

*Article*

Chronic Kidney Disease Prediction using Machine Learning

A.VINAY TEJA – 2203A52223

Section :AIML -A

\*SR UNIVERSITY

**Abstract:** Chronic Kidney Disease (CKD) can be a serious health issue, but catching it early can make a big difference. We used machine learning models to create a way to predict CKD before it gets worse. We tried different Machine learning models like Logistic Regression, Perceptron Learning, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM).The dataset contains various features like age, gender, blood pressure and so on…and the best part is that all models are trained and tested well and have a good accuracy. so, it can help

This means we can help people and doctors know if someone is at risk for CKD. If we catch it early, we can take action to keep people healthy. This project shows how machine learning and technology can help with healthcare and improve our chances of finding CKD early.

**Keywords:** Logistic Regression ; Support Vector Machine (SVM) ; Perceptron Learning; K-Nearest Neighbors (KNN) ;

**Introduction**

Kidneys are important for our health, but sometimes they don't work as well as they should. This can lead to a condition called Chronic Kidney Disease (CKD), which can be quite serious. Actually the primary function of a kidney is to purify/filter impurities from the blood. By this kidneys loss this filtering ability. The key is to find CKD early, so we can do something about it. In this project, we used machine learning models to help with this.

Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed for every specific task.

We took various features in a dataset like age, gender, blood pressure, and lab test results to help the model learn. The exciting part is that the model learned with a good accuracy to help at predicting CKD. This means we can use it to help people and doctors know if there's a risk of CKD. Early detection can make a big difference in keeping people healthy. So, in this project, how computers can be a be a big help in healthcare, especially when it comes to finding CKD early.

## Methods

### objective

### The primary objective of Chronic Kidney Disease (CKD) management is to slow down the progression of kidney damage, reduce symptoms, and minimize complications, ultimately improving the quality of life for individuals affected by CKD.

### Research Design

Designing a research study for chronic kidney disease prediction using machine learning involves several key components and considerations. Below, I outline a research design that encompasses these essential elements:

* + - 1. Data Collection
      2. Data Preprocessing
      3. Data Splitting
      4. Model Selection
      5. Model Training

## Data and Methodology

### Data Description:

To create a dataset for chronic kidney disease prediction using machine learning, you need to provide a description of the data. A well-documented dataset description helps researchers and data analysts understand the data's content and structure.

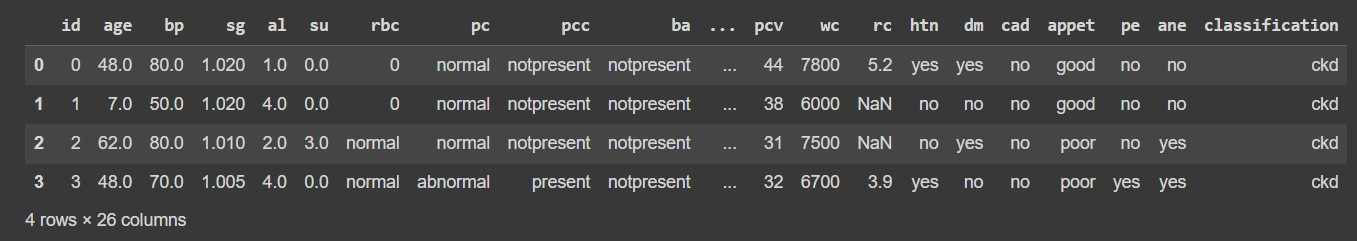
The features in the dataset are:

The dataset contains a wide range of medical information about patient’s age and blood pressure is the crucial parameter in order to know the flow of the blood. Specific gravity is the measure of urine concentration. Red blood cells and pus cells are the indicators of the function issues of the kidney. Pus cell clumps or bacteria suggests the infection. Blood glucose random is helpful for managing diabetes. While blood urea and serum creatinine are the markers of kidney function. Sodium and potassium levels are important. Haemoglobin levels indicate the overall blood health. Appetite and pedal edema are responsible for well-being of fluid retention. Anaemia is a condition caused by low blood cells

**Target Variable:**

**Chronic Kidney Diagnosis:** A binary variable indicating whether the patient has been with chronic kidney disease. The value 1 indicates the presence of Chronic Kidney Disease (CKD) , and the value 0 indicates the absence of chronic Kidney Disease (CKD).

1. **DATA COLLECTION:**



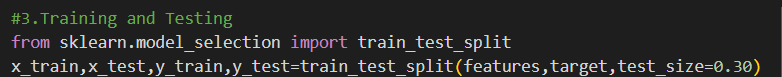
1. **DATA PREPROCESSING:**

Data preprocessing involves cleaning, transformingand organizing the rawdata into suitable format in order to analyze the model. The proper data preprocessing improve the quality of the results Read the data set and identify the Null values if there are any null values fill them with zeroes/fill with mean value/eliminate the column.



1. **DATA SPLITTING:**

Data splitting is a common practice in data analysis and machine learning. It involves the process of dividing a dataset into multiple subsets for various purposes, this is used to train, test and validate the machine learning models. These are essential for assessing the model's performance, preventing overfitting, and making decisions.

****

1. **Training Set**: This portion of the dataset is used to train the machine learning model. It's the part of the data that the model uses to learn patterns, relationships, and associations between the input features and the target variable. The training set is usually larger than the test set, often containing around 70-80% of the data.

1. **Testing Set**: This subset is kept separate from the training data and is used to evaluate the model's performance and generalization. The model has never seen the data in the test set during training. It helps you estimate how well your model will perform on new, unseen data. Typically, the test set contains around 20-30% of the data, but, again, this can vary.
2. **MODEL SELECTION:**

Model selection is the process of choosing the most appropriate machine learning or statistical model to solve a specific problem. Model selection involves considering various factors, including the dataset we have and the type of problem to solve. Generally there are two types (i.e., classification or regression) . so in order the model that best fits our dataset is classification.

Classification has many models

* Logistic Regression
* Perceptron Learning
* Support vector machine (SVM)
* K-Nearest Neighbors (KNN)

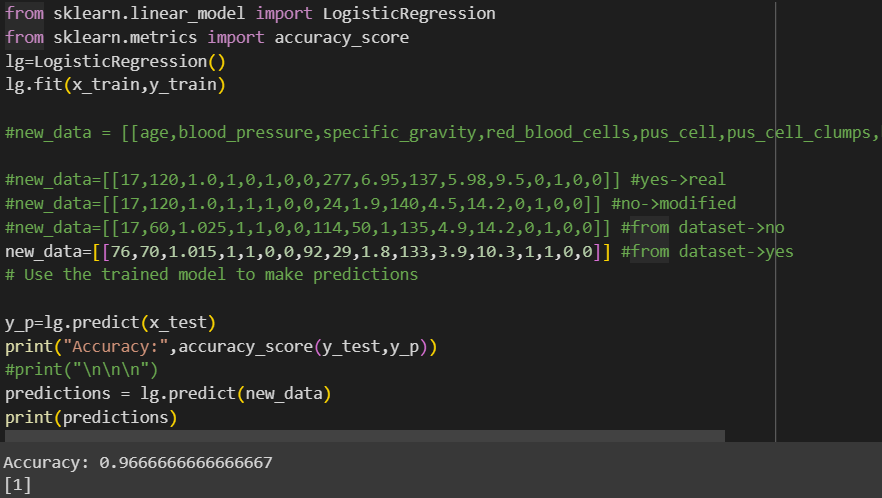
1. **MODEL TRAINING:**

Model training is crucial step in machine learning models. It Is the step where machine learning model learns from the data and adapts to its parameters in order to make accurate predictions. We train all the above mentioned models with training set and measure accuracy for each model to find out the best model in order to predict CKD

1. **LOGISTIC REGRESSION:**

**Accuracy:**

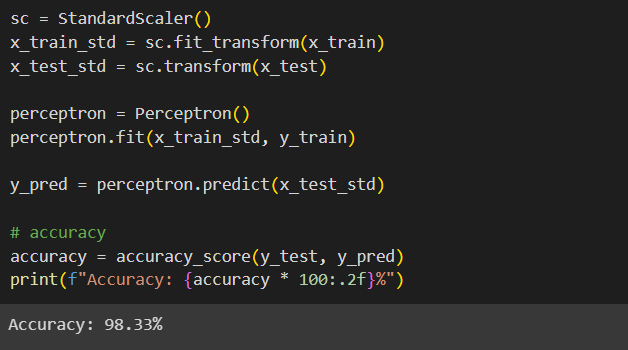
Logistic Regression is statistical and machine learning model used for binary classification tasks. It is commonly used when the dependent variable is binary or in the case of it has only two outcomes i.e.,([yes/n],[0/1]). Moreover it predicts whether the given input that belongs to one of the two classes or not.

****

1. **Perceptron Learning:**

**Accuracy:97%**

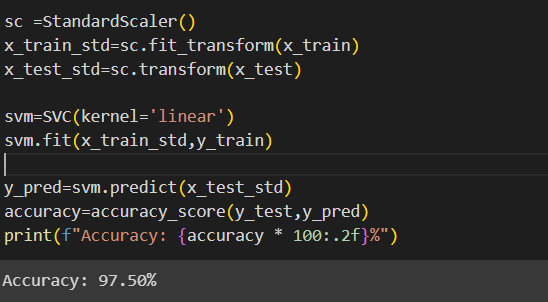
Perceptron learning is the simplest form of neural networks and it is considered to be the building blocks of complex neural networks. Perceptron learning comes under supervised learning algorithm used for binary classification tasks. Basically it takes set of inputs and multiplies them with corresponding weights and sums the result and apply to an activation function to produce output in the form of 0 or 1.



1. **SUPPORT VECTOR MACHINE (SVM):**

**Accuracy:97.5%**

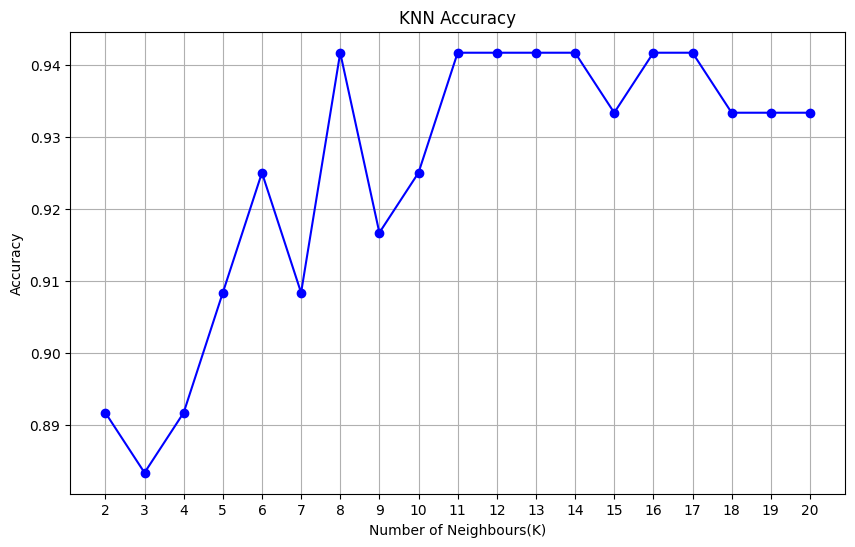
Support vector Machine (SVM) is a powerful machine learning algorithm which can do both classification and regression tasks. It’s primary thing is to find a hyperplane Which is used separate data points into distinct classes. Support vector Machine is used in various applications like text classification, image classification and this are able to handle high dimensional data. Same like Logistic Regression and perceptron learning we find accuracy for the **SVM.**

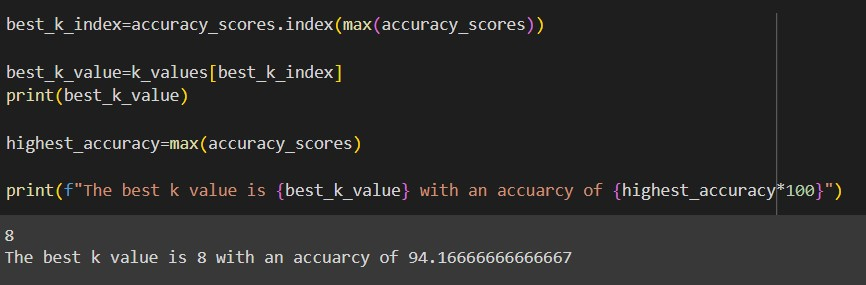
****

**iv). K-nearest neighbours (KNN):**

**Accuracy:94%**

In the analysis, the K-Nearest Neighbors (KNN) algorithm is applied to the employee attrition dataset. KNN is a supervised machine learning algorithm used for classification tasks. The analysis uses bootstrapping, resampling the dataset multiple times (100 times), to assess the performance of KNN in different data subsets. For each iteration, it randomly selects a subset of the data for training and testing.



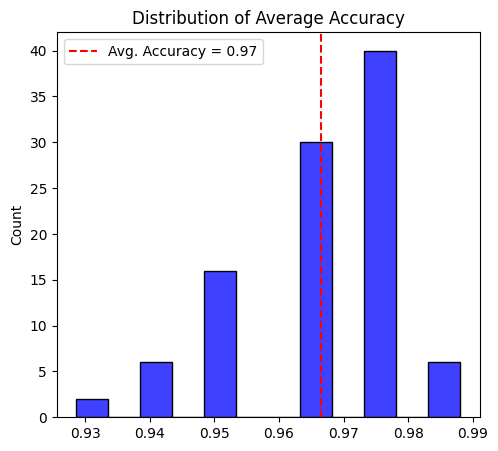


* Now we implement bootstrapping method with respect to **KNN**, **SVM**, **PERCEPTRON** **LEARNING** and **LOGISTIC REGRESSION**

**BOOTSTRAPPING :** Bootstrapping is resampling technique used in statistics and machine learning to estimate the sampling distribution of a statistic.

1. Bootstrapping with respect to **Logistic Regression**:

When applied Bootstrapping to Logistic Regression the average accuracy we got for 100 samples is 97% where the fluctuations are between 93% to 97%. Here the accuracy for the model shows the robustness and stability of the model and therefore shows high average accuracy



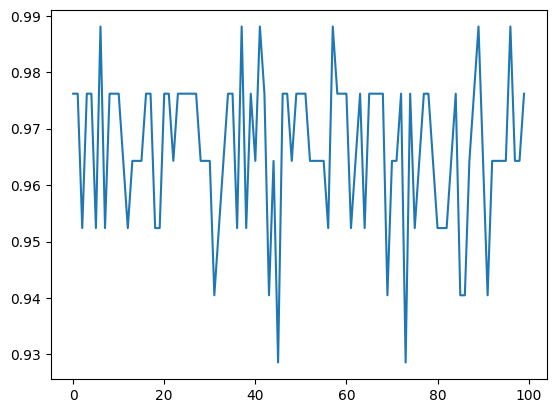
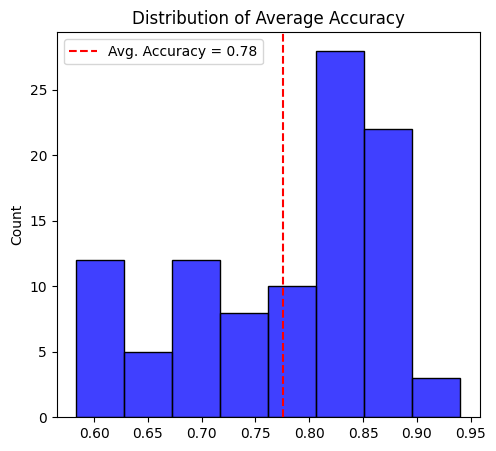


Figure i. Logistic Regression

1. Bootstrapping with respect to **Perceptron Learning**:

When applied Bootstrapping to Perceptron Learning the average accuracy we got for 100 samples is 78% where the fluctuations are between 60% to 95%. Here the accuracy for the model is relatively good and the wide range of fluctuations suggests that it may vary in various subsets. And the wide range of fluctuations indicate the complexity and might leads to non-linearity.



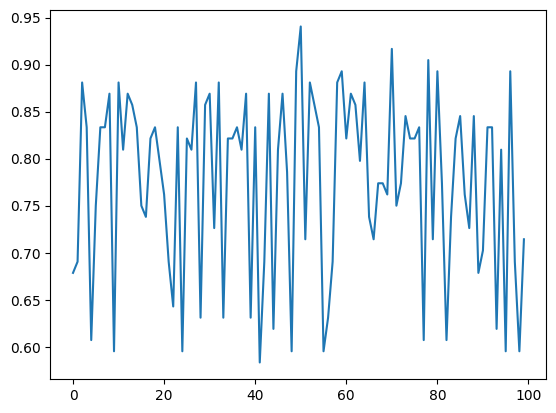
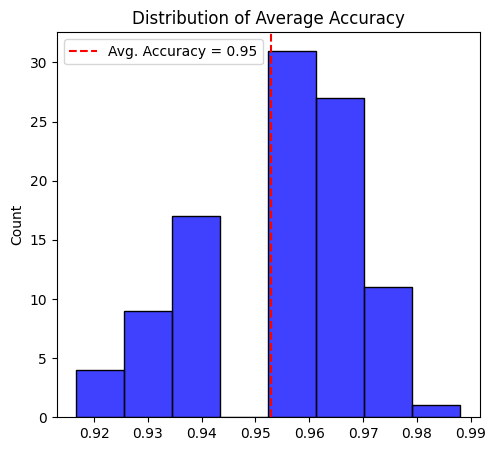


Figure ii) Perceptron Learning

1. Bootstrapping with respect to **SVM**:

When applied Bootstrapping to SVM the average accuracy we got for 100 samples is 95% where the fluctuations are between 92% to 98%. Here the accuracy for the model is relatively good and indicates a strong sign it is the best performing model in terms of accuracy and delivering accurate prediction the wide range of fluctuations suggests that it is consistent in various subsets.



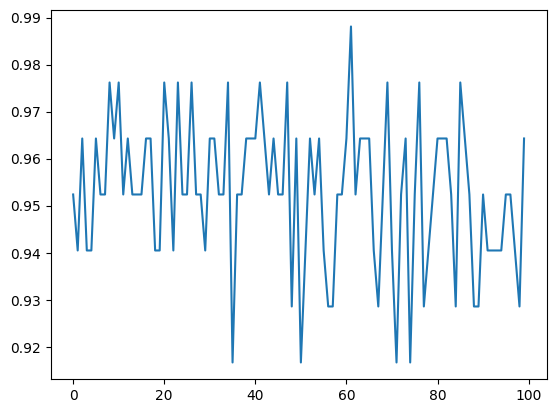
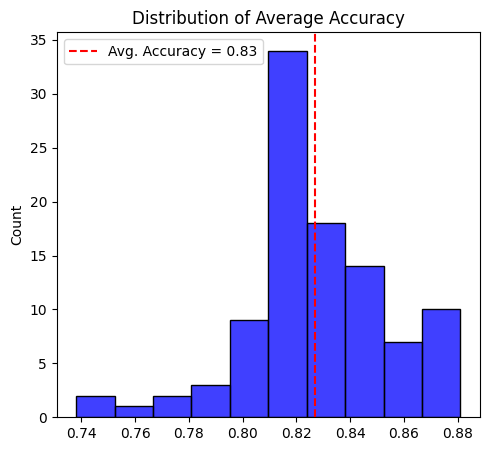


Figure iii) support vector Machine

1. Bootstrapping with respect to **KNN**:

When applied Bootstrapping to KNN the average accuracy we got for 100 samples is 83% where the fluctuations are between 74% to 88%. Here the accuracy for the model is relatively good and indicates a strong sign that it is a best performing model in terms of accuracy and delivering accurate prediction and the average accuracy for KNN is a good sign of predicting target variable. While the fluctuations range moderate across the various samples



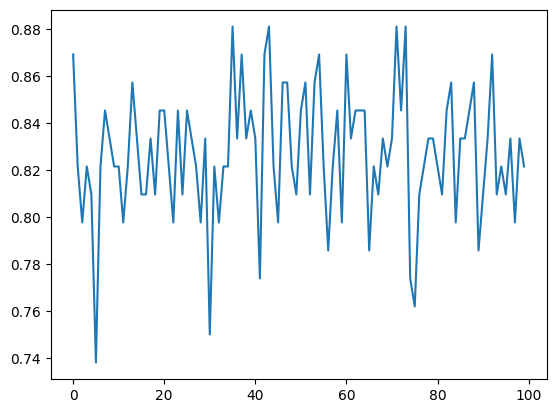


Figure iv) KNN

1. **Conclusion:**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 97% |
| Perceptron Learning | 78% |
| Support Vector Machine (SVM) | 95% |
| K-Nearest Neighbors(KNN) | 83% |

Logistic regression stands out of all models with an accuracy of 97% and perceptron Learning is least performed with an accuracy of 78% and SVM with 95% and KNN with 83%. So Logistic regression performs well in finding out Chronic Kidney Disease (CKD)

**Summary:**

In summary, the Logistic Regression i.e., fig.1 has more accuracy indicating relatively better predictive performance compared to the other models listed in this context of predicting chronic Kidney Disease.

**References**

1.case study 1:[cross ref](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9874070/)

2. Case study 2: [cross ref](https://link.springer.com/article/10.1007/s44174-022-00027-y)

3. Case study 3: [cross ref](https://www.hindawi.com/journals/cin/2023/9266889/)

4. Case study 4: [cross ref](https://ieeexplore.ieee.org/document/9185249)

5. data set : [cross ref](https://www.kaggle.com/datasets/mansoordaku/ckdisease)

6. Capstone project link : [cross ref](https://github.com/vinay-reddy31/statml_2203A52223.git)

**Bootstrap Accuracies of all models:**

