Iris Flower Classification

Name: Arjun Bhardwaj

Registration number: 11914887

Section: KM016

Abstract

Machine learning, as a powerful approach to achieve Artificial Intelligence, has been widely used in pattern recognition, a very basic skill for humans but a challenge for machines. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

1. Introduction:

Since the computer was invented, it has begun to affect our daily life. It improves the quality of our lives, it makes our life more convenient and more efficient. A fascinating idea is to let a computer think and learn as a human. Basically, machine learning is to let a computer develop learning skills by itself with given knowledge. Pattern recognition can be treated like computer being able to recognize different species of objects. Therefore, machine learning has close connection with pattern recognition. In this project, the object is the Iris flower. The data set of Iris contains three different classes: Setosa, Versicolour, and Virginica. The designed recognition system will distinguish these three different classes of Iris.

Iris flower has three species; setosa, versicolor, and virginica, which differs according to their measurements. Now assume that you have the measurements of the iris flowers according to their species, and here your task is to train a machine learning model that can learn from the measurements of the iris species and classify them

Iris Versicolor **Iris Setosa** **Iris Virginica**

Here you can see above these the pictures of three species of iris and I'm building this machine learning model for the recognition of these species by our computer system.

1.1 Objective:

The objective of this machine learning project is to predict the species of the flowers according to the characteristics of the Iris dataset. This model is trained to learn patterns from the data set based on those features.

Major Objectives of this Machine Leaning Model:

- To classify different species of iris flower
- To train & test the model
- Improve the dataset for the optimal output
- To test the accuracy of different models like Logistic Regression, KNN Neighbours, Naïve Bayes & Decision Tree

The major objective of this model is to classify different species of iris flower with the help of different machine learning algorithms, and test their accuracy to get the best result possible. Because these type of classification will help to include the machine learning in our daily life, more the machine systems will learn about our surroundings and able to classify between them more they were ready for usage in future tech environment what I can see the use of this type of model is in the farming sector where we can classify different species of plants with the help of machine and with this ability they can automatically remove the unwanted plants like weeds, also it can be used to sour the particular species plant all this may be sounds little bit ambitious right now but all this possible one day with these small initial steps like this model.

2. Dataset Description:

The Iris dataset was used in R.A. Fisher's classic 1936 paper, The Use of Multiple Measurements in Taxonomic Problems, and can also be found on the UCI Machine Learning Repository.

It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

The columns in this dataset are:

- Id
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm
- Species

Description of Different Features in Dataset:

Septal Length:

The sepal is a leaf -shaped structure found in iris flower, or angiosperms. It is found on the outermost part of the iris flower, and like a petal, a sepal is considered to be a modified leaf. Its length parameter is one of the feature.

Septal Width:

The sepal is a leaf -shaped structure found in iris flower, or angiosperms. It is found on the outermost part of the iris flower, and like a petal, a sepal is considered to be a modified leaf. Its Width parameter is one of the feature.

Petal Length:

Petals are modified leaves that surround the reproductive parts of iris flowers. They are often brightly colored or unusually shaped to attract pollinators. Its length parameter is one of the feature.

Petal Width:

Petals are modified leaves that surround the reproductive parts of iris flowers. They are often brightly colored or unusually shaped to attract pollinators. Its Width parameter is one of the feature.

Species:

There are three different species of iris flower present in the dataset setosa, versicolor, and virginica.

Dataset Sample:

```
    Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  \
0    1            5.1           3.5            1.4           0.2
1    2            4.9           3.0            1.4           0.2
2    3            4.7           3.2            1.3           0.2
3    4            4.6           3.1            1.5           0.2
4    5            5.0           3.6            1.4           0.2
..  ...            ...           ...            ...           ...
145  146            6.7           3.0            5.2           2.3
146  147            6.3           2.5            5.0           1.9
147  148            6.5           3.0            5.2           2.0
148  149            6.2           3.4            5.4           2.3
149  150            5.9           3.0            5.1           1.8

            Species
0       Iris-setosa
1       Iris-setosa
2       Iris-setosa
3       Iris-setosa
4       Iris-setosa
..              ...
145  Iris-virginica
146  Iris-virginica
147  Iris-virginica
148  Iris-virginica
```

.

3. Approach:

In this approach four different models with different algorithms of machine learning are applied training dataset to predict the results. 75% of the data from the dataset is used to train the system and results are predicted by using 25% of the test data. Features are selected from the dataset to improve the results. After that all models accuracy has to be check & compared and on that of that get the best result possible.

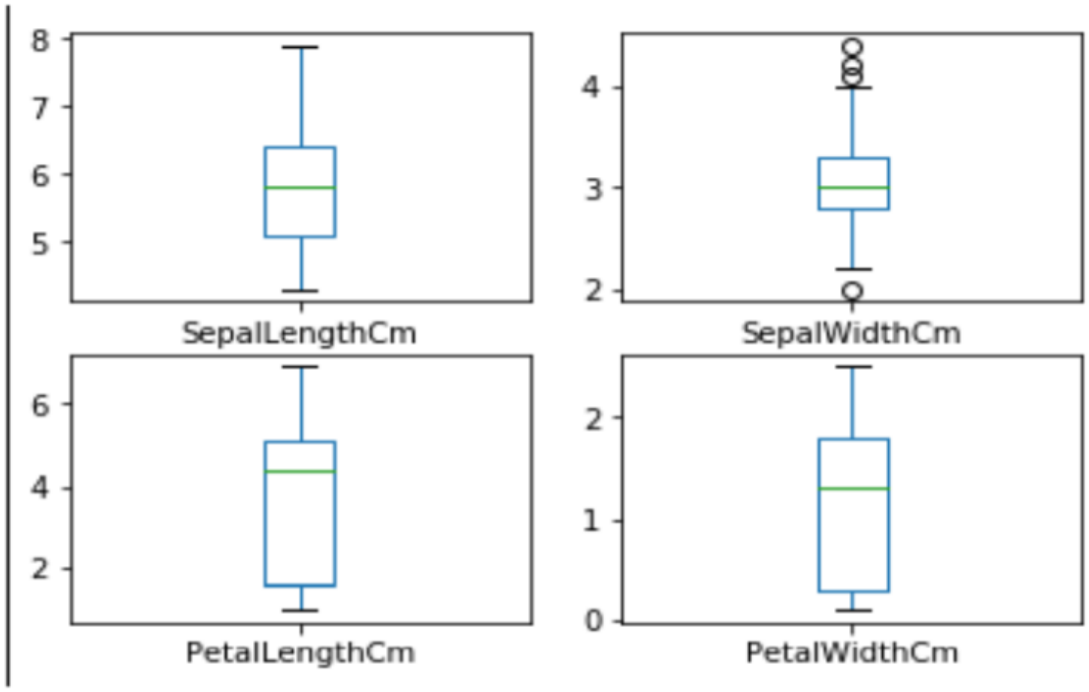
4. Features & Correlation:

Features Used:

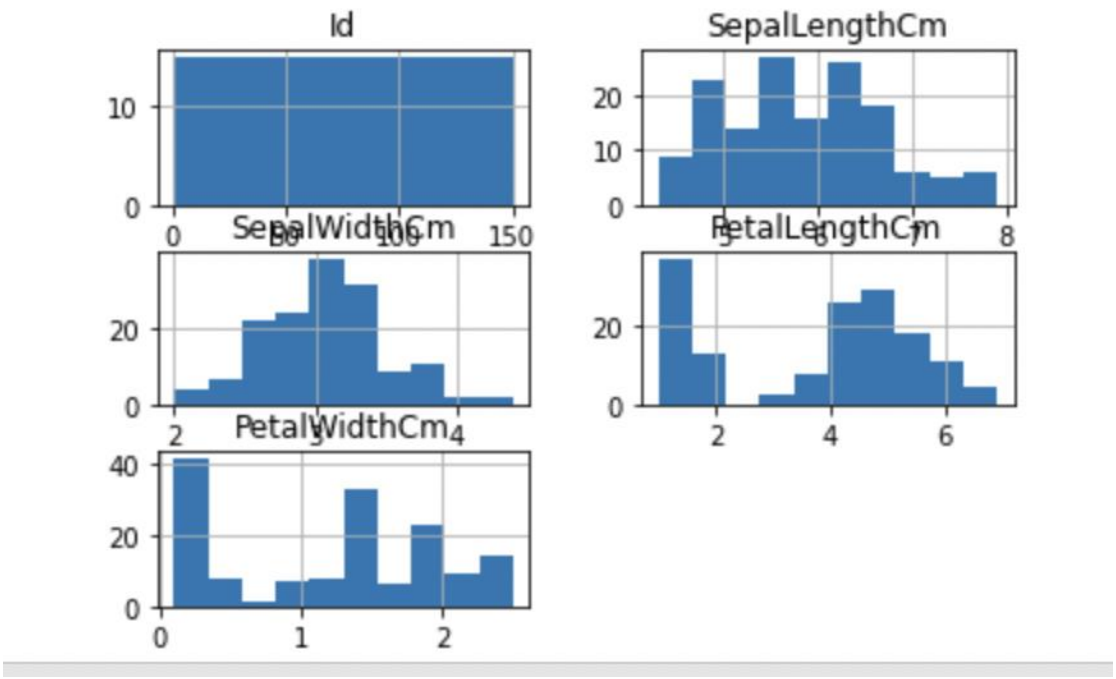
1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

4.1. Checking for outliers:

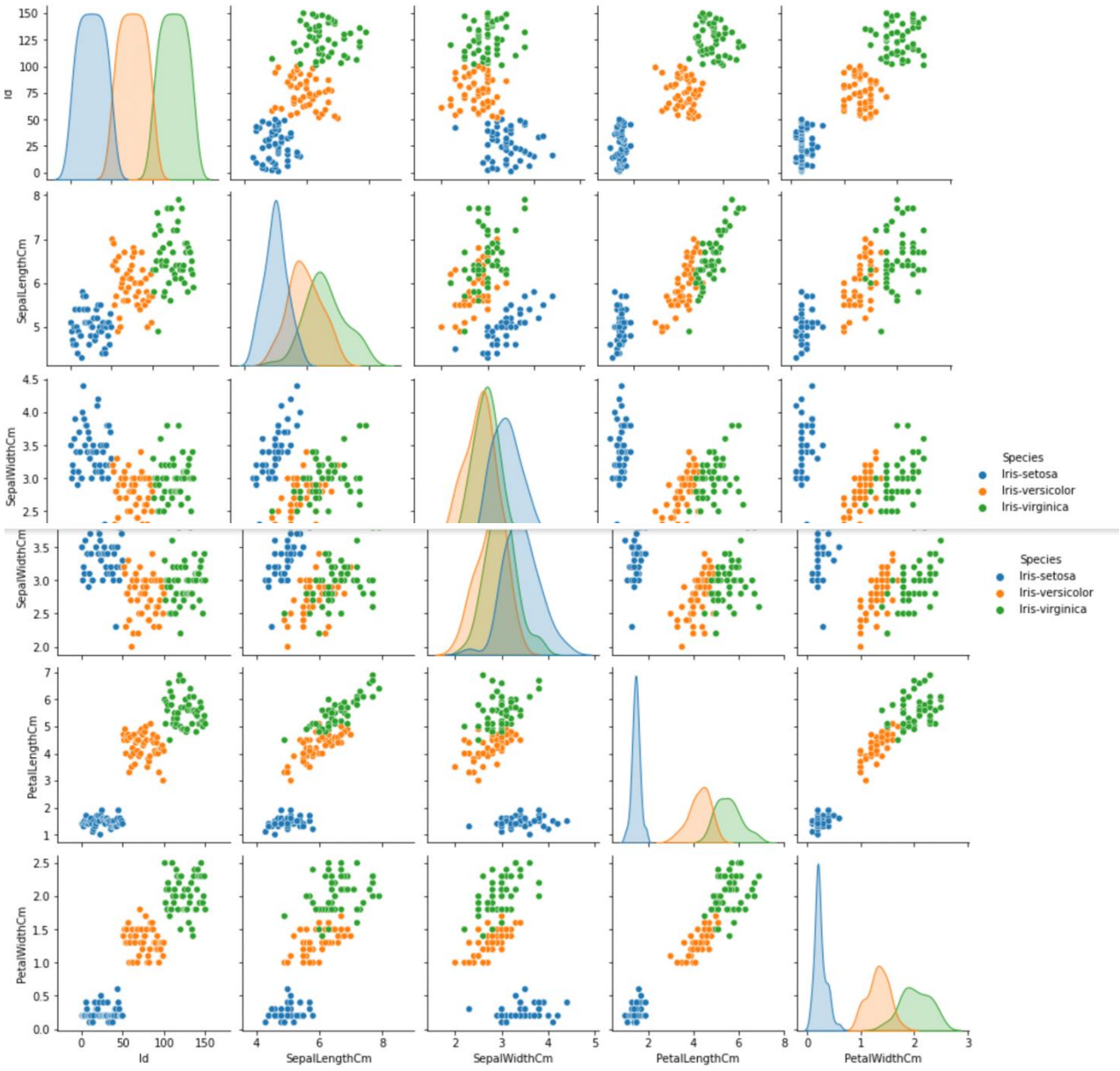
Outliers are nothing but data points that differ significantly from other observations. They are the points that lie outside the overall distribution of the dataset. Outliers, if not treated, can cause serious problems in statistical analyses.



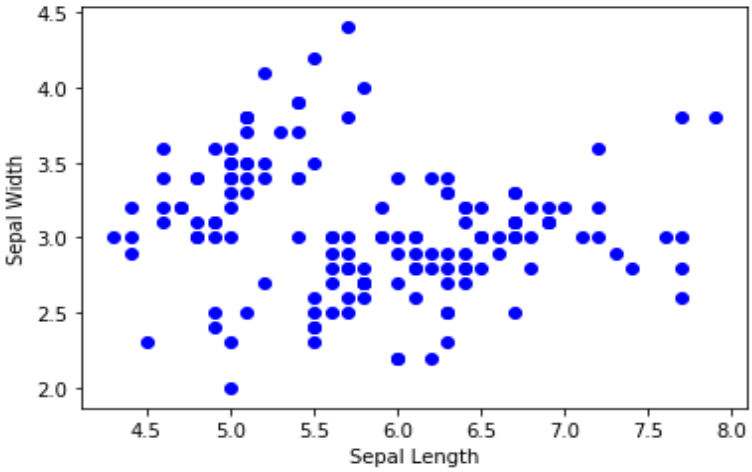
Box and whisker plots(Give idea about distribution of input attributes)



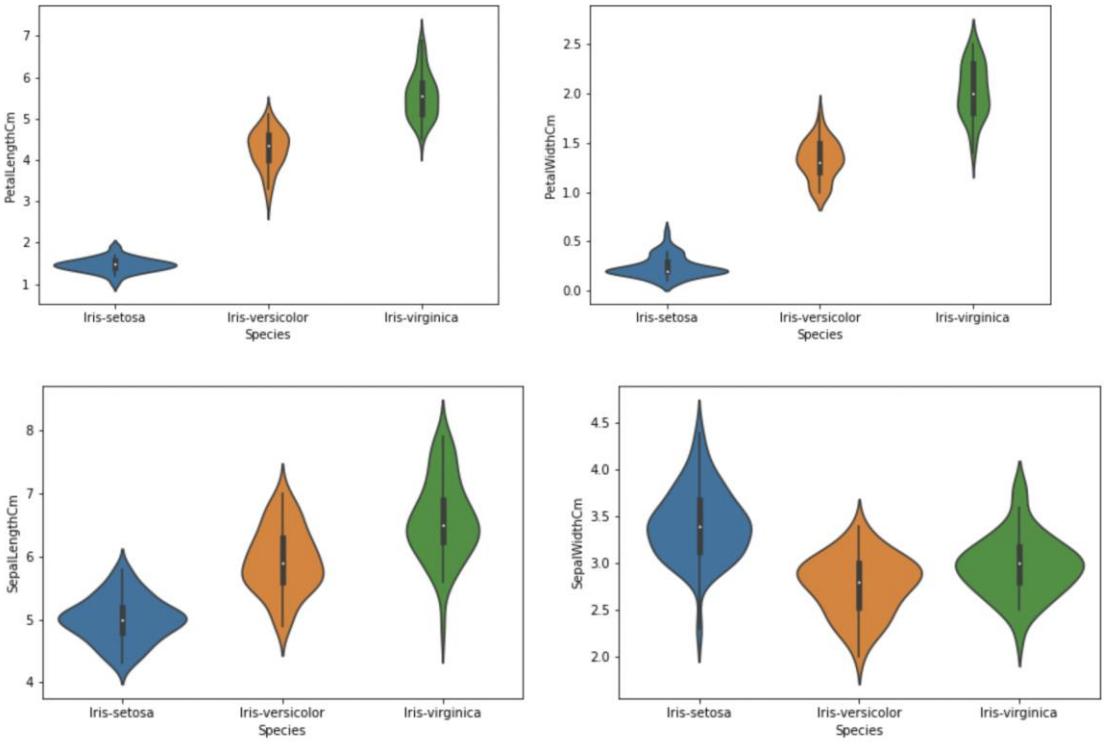
Plotting Histogram

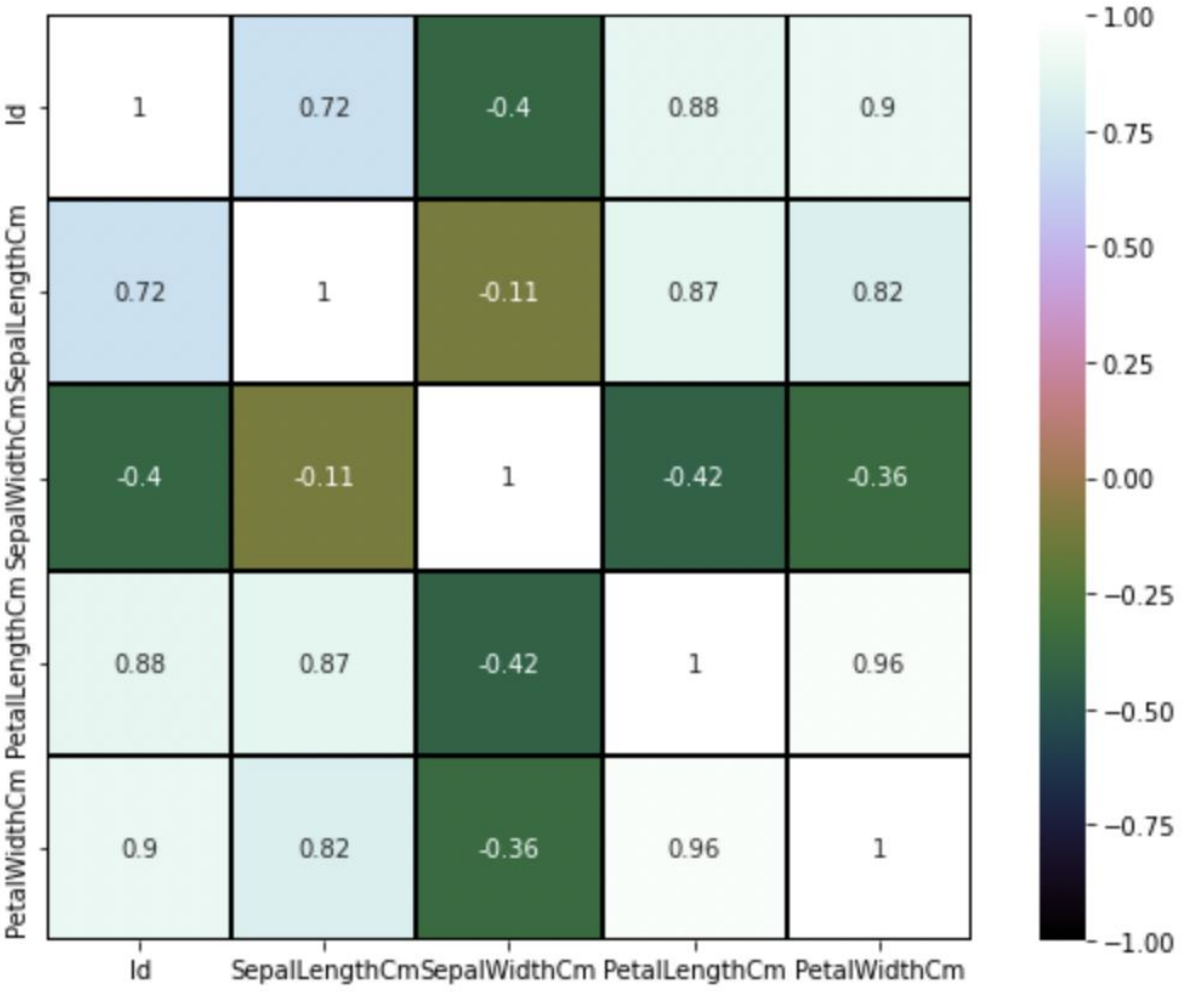

Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Species
Iris-setosa
Iris-versicolor
Iris-virginica

In [75]:
```
plt.xlabel("Sepal Length")
plt.ylabel("Sepal Width")
plt.scatter(X,Y,color='b')
plt.show()
```



Plotting Scatter Graph Between Sepal Length and Sepal Width:





Heat Map of data set

4.1. Correlation in all features in dataset:

This is the correlation among all the feature in the dataset.

```
                      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  \
Id              1.000000       0.716676     -0.397729       0.882747
SepalLengthCm   0.716676       1.000000     -0.109369       0.871754
SepalWidthCm   -0.397729      -0.109369      1.000000      -0.420516
PetalLengthCm   0.882747       0.871754     -0.420516       1.000000
PetalWidthCm    0.899759       0.817954     -0.356544       0.962757

                PetalWidthCm
Id                  0.899759
SepalLengthCm       0.817954
SepalWidthCm       -0.356544
PetalLengthCm       0.962757
PetalWidthCm        1.000000
```

5. Observations:

1. Using Sepal_Lenght & Sepal_Width features, we can only distinguish Setosa flower from others.

   2. Seperating Versicolor & Virginica is much harder as they have considerable overlap.

   3. Hence, Sepal_Lenght & Sepal_Width features only work well for Setosa.

6. Machine Learning Algorithms/ Models Used:

1. Logistic Regression: The logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

2. Support Vector Machine: Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well

as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future

3. Gaussion Naive Bayes(NB) : Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features (see Bayes classifier)

4. K-Nearest Neighbour(KNN) : K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique, It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

5. Deision Tree: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility

7. Machine Learning Algorithms/ Models Evaluation:

Description of Model Evaluation Parameters:

Correlation(r): The correlation coefficient is a measure of linear association between the predicted numeric target value and the actual numeric value. Value of the correlation coefficient always lie between -1 and +1. A correlation coefficient of +1 means that two variables are perfectly related in a positive linear manner, a correlation coefficient of -1 means that two variables are perfectly related in a negative linear manner, and a correlation coefficient of 0 means that there is no linear relationship

present between the two variables. The correlation between two x and y variables are calculated by using equation 2

$$\Sigma$$
$$(x - \bar{x})(y - \bar{y})$$

$$\text{Corr}(r) = \sqrt{\Sigma}$$
$$(x - \bar{x})2\ (y - \bar{y})2$$

Accuracy: The prediction accuracy of each machine learning regression method is used to evaluate the overall match between actual and predicted values.

8. Result Analysis:

8.1 Performance comparison

Performance comparison of all the models is shown below. It has been found that Logistic Regression and SVM have highest Accuracy and Decision tree have lowest accuracy

```
results = pd.DataFrame({
    'Model': ['Logistic Regression','Support Vector Machines', 'Naive Bayes','KNN' ,'Decision Tree'],
    'Score': [0.947,0.947,0.947,0.947,0.921]})

result_df = results.sort_values(by='Score', ascending=False)
result_df = result_df.set_index('Score')
result_df.head(9)
```

Score	Model
0.947	Logistic Regression
0.947	Support Vector Machines
0.947	Naive Bayes
0.947	KNN
0.921	Decision Tree

9. Conclusion:

In this Project we used the model for prediction of different species of iris flower, The system is trained with 75%of the data set and 25% of the dataset is used for testing. We have five different models / algorithm for this purpose. Out of

four two models logistic regression & SVM shows highest accuracy of 94.7 percent and the decision tree shows the lowest accuracy. Also projects will opens the different doors in the field machine learning this project represents that machine can learn about our surroundings and can make decision according to that which is the biggest goal of this decade.

11. References

‘Clustering - K-means demo’, K-means-Ineractive demo, Available at: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html. Consulted 22 AUG 2013

Bache, K.& Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Bishop, C. 2006. Pattern Recognition and Machine Learning. New York: Springer, pp.424-428.

Fisher, R.A. 1936. UCI Machine Learning Repository: Iris Data Set. Available at: http://archive.ics.uci.edu/ml/datasets/Iris. Consulted 10 AUG 2013

Improved Outcomes Software.,2004. K-Means Clustering Overview, Available at: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/KMeans_Clustering_Overview.htm. Consulted 22 AUG 2013

Mitchell, T. 1997. Machine learning. McGraw Hill.

Mjolsness, E. & Decoste, D. 2001. Machine learning for science: state of the art and future prospects.Science, 293 (5537), pp. 2051--2055.

Pedregosa, F.& Varoquaux, G. 2.11., Scikit-learn: machine learning in Python — Scipy lecture notes, Available at: http://scipylectures.github.io/advanced/scikit-learn/. Consulted 22 AUG 2013

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,
Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Others.,2011. Scikitlearn: Machine Learning in Python., JMLR 12, pp. 2825-2830.
Python Software Foundation., 2013. General Python FAQ — Python v2.7.5 documentation. Available at: http://docs.python.org/2/faq/general.html. Consulted 20 AUG 2013