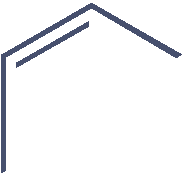
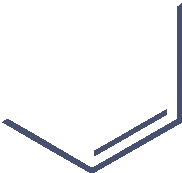
## 



*Review*

Machine learning models for early diagnosis of diabetes

Paka Neha – 2203A52168

1 Affiliation : 2203A52168@sru.edu.in

2 Section : AIML AA

**\*** SR UNIVERSITY



**Abstract:** Diabetes is a common disease that affects about 100 people, at least 30 of whom have diabetes. People with diabetes often have common symptoms. The goal of this program is to diagnose diabetes or non-diabetes based on these symptoms. These predictions can provide individuals with valuable insights to better understand diabetes and its causes.

In this work, we use different algorithms to predict the outcome (diabetes or not). While the results may not be entirely accurate, they can provide valuable predictions. Like the saying ‘prevention is better than cure’. Knowing the symptoms gives you the opportunity to prevent or manage your diabetes.

In this work, we will use several algorithms such as logistic regression, perceptron, SVM (Support Vector Machine) for classification, and KNN (K-Nearest Neighbors) for product prediction Optimal approach of diabetes prediction is determined by the algorithm to achieve the highest accuracy. It is important to note that the optimal method may be different for each individual data set.

**Key words:** Logistic Regression, Support Vector Machine, Perceptron, K-Nearest Neighbors, Bootstrapping, Accuracy, Precision, Recall, F1-score, Model Selection, Research, methodology, Analysis, Model Interpretation

**Citation:** Sidani, S.; Patel, K.D. Interprofessional Education in Diabetes Care—Findings from an Integrated Review. *Diabetology* **2023**, *4*, 356–375. [https://doi.org/](https://doi.org/10.3390/diabetology4030030) [10.3390/diabetology4030030](https://doi.org/10.3390/diabetology4030030)

Academic Editor: Giancarlo Tonolo Received: 15 March 2023

Revised: 27 July 2023

Accepted: 7 August 2023

Published: 28 August 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license ([https://](https://creativecommons.org/licenses/by/4.0/) [creativecommons.org/licenses/by/](https://creativecommons.org/licenses/by/4.0/) 4.0/).

# 

# Introduction

Diabetes is a chronic medical condition that affects how the body handles glucose, a type of sugar used for energy. There are two main options: Option 1 and Option 2.

Type 1 diabetes usually develops in childhood or early adulthood. In this way, the immune system mistakenly attacks and destroys the insulin-producing pancreatic cells. As a result, the body cannot produce insulin, requiring daily insulin injections or an insulin pump.

Type 2 diabetes, which is very common in adults, is generally caused by insulin resistance and decreased insulin production. Factors such as genetics, obesity, and lifestyle choices play an important role. Managing type 2 diabetes typically includes dietary changes, regular exercise, and in some cases, oral medication or insulin.

Diabetes causes sugar levels to rise, which can lead to health problems if not controlled properly. Symptoms include excessive thirst, frequent urination, fatigue, and slow wound healing.

*Diabetology* **2023**, *4*, 356–375. <https://doi.org/10.3390/diabetology4030030> <https://www.mdpi.com/journal/diabetology>

Mumucchy chitam avichalyam, exertion, aggregate, aggregate, aggregate, aggregate, aggregate, appropriation, decomposition, kidney, kidney, kidney, kidney, heredity, heredity, heredity, steadfastness, subject, method, method, method, method, method , Before the weapon, Genetics that meet the unique needs of each of their individuals.

**Current Generation Use:** This project helps us to detect diabetes early by detecting few vital symptoms. In today’s educated and informed world, these signs are easy to spot. The project has two main goals: helping those who may develop diabetes to reduce their risk and providing them with advice on how to manage the condition. It also educates health enthusiasts on how to take care of themselves and reduce their chances of developing diabetes.

Through this project, people can learn how to live healthier lives. They can choose healthy foods, exercise regularly, and take steps to prevent diabetes. This business promotes a culture of self-care and takes care of our health. It shows how we can use knowledge and technology to improve public health in our modern world**.**

**Literature review:** A literature review on diabetes using computer programs shows that many researchers are trying to find ways to predict future diabetes In order to make these predictions they use computer techniques use, such as decision trees and mouse-assisted devices.

To filter these predictions, they look at a lot of data, including medical records and data from the devices you wear. They use numbers to assess the quality of their predictions, and to do this they use numbers such as sensitivity, specificity, accuracy, and AUC-ROC.

Researchers have done a lot of research to see if their predictions work. Big data and computational techniques are used to evaluate the best results in diabetes prognosis.

However, some problems remain, such as not always having the best information and why they are always forecasting. In the future, they want to use more data, make predictions easier to understand, and target people’s health in real time to help them stay healthy and avoid diabetes.

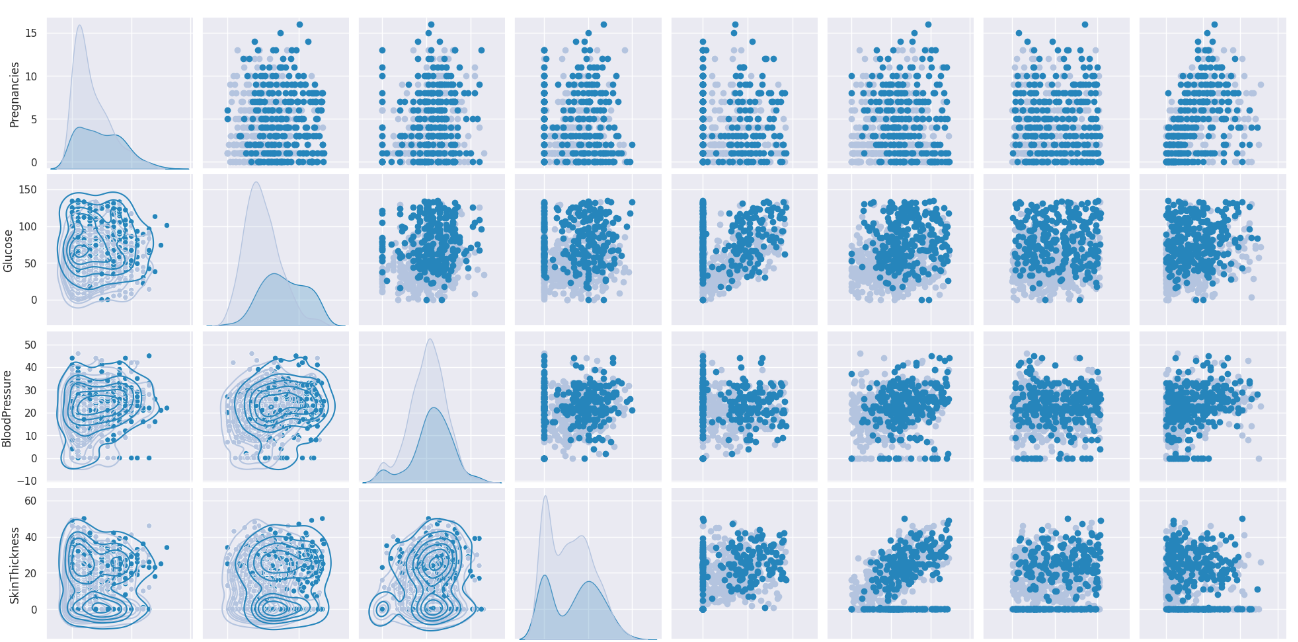
**Few Results on same project:**

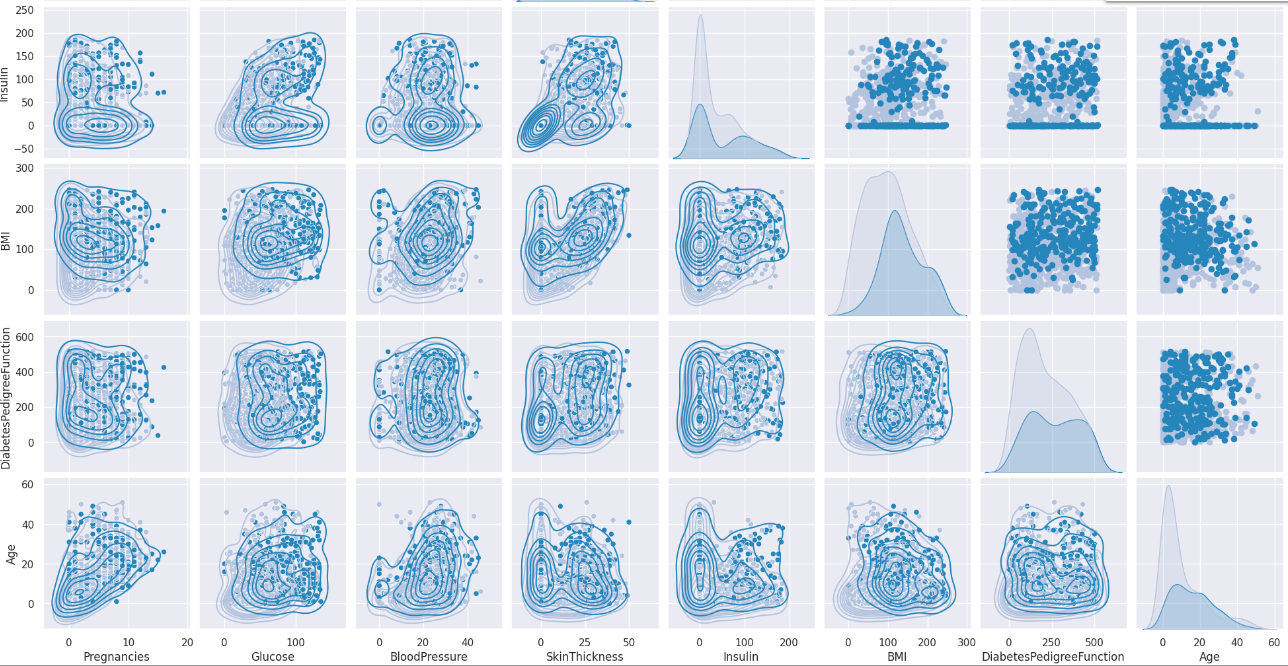
|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| **Project 1:**   * SVM (Support Vector Machine)       **Project 2:**   * Logistic Regression * KNN (k-nearest neighbor) * SVM (support Vector Machine) | 0.7401574803149606  75.000000000000000000  73.000000000000000000  78.000000000000000000 |

1. **Methodology:**
   1. **Dataset:** This data set consists of 768 rows and 9 columns.

* **Target Variable:** Outcome (diabetic or not)
* **Features:** Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI Diabetes Pedigree Function, Age.
  + - * 1. **Pregnancies:** Number of pregnancies experienced.
        2. **Glucose:** Blood sugar level measurement.
        3. **Blood Pressure**: Systolic and diastolic pressure.
        4. **Skin Thickness:** Subcutaneous fat layer thickness.
        5. **Insulin:** Hormone for blood sugar control.
        6. **BMI:** Body mass index calculation.
        7. **Diabetes Pedigree Function:** Genetic diabetes risk assessment.
        8. **Age:** Patient's age in years.
        9. **Outcome:** Presence of diabetes indicator.

* **Data Visualization:**

**

****

**Fig 1: visualization of the data set**

**Preprocessing:**

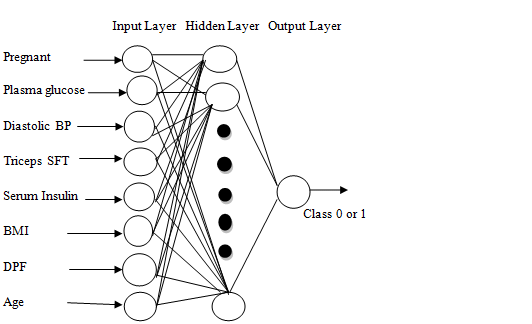
* **Normalization:** This data set requires normalization i.e making the values in the form of 0’s and 1’s .This process includes many methods min max scalar is one of them.
* **Filling NaN values:** The data set contains few missing values these are filled with 0’s by implementing the below code.
* **Drop:** When we want to eliminate any column we use this process**.**
  1. **Implementation:**

**Logistic Regression:** Logistic regression resembles a selection tool in many areas. Imagine you want to know if an email is spam or not. Logistic regression helps you with that. It calculates the probability that something is in one of two categories such as 0 or 1, yes or no, true or false. Then, it makes a decision based on the rule, e.g., if the probability is greater than 0.5, it says "yes", otherwise it says "no". People use the tool to diagnose diseases in medicine, filter spam in email systems, and analyze credit risk in the financial sector. It is a simple but practical way of making decisions with only two possible outcomes.

**Perceptron:** A perceptron is a supervised machine learning tool that can be used to quantify the adequacy of predictions. It basically consists of four steps and is commonly used for binary classification, which determines whether a given input belongs to a particular class or not In the first step, it fetches input data from the dataset, and afterwards is initialized with random weighting.

In the next step, the bias setting is applied, and then the activation function is applied. The result of the activation function is the final prediction. The default activation function is sigmoid, and results range from 0 to 1. The perceptron calculates the accuracy of the data set and performs a supervised learning process

Although it includes a hidden layer, it is also known as a monolayer neural network. Please note that some grammatical errors have been corrected in this answer.



**Support Vector Machine (SVM):** A support vector machine (SVM) is a supervised learning algorithm for assessing the accuracy of a model. In this method, the dependent variable and the target variable are plotted on a graph. Of all the lines, one of them is considered the best fit line, and is known as the "maximum margin overplane", not the "maximum". Points adjacent to this hyperplane are known as "support vectors,". When the data set contains three variables, the result is usually a 2D model. SVMs are powerful tools for classification and regression applications, and they excel in finding optimal decision boundaries by maximizing the differences between classes

**K-Nearest Neighbor:** k-Nearest Neighbors (KNN) is a simple but effective technique that can be used to predict diabetes by classification

KNN is a non-parametric algorithm that classifies data points based on their similarity to other data points in the training set. It uses distance measures (e.g. Euclidean distance) to measure similarity between data points.

Set the value of "k", which represents the number of nearest neighbors to consider. Choosing the right "k" value is important; You can experiment with different values ​​and use techniques like cross-validation to find the best results.

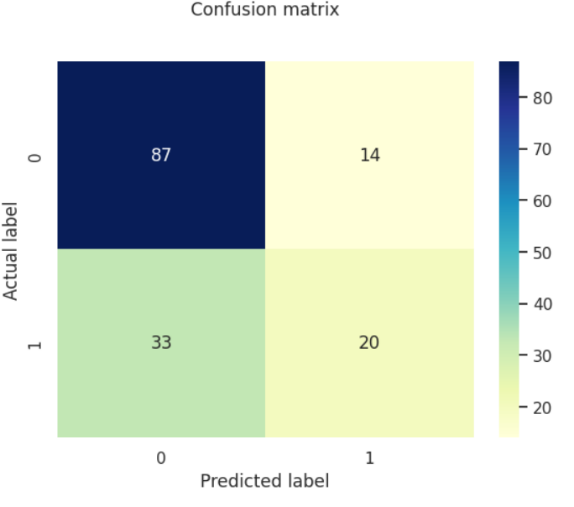
KNN models define decision boundaries based on the similarity of data points. Visualizing these boundaries can help to understand how the model separates the two groups (diabetic and nondiabetic).

The trained KNN model can be incorporated into clinical decision support systems to help healthcare professionals assess an individual's risk of diabetes and determine the need for further screening or intervention.

KNN is a versatile and easy-to-interpret system that can be used for diabetes prediction. However, choosing an appropriate "k" value and dealing with class imbalances are important considerations for good results in practice.

**Confusion matrix for KNN:** In K-Nearest Neighbors (KNN) classification, you can construct a confusion matrix to evaluate the performance of your model. The confusion matrix is ​​a table that helps you understand how well the model classifies observations as true positives, true negatives, false positives, and false negatives. You can use libraries in Python to calculate and build the confusion matrix, such as scikit-learn.

The confusion matrix is ​​a valuable tool to evaluate the performance of classification models, as it provides insights into accuracy, precision, recall, and F1-score among other metrics Helps you understand how the model distinguishes good learning from sins well below and any biases or issues in its forecasts or if so . . . . . . . . . . . .



**2.4.5. Boot strap:** Bootstrapping provides a way to measure the variability and robustness of your models. It helps in assessing how the performance of the model may change when applied to different subsets of data. Additionally, bootstrapping can be particularly useful when dealing with limited data or when you want to quantify the uncertainty associated with your predictions.

Although bootstrapping offers advantages, it is important to thoroughly examine and validate the results of the bootstrapped model and ensure that the chosen method of collection (e.g., polling, averaging) is you the specific nature of the diabetes prognosis problem corresponds.

1. **Bootstrapping for Logistic Regression:**

Bootstrap samples are generated from the training data set by replacing randomly selected data points.

Training multiple logistic regression models on these bootstrapped samples yields a range of models.

The final prediction is usually obtained by pooling predictions from these models, which can reduce overfitting and improve overall model effectiveness.

1. **Bootstrapping for SVM Classification:**

Bootstrapping can be used to generate multiple subsamples of the training data for SVMs.

Each subset is used to train the SVM model separately.

Combining predictions from these models, such as by analysis or averaging, can lead to more robust predictions, especially when dealing with noisy or unbalanced data

1. **Bootstrapping for Perceptron:** When using Perceptron, bootstrapping can be incorporated by sampling repeated data sets.

Several perceptron models are trained on these resampled data sets.

A final prediction can be made by considering the output of each perceptron, which can improve the performance of the model.

1. **Bootstrapping for k-Nearest Neighbors (KNN):**

Bootstrapping in KNN can help in assessing the stability and improvement of the robustness of the model.

Several KNN models can be trained on bootstrapped datasets.

The final prediction can be determined using the majority of votes or a weighted set of predictions from these KNN models.

**RESULT ANALYSIS:**

The result accuracies we finally got are:

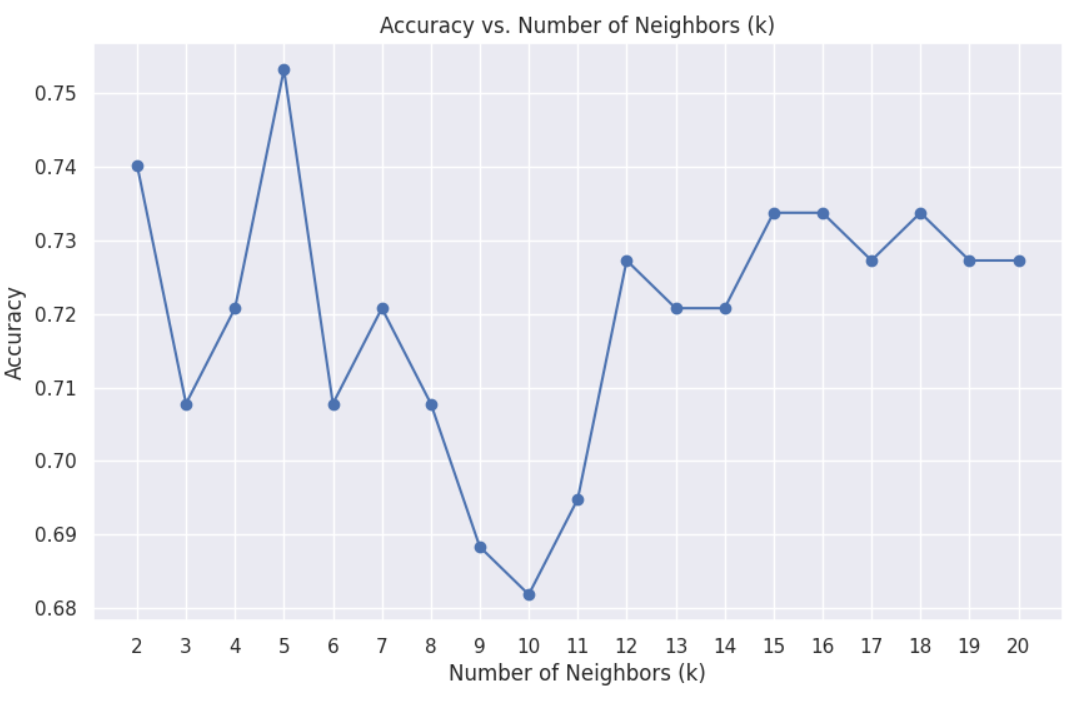
Logistic Regression:0.7467%

SVM:0.7597%

Perceptron Learning:34.42%

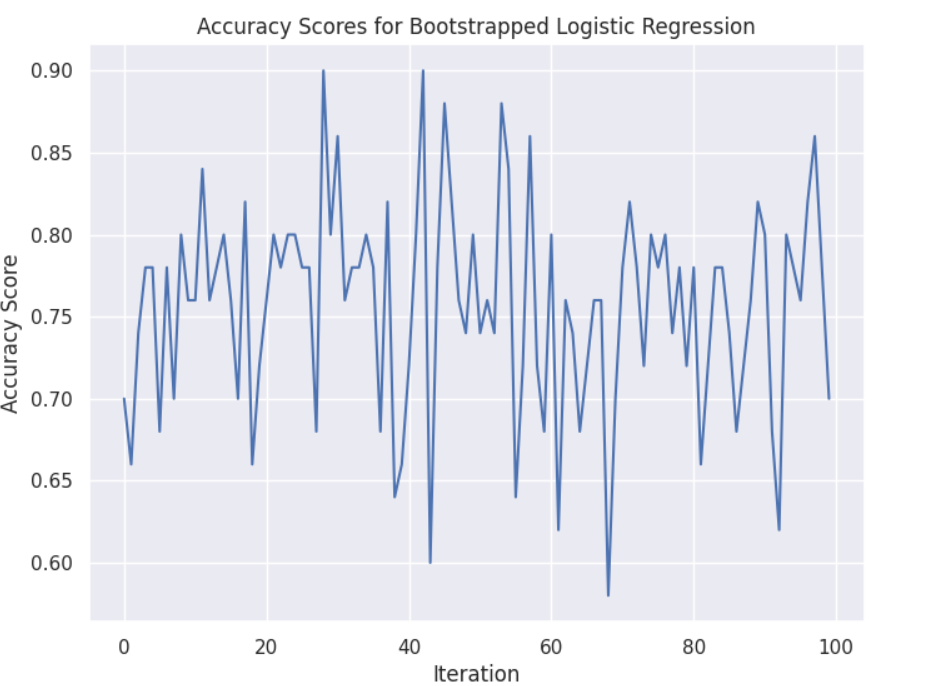
KNN: k= 5 is the most accurate value for the given dataset.

Fig 2 shows the graph for KNN k values accuracy

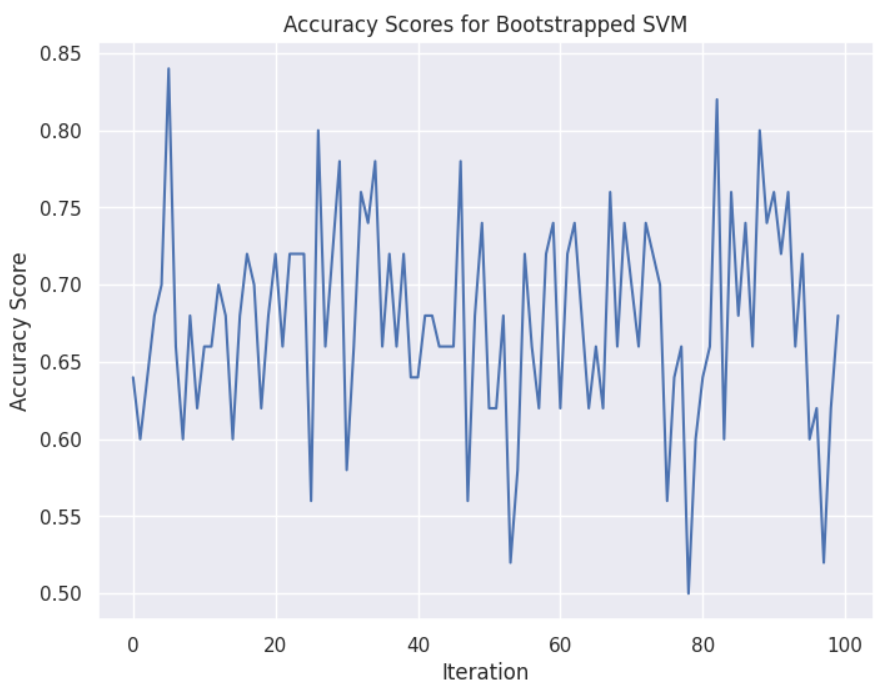
****

finally after bootstrapping is applied we got the following graphs for the 4 models:

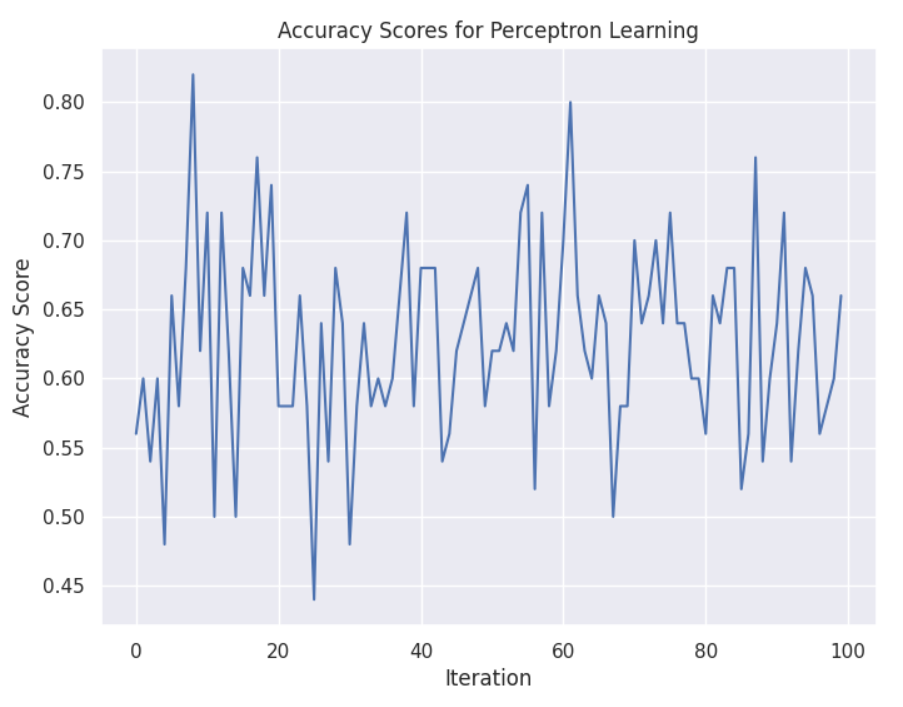
Fig-3 Logistic Regression, Fig-4 SVM, Fig-5 Perceptron Learning, Fig-6 KNN

****

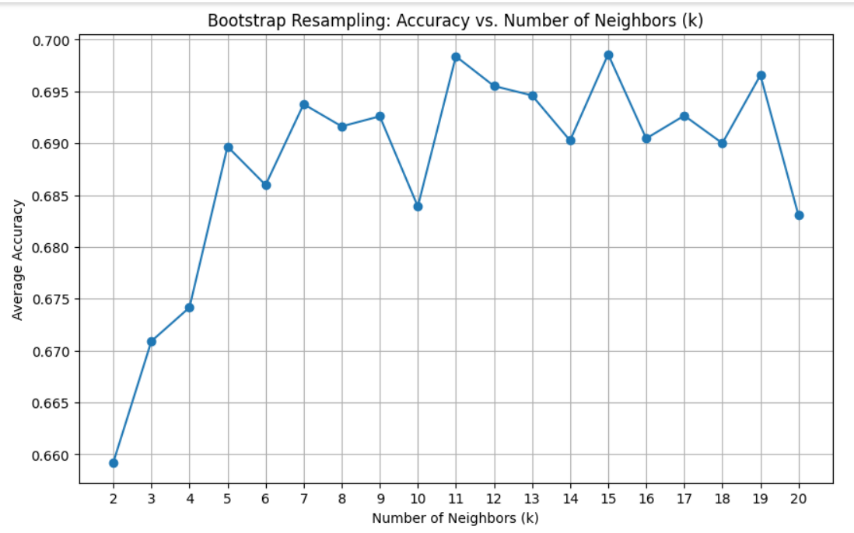
**Fig 3**



**Fig 4**

****

**Fig 5**



**Fig-6**

**CONCLUSION:** The algorithm that provides the highest accuracy is considered the best fit for this data set. For this particular data structure, KNN outperforms other algorithms, making it the most efficient method.

It is important to note that the optimal algorithm may be different for different types of data, ie. the best method may change depending on the specific characteristics and requirements of each data system.

**ACKNOWLEDGEMENTS:**

I thanks SR University, Hasanparthy , Faculty for collecting this.

**Reference:**

1. <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>
2. [https://archive.ics.uci.edu/ml/machine-learning-databases/00000/UCIMachine Learning Repository Diabetes Data Set.htm](https://archive.ics.uci.edu/ml/machine-learning-databases/00000/UCIMachine%20Learning%20Repository%20Diabetes%20Data%20Set.htm)
3. <https://www.ncbi.nlm.nih.gov/books/NBK362855/>

1. <https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/>