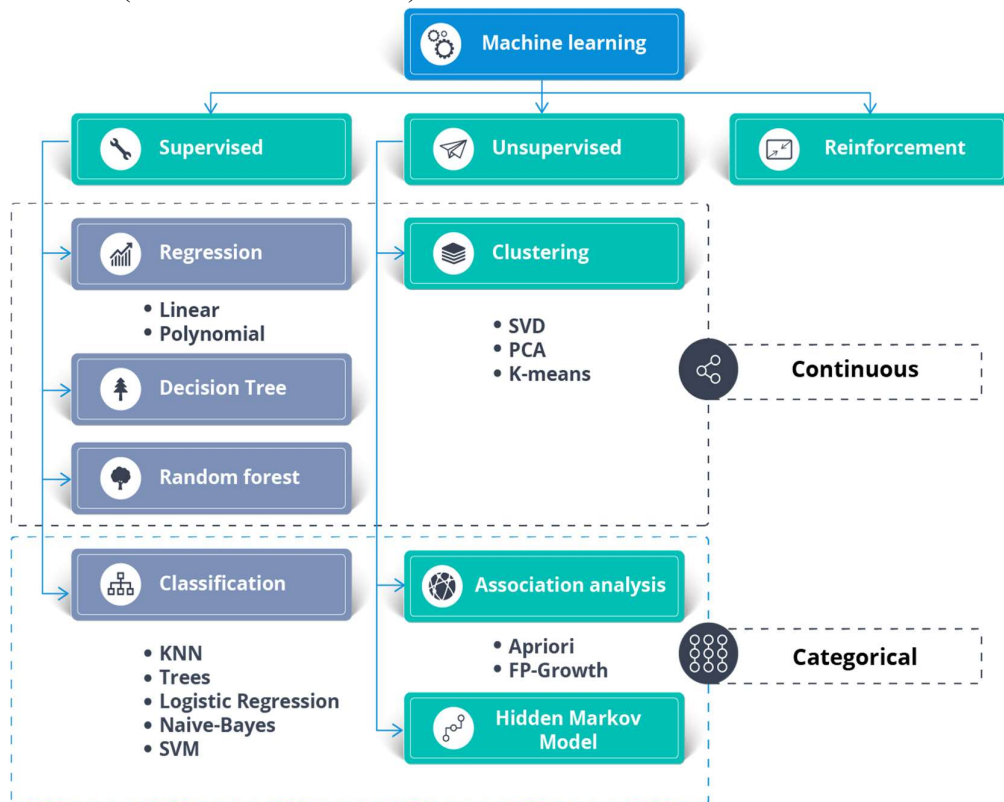- ✓ Association Analysis
- ✓ HMM(Hidden Markov Model)

Fig 1: Machine Learning Algorithms

Supervised Learning

Deep classification is among the most basic methods for machine learning. Learning by design is the process of creating and following certain educational objectives in advance. Earlier on in its learning process, the computer uses data and math to understand what it must know. In order to obtain the required basic data, we want to progressively finish the essential educational materials in a managed environment. Unlike other learning approaches, supervised learning has the capacity to fully harness the inherent general learning capabilities of the machine. When system learning is complete, it may be effective for tackling regression or classification issues that are highly codified. Currently, more traditional learning strategies, like as BN, SVN, KNN, and others, are the most popular. Since the whole learning process does have a goal, deep learning is more regular and the learning material is more regular.

RFT- (Reinforcement) Learning

Learning algorithms are a subset of software development that include both autonomous and supervised learning. As its name suggests, reinforcement learning is a rigorous method of topic mastery. The information from the prior year will be used in the application process. By analyzing feedback information from a particular component, a control signal of data acquisition is created. Lastly, reinforcement learning is a learning approach that extends the application of statistical and dynamic learning via the collection of extra data. This approach is often used to address the difficulty of robot control.

Normally used algorithms in ML

DT-Decision Tree Algorithm

It is based on the theory of relevant data is digested in a forest structure, starting with the collection's root node and proceeding outwards to the intersection of all nodes. A comprehensive analysis of real-world cases. The decision number method will continue to cut down on paths to make data more reliable and easier to understand. The formula is a way to do math that starts at the top and moves down. The content of the cluster is analyzed to identify which qualities are the most important, and the node is then enlarged to accommodate more than two. In data analysis, for example, you may name the most informative decision tree "tree A" before deciding on the maximum allowable branch split. If the maximum allowed is five, for instance, the bigger tree A in the classification will quit splitting and instead use the pruning strategy to study the larger tree model, so improving the data and producing more scientifically sound conclusions.
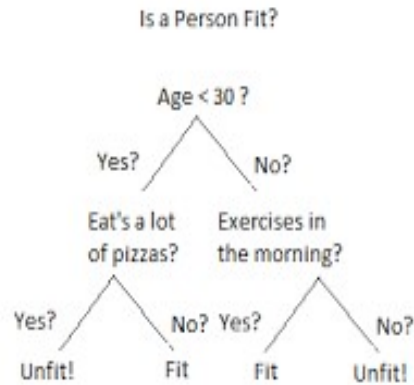
Fig 2: Decision Tree Algorithm

RFA Algorithm

Based on ML technique, the random forest strategy may be utilized to further process data calculations. In reality, the random forest method will be advantageous for managing unreliable data. Multiple decision trees will be developed in parallel as part of the inquiry into the data analysis technique, and the integrated approach will be employed in the regression phase.This distribution is in a condition of random distribution, with each set maintaining its independence. Voting will be conducted on classification data to choose the best solutions. Following a vote, the classification receiving the most votes provides the vector value xi, whose contents are classed to generate the mean score for each condition and serve as the foundation for the final decision.

ANN

The term "artificial neural network" refers to a system that can do tasks that previously required a huge number of neurons by linking individual "data neurons" across the internet in the same manner that humans can remember and learn. However, the method for artificial neural networks is based on this developing data analysis method. The digital units in the designated neurons are genuine to a high degree, and the data may be exported accurately. It's comparable to the forward and backward movements the body makes during running. Self-organizing NN (SOM), MLFN, and ART are important types of artificial neural networks at the present time. Setting the weighting coefficient and output threshold beforehand may facilitate the data analysis and computation processes. When the aggregate of the calculations hits this threshold, a predetermined value is broadcast to the outside world, providing structure to an otherwise chaotic numerical analytic process.

SVM - Support Vector Machine Algorithm

SVM is a common machine learning approach that is used frequently throughout the operation. In order to do the predetermined data analysis, the algorithm often leverages the vector machine method. During this phase, the SVM algorithm will automatically assess the data being processed in order to maximize them. To improve the scientific trustworthiness of the end analysis of the data findings, it is important to collect and study many sets of samples to find the border value's sample data. In the case of H (d), for example, the data is first handled

in a central location using SVM technology before being sent to all the other locations. Second, the edge of the H (d) plane is set by choosing the longest run in the plane. Last but not least, the output vector is computed by evaluating the vector information included in the H (d) plane, resulting in a more precise data analysis.

Naïve Bayes

The word "naive" refers to the idea that the outward appearance of one trait has no effect on the outward appearance of any other traits. For example, you can tell an apple is an apple because it is red, round, and tasty. So, instead of asking another person, we can tell it's an apple by any one of its characteristics. Bayes got its name from the Bayes' which it is based on..

Proposed Methodology

In this proposed method as combined algorithms like support vector machine, decision tree, naïve bayes and artificial neural network give better accuracy for given medical data set. The data set has taken from Kaggle database and UCI repository.
Preprocessing method has to be done a future extraction ,selection and related dataset. After that we apply algorithm for getting result using given dataset.

Algorithm

Step 1: Dataset has taken from different database
Step 2: Preprocessing method can be used for the given data set
Step 3: Apply algorithm for after preprocessing dataset
Step 4: Combined algorithm ANN+NB+SVM used for given dataset
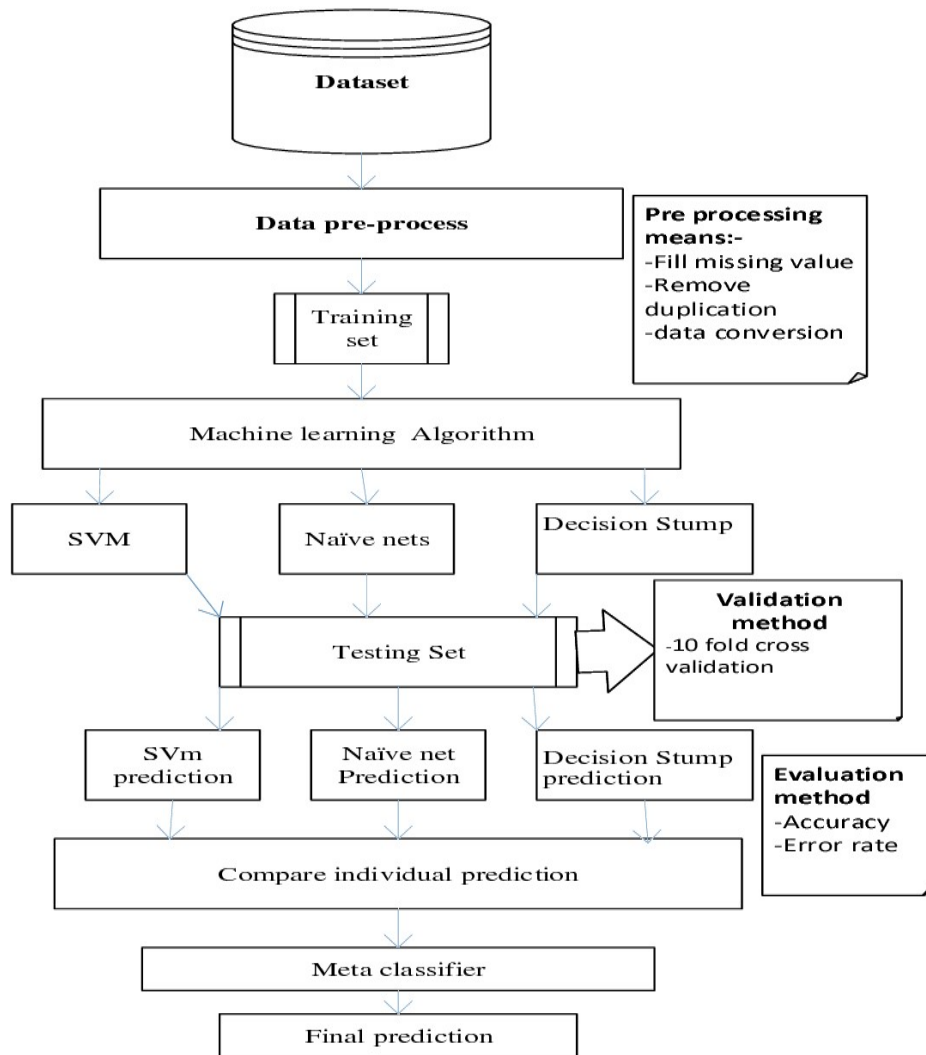Step 5: After algorithm proceed using prediction
Step 6: Result

FIG 3: COMBINED ALGORITHM PERFORMANCE

RESULT AND DISCUSSION

In this experiment, six different machine-learning algorithms were examined. We could use NB, ANN, and SVM. Each of the two data sets, both for training and for testing, encompassed 70% of the total data and 30% of the test data. The three approaches were tested on the same dataset to compare their performance. Equation 1 may be used to avoid accuracy. The accuracy of the algorithm is how much of time it is right.

$$\text{ACCURACY} = (\text{TRUE POSITIVE} + \text{TRUE NEGATIVE}) / (P + N)$$

All that was anticipated true positives and true negatives divided by total positives and negatives. The percentages of valid predictions (TP, TN, FN, and FP) produced by each algorithm are shown in Table 1. TP refers to both diagnosed and predicted diabetic patients. FN validated a diabetes diagnosis despite a poor prognosis. Diabetes was predicted, but there was none. Neither the actual nor anticipated diabetes happened. Tab 1 displays many algorithm comparisons.

S. No	Algorithm	Prediction (%)
1	DT	61
2	LR	67
3	RF	87
4	NB	81
5	KNN	83
6	SVM	87
7	SVM+NB	88
8	SVM+ANN+NB	92

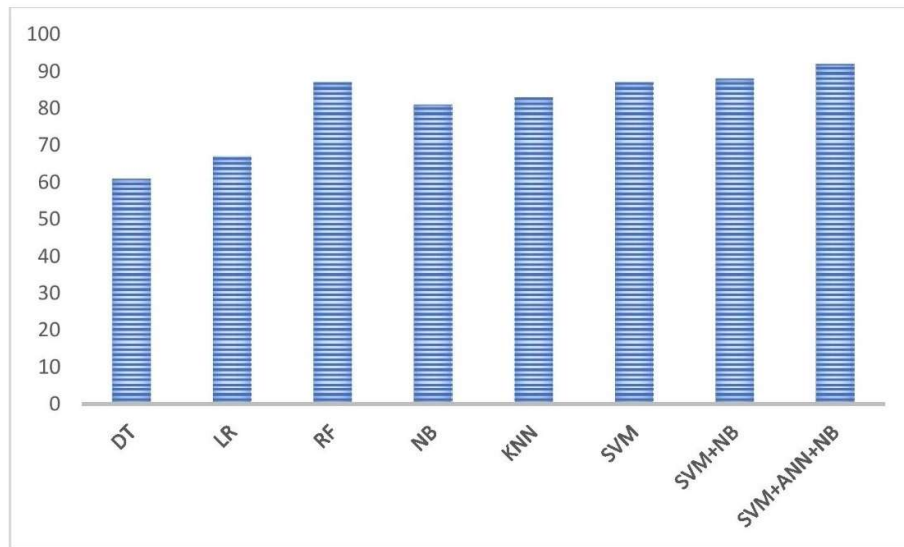**Accuracy chart for combined algorithms**

FIG 4 : ACCURACY CHART

CONCLUSION

In conclusion, machine learning is still in its infancy, relies heavily on supervised learning, and has not yet conquered limited AI. Experts need to strengthen both the theoretical foundations of artificial intelligence and its current practical applications. There is immense promise in machine learning, and it is important that it be fostered in the appropriate scientific and technological domains. The present study uses a method from machine learning based on a combination of support vector machines (SVMs), neural networks (NNs), and artificial neural networks (ANNs) to estimate the likelihood that a patient has diabetes. To build the aforesaid method and improve accuracy, we used 70% training data and 30% testing data.

Reference:

1. Li Kanghua, Jiang Shan. Machine Learning and Cultural Production Reform Based on the Perspective of the Development of AI Technology, Journal of Xiangtan University (Philosophy and Social Sciences), 2020, 44(01):74-79.
2. Wei jin(2020)Research on Machine Learning and Its Algorithms and Development, journal of physics conferences

3. Zhang Run, Wang Yongbin. Research on machine learning and its algorithm and development, Journal of Communication University of China (Natural Science Edition), 2018, 23 (02):10-18 + 24.

4. W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140-0

5. M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine learning techniques in cognitive radios," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1136–1159, Oct. 2012.

6. P. Harrington, "Machine Learning in action", Manning Publications Co., Shelter Island, New York, 2012

7. E. Ahmed et al., "The role of big data analytics in Internet of Things," Comput. Networks, vol. 129, no. December, pp. 459–471, 2017

8. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.

9. L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350–361, May 2017.

10. J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

11. K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings," IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294–1307, Jul. 2016

12. K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization Based on Social Big Data Analysis in the Vehicular Networks," IEEE Trans. Ind. Informatics, vol. 13, no. 4, pp. 1932– 1940, Aug. 2017.

13. Reddy, C. S., Yookesh, T. L., & Kumar, E. B. (2022). A Study On Convergence Analysis Of Runge-Kutta Fehlberg Method To Solve Fuzzy Delay Differential Equations. JOURNAL OF ALGEBRAIC STATISTICS, 13(2), 2832-2838.

14. Kumar, E. B., & Thiagarasu, V. (2017, October). Color channel extraction in RGB images for segmentation. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES) (pp. 234-239). IEEE.

15. Yookesh, T. L., et al. "Efficiency of iterative filtering method for solving Volterra fuzzy integral equations with a delay and material investigation." Materials today: Proceedings 47 (2021): 6101-6104.