**AI Based Diabetes Prediction System**

**Project Title : Diabetes Prediction System**

**Phase 4: Development part2**

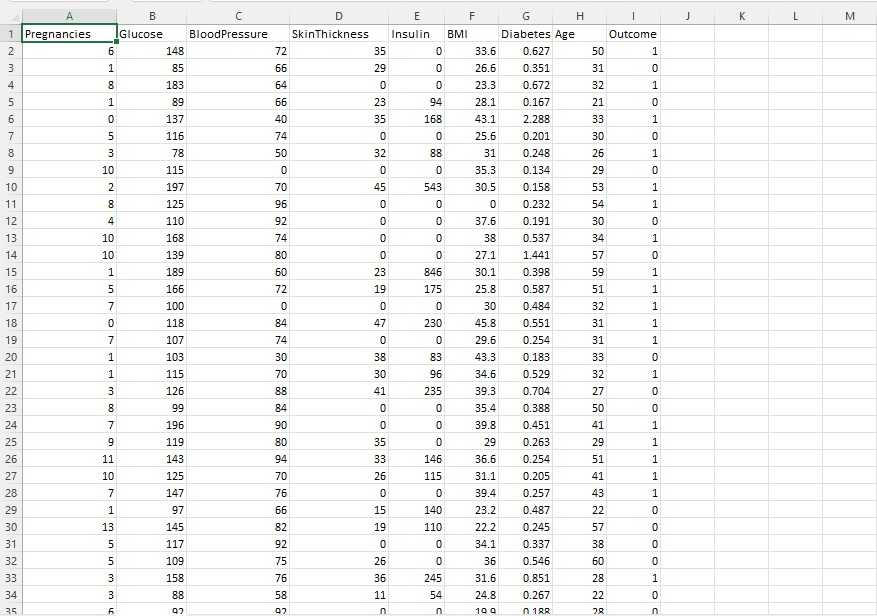


**Introduction:**

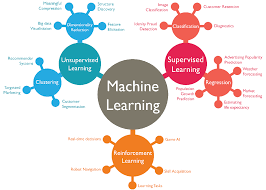
**Machine learning algorithms are mathematical model mapping methods used to learn or uncover underlying patterns embedded in the data. Machine learning comprises a group of computational algorithms that can perform pattern recognition, classification, and prediction on data by learning from existing data (training set). The most common machine learning approaches in biology are support vector machines (SVM) and artificial neural networks (ANN) . ANN model with deep neuron layers can be used to predict sequence specificities of DNA- and RNA-binding proteins, noncoding variants, alternative splicing, and quantitative structure–activity relationship (QSAR) of drugs . Deep learning models have outperformed other machine learning methods in identifying more complex features from data [102]. To achieve complex results, deep learning techniques require a higher volume of data and computational time, compared to other machine learning algorithms. Omic data such as genome, transcriptome, epigenome, proteome, and metabolome may be integrated into a single model, which has large dimensions, and requires extensive time to build an appropriate model. Data collection can be minimized by reducing the dimension of input data, which can be done before or after data integration with principal component analysis (PCA), or after data integration with feature selection algorithms . Mode of action by network identification (MNI) combines reverse engineering network modeling with machine learning to decipher regulatory interactions. MNI uses a training set of multidimensional omic data to identify genetic components and network that correspond to a specific state. MNI, using a set of ordinary differential equations, directed graph relating the amounts of biomolecules to each other can be generated. For example, when transcriptomic data are used as training data, regulatory influences between genes can be inferred. In addition to MNI, another network-based system CellNet classifies cellular states based on the status of gene regulatory network . Both MNI and CellNet utilize machine learning integrated reverse engineering methods.**

**Dataset Link :** <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>

**Dataset:**



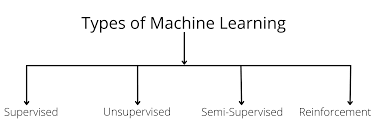
**Machine Learning Algorithms :**



**Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own. Different algorithms can be used in machine learning for different tasks, such as simple linear regression that can be used for prediction problems like stock market prediction, and the KNN algorithm can be used for classification problem.**

**In this topic, we will see the overview of some popular and most commonly used machine learning algorithms along with their use cases and categories.**

**Types of Machine Learning Algorithms :**



**1) Supervised Learning Algorithm :**

**Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.**

**In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.**

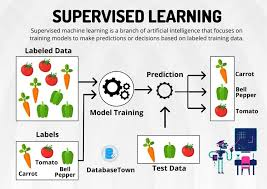
**Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).**

**In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.**

**How Supervised Learning Works?**

**In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.**

**The working of Supervised learning can be easily understood by the below example and diagram:**



**Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.**

**If the given shape has four sides, and all the sides are equal, then it will be labelled as a Square.**

**If the given shape has three sides, then it will be labelled as a triangle.**

**If the given shape has six equal sides then it will be labelled as hexagon.**

**Now, after training, we test our model using the test set, and the task of the model is to identify the shape.**

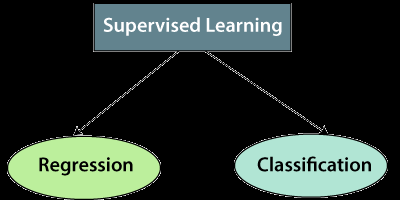
**The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.**

**Steps Involved in Supervised Learning:**

* **First Determine the type of training dataset**
* **Collect/Gather the labelled training data.**
* **Split the training dataset into training dataset, test dataset, and validation dataset.**
* **Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.**
* **Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.**
* **Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.**
* **Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.**

**Types of supervised Machine learning Algorithms:**

**Supervised learning can be further divided into two types of problems:**

****

**1. Regression**

**Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:**

* **Linear Regression**
* **Regression Trees**
* **Non-Linear Regression**
* **Bayesian Linear Regression**
* **Polynomial Regression**

**2. Classification**

**Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.**

**Advantages of Supervised learning:**

* **With the help of supervised learning, the model can predict the output on the basis of prior experiences.**
* **In supervised learning, we can have an exact idea about the classes of objects.**
* **Supervised learning model helps us to solve various real-world problems such as fraud detection, spam filtering, etc.**
* **Disadvantages of supervised learning:**
* **Supervised learning models are not suitable for handling the complex tasks.**
* **Supervised learning cannot predict the correct output if the test data is different from the training dataset.**
* **Training required lots of computation times.**
* **In supervised learning, we need enough knowledge about the classes of object.**

**2) Unsupervised Learning Algorithm**

**In the previous topic, we learned supervised machine learning in which models are trained using labeled data under the supervision of training data. But there may be many cases in which we do not have labeled data and need to find the hidden patterns from the given dataset. So, to solve such types of cases in machine learning, we need unsupervised learning techniques.**

**What is Unsupervised Learning?**

**As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as:**

**Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.**

**Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.**

**Example: Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.**

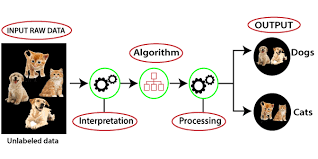


**Why use Unsupervised Learning?**

* **Below are some main reasons which describe the importance of Unsupervised Learning:**
* **Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.**
* **Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.**
* **In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.**

**Working of Unsupervised Learning**

**Working of unsupervised learning can be understood by the below diagram:**

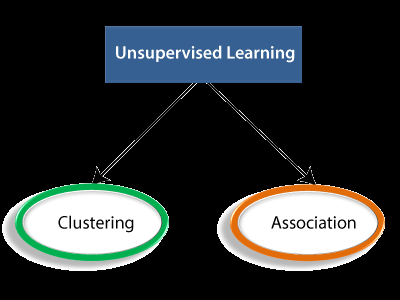


**Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.**

**Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.**

**Types of Unsupervised Learning Algorithm:**

**The unsupervised learning algorithm can be further categorized into two types of problems:**



**Clustering: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.**

**Association: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.**

**Unsupervised Learning algorithms:**

**Below is the list of some popular unsupervised learning algorithms:**

* **K-means clustering**
* **KNN (k-nearest neighbors)**
* **Hierarchal clustering**
* **Anomaly detection**
* **Neural Networks**
* **Principle Component Analysis**
* **Independent Component Analysis**
* **Apriori algorithm**
* **Singular value decomposition**

**Advantages of Unsupervised Learning**

**Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.**

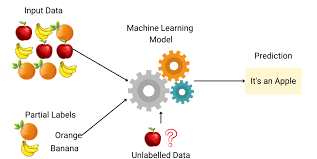
**Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.**

**Disadvantages of Unsupervised Learning**

**Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.**

**The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.**

**3) Semi-Supervised Learning**



**Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.**

**Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.**

**To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced. The main aim of semi-supervised learning is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.**

**We can imagine these algorithms with an example. Supervised learning is where a student is under the supervision of an instructor at home and college. Further, if that student is self-analysing the same concept without any help from the instructor, it comes under unsupervised learning. Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of an instructor at college.**

**Advantages :**

* **It is simple and easy to understand the algorithm.**
* **It is highly efficient.**
* **It is used to solve drawbacks**

**Disadvantages :**

* **Iterations results may not be stable.**
* **We cannot apply these algorithms to network-level data.**
* **Accuracy is low.**

**4) Reinforcement Learning**



**Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance. Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.**

**In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.**

**The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.**

**Due to its way of working, reinforcement learning is employed in different fields such as Game theory, Operation Research, Information theory, multi-agent systems.**

**A reinforcement learning problem can be formalized using Markov Decision Process(MDP). In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.**

**Categories of Reinforcement Learning**

**Reinforcement learning is categorized mainly into two types of methods/algorithms:**

**Positive Reinforcement Learning: Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.**

**Negative Reinforcement Learning: Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.**

**Real-world Use cases of Reinforcement Learning**

**Video Games:**

**RL algorithms are much popular in gaming applications. It is used to gain super-human performance. Some popular games that use RL algorithms are AlphaGO and AlphaGO Zero.**

**Resource Management:**

**The "Resource Management with Deep Reinforcement Learning" paper showed that how to use RL in computer to automatically learn and schedule resources to wait for different jobs in order to minimize average job slowdown.**

**Robotics:**

**RL is widely being used in Robotics applications. Robots are used in the industrial and manufacturing area, and these robots are made more powerful with reinforcement learning. There are different industries that have their vision of building intelligent robots using AI and Machine learning technology.**

**Text Mining**

**Text-mining, one of the great applications of NLP, is now being implemented with the help of Reinforcement Learning by Salesforce company.**

**Advantages and Disadvantages of Reinforcement Learning :**

**Advantages :**

* **It helps in solving complex real-world problems which are difficult to be solved by general techniques.**
* **The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.**
* **Helps in achieving long term results**.

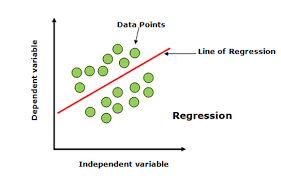
**Disadvantage :**

* **RL algorithms are not preferred for simple problems.**
* **RL algorithms require huge data and computations.**
* **Too much reinforcement learning can lead to an overload of states which can weaken the results.**

**List of Popular Machine Learning Algorithm :**

* **Linear Regression Algorithm**
* **Logistic Regression Algorithm**
* **Decision Tree**
* **SVM**
* **Naïve Bayes**
* **KNN**
* **K-Means Clustering**
* **Random Forest**
* **Apriori**
* **PCA**

**1. Linear Regression**



**Linear regression is one of the most popular and simple machine learning algorithms that is used for predictive analysis. Here, predictive analysis defines prediction of something, and linear regression makes predictions for continuous numbers such as salary, age, etc.**

**It shows the linear relationship between the dependent and independent variables, and shows how the dependent variable(y) changes according to the independent variable (x).**

**It tries to best fit a line between the dependent and independent variables, and this best fit line is knowns as the regression line.**

**The equation for the regression line is:**

**y= a0+ a\*x+ b**

**Here, y= dependent variable**

**x= independent variable**

**a0 = Intercept of line.**

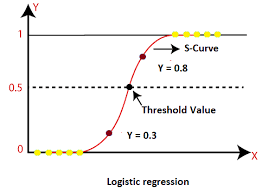
**Linear regression is further divided into two types:**

**The below dSimple Linear Regression: In simple linear regression, a single independent variable is used to predict the value of the dependent variable.**

**Multiple Linear Regression: In multiple linear regression, more than one independent variables are used to predict the value of the dependent variable.**

**iagram shows the linear regression for prediction of weight according to height:**

**2. Logistic Regression**

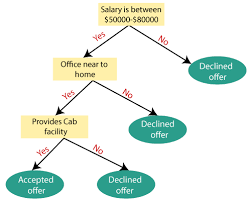


**Logistic regression is the supervised learning algorithm, which is used to predict the categorical variables or discrete values. It can be used for the classification problems in machine learning, and the output of the logistic regression algorithm can be either Yes or NO, 0 or 1, Red or Blue, etc.**

**Logistic regression is similar to the linear regression except how they are used, such as Linear regression is used to solve the regression problem and predict continuous values, whereas Logistic regression is used to solve the Classification problem and used to predict the discrete values.**

**Instead of fitting the best fit line, it forms an S-shaped curve that lies between 0 and 1. The S-shaped curve is also known as a logistic function that uses the concept of the threshold. Any value above the threshold will tend to 1, and below the threshold will tend to 0.**

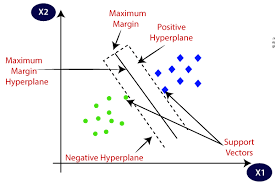
**3. Decision Tree Algorithm**



**A decision tree is a supervised learning algorithm that is mainly used to solve the classification problems but can also be used for solving the regression problems. It can work with both categorical variables and continuous variables. It shows a tree-like structure that includes nodes and branches, and starts with the root node that expand on further branches till the leaf node. The internal node is used to represent the features of the dataset, branches show the decision rules, and leaf nodes represent the outcome of the problem.**

**Some real-world applications of decision tree algorithms are identification between cancerous and non-cancerous cells, suggestions to customers to buy a car, etc.**

**4. Support Vector Machine Algorithm**

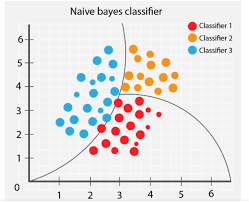


**A support vector machine or SVM is a supervised learning algorithm that can also be used for classification and regression problems. However, it is primarily used for classification problems. The goal of SVM is to create a hyperplane or decision boundary that can segregate datasets into different classes.**

**The data points that help to define the hyperplane are known as support vectors, and hence it is named as support vector machine algorithm.**

**Some real-life applications of SVM are face detection, image classification, Drug discovery, etc. Consider the below diagram:**

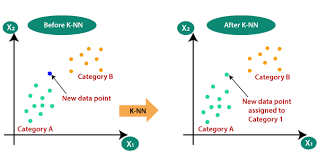
**5. Naive Bayes Algorithm:**



**Naïve Bayes classifier is a supervised learning algorithm, which is used to make predictions based on the probability of the object. The algorithm named as Naïve Bayes as it is based on Bayes theorem, and follows the naïve assumption that says' variables are independent of each other.**

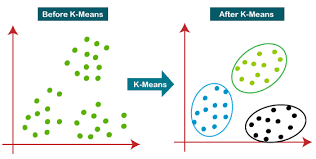
**The Bayes theorem is based on the conditional probability; it means the likelihood that event(A) will happen, when it is given that event(B) has already happened. The equation for Bayes theorem is given as:**

**6. K-Nearest Neighbour (KNN)**



**K-Nearest Neighbour is a supervised learning algorithm that can be used for both classification and regression problems. This algorithm works by assuming the similarities between the new data point and available data points. Based on these similarities, the new data points are put in the most similar categories. It is also known as the lazy learner algorithm as it stores all the available datasets and classifies each new case with the help of K-neighbours. The new case is assigned to the nearest class with most similarities, and any distance function measures the distance between the data points. The distance function can be Euclidean, Minkowski, Manhattan, or Hamming distance, based on the requirement.**

**7. K-Means Clustering**



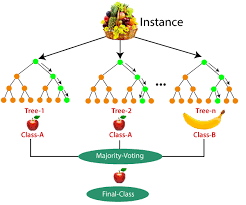
**K-means clustering is one of the simplest unsupervised learning algorithms, which is used to solve the clustering problems. The datasets are grouped into K different clusters based on similarities and dissimilarities, it means, datasets with most of the commonalties remain in one cluster which has very less or no commonalities between other clusters. In K-means, K-refers to the number of clusters, and means refer to the averaging the dataset in order to find the centroid.**

**It is a centroid-based algorithm, and each cluster is associated with a centroid. This algorithm aims to reduce the distance between the data points and their centroids within a cluster.**

**This algorithm starts with a group of randomly selected centroids that form the clusters at starting and then perform the iterative process to optimize these centroids' positions.**

**It can be used for spam detection and filtering, identification of fake news, etc.**

**8. Random Forest Algorithm**



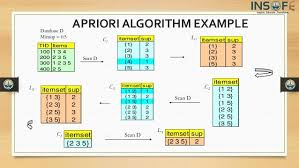
**Random forest is the supervised learning algorithm that can be used for both classification and regression problems in machine learning. It is an ensemble learning technique that provides the predictions by combining the multiple classifiers and improve the performance of the model.**

**It contains multiple decision trees for subsets of the given dataset, and find the average to improve the predictive accuracy of the model. A random-forest should contain 64-128 trees. The greater number of trees leads to higher accuracy of the algorithm.**

**To classify a new dataset or object, each tree gives the classification result and based on the majority votes, the algorithm predicts the final output.**

**Random forest is a fast algorithm, and can efficiently deal with the missing & incorrect data.**

**9. Apriori Algorithm**

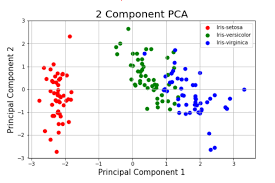


**Apriori algorithm is the unsupervised learning algorithm that is used to solve the association problems. It uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rule, it determines how strongly or how weakly two objects are connected to each other. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.**

**The algorithm process iteratively for finding the frequent itemsets from the large dataset.**

**The apriori algorithm was given by the R. Agrawal and Srikant in the year 1994. It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions in patients.**

**10. Principle Component Analysis**



**Principle Component Analysis (PCA) is an unsupervised learning technique, which is used for dimensionality reduction. It helps in reducing the dimensionality of the dataset that contains many features correlated with each other. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. It is one of the popular tools that is used for exploratory data analysis and predictive modeling.**

**PCA works by considering the variance of each attribute because the high variance shows the good split between the classes, and hence it reduces the dimensionality.**

**Some real-world applications of PCA are image processing, movie recommendation system, optimizing the power allocation in various communication channels.**

**program:**

**from sklearn.metrics import accuracy\_score**

**print("Logistic Regression is %f percent accurate" % (accuracy\_score(LR\_pred, Y\_test)\*100))**

**print("KNN is %f percent accurate" % (accuracy\_score(KNN\_pred, Y\_test)\*100))**

**print("Naive Bayes is %f percent accurate" % (accuracy\_score(NB\_pred, Y\_test)\*100))**

**print("Linear SVMs is %f percent accurate" % (accuracy\_score(LSVM\_pred, Y\_test)\*100))**

**print("Non Linear SVMs is %f percent accurate" % (accuracy\_score(NLSVM\_pred, Y\_test)\*100))**

**print("Decision Trees is %f percent accurate" % (accuracy\_score(DT\_pred, Y\_test)\*100))**

**print("Random Forests is %f percent accurate" % (accuracy\_score(RF\_pred, Y\_test)\*100))**

**Output:**

**Logistic Regression is 84.357542 percent accurate**

**KNN is 75.977654 percent acccurate**

**Naive Bayes is 82.681564 percent accurate**

**Linear SVMs is 35.754190 percent accurate**

**Non Linear SVMs is 78.212291 percent accurate**

**Decision Trees is 80.446927 percent accurate**

**Random Forests is 84.916201 percent accurate**

**Conclusion:**

**Now, you know that Machine Learning is a technique of training machines to perform the activities a human brain can do, albeit bit faster and better than an average human-being. Today we have seen that the machines can beat human champions in games such as Chess, AlphaGO, which are considered very complex. You have seen that machines can be trained to perform human activities in several areas and can aid humans in living better lives.**

**Machine Learning can be a Supervised or Unsupervised. If you have lesser amount of data and clearly labelled data for training, opt for Supervised Learning. Unsupervised Learning would generally give better performance and results for large data sets. If you have a huge data set easily available, go for deep learning techniques. You also have learned Reinforcement Learning and Deep Reinforcement Learning. You now know what Neural Networks are, their applications and limitations.**

**Finally, when it comes to the development of machine learning models of your own, you looked at the choices of various development languages, IDEs and Platforms. Next thing that you need to do is start learning and practicing each machine learning technique. The subject is vast, it means that there is width, but if you consider the depth, each topic can be learned in a few hours. Each topic is independent of each other. You need to take into consideration one topic at a time, learn it, practice it and implement the algorithm/s in it using a language choice of yours. This is the best way to start studying Machine Learning. Practicing one topic at a time, very soon you would acquire the width that is eventually required of a Machine Learning expert.**