**A comparative study of classification models: Support Vector Machine, Decision Tree,**

**K-Nearest Neighbors**

# Project report in partial fulfillment of the requirement for the award of the degree of

# Bachelor of Technology

# In

# COMPUTER SCIENCE AND TECHNOLOGY & COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

**Submitted By**

|  |  |  |  |
| --- | --- | --- | --- |
| Rounak Choudhury |  |  | Enrollment No. 12021002001053 |
| Anindita Saha |  |  | Enrollment No. 12021002023093 |
| Rajanya Sinha Roy |  |  | Enrollment No. 12021002022020 |
| Pritha Paul |  |  | Enrollment No. 12021002012001 |
| Ativi Kumar Das |  |  | Enrollment No. 12021002023023 |
|  |  |  | |

**Under the guidance of**

Prof. Stobak Dutta

Department of Computer Science and Technology & Computer Science and Information Technology

(CST & CSIT)



UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

# CERTIFICATE

This is to certify that the project titled **A comparative study of classification models: Support Vector Machine, Decision Tree, K-Nearest Neighbors s**ubmitted by **Rounak Choudhury(University Roll No. 12021002001053), Anindita Saha(University Roll No. 12021002023093)**, **Rajanya Sinha** **Roy(University Roll No. 12021002022020), Pritha Paul(University Roll No. 12021002012001)** and **Ativi Kumar Das(University Roll No. 12021002023023)** students of UNIVERSITY OF ENGINEERING and MANAGEMENT, KOLKATA, in partial fulfillment of requirement for the degree of Bachelor of Computer Science and Technology and Computer Science and Information technology is a bonafide work carried out by them under the supervision and guidance of **Prof. Stobak Dutta** during 5th Semester of academic session of 2021 - 2025. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of guide

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of Head of Department

# ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us. We are sincerely grateful to our guide Prof. Stobak Dutta of the Department of CST & CSIT, UEM, Kolkata, for her wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now. Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Rounak Choudhury

Anindita Saha

Rajanya Sinha Roy

Pritha Paul

Ativi Kumar Das

# TABLE OF CONTENTS

## ABSTRACT................................................................................................................ 6

## CHAPTER – 1: INTRODUCTION.......................................................................... 7

## CHAPTER – 2: LITERATURE SURVEY

## 2.1 Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer................................................................................................. 8

## 2.2 Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis.......................................................................................................................... 8

## 2.3 PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES……………………………………………………………. 8

## 2.4 A hybrid CNN-KNN model for MRI brain tumor classification……………………………………………………………….. 8

## 2.5 Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm and support vector machine in an electronicnose……………………………………………................................. 8

## 2.6 Classification of hyperspectral remote sensing images with support

## vector machines…………………………………………………………….9

## 2.7 MSVM-kNN: Combining SVM and k-NN for Multi-class Text Classification

## ……………………………………………………………………………… 9

## 2.8 Machine Learning for Smartphone-Based Early Detection of Diabetic Disease in Pima Indians Diabetes Database……………………………………………….. 9

## 2.9 Power Transformer Fault Classification Based on Dissolved Gas Analysis by Implementing Bootstrap and Genetic Programming………………………………………………………………9

## 2.10 An Efficient Prediction of Breast Cancer Data using Data Mining Techniques…………………………………………... 10

**2.11** **Early Prediction of Diabetes Mellitus Using Machine Learning…..10**

**2.12 An experimental study on upper limb position invariant EMG signal classification**

**Based on deep neural networks……………………………………………10**

**2.13** **Tree Species Discrimination in Tropical Forests Using Airborne Imaging Spectroscopy…………………………………………………………………….10**

**2.14 Diabetes Prediction using Machine Learning Techniques…………10**

**2.15 Prediction of Diabetes using Machine Learning…………………………11**

**2.16 Performance Analysis of Machine Learning Approaches in Diabetes Prediction…………………………………………………………………………11**

**2.17 Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM………………………………………………………………………………...11**

**2.18 Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An   
Ensemble Approach………………………………………………………………11**

## CHAPTER – 3: PROBLEM STATEMENT …………………………………12

## CHAPTER – 4: PROPOSED SOLUTION ……………………………….13-14

## CHAPTER – 5: EXPERIMENTAL SETUP AND RESULT ANALYSIS…..15-23

## CHAPTER – 6: CONCLUSION & FUTURE SCOPE……………………....24-25

## BIBLIOGRAPHY .........................................................................................26-28

# ABSTRACT This report presents a comparative study of different machine learning models, namely Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees (D-Tree), applied to a diabetic dataset. The objective of this project is to analyze and compare the performance of these models in predicting accuracy scores. The project begins with exploratory data analysis (EDA) of the diabetic dataset. EDA helps in understanding the structure, patterns, and relationships within the data. It involves data preprocessing, feature engineering, and visualization techniques to gain insights into the dataset. After performing EDA, the dataset is divided into training and testing sets. The models, SVM, K-NN, and D-Tree, are then trained on the training set and evaluated on the testing set. The performance of each model is measured using accuracy scores. The results are analyzed in an analytical manner to compare the performance of the different models. The accuracy scores of each model are examined, and their strengths and weaknesses are identified. This analysis provides valuable insights into the effectiveness of each model in predicting diabetic outcomes. The report concludes with key findings and recommendations based on the comparative analysis of the models. It highlights the advantages and limitations of each model and provides insights into their applicability in diabetic prediction tasks. Overall, this project contributes to the field of machine learning by providing a comparative study of SVM, K-NN, and D-Tree models on a diabetic dataset. The findings and analysis presented in this report can serve as a valuable reference for researchers and practitioners in the field, helping them make informed decisions when choosing the appropriate model for similar classification tasks.

# INTRODUCTION

Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions. There are many types of machine learning models, such as Logistic Regression, Random Forest, Decision Trees (D-Tree), and others. Each model has its own advantages and disadvantages, depending on the problem domain, the data characteristics, and the evaluation criteria. However, there is a lack of comprehensive and accessible literature that compares and contrasts these models and explains how they work and what they can achieve.  
  
In this project, we aim to fill this gap by conducting an analytical study of the results and scores of different machine learning models on various datasets. We will compare the models based on their accuracy, precision, recall, F1-score, and other metrics. We will also explain the underlying principles and assumptions of each model, as well as their strengths and limitations. We will follow these steps in our project:  
  
- Review the existing literature on machine learning models and their applications.  
- Select a dataset for a two class Problem  
- Preprocess the data and perform exploratory data analysis to understand the data distribution and features.  
- Apply different machine learning models to each dataset and tune their hyperparameters.   
- Evaluate the performance of each model using various metrics and compare them using statistical tests or visualizations.  
- Interpret the results and discuss the implications and challenges of using different machine learning models.  
  
By doing this project, we hope to provide a clear and concise overview of the main machine learning models and their capabilities and limitations. We also hope to inspire further research and innovation in this field.

**LITERATURE SURVEY**

1. **IN LITERATURE [1]**- Implemented a breast tumor classification model using Decision Tree and K-Nearest Neighbors algorithms in Python with scikit-learn. Loaded and preprocessed data, trained models, and evaluated accuracy, classification report, and confusion matrix. Emphasized data preprocessing, parameter tuning, and considering additional metrics for optimal performance. Breast tumor classification, utilizing Decision Tree and K-Nearest Neighbors on the Wisconsin Diagnostic Breast Cancer dataset. Employed Principal Component Analysis for feature selection. Comparative analysis reveals KNN's superior performance over the decision-tree classifier in distinguishing between benign and malignant cases, assessed through standard metrics.
2. **IN LITERATURE [2]**- Proposed a vertical sample-based KNN/LSVM classification method for high-dimensional gene expression data, optimizing weights with genetic algorithms. Achieved high accuracy and efficiency by combining majority voting, boundary approach, and P-tree representation. Demonstrated robustness to noise and aims to apply the approach to large-scale time series data for scalable analysis.
3. **IN LITERATURE [3]**- Utilized Machine Learning in healthcare to protect against heart diseases and locomotor disorders. Employed an online UCI dataset with 303 rows and 76 properties, selecting 14 for testing. Applied supervised learning methods (Naive Bayes, SVM, Logistic regression, Decision Tree, Random Forest, KNN) with KNN (eight neighbors) showing superior effectiveness, sensitivity, precision, accuracy, and F1-score compared to other methods.
4. **IN LITERATURE [4]**- Proposed a hybrid CNN-KNN model for MRI brain tumor classification, integrating CNN for feature extraction and KNN for prediction. Achieved a 96.25% accuracy on BraTS datasets, outperforming traditional classifiers like CNN, SVM, and ensemble models. The hybrid model combines the advantages of CNN and KNN, providing automatic feature extraction, reduced complexity, and promising results. Presented a Hybrid CNN-KNN model for MRI brain tumor classification using BraTS datasets. The model automatically extracts non-handcrafted features via CNN, enhancing efficiency. Evaluated against various classifiers, it combines the strengths of CNN and KNN, offering promising results with 96.25% accuracy. The hybrid model reduces complexity and proves advantageous over traditional classifiers, showcasing potential for MRI tumor classification.
5. **IN LITERATURE [5]**- Studied an electronic nose (e-nose) for n-butanol concentration sensing using 12 metal oxide gas sensors. Employed multiclass SVM and k-NN algorithms, enhancing their performance with a decision tree structure for sensor subset selection. Achieved improved classification success, demonstrating a rise from 87% to 93% (k-NN) and 86% to 96% (SVM) through the proposed approach. Investigated e-nose for n-butanol concentration using 12 gas sensors. Applied multiclass SVM and k-NN with a decision tree for sensor subset selection, improving classification success. Results showed enhanced accuracy, rising from 87% to 93% (k-NN) and 86% to 96% (SVM) using the proposed approach.

# IN LITERATURE [6]- This paper explores hyperspectral remote sensing image classification using support vector machines (SVMs). It presents theoretical and experimental analyses to evaluate SVMs in hyperdimensional feature spaces, comparing their effectiveness with conventional feature-reduction methods and assessing performance in various hypersubspaces. The study includes comparisons with radial basis function neural networks and K-nearest neighbor classifiers. Additionally, the application of binary SVMs to multiclass problems is examined through four strategies, with performance indicators like classification accuracy, computational time, parameter stability, and multiclass architecture complexity. Results on a real hyperspectral dataset suggest that SVMs offer a valid and effective alternative to traditional pattern recognition approaches.

1. **IN LITERATURE [7]-** Surface electromyography (sEMG) signal classification is crucial for man-machine interfaces, particularly in prosthetic control with multiple degrees of freedom. This article explores a Deep Neural Network (DNN) based classification system for upper limb position invariant myoelectric signals, comparing it with existing tools. The study utilizes sEMG data from eleven subjects across five limb positions, employing time domain power spectral descriptors (TDPSD) as features for a fully connected feed-forward DNN model, eliminating feature dimensionality reduction to enhance overall simplicity.

1. **IN LITERATURE [8]-** Our research aims to detect diabetes early, crucial for preventing complications like kidney and heart failure. We employ an ensemble of machine learning algorithms on diverse datasets, focusing on features extraction for accurate prediction. Results, analyzed in Jupyter Notebook, reveal Random Forest as the most accurate, with an 89.58% classification accuracy, surpassing other methods. Future enhancements could incorporate medical cell image databases and leverage emerging technologies like IoT, cloud computing, and blockchain for a secure and portable automatic diabetes detection device.

1. **IN LITERATURE [9]-** There are various application of hand written character recognition. The optical character recognition of the documents is comparing with handwriting documents by a human. This OCR is used for translation purposes of characters from various typesof files such as image, word document files. This research article consisting of various approaches to compare and recognize the handwriting characters from the image documents. Thus SVM (Support Vector Machine) has increased the OCR system performance by providing good result of 91% accuracy to recognize characters from documents.Breast cancer is a leading cause of women's mortality worldwide, necessitating early detection. This paper compares data mining classifiers on the Wisconsin Breast Cancer (WBC) dataset, focusing on classification accuracy. Experimenting with six techniques, Support Vector Machine (SVM) emerges with higher prediction accuracy than others. The study underscores SVM's suitability in breast cancer prediction, recommending its use for similar classification challenges due to its precision, computational efficiency, and superior performance compared to other classifiers.

1. **IN LITERATURE [10]-** Diabetes mellitus, a harmful disease resulting from insulin resistance, impacts vital organs and requires early detection. Leveraging machine learning for predictive analysis in healthcare, this study develops a model for early diabetes prediction using four classification algorithms: Linear Discriminant Analysis, K-nearest neighbor, Support Vector Machine, and Random Forest. The Pima Indian Diabetes Database is employed for experimental analysis, demonstrating that Random Forest achieves the highest accuracy of 87.66%, surpassing other algorithms in predicting early-stage diabetes, contributing to personalized healthcare and improved patient diagnosis.
2. **IN LITERATURE [11]-** Surface electromyography (sEMG) signal classification is vital for man-machine interfaces controlling prosthetic devices. This article presents a detailed exploration of a Deep Neural Network (DNN) based classification system for myoelectric signals, focusing on upper limb positions. Using a dataset from eleven subjects, the fully connected DNN model, trained on time domain power spectral descriptors, achieves an average accuracy of 98.88%. Customizing the DNN for each subject enhances results, outperforming k-Nearest Neighbour, Random Forest, and Decision Tree classifiers. The study highlights the DNN's competitive performance, requiring minimal feature engineering compared to state-of-the-art SVM models.
3. **IN LITERATURE [12]-**Diabetes mellitus, marked by insulin resistance, poses severe health risks if not detected early. Leveraging machine learning for predictive analysis, this study aims to predict diabetes using four classification algorithms: Linear Discriminant Analysis, K-nearest neighbor, Support Vector Machine, and Random Forest. Utilizing the Pima Indian Diabetes Database, the proposed model achieves close results to clinical outcomes, aiding personalized patient diagnosis. Results indicate Random Forest's superior accuracy of 87.66%, outperforming other algorithms, emphasizing its potential in early-stage diabetes prediction for improved healthcare outcomes.
4. **IN LITERATURE [13]-**In a Hawaiian tropical forest, we utilize airborne hyperspectral imagery from the Carnegie Airborne Observatory-Alpha system to identify canopy species through supervised classification. Comparing nonparametric methods (support vector machines, artificial neural network, k-nearest neighbor) and parametric methods (discriminant analysis), we find regularized discriminant analysis, linear discriminant analysis, and support vector machines advantageous. Improved support vector machine optimization is suggested. Combining spectral and spatial information enhances accuracy, demonstrated through regularized discriminant analysis. The study affirms the feasibility of species mapping in tropical forests using high-fidelity imaging spectroscopy, emphasizing the importance of classifier choice and optimization.
5. **IN LITERATURE [14]-**Diabetes, resulting from elevated glucose levels, poses serious health risks if untreated, affecting the heart, kidneys, blood pressure, eyes, and other organs. Early prediction is crucial for effective control. This project employs various machine learning techniques, including K-Nearest Neighbor, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Random Forest, on a dataset to predict diabetes. Random Forest demonstrates superior accuracy, indicating its effectiveness in early diabetes prediction compared to other models.
6. **IN LITERATURE [15]-**Machine learning, a branch of artificial intelligence, has significantly impacted medical sciences, particularly in disease diagnosis. Diabetes, a prevalent global issue, is linked to elevated blood sugar levels. Unhealthy lifestyles contribute to its rapid growth, causing various complications. Early detection is crucial to prevent severe consequences. This research focuses on classifying individuals with diabetes using algorithms such as Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, and Decision Tree. The experiment reveals that the KNN algorithm outperforms others, achieving an accuracy of 85%.
7. **IN LITERATURE [16]-** Explored Chronic Kidney Disease (CKD) prediction using data mining. Emphasized early detection's importance to prevent complications. Compared Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) classifiers for accuracy, precision, and execution time. Results favoured KNN, demonstrating its superior performance over SVM in CKD prediction.
8. **IN LITERATURE [17]**- Addressing the rising prevalence of Diabetes Mellitus (DM), this study employs Machine Learning Techniques (MLTs) for early diagnosis and prediction. Utilizing Pima Indian Diabetes Dataset and 130\_US hospital diabetes data sets, the system employs Random Forest, KNN, Naïve Bayes, and J48 techniques, achieving an impressive 93.62% accuracy for PIDD and 88.56% for the 130\_US hospital dataset.
9. **IN LITERATURE [18]-** The research explores machine learning (ML) techniques, including Logistic Regression, Decision Tree, XGBoost, SVM, KNN, and Random Forest, on the PIMA Indian Diabetes Dataset for early diabetes prediction. XGBoost outperformed others with 80.73% accuracy, followed by SVM at 80.21%. The study suggests ML's potential in aiding early diabetes detection, paving the way for future comparisons with Deep Learning techniques.

# PROBLEM STATEMENT

# The field of machine learning has witnessed remarkable advancements, leading to the development of various models for solving classification problems. Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees (D-Tree) are among the most widely used models in the machine learning community. However, there is a lack of comprehensive comparative papers that provide a detailed understanding of these models and their performance in classification tasks. The absence of such comparative studies hinders researchers and practitioners from making informed decisions about which model to choose for their specific problem. This knowledge gap poses a challenge as it limits the ability to select the most suitable model and understand the nuances of the models' performance on different datasets. To address this problem, our project aims to conduct an analytical study comparing SVM, K-NN, and D-Tree models. We intend to delve into the workings of these models and evaluate their performance on a diabetic dataset. By doing so, we aim to provide a comprehensive understanding of these models and their effectiveness in classifying diabetic outcomes. The absence of comparative papers that highlight the functionality and performance of SVM, K-NN, and D-Tree models raises several questions. Such questions include how SVM, K-NN, and D-Tree models differ in terms of their underlying principles and algorithms, what strengths and weaknesses each model exhibits in handling classification tasks, how these models perform in predicting diabetic outcomes, whether there are significant differences in the accuracy scores achieved by these models on the diabetic dataset, and which model proves to be the most effective in classifying diabetic outcomes based on the dataset used in this project. By undertaking this project, we aim to answer these questions and contribute to the understanding of these models' behaviour. Our findings will help researchers and practitioners gain insights into the performance of SVM, K-NN, and D-Tree models, enabling them to make well-informed decisions when selecting the most appropriate model for classification tasks. Overall, our project seeks to bridge the gap in comparative knowledge by providing an analytical study of SVM, K-NN, and D-Tree models. This will aid in enhancing the understanding of these models and their applicability in the context of diabetic prediction, ultimately advancing the field of machine learning.

**PROPOSED SOLUTION**

To address the lack of comprehensive comparative papers on the functionality and performance of Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Decision Trees (D-Tree) models, we have devised a systematic approach for conducting an analytical study. The following steps outline our solution to the problem:  
  
~ Data Collection: We collected a diabetic dataset from credible and reliable sources. The dataset contains relevant features such as glucose level, blood pressure, and BMI, along with corresponding labels indicating the presence or absence of diabetes. This dataset serves as the foundation for our project.  
  
~ Exploratory Data Analysis (EDA): We performed EDA to gain insights into the dataset. This involved data preprocessing, handling missing values, and exploring statistical measures such as mean, standard deviation, and correlation. Visualizations like histograms, scatter plots, and box plots were used to understand the distribution and relationships between variables.  
  
~ Model Selection: Based on the problem statement, we chose SVM, K-NN, and D-Tree models for comparison. These models were selected due to their popularity and effectiveness in classification tasks. Each model has its own underlying principles and algorithms.  
  
~ Data Preprocessing: Before training the models, we preprocessed the dataset. This involved scaling the numerical features to a standard range and encoding categorical variables if present. Data normalization and feature scaling ensure fair comparison and prevent any bias towards specific models.  
  
~ Model Training and Evaluation: We split the preprocessed dataset into training and testing sets. Each model was trained on the training set using appropriate algorithms and hyperparameters. Cross-validation techniques, such as k-fold cross-validation, were employed to ensure robustness of our results. The trained models were then evaluated on the testing set to obtain accuracy scores.  
  
~ Comparative Analysis: We compared the accuracy scores achieved by the SVM, K-NN, and D-Tree models. Statistical analysis was performed to determine if there were significant differences in the performance of these models on the diabetic dataset. We also analyzed the strengths and weaknesses of each model in handling the specific classification task.  
  
~ Result Interpretation: We interpreted the results obtained from the comparative analysis. We discussed the performance of each model and highlighted their advantages and limitations. We identified the model that achieved the highest accuracy score on the diabetic dataset and provided insights into the reasons behind its performance.  
  
~ Conclusion and Recommendations: Based on the comparative analysis, we drew conclusions about the effectiveness of SVM, K-NN, and D-Tree models in classifying diabetic outcomes. We provided recommendations on the most suitable model for diabetic prediction based on the dataset used in this project.  
  
By following this solution, we aimed to bridge the knowledge gap surrounding the functionality and performance of SVM, K-NN, and D-Tree models. Our systematic approach allowed us to conduct an analytical study, providing valuable insights into these models and their applicability in the domain of diabetic prediction. We also gained insights into the working mechanisms and performance trade-offs of each model. Our project contributes to the field of machine learning by providing a clear and detailed comparison of the most popular and widely used models in various domains and applications.

Our project addresses the dearth of comprehensive comparative studies elucidating the functionalities, behaviours, and performance metrics of Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Decision Trees in the realm of machine learning.

**EXPERIMENTAL SETUP**

The Diabetes Health Indicators Dataset contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes. The dataset contains 253680 instances of patients with 21 features consist of some demographics, lab test results, and answers to survey questions for each patient. The target variable for classification is whether a patient has diabetes, is pre-diabetic, or healthy. The CDC funded the creation of the dataset. Each row represents a person participating in this study. There are no Missing Values in these dataset. The dataset represents two class problem consisting of categorical and integer features. Data pre-processing was performed with basic exploratory data analysis. **Diabetes\_binary** is the targeted binary input. It defines **0=no diabetes and 1=prediabetes and diabetes**. **High BP** is another feature which describes **0=no High BP and 1= High BP**. Then we have **HighChol** feature which describes **0=no high cholesterol, 1=high cholesterol**. **CholCheck** is an another feature which determines **0= no cholesterol check in 5years ,1= yes cholesterol check in 5 years**. **BMI** is a feature which is used to measure Body Mass Index. Another variable name **Smoker** feature determines if you have **smoked at least 100 cigarettes in your entire life**. Thus **0 determines no and 1 determines yes**. **Stroke** feature determines **0= no ,1= yes**. The **Heart Disease or Attack** feature determines coronary heart disease (CHD)or myocardial infarction (MI) where **0=no and 1= yes**. **PhysActivity** determines the **physical activity in past 30 days –not including job where 0 = no and 1 = yes**. **Fruits** Feature determines **Consume Fruit 1 or more times per day where 0 = no 1 = yes. Veggies** Feature specifies **Consume Vegetables 1 or more times per day where 0 = no 1 = yes**. **HvyAlcoholConsump** Feature determine**s heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) where 0 = no 1 = yes**. **AnyHealthcare** feature shows if **anyone have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc, where 0 = no 1 = yes**. **NoDocbcCost** feature determines **whether there was time in the past 12 months when you needed to see a doctor but could not because of cost where 0 = no 1 = yes**. **GenHlth** Feature determines your general your health in a **scale 1-5 where 1 = excellent, 2 = very good, 3 = good ,4 = fair ,5 = poor**. **MentHlth** Feature shows about your mental health, which **includes stress, depression, and problems with emotions on a scale 1-30 days**. **PhysHlt** feature nowthinks about your physical health, which includes **physical illness and injury, for how many days during the past 30 days was your physical health not good on a scale 1-30 days**. **DiffWalk** Feature determines if there have any **serious difficulty walking or climbing stairs where 0 = no 1 = yes**. **Sex** Feature determines sex of a person **where 0 = female 1 = male.** **Age** feature specifies age **13-level age category 1 = 18-24 9 = 60-64 13 = 80 or older. Education** Feature determines **Education Level scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate).** **Income** feature determines **Income feature scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more**. The dataset is splitted into 70 :30 for training and testing sets.

**Support Vector Machine:**

The scikit-learn SVM algorithm is a prominent choice for both classification and regression tasks in machine learning. It operates on the principle of determining an optimal decision boundary, termed a hyperplane, between distinct classes. This hyperplane aims to maximize the margin between the closest points of each class, a process known as maximum margin classification. The algorithm identifies support vectors—data points closest to the decision boundary—crucial in defining the hyperplane's position and orientation. By maximizing the margin between these support vectors, the SVM algorithm enhances model generalization.

Support Vector Machine have **several advantages** over other machine learning algorithms, such as:

• Decision Boundary and Support Vectors Plot: Output graphs, such as the decision boundary, visually showcase class separation, aiding in evaluating the model's classification accuracy. The support vectors plot illustrates the distribution of these crucial points across classes, offering insights into their impact on the decision boundary and the overall model performance.

• High-Dimensional Data Handling: SVM excels in managing datasets with numerous features without compromising performance. It maintains robustness against overfitting, a desirable trait for tasks where generalization to unseen data is critical.

• Kernel Functions Flexibility: Supports various kernel functions (linear, polynomial, RBF, sigmoid), allowing customization based on data characteristics. This versatility enables handling both linear and non-linear classification tasks effectively.

• Types of Classifications: scikit-learn's SVM offers svc (Support Vector Classifier) and svr (Support Vector Regressor). Additionally, multi-class SVM classifiers like NuSVC and LinearSVC extend SVM's binary classification approach to handle multiple classes efficiently and accurately.

On the other hand, Support Vector Machine also have **some limitations** compared to other algorithms:

• Computational Complexity: Training time for SVM, especially with large datasets, can be considerably longer compared to other algorithms, posing a challenge in real-time applications or scenarios requiring swift model deployment.

• Sensitivity to Data Characteristics: SVM's performance might degrade with noisy or overlapping classes as it heavily relies on clear class separation for optimal results.

In summary, while scikit-learn SVM excels in handling high-dimensional data, robustness against overfitting, and flexibility in kernel choices, its computational complexity and sensitivity to data characteristics warrant careful consideration in application, particularly in scenarios involving large or noisy datasets.

Top of Form

**Decision Tree:**

D tree, short for decision tree, is a popular machine learning algorithm used for both classification and regression tasks. It is a tree-like model where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents an outcome or prediction. Decision trees are easy to understand and interpret, making them a versatile tool in the field of data analysis and predictive modelling.

Decision trees consist of nodes, branches, and leaf nodes. The root node is the topmost node in the tree, representing the starting point of the decision-making process. Each internal node corresponds to a feature or attribute, and the branches represent the possible values or decisions that can be made based on that feature. Finally, the leaf nodes provide the final predictions or outcomes.

Decision trees can handle both categorical and numerical data. When dealing with categorical features, the algorithm splits the data based on the different categories. For numerical features, the algorithm selects a threshold value and splits the data into two groups based on whether the feature value is above or below the threshold. The process of constructing a decision tree involves recursively splitting the data based on the selected features and values until a stopping criterion is met.

Decision trees can handle both categorical and numerical data. When dealing with categorical features, the algorithm splits the data based on the different categories. For numerical features, the algorithm selects a threshold value and splits the data into two groups based on whether the feature value is above or below the threshold. The process of constructing a decision tree involves recursively splitting the data based on the selected features and values until a stopping criterion is met.

The working mechanism of decision trees can be summarized in three steps: feature selection, splitting, and prediction.

First, the algorithm evaluates different features and selects the one that provides the best split. This is often done using metrics such as Gini index or information gain. The selected feature becomes the root node of the tree.

Next, the algorithm splits the data based on the selected feature. If the feature is categorical, the data is divided into different subsets based on the categories. If the feature is numerical, the data is split based on the threshold value.

Finally, the process is repeated recursively on each subset of data until a stopping criterion is met. This criterion could be a maximum tree depth, a minimum number of samples per leaf, or any other user-defined condition. Once the tree is constructed, predictions can be made by traversing the tree from the root node to the appropriate leaf node based on the feature values of the input data. The algorithm evaluates the available features and selects the one that provides the best split.

Decision trees have **several advantages** over other machine learning algorithms, such as:

* Interpretability: Decision trees are easy to understand and interpret, making them a valuable tool for gaining insights from data. The decision rules and splits in the tree can be visualized and explained in a straightforward manner.
* Handling Nonlinear Relationships: Decision trees can handle nonlinear relationships between features and target variables without the need for complex transformations or feature engineering.
* Handling Missing Data: Decision trees can handle missing data by simply skipping over the missing values during the splitting process. This eliminates the need for imputation techniques.
* Versatility: Decision trees can be used for both classification and regression tasks, making them a versatile tool in the field of data analysis. Decision trees can handle both categorical and numerical data, and they can be easily extended to handle multiclass classification and multi-output regression.
* Robustness: Decision trees are robust to outliers and noise in the data. Outliers and noisy data points have minimal impact on the overall structure of the tree, making decision trees a reliable choice for analyzing real-world datasets.

On the other hand, decision trees also have **some limitations** compared to other algorithms:

* Overfitting: Decision trees are prone to overfitting, especially when the tree becomes too deep and complex. This can lead to poor generalization on unseen data.
* Instability: Decision trees are sensitive to small changes in the data, which can result in different tree structures. This instability can make decision trees less robust than other algorithms.
* Bias towards Features with More Categories: Decision trees tend to favor features with more categories or levels during the splitting process, which can lead to biased predictions.
* Lack of Global Optimization: Decision trees make local decisions at each node without considering the global structure of the data. This lack of global optimization can result in suboptimal splits and suboptimal overall performance.
* Difficulty in Capturing Complex Relationships: While decision trees can handle nonlinear relationships, they may struggle to capture complex relationships that require high-order interactions or transformations. In such cases, other algorithms like neural networks or support vector machines may be more suitable.

**K Nearest Neighbors :**

The K Nearest Neighbors (KNN) algorithm is a powerful machine learning technique that is widely used for classification and regression tasks. It is a non-parametric algorithm that makes predictions based on the similarity of data points. The KNN algorithm is particularly effective when dealing with large datasets and can provide accurate results even in the presence of noisy or incomplete data.

The algorithm classifies a new data point by examining its K nearest neighbors in the training dataset. The value of K is a user-defined parameter that determines the number of neighbors to consider. Once the K nearest neighbors are identified, the algorithm assigns the new data point to the majority class among its neighbors.

The KNN algorithm works by measuring the distance between data points in a multi-dimensional feature space. The most commonly used distance metric is the Euclidean distance, which calculates the straight-line distance between two points. However, other distance metrics such as Manhattan distance and Minkowski distance can also be used.

To classify a new data point, the algorithm calculates the distance between the new point and all the points in the training dataset. It then selects the K nearest neighbors based on the calculated distances. The algorithm assigns the new data point to the class that is most frequent among its K nearest neighbors.

The **working mechanism** of the KNN algorithm can be summarized in the following steps:

* Load the training data into memory.
* For each new data point, calculate the distance between the new point and all the points in the training dataset.
* Select the K nearest neighbors based on the calculated distances.
* Assign the new data point to the class that is most frequent among its K nearest neighbors.

The KNN algorithm offers **several advantages** that make it a popular choice among data scientists and machine learning practitioners. Some of the key advantages include:

* Simplicity: The KNN algorithm is easy to understand and implement. It does not make any assumptions about the underlying data distribution, making it suitable for a wide range of problems.
* Non-parametric: KNN is a non-parametric algorithm, which means that it does not make any assumptions about the functional form of the data. This makes it more versatile and flexible compared to parametric algorithms.
* Robust to noisy data: KNN can handle noisy or incomplete data effectively. It does not rely on the assumption of linear separability and can work well with complex decision boundaries.
* No training phase: Unlike many other machine learning algorithms, KNN does not require an explicit training phase. The algorithm simply stores the training data in memory and uses it for classification or regression tasks.

## The few disadvantages of KNN algorithm includes:

* Computational complexity: KNN can be computationally expensive, especially when dealing with large datasets. The algorithm needs to calculate the distance between the new data point and all the points in the training dataset, which can be time-consuming.
* Sensitivity to feature scaling: KNN is sensitive to the scale of the features. If the features have different scales, the algorithm may give more weight to features with larger values. It is important to normalize or standardize the features before applying the KNN algorithm.
* Curse of dimensionality: KNN can suffer from the curse of dimensionality, which refers to the problem of having a high-dimensional feature space. As the number of dimensions increases, the distance between points becomes less meaningful, making it harder for KNN to find meaningful neighbors.
* Computational complexity: As mentioned earlier, KNN can be computationally expensive, especially when dealing with large datasets. The algorithm needs to calculate the distance between the new data point and all the points in the training dataset, which can be time-consuming.
* Sensitivity to feature scaling: KNN is sensitive to the scale of the features. If the features have different scales, the algorithm may give more weight to features with larger values. It is important to normalize or standardize the features before applying the KNN algorithm.
* Lack of interpretability: KNN is often considered a black box algorithm, making it difficult to interpret the results and understand the underlying decision-making process. This lack of interpretability can be a limitation in certain applications where explainability is crucial.

To **improve the accuracy and performance** of the KNN algorithm:

* Feature selection: By selecting relevant features and discarding irrelevant ones, the dimensionality of the feature space can be reduced, leading to improved performance and reduced computational complexity.
* Feature scaling: Scaling the features to a common range can help to mitigate the sensitivity of KNN to feature scaling. Techniques like normalization and standardization can be used to ensure that all features are on a similar scale.
* Cross-validation: Cross-validation is a technique used to evaluate the performance of a model on unseen data. By dividing the dataset into multiple subsets and training the model on different combinations of these subsets, the performance of the KNN algorithm can be assessed more accurately.
* Choosing the optimal value of K: The value of K has a significant impact on the performance of the KNN algorithm. It is important to choose a suitable value of K that balances between overfitting and underfitting. This can be done through cross-validation or other performance evaluation techniques.
* Handling imbalanced datasets: If the dataset is imbalanced, meaning that some classes have significantly fewer samples than others, the KNN algorithm may be biased towards the majority class. Techniques like oversampling or undersampling can be used to balance the dataset and improve the performance of the algorithm.
* Handling missing data: KNN does not handle missing data by default. It is important to handle missing values in the dataset before applying the KNN algorithm. Techniques like imputation or deletion can be used to handle missing data effectively.

## The few advantages of KNN algorithm includes:

* Versatility: KNN can be used for both classification and regression tasks, making it a versatile algorithm that can be applied to a wide range of problems.
* No assumptions about data distribution: KNN does not make any assumptions about the underlying data distribution, making it suitable for both linear and non-linear problems.
* Incremental learning: KNN supports incremental learning, which means that it can update the model with new data points without retraining the entire model. This makes it suitable for online learning scenarios where new data is continuously available.

# RESULT AND ANALYSIS

Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions. There are many types of machine learning models, such as support vector machines (SVM), Random Forest, decision trees (D-TREE), and others. Each model has its own advantages and disadvantages, depending on the problem domain, the data characteristics, and the evaluation criteria. However, there is a lack of comprehensive and accessible literature that explains how these models work and how they compare with each other.  
  
In this project, we aim to fill this gap by conducting a comparative study of three popular machine learning models: SVM, KNN, D-TREE. We use a diabetic dataset from the UCI Machine Learning Repository, which contains 2,53,680 instances of patients with 21 attributes, such as age, blood pressure, glucose level, etc. The target variable is whether the patient has diabetes or not. We apply exploratory data analysis (EDA) to understand the data distribution, the correlation among variables, and the potential outliers. We then preprocess the data by scaling, encoding, and splitting it into training and testing sets.  
  
We train each model using the training set and evaluate its performance using the testing set. We use accuracy as the main metric to measure how well the model predicts the correct class label for each instance. We also use other metrics, such as precision, recall, and F1-score, to assess the model's ability to identify positive (diabetic) and negative (non-diabetic) cases. We compare the results of the four models using tables and graphs, and analyze their strengths and weaknesses.  
  
We find that the accuracy score for training data D-TREE with 98.82% followed by, KNN achieves the accuracy of 86.56%, followed by SVM with 86.17%. However, the highest accuracy score for testing data KNN achieves 86.17%, SVM has 85.81% and D-Tree has 79.98%.

**Decision Tree (D-Tree):**

• Precision (89): The D-Tree model demonstrates a precision of 89%, signifying that 89% of the instances classified as positive by the model were indeed correct.

• Recall (88): With a recall score of 88%, the D-Tree model identifies 88% of the actual positive instances, showcasing its capability to capture relevant cases effectively.

• F1-Score (88): The harmonic mean of precision and recall yields an F1-Score of 88%, reflecting a balanced performance between precision and recall for the D-Tree model.

**Support Vector Machine (SVM):**

• Precision (86): SVM displays a precision of 86%, denoting that 86% of the positive predictions made by the model are accurate.

• Recall (100): An impressive recall score of 100% indicates that the SVM model successfully captures all actual positive instances without missing any.

• F1-Score (92): With an F1-Score of 92%, SVM achieves a balanced performance, considering both precision and recall.

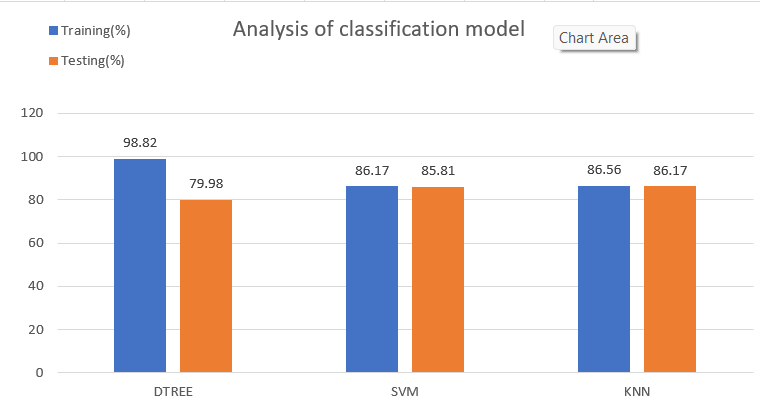
**K-Nearest Neighbors (KNN):**

• Precision (87): KNN demonstrates a precision of 87%, indicating an accuracy of 87% in identifying positive instances.

• Recall (99): With a recall score of 99%, KNN identifies 99% of the actual positive instances, showcasing its ability to capture most relevant cases.

• F1-Score (93): The KNN model shows a well-balanced F1-Score of 93%, suggesting a harmonious blend of precision and recall in its predictive capabilities.

We conclude that there is no single best model for every problem, and that choosing a suitable model depends on various factors, such as the data quality, the problem complexity, and the evaluation criteria. We hope that our project can provide a useful reference for anyone who wants to learn more about machine learning models and their comparison.



*Fig.1: Analysis of Classification Models*

# CONCLUSION

# In this project, we have compared the performance of three different machine learning models: support vector machine (SVM), k-nearest neighbor (KNN), decision tree (D-Tree). We have applied these models to one diabetic dataset. We have evaluated the models using various metrics, such as accuracy, precision, recall, and F1-score. Our results show that there is no single best model for all datasets. Each model has its own strengths and weaknesses, depending on the characteristics of the data, such as the number of features, the number of classes, the distribution of the data, and the presence of noise or outliers. Therefore, it is important to choose the appropriate model for each specific problem domain. Some general observations from our analysis are: - SVM tends to perform well on datasets with high-dimensional features and clear boundaries between classes. However, it can be sensitive to the choice of kernel function and hyperparameters, and it can be computationally expensive to train and test. - KNN is a simple and intuitive model that can handle non-linear and complex data. However, it can be affected by the choice of distance metric and the number of neighbors, and it can be slow to predict new instances due to the need to search for the nearest neighbors. - D-Tree is a transparent and interpretable model that can capture non-linear relationships and interactions between features. However, it can be prone to overfitting and underfitting, depending on the depth and complexity of the tree, and it can be unstable due to the sensitivity to small changes in the data. We hope that our project has provided some useful insights into the working and performance of these machine learning models. We also hope that our project has contributed to the existing literature on comparative studies of machine learning models, which are not very abundant in the current times. We believe that such studies are valuable for both researchers and practitioners who want to understand the strengths and limitations of different machine learning approaches and apply them effectively to real-world problems.

Compared to decision trees and random forests, SVM can handle high-dimensional data more effectively. It can also handle non-linear classification tasks using kernel functions, which makes it more versatile than linear models.

On the other hand, algorithms like logistic regression and naive Bayes are computationally less expensive than SVM. They can be faster for training and prediction, especially for large datasets.

# FUTURE SCOPE

The project work that we have done in the field of machine learning is a valuable contribution to the literature of comparative analysis of different models such as SVM, K-NN, and D-Tree. We have provided a clear and detailed explanation of how these models work, what are their advantages and disadvantages, and how they perform on various datasets. We have also presented the results of our experiments and the scores of the models on different metrics.  
  
However, there is still room for improvement and further research in this area. Some of the possible directions for future work are:  
  
- We could use more datasets from different domains and applications to test the generalizability and robustness of the models.  
- We could explore other models that are not covered in this project, such as neural networks, random forests, or ensemble methods, and compare them with the ones we have used.  
- We could fine-tune the hyperparameters of the models to optimize their performance and avoid overfitting or underfitting.  
- We could implement some feature engineering techniques to enhance the quality and relevance of the input data for the models.  
- We could conduct a deeper analysis of the results and the scores, such as using statistical tests, visualizations, or error analysis, to gain more insights into the strengths and weaknesses of the models.  
  
These are some of the ideas that we have in mind for the next version of this project. We hope that our work will inspire other researchers and practitioners to explore this fascinating field of machine learning and to develop more efficient and effective models for various tasks and problems.

# 

# BIBLIOGRAPHY

[1] Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer

Harikumar Rajaguru and Sannasi Chakravarthy S R.

[2] Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis

Fei Pan a, Baoying Wang a, Xin Hu b, William Perrizo a.

[3] PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES.

Ramesh TR, Umesh Kumar Lilhore, Poongodi M, Sarita Simaiya, Amandeep Kaur, Mounir Hamdi

Malaysian Journal of Computer Science

https://doi.org/10.22452/mjcs.sp2022no1.10[132-148]

[4] A hybrid CNN-KNN model for MRI brain tumor classification

Srinivas B., Prof.Gottapu Sasibhushana Rao

International Journal of Recent Technology and Engineering (IJRTE) · June 2019

DOI: 10.35940/ijrte.B1051.078219

[5] Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm and support vector machine in an electronic nose

Selda Güney, Ayten Atasoy

Sensors and Actuators B: Chemical

https://doi.org/10.1016/j.snb.2012.03.047

[6] Classification of hyperspectral remote sensing images with support vector machines

F. Melgani and L. Bruzzone

IEEE Transactions on Geoscience and Remote Sensing

doi: 10.1109/TGRS.2004.831865, vol. 42, no. 8, pp. 1778-1790, Aug. 2004

[7] MSVM-kNN: Combining SVM and k-NN for Multi-class Text Classification

P. Yuan, Y. Chen, H. Jin and L. Huang

IEEE International Workshop on Semantic Computing and Systems, Huangshan, China

doi: 10.1109/WSCS.2008.36., 2008, pp. 133-140

[8] Power Transformer Fault Classification Based on Dissolved Gas Analysis by Implementing Bootstrap and Genetic Programming

A. Shintemirov, W. Tang and Q. H. Wu

IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Jan. 2009

doi: 10.1109/TSMCC.2008.2007253., vol. 39, no. 1, pp. 69-79

[9] An Efficient Prediction of Breast Cancer Data using Data Mining Techniques.

A. Shintemirov, W. Tang and Q. H. Wu

"Power Transformer Fault Classification Based on Dissolved Gas Analysis by Implementing Bootstrap and Genetic Programming,"

IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Jan. 2009

doi: 10.1109/TSMCC.2008.2007253., vol. 39, no. 1, pp. 69-79

[10] Early Prediction of Diabetes Mellitus Using Machine Learning.

G. Tripathi and R. Kumar

2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020

doi: 10.1109/ICRITO48877.2020.9197832, pp. 1009-1014

[11] An experimental study on upper limb position invariant EMG signal classification based on deep neural network

Anand Kumar Mukhopadhyay a, Suman Samui b

A) Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India

B) Advanced Technology Development Centre, Indian Institute of Technology, Kharagpur, India

Received 30 May 2018, Revised 29 April 2019, Accepted 17 August 2019, Available online 3 September 2019, Version of Record 3 September 2019.

[12] Tree Species Discrimination in Tropical Forests Using Airborne Imaging Spectroscopy

J. -B. Feret and G. P. Asner

IEEE Transactions on Geoscience and Remote Sensing, Jan. 2013

doi: 10.1109/TGRS.2012.2199323, vol. 51, no. 1, pp. 73-84

[13] Machine Learning for Smartphone-Based Early Detection of Diabetic Disease in Pima Indians Diabetes Database

Mitushi Soni

Dept of Computer Science and Engineering

Shri G.S. Institute of Technology and Science, Indore, India

Dr. Sunita Varma

Dept of Information Technology

Shri G.S. Institute of Technology and Science, Indore, India.

[14] Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM.

M. A. B. Siddique, S. Sakib, M. M. R. Khan, A. K. Tanzeem, M. Chowdhury, and N. Yasmin, “Deep Convolutional Neural

Networks Model-based Brain Tumor Detection in Brain MRI Images,” in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020

doi: 10.1109/I-SMAC49090.2020.9243461, pp. 909–914

[15] Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An

Ensemble Approach

[16] Performance Analysis of Machine Learning Approaches in Diabetes Prediction

M. F. Faruque, I. H. Sarker, and others

International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–4.

[17] Diabetes Prediction using Machine Learning Techniques

M. F. Faruque, I. H. Sarker, and others

“Performance analysis of machine learning techniques to predict diabetes mellitus,”

International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–4.

[18] Prediction of Diabetes using Machine Learning

S. Sakib, M. A. B. Siddique, and M. A. Rahman

“Performance Evaluation of t-SNE and MDS Dimensionality Reduction Techniques with KNN, ENN and SVM Classifiers,” in 2020

IEEE Region 10 Symposium (TENSYMP), 2020, pp.[ 5–8]

[19] Dataset- <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>

[20] Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurelien Geron