For b = 1, …., B:

1.Sample, with replacement, n training examples from X, Y, call these Xu, By.

2.Train a classification or regression tree fb on Xu, By.

After training, predictions for unseen samples *x′* can be made by averaging the predictions from all the individual regression trees on *x′*:

f ^ = 1 B ∑ b = 1 B f b ( x ′ ) {\displaystyle {\hat {f}}={\frac {1}{B}}\sum _{b=1}^{B}f_{b}(x')}

or by taking the majority vote in the case of classification tree.

## K- NEAREST NEIGHBORS

In pattern recognition, the k- nearest neighbours' algorithm (k-nan) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

- · In KNN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k=1, then the object is simply assigned to the class of that single nearest neighbour.
- · In KNN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbours.
- · KNN is a type of instance-based learning or lazy learning, where the function is only approximated locally and computation is deferred until classification. The KNN algorithm is among the simplest of all machine learning algorithms.

## DECISION TREE

A decision tree is a decision support tool that uses a tree-like model of their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operation research, specifically in decision analysis, to help identify a strategy most likely to reach goal, but are also a popular tool in machine learning

A decision tree consists of three types of nodes:

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

## SUPPORT VECTOR MACHINE

In machine learning **support-vector machines** (**SVMs**, also **vector networks**) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model linear that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

## LOGISTIC REGRESSION

In statistics, the **logistic model** (or **logit model**) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from ***log****istic* *un****it***, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probity model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each dependent variable having its own parameter.

## 5. DATASET SELECTION (DIABETES DATASET)

| S No | Attribute | Type |
|---|---|---|
| 1. | Number of times pregnant | Numeric |
| 2. | Plasma glucose concentration | Numeric |
| 3. | Blood pressure (Diastolic) | Numeric |
| 4. | Triceps skin fold thickness(mm) | Numeric |
| 5. | 2-Hour Serum Insulin | Numeric |
| 6. | Body mass index(kg/m^2) | Numeric |

| | | |
|---|---|---|
| 7. | Diabetes pedigree function | Numeric |
| 8. | Age(years) | Numeric |
| 9. | Class Variable ( True or False) | Nominal |

**Table 1. Dataset Information**

## 6.COMPARISON AND RESULTS

| Reference | Proposed model / Method | Dataset Used | Purpose | Accuracy Achieved (%) |
|---|---|---|---|---|
| N. Gupta et.al (2013) | Decision Tree | PIMA Indian Diabetes Data | To predict diabetes | 81.33% |
| P. Yasodha, M.Kannan (2011) | Bayes Net | A hospital repository | To predict diabetes | 66.2% |
| A.Iyer et.al (2015) | Decision Tree | PIMA Indian Diabetes Data | To predict diabetes | 74.8% |
| K.Rajesh, V.Sangeetha (2015) | Decision Tree | PIMA Indian Diabetes Data | To predict diabetes | 87% |
| Lee (2014) | Decision Tree | National Health and Nutrition Examination Survey | To predict diabetes | 67% |

| | | | | |
|---|---|---|---|---|
| Chick et al (2012) | KNN | PIMA Indian Diabetes Data | To predict diabetes | 89.10% |
| Our Proposed Framework | Decision Trees<br>Naïve Bayes<br>KNN(K=1)<br>KNN(K=3) | PIMA Indian Diabetes Data | To model diabetic prediction | 94.44%<br>79.84%<br>93.79%<br>76.79% |

**Table 2. Comparison of algorithms and its accuracy**

| Classifiers | Without Bootstrapping (Accuracy rate %) | After bootstrapping (Accuracy rate %) |
|---|---|---|
| Logistic regression with SVM | 71.45% | 74.89% |
| Decision tree(J48) | 78.43% | 94.4% |
| k-NN(k=1) | 69.93% | 93.79% |
| k-NN(k=3) | 72.2% | 76.69% |

**Table 3.Bootstrapping Accuracy Rate**

## 7. CONCLUSION

There are Various data mining method and its application were studied or reviewed. Application of machine learning algorithm were applied in different medical data sets including machine Diabetes dataset. Machine learning methods have different power in different data set. We obtained 768record diabetes data set from UCI. the comparison of individual algorithm and the proposed method is done on this study. We applying 10 cross validation us for evaluation of the performance of these machine learning classification methods purpose. In this study the proposed method provides high accuracy with accuracy value of 90.36% and decision Stump provided less accuracy than other by providing 83.72% accuracy. Therefore, using ensemble method used to provide better prediction performance or accuracy than single one.

## 8. FUTURE WORK

In this study we concentrated only Diabetes disease for future it can be extended to apply this method in another diseases Small amount sample data used on this study.it can be apply in large amount of data for future extension .on this study also only a single data set used therefore for future multiple data set can be used for prediction .in this study only limited base classifier used .for future it is possible to use another base classifier like ANN, Naive Bayes, KNN, Random tree ,and other.

## REFERENCES

**[1]** Yashoda and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in Waikato", International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2011.

**[2]** A. Ayer, J. S and R. Sumbala, "Diagnosis of Diabetes Using Classification Mining Techniques", IJDKP, vol. 5, no. 1, pp. 01-14, 2015.

**[3]** NIyati Gupta, A. Rawal, and V. Narasimhan, "Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data", IOSR Journal of Computer Engineering, vol. 11, no. 5, pp. 70-73, 2013.

**[4]** M. Chicheme. Said, and N. Setout, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System 2 (AIRS2) with fuzzy K-nearest neighbour," Journal of medical systems, vol.36, no.5, pp. 27212729, 2012.

**[5]** K. Sharmila and S. Manickam, "Efficient Prediction and Classification of Diabetic Patients from big data using R," International Journal of Advanced Engineering Research and Science, vol. 2, Sep 2015.

**[6]** S. Sadhana and S. Savitha, "Analysis of Diabetic Data Set Using Hive and R," International Journal of Emerging Technology and Advanced Engineering, vol. 4, July 2014.

**[7]** Sassanian and G. Hari Sekaran, "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients," International Journal of Science and Research, vol. 4, April 2015.

**[8]** W. Raghunath and V. Raghunath, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2, no. 1, p. 3, 2014.

**[9]** S. Hay, D. George, C. Moyes and J. Brownstein, "Big Data Opportunities for Global Infectious Disease Surveillance", Plops Med, vol. 10, no. 4, p. e1001413, 2013.

**[10]** G. Weber, K. Mandl and I. Keohane, "Finding the Missing Link for Big Biomedical Data", JAMA, 2014.

**[11]** M. Barrett, O. Humbled, R. Hiatt and N. Adler, "Big Data and Disease Prevention: From Quantified Self to Quantified Communities", Big Data, vol. 1, no. 3, pp. 168-175, 2013.

**[12]** S. Rao, S. Suma and M. Sunitha, "Security Solutions for Big Data Analytics in Healthcare", 2015 Second International Conference on Advances in Computing and Communication Engineering, 2015. [16] D. Peter Augustine, "Leveraging big data Analytics and Hadoop in developing India's healthcare services," International Journal of Computer Applications, vol. 89, no. 16, pp. 44–50, 2014.

**[13]** Sivaram, M., B. DurgaDevi, and J. Anne Steffi. "Steganography of two lsb bits." International Journal of Communications and Engineering 1.1 (2012): 2231-2307.

**[14]** Sivaram, M., et al. "Exploiting the Local Optima in Genetic Algorithm using Tabu Search." Indian Journal of Science and Technology 12 (2019): 1.

**[15]** Mohammed, Amin Salih, et al. "DETECTION AND REMOVAL OF BLACK HOLE ATTACK IN MOBILE AD HOC NETWORKS USING GRP PROTOCOL." International Journal of Advanced Research in Computer Science 10.6 (2018).

**[16]** Viswanathan, M., et al. "Security and privacy protection in cloud computing." Journal of Advanced Research in Dynamical and Control Systems (2018): 1704-1710.

**[17]** Nithya, S., et al. "Intelligent based IoT smart city on traffic control system using raspberry Pi and robust waste management." Journal of Advanced Research in Dynamical and Control Systems, Pages (2018): 765-770.

**[18]** Dhivakar, B., et al. "Statistical Score Calculation of Information Retrieval Systems using Data Fusion Technique." Computer Science and Engineering 2.5 (2012): 43-5.

**[19]** Mohammed, Amin Salih, Shahab Wahhab Kareem, and M. Sivaram. "Time series prediction using SRE-NAR and SRE-ADALINE." (2018): 1716-1726.

**[20]** Abraham, Steffin, Tana Luciya Joji, and D. Yuvaraj. "Enhancing Vehicle Safety with Drowsiness Detection and Collision Avoidance."

International Journal of Pure and Applied Mathematics 120.6 (2018): 2295-2310.

**[21]** Porkodi, V., et al. "Survey on White-Box Attacks and Solutions." Asian Journal of Computer Science and Technology 7.3 (2018): 28-32.

**[22]** Malathi, N., and M. Sivaram. "An Enhanced Scheme to Pinpoint Malicious Behavior of Nodes In Manet's." (2015).

**[23]** Sivaram, M. "Odd and even point crossover based Tabu ga for data fusion in Information retrieval." (2014).

**[24]** Sivaram, M., et al. "Emergent News Event Detection from Facebook Using Clustering."

**[25]** Punidha, R. "avithra K, Swathika R, and Sivaram M,"Preserving DDoS Attacks sing Node Blocking Algorithm." International Journalof Pure and Applied Mathematics, Vol. 119, o. 15, 2018." 633-640.

**[26]** Batri, K., and M. Sivaram. "Testing the impact of odd and even point crossover of genetic algorithm over the data fusion in information retrieval." European Journal of Scientific Research (2012).

**[27]** Mohamme, Sivaram Yuvaraj Amin Salih, and V. Porkodi. "Estimating the Secret Message in the Digital Image." International Journal of Computer Applications 181.36 (2019): 26-28.

**[28]** Manikandan, V., et al. "PRIVACY PRESERVING DATA MINING USING THRESHOLD BASED FUZZY CMEANS CLUSTERING." ICTACT Journal on Soft Computing 9.1 (2018).

**[29]** Obulatha-II-ME-CSE, Miss O. "Position Privacy Using LocX."

**[30]** Sivaram, M., et al. "The Real Problem Through a Selection Making an Algorithm that Minimizes the Computational Complexity."

**[31]** Sivaram, M., et al. "DETECTION OF ACCURATE FACIAL DETECTION USING HYBRID DEEP CONVOLUTIONAL RECURRENT NEURAL NETWORK."

**[32]** V. Porkodi, Dr.D. Yuvaraj, Dr. Amin Salih Mohammed, V. Manikandan and Dr.M. Sivaram. "Prolong the Network Lifespan of WirelessSensor Network" (2018): 2034-2038.

**[33]** M, Sivaram, ENABLING ANONYMOUS ENDORSEMENT IN CLOUDS WITH DECENTRALIZED ACCESS CONTROL (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=

**[34]** M, Sivaram, INTEGER WAVELET TRANSFORM BASED APPROACH FOR HIGH ROBUSTNESS OF AUDIO SIGNAL TRANSMISSION (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=

**[35]** M, Sivaram, HEALTHCARE VISIBLE LIGHT COMMUNICATION (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=

**[36]** M, Sivaram, Preserving DDoS Attacks Using Node Blocking Algorithm (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=

health data efficiently.

Based on the paper "Diabetes Disease Prediction Using Machine Learning Algorithms" by Arwatki Chen Lyngdoh et al., the research focuses on analyzing five supervised machine learning algorithms for diabetes prediction. The study achieved a stable and highest accuracy of 76% with the KNN classifier possibly due to its effectiveness in handling the dataset used, while other classifiers also showed stable accuracy above 70%. The paper provides an insight into why specific machine learning classifiers yield varying levels of accuracy and stability. This analysis is crucial for understanding the effectiveness of different algorithms in diabetes prediction, offering a valuable comparison and contrast of these approaches for your report's background section.

The paper "Predicting Diabetes Mellitus With Machine Learning Techniques" by Quan Zou et al. focuses on the application of decision tree, random forest, and neural network algorithms for predicting diabetes mellitus. The researchers used a dataset from hospital physical examinations in Luzhou, China, which included 14 attributes. They employed principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) for feature selection. The study found that random forest achieved the highest accuracy (ACC = 0.8084) when all attributes were used, demonstrating the effectiveness of this algorithm in diabetes prediction. This superior performance is likely due to Random Forest's ability to manage the complexity and interdependencies within the dataset's 14 attributes. It effectively handles overfitting, a common challenge in machine learning, making it more reliable for such medical predictions.

The paper "Classification and prediction of diabetes disease using machine learning paradigm" by Md. Maniruzzaman et al. investigates the application of machine learning classifiers for diabetes prediction. The study used logistic regression for feature selection identifying the most significant predictors for diabetes and compared four classifiers: Naïve Bayes, Decision Tree, Adaboost, and Random Forest. Random Forest outperformed other classifiers with a classification accuracy of 94.25% in a tenfold cross-validation protocol. This superior performance of Random Forest is attributed to its robustness in handling complex datasets, emphasizing its effectiveness in medical predictions like diabetes.

The paper "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster" by N. Yuvaraj and K. R. SriPreethaa examines the use of machine learning algorithms for diabetes prediction, utilizing a Hadoop-based cluster. A Hadoop-based cluster is a network of computers using Apache Hadoop, a framework for processing and analyzing large datasets. It's designed to efficiently handle big data by distributing the workload across multiple machines, making it ideal for complex tasks like healthcare analytics, where large volumes of data are common. The study compares various algorithms like Neural Networks, Support Vector Machine, Decision Tree, Naive Bayes, and Random Forest, focusing on how they perform with large healthcare datasets. This research provides insights into how these algorithms perform under the constraints and capabilities of a Hadoop cluster, offering a different perspective on algorithm efficiency and scalability in handling large volumes of healthcare data.

The paper "Diabetes Mellitus Prediction using Classification Techniques" by Abdulhakim Salum Hassan et al. explores the use of various classification techniques, like Decision Tree, K-Nearest Neighbors, and Support Vector Machines (SVM), in diabetes prediction. Their findings reveal that SVM outperforms the other two, with an accuracy of 90.23%. This high performance is indicative of SVM's robustness in handling the dataset's complexities. Decision Tree and KNN, while effective, did not match the precision offered by SVM. This result highlights SVM's potential as a powerful tool in medical diagnostics, especially in identifying diabetes risks based on various health parameters. The research offers valuable insights into the comparative effectiveness of these machine learning methods, highlighting SVM's superior performance in this context.

## Experiments and Results

In our quest to harness machine learning for diabetes prediction, we meticulously crafted a model using the Decision Tree algorithm. This section outlines the step-by-step methodology employed in our experiment, along with detailed results and visualizations that elucidate our model's performance.

Data Acquisition and Preprocessing: Our dataset was a comprehensive collection of medical records detailing key physiological parameters. The preprocessing phase was critical, involving data cleaning, handling missing values, and normalizing features to a uniform scale to ensure consistent model input. We also employed techniques to address class imbalance, such as synthetic minority oversampling, to provide our model with a balanced representation of outcomes.

Feature Selection and Decision Tree Construction: Feature selection was paramount, as irrelevant or redundant features could skew our model's predictions. We used a combination of statistical tests and domain knowledge to select features that have substantial influence on diabetes outcomes, such as glucose levels and BMI. The Decision Tree was then constructed, with each node representing a feature and each branch representing a decision threshold, culminating in leaves that denote the predictive outcome.

Visual Analysis of the Decision Process: The visual analysis began with plotting the tree structure, offering an intuitive map of the decision paths. Each split was based on the feature that most effectively divided the dataset into subsets, each progressively purer in a particular class. The depth of the tree was fine-tuned to prevent overfitting, striking a balance between model complexity and generalizability.

Confusion Matrix and Interpretation: The confusion matrix provided a straightforward representation of the model's predictions versus the actual outcomes. By dissecting the true positives, true negatives, false positives, and false negatives, we gained insights into the model's sensitivity (true positive rate) and specificity (true negative rate). The matrix revealed that while our model was adept at identifying non-diabetic instances, it required further refinement to improve its detection of diabetic cases, as indicated by the number of false negatives.

Detailed Classification Metrics: The classification report shed light on the precision, recall, and F1-scores for each class. These metrics provided a more granular view of the model's performance beyond mere accuracy. For instance, the F1-score, a weighted average of precision and recall, highlighted the trade-offs between the model's ability to identify all relevant instances and its precision in doing so.

Assessment of Model Accuracy: Our model's accuracy was quantified at 66.88%, a measure of its overall correctness across all predictions. To assess the reliability of this metric, we employed k-fold cross-validation, which further corroborated the model's robustness and its aptitude for generalization beyond the training data.

Visualization of Outcomes: Finally, the significance of data visualization in our experiment cannot be overstated. From histograms to scatter plots, each graphical representation served to make the abstract tangible, linking numerical findings to visual patterns. These visual tools not only facilitated a deeper understanding of the dataset's characteristics but also allowed us to communicate complex results in an accessible manner.

Reflection on Results: The experiments conducted and the results obtained provide a transparent look at the predictive prowess of a Decision Tree in the context of diabetes. While the accuracy and precision are commendable, the model's performance in correctly identifying diabetic instances opens a dialogue for improvement. The visual and numerical analyses together form a compelling narrative on the utility and limitations of our model, setting a foundation for further discussion on its potential impact in healthcare diagnostics.

## Discussion

The deployment of the Decision Tree model in predicting diabetes is a testament to the transformative potential of machine learning in healthcare. This model's ability to discern patterns in medical data and predict outcomes holds promise for advancing early detection methods for diabetes, which is a critical step in proactive healthcare management.

Impact on Early Detection and Preventive Care: Early detection of diabetes is crucial as it allows for timely intervention that can mitigate the disease's progression and associated complications. Our model's practicality comes from its interpretability; healthcare professionals can easily comprehend and explain the model's decisions, enhancing patient-provider communication. This clarity is vital for preventive care strategies, as patients are more likely to engage in their health management when they understand the rationale behind medical advice.

Enhancing Clinical Decision-Making: In real-world healthcare settings, the model could act as an assistive tool for clinicians, providing a preliminary risk assessment that can be refined through further tests and examinations. It is designed to support, rather than replace, clinical judgment, offering a data-informed perspective that complements traditional diagnostic methods.

Integration Challenges and Opportunities: Integrating the model into existing healthcare infrastructures poses both challenges and opportunities. A primary challenge is ensuring compatibility with electronic health record (EHR) systems, requiring the model to be adaptable to the diverse data formats and