# VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018



**A Mini Project Report on**

***“MEDICAL EXPENDITURE PREDICTION SYSTEM”***

*Submitted in partial fulfillment of the requirements as a part of the*

# AI/ML INTERNSHIP (NASTECH)

*For the award of degree of*

## Bachelor of Engineering in

**Information Science and Engineering**

Submitted by

## SUMANTH S SALONI KUMARI

**1RN18IS088 1RN18IS090**

### Internship Project Coordinators

**Dr. R Rajkumar Dr. Prakasha S**

**Associate Professor Associate Professor**

**Dept. of ISE, RNSIT Dept. of ISE, RNSIT**

**Department of Information Science and Engineering RNS Institute of Technology**

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post, Bengaluru – 560 098

**2020 -2021**

**RNS Institute of Technology**

**Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post, Bengaluru – 560 098**

**DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING**

**CERTIFICATE**

This is to certify that the mini project report entitled ***MEDICAL EXPENDITURE PREDICTION SYSTEM*** has been successfully completed by **SUMANTH S** bearing USN **1RN18IS088** and **SALONI KUMARI** bearing USN **1RN18IS090** , presently VII semester students of **RNS Institute of Technology** in partial fulfillment of the requirements as a part of the ***AI/ML Internship (NASTECH)*** for the award of the degree of ***Bachelor of Engineering in Information Science and Engineering*** under **Visvesvaraya Technological University, Belagavi** during academic year **2021 – 2022**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report and deposited in the departmental library. The mini project report has been approved as it satisfies the academic requirements as a part of Internship.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dr. R Rajkumar** | **Dr. Prakasha S** |  | **Dr. Suresh L** |
| Coordinator | Guide |  | Professor and HoD |
| Associate Professor | Associate Professor |  |  |

|  |  |
| --- | --- |
| **External Viva** | |
| **Name of the Examiners** | **Signature with date** |
| 1. |  |
| 2. |  |

# ABSTRACT

The Iris flower data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems.

It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.

The data set consists of 50 samples from each of three species of Iris (Iris Setosa, Iris virginica, and Iris versicolor)

This dataset became a typical test case for many statistical classification techniques in machine learning such as support vector machines

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other

# ACKNOWLEDGMENT

The fulfillment and rapture that go with the fruitful finishing of any assignment would be inadequate without the specifying the people who made it conceivable, whose steady direction and support delegated the endeavors with success.

We would like to profoundly thank **Management** of **RNS Institute of Technology**

for providing such a healthy environment to carry out this AI/ML Internship Project.

We would like to express our thanks to our Principal **Dr. M K Venkatesha** for his support and inspired us towards the attainment of knowledge.

We wish to place on record our words of gratitude to **Dr. Suresh L,** Professor and Head of the Department, Information Science and Engineering, for being the enzyme and master mind behind our Internship Project.

We would like to express our profound and cordial gratitude to my Internship Project Coordinators**, Dr. R Rajkumar,** Associate Professor, Department of Information Science and Engineering for their valuable guidance, constructive comments, continuous encouragement throughout the Internship Project and guidance in preparing report.

I would like to express my profound and cordial gratitude to my Faculty in charge **Dr. Prakasha S**, Associate Professor, Department of Information Science and Engineering for his/her valuable guidance in preparing Project report.

We would like to thank all other teaching and non-teaching staff of Information Science & Engineering who have directly or indirectly helped us to carry out the Internship Project.

Also, we would like to acknowledge and thank our parents who are source of inspiration and instrumental in carrying out this Mini Project Work.

**SUMANTHM S SALONI KUMARI**

**USN:1RN18IS088 USN:1RN18IS090**

# TABLE OF CONTENTS

**Abstract iii**

Acknowledgement iv

Table of Content v

[List of Figures](#_TOC_250004) vi

1. INTRODUCTION 01
   1. [ORGANIZATION/ INDUSTRY](#_bookmark2) 01
      1. Company Profile 01
      2. Domain/ Technology (Data Science/Mobile computing/…) 01
      3. Department/ Division / Group 02
   2. PROBLEM STATEMENT 02
      1. [Existing System and their Limitations 02](#_TOC_250001)
      2. [Proposed Solution 02](#_TOC_250000)
      3. Problem formulation 02
2. **REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES** 03
   1. Hardware & Software Requirements 03
   2. Tools/ Languages/ Platform 03
3. **DESIGN AND IMPLIMENTATION** 04
   1. Architecture/ DFD/Sequence diagram/Class diagrams /Flowchart 04
   2. Problem statement 05
   3. Algorithm/Methods/ Pseudo code 07
   4. Libraries used / API’S 07
4. **OBSERVATIONS AND RESULTS** 08
   1. Testing 08
   2. Results & Snapshots 08
5. **CONCLUSION AND FUTURE WORK** 14
   1. Conclusion 14
   2. Future Enhancement 14
6. **REFERENCES** 15

# LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **Figure. No.** | **Descriptions** | **Page** |
| Figure. 3.1 | Gradient boosting regression Model | 04 |
| Figure. 3.2 | Description of the Dataset | 06 |
| Figure. 4.1 | Reading CSV file | 08 |
| Figure. 4.2 | Price prediction model | 09 |
| Figure. 4.3 | Data Analysis | 09 |
| Figure. 4.4 | Price range for smokers vs non-smokers | 10 |
| Figure. 4.5 | Linear Regression model | 10 |
| Figure. 4.6 | Polynomial Regression model | 11 |
| Figure. 4.7 | Random Forest Regression model | 11 |
| Figure. 4.8 | Gradient Booster Regression model | 12 |
| Figure. 4.9  Figure. 4.9.1 | Visualizing Accuracies of different algorithms  Result | 12  13 |

**Chapter 1**

**INTRODUCTION**

### ORGANIZATION/INDUSTRY

### COMPANY PROFILE

NASTECH is formed with the purpose of bridging the gap between Academia and Industry Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

* + 1. **DOMAIN/TECHNOLOGY**

The domain chosen for our project is AI/ML. Machine learning, the fundamental driver of AI, is possible through algorithms that can learn themselves from data and identify patterns to make predictions and achieve your predefined goals, rather than blindly following detailed programmed instructions, like in traditional computer programming. This technology allows the machine to perceive, learn, reason and communicate through observation of data, like a child that grows up and acquires knowledge from examples. Machines also have the advantage of not being limited by our inherent biological limitations. With machine learning, manufacturing companies have increased production capacity up to 20%, while lowering material consumption rates by 4%.

Nowadays, the revolutionary AI technology evolved from rule-based expert systems to machine learning and more advanced subcomponents such as deep learning (learning representations instead of tasks), artificial neural networks (inspired by animal brains) and reinforcement learning (virtual agents rewarded if they made good decisions).

The AI can master the complexity of the intertwining industrial processes to enhance the whole flow of production instead of isolated processes. This enormous cognitive capacity gives the AI the ability to consider the spatial organization of plants and the timing constraints of live production. Another key advantage is the capability of AI algorithms to think probabilistically, with all the subtlety this allows in edge cases, instead of traditional rule-based methods that require rigid theories and a full comprehension of problems.

### Department

R.N.Shetty Institute of Technology (RNSIT) established in the year 2001, is the brain-child of the Group Chairman, Dr. R. N. Shetty. The Murudeshwar Group of Companies headed by Sri. R. N. Shetty is a leading player in many industries viz construction, manufacturing, hotel, automobile, power & IT services and education. The group has contributed significantly to the field of education. A number of educational institutions are run by the

R. N. Shetty Trust, RNSIT being one amongst them. With a continuous desire to provide quality education to the society, the group has established RNSIT, an institution to nourish and produce the best of engineering talents in the country. RNSIT is one of the best and top accredited engineering colleges in Bengaluru.

### PROBLEM STATEMENT

### Existing System and their Limitations

A manual method is currently used in the market to predict the insurance price. The problem with this is that it is very time consuming and tedious process for the insurance company to manually provide this for every customer. To overcome this, insurance companies tend to hire an agent which again increases the cost of the process.

Moreover, there is a chance that the agent might be prone to manual error or bribery.

### Proposed Solution

To eliminate the drawback of manual method, Machine learning algorithms can be used by insurance companies to provide a fast, easy and customer friendly approach for the problem. Also, the new system will be cost and time efficient. This will have simple operations.

### Program formulation

The proposed system works on Gradient Boosting Regression Algorithm. This algorithm takes into account all the different conditions on which the insurance can be provided and gives a highly accurate estimate.

### Chapter 2

**REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES**

* 1. **Hardware and Software Requirements**
     1. **Hardware Requirements:**
        + Processor: Pentium IV or above
        + RAM: 4 GB or more
        + Hard Disk: 2GB or more
     2. **Software Requirements:**
        + Operating System: Windows 7 or above
        + IDE: Google Colab

### Tools/Languages/Platforms

* Python

### Chapter 3

**DESIGN AND IMPLIMENTATION**

* 1. **Architecture/ DFD/Sequence diagram/Class diagrams /Flowchart**

**K-Means** K-means clustering algorithm **tries to group similar items in the form of clusters**. The number of groups is represented by K. ... If you will notice here then you will find that they are forming a group or cluster, where each of the vegetables is kept within their kind of group forming the clusters.

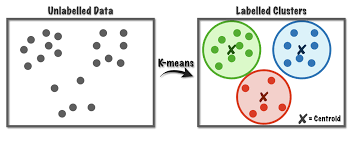


Figure 3.1 K-Means Model

### Problem Statement

The goal of this statistical analysis is to help us understand the relationship between insurance prices and how these variables are used to predict insurance price.

Gradient boosting regression Model has been used in terms of minimizing the difference between predicted and actual rating.

The following features have been used:

* + 1. Sepal length: describes the length of the sepal of the flower in cm
    2. Sepal width: describes the width of the sepal of the flower in cm.
    3. Petal length: describes the length of the petal of the flower in cm.
    4. Petal width: describes the width of the petal of the flower in cm.
    5. Species: can be of three types iris versiclolor iris setosa iris virginica .

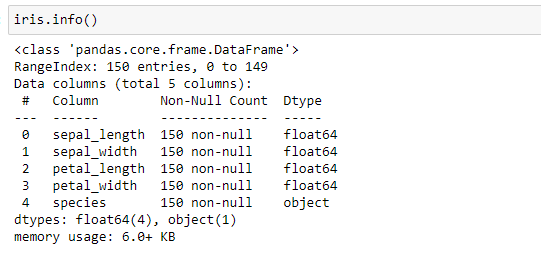
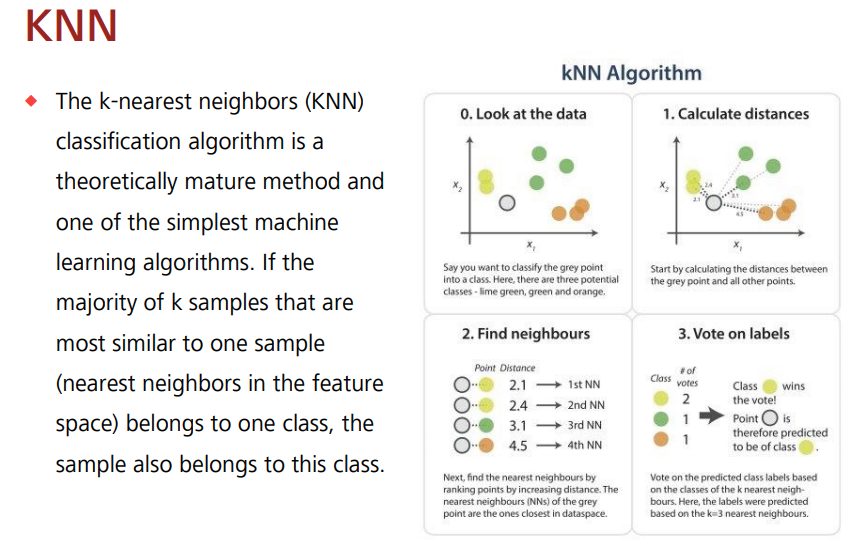


Figure 3.2 Description of the Dataset

The above fig3.2, shows the description of the dataset.

### 3.3 Algorithm



### Libraries

* + - Pandas
    - Numpy
    - Plotly
    - sklearn
    - Seaborn
    - Matplotlip

**Chapter 4**

**OBSERVATION AND RESULTS**

### Testing

**Evaluation on Test Data**

knn = KNeighborsClassifier(n\_neighbors=5, p=2, metric='minkowski')

knn.fit(x\_train, y\_train)

print('The accuracy of the knn classifier is {:.2f} out of 1 on training data'.format(knn.score(x\_train, y\_train)))

print('The accuracy of the knn classifier is {:.2f} out of 1 on test data'.format(knn.score(x\_test, y\_test)))Visualizing Our predictions based on different algorithms

predict = pd.DataFrame(data = models, columns=['Model','MAE', 'MSE', 'RMSE', 'Variance Score'])

**Perfect predictions**

k\_range = list(range(1,26))

scores = []

for k in k\_range:

knn = KNeighborsClassifier(n\_neighbors=k)

knn.fit(x, y y\_pred = knn.predict(x)

scores.append(metrics.accuracy\_score(y, y\_pred)

plt.plot(k\_range, scores)

plt.xlabel('Value of k for KNN')

plt.ylabel('Accuracy Score')

plt.title('Accuracy Scores for Values

### Results & Snapshots

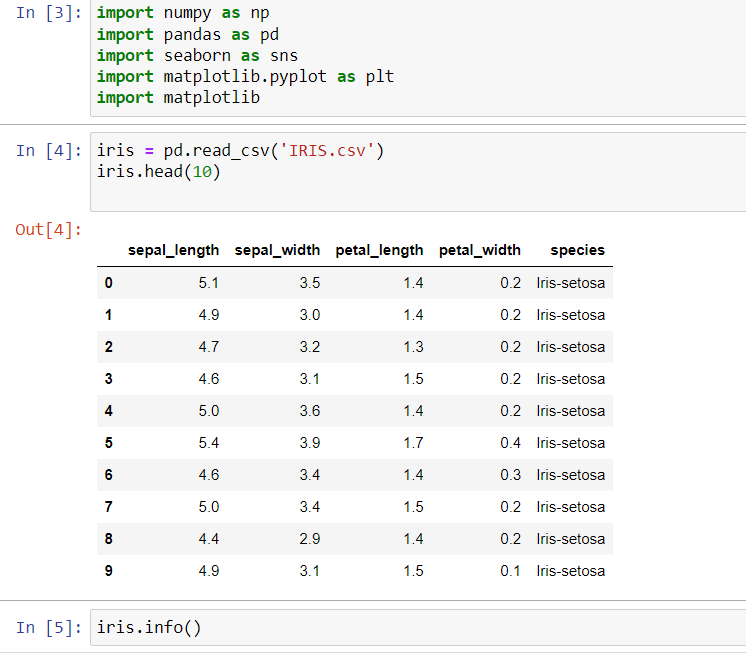
****

Figure 4.1 Reading CSV File

In the above fig 4.1, we are first importing all the modules required and then reading the dataset.csv file.

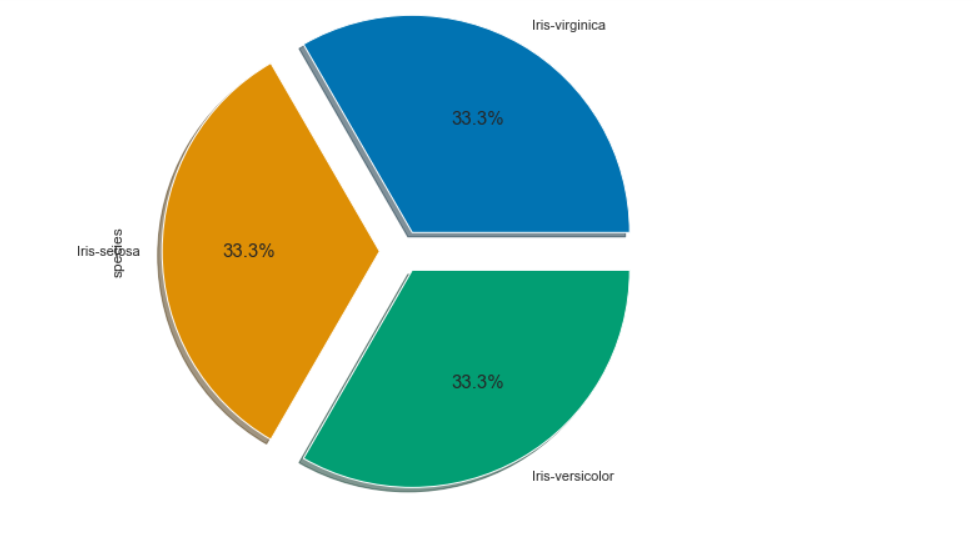
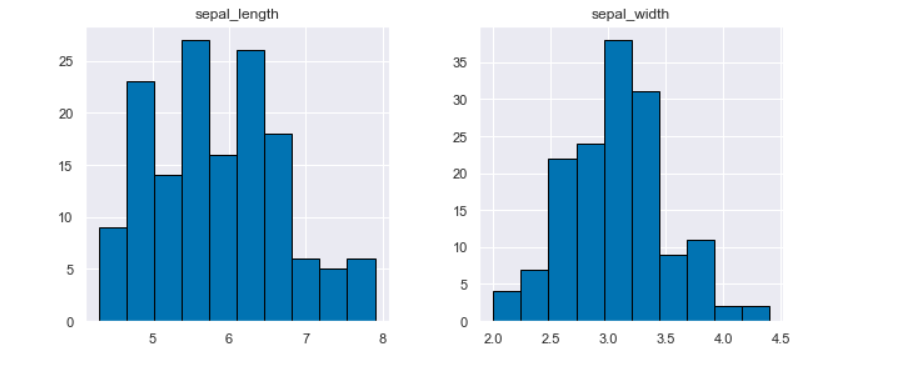


Figure 4.2 iris prediction model

In the above fig4.2,iris versicolor is 33.3% ,iris virginica is 33.3% and iris setosa is 33.3%.



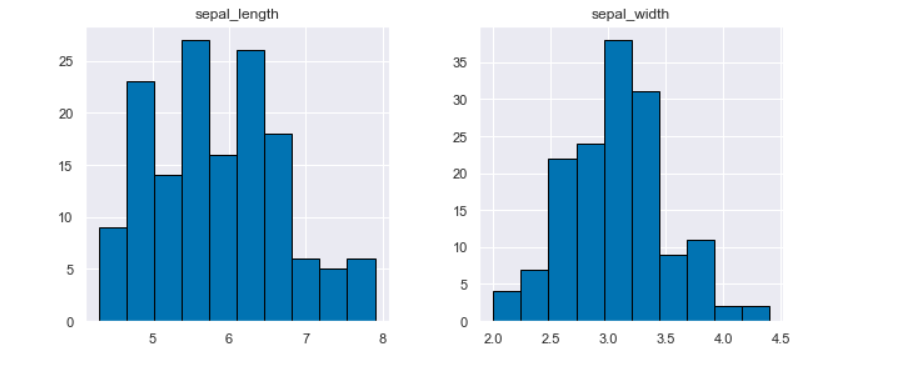


Figure 4.3 Data Analysis

In the above fig4.3comparison of sepal length,sepal width ,petal length and petal width .

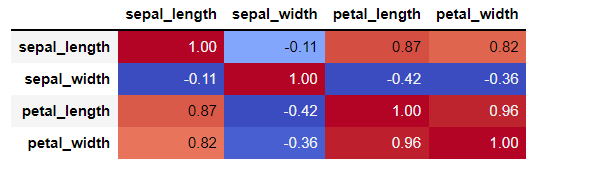
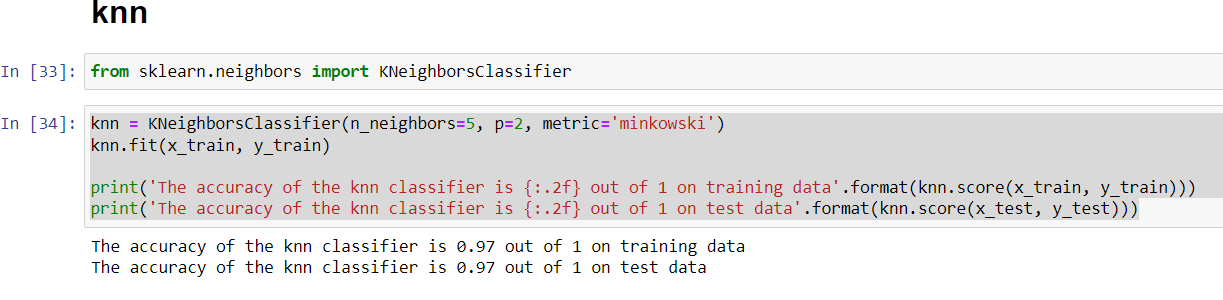


Figure 4.4 sepal length correlation

In the above fig4.4we are viewing the correlation of the all features with sepal length.



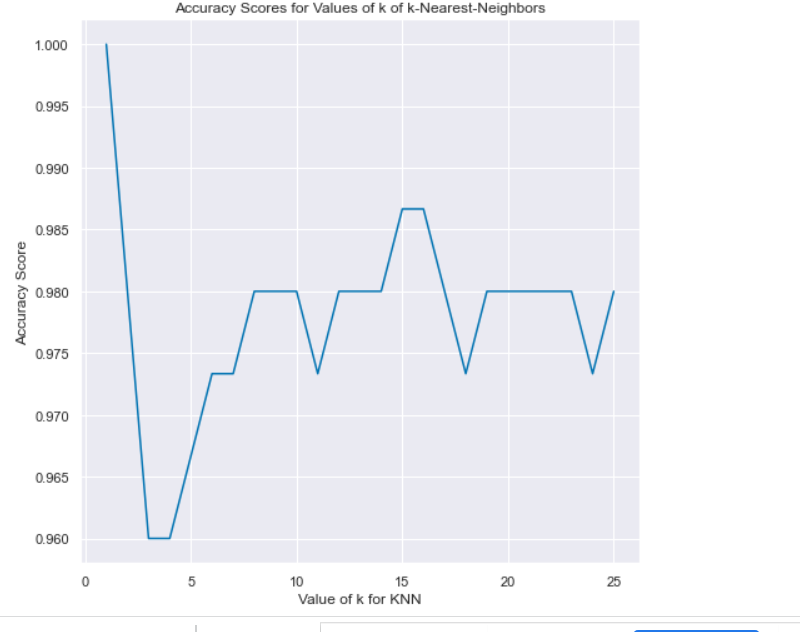


Figure 4.5 knn

In the above fig4.5KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

.

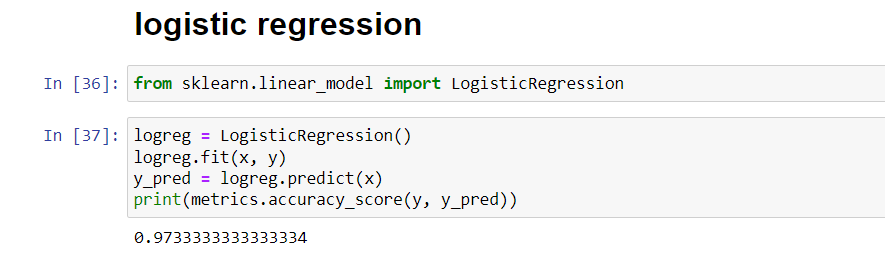


Figure 4.6 logisticRegression model

In the above fig4.5, Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable,

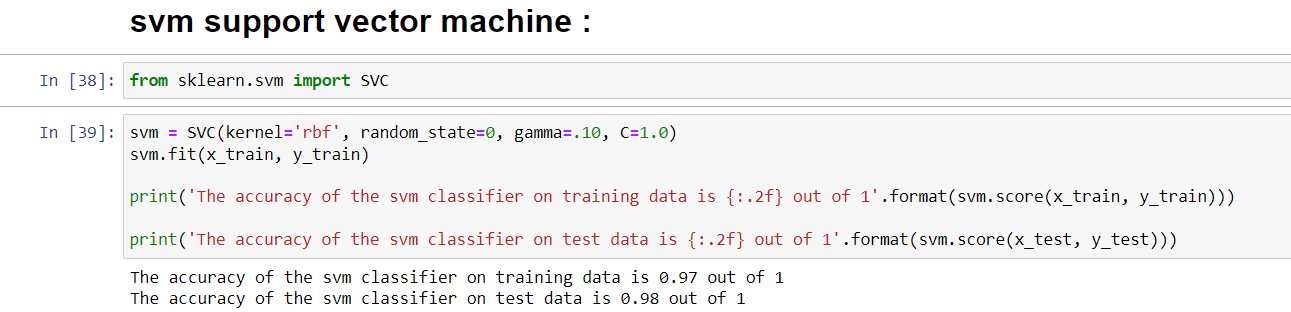
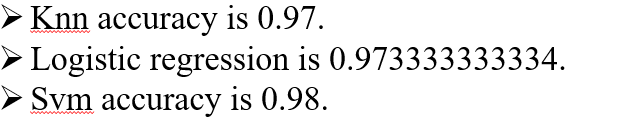


Figure 4.7 svm model

In the above fig4.7, Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection

**RESULT**



### Chapter 5

**CONCLUSION AND FUTURE ENHANCEMENT**

* 1. **Conclusion**

The proposed model is the best substitute for the manual method where a third party is involved as the middleman and is potentially vulnerable along with it being cheaper for the end customers.

Based on the results, it can be concluded that such ML-driven predictions are easily comprehendible and significant from a data-analytics point of view.

When correctly implemented, a high rate of accuracy can be achieved.

### Future Enhancement

* To make the interface more informative and user-friendly by implementing better GUI designs.
* Using bigger training data sets to get a more accurate estimate of the prices.
* Implementing other machine learning algorithms which can improve the accuracy of the model.

## Chapter 6

**REFERENCE**

1. Dual Spatial Attention Network for Iris Recognition

**DOI:**10.1109/WACV48630.2021.00093

2)A novel fuzzy clustering method based on GA, PSO and Subtractive Clustering

Year: 2020, Pages: 331-337

DOI Bookmark: 10.1109/CSCI51800.2020.00063

3)A novel fuzzy clustering method based on GA, PSO and Subtractive Clustering

Year: 2020, Pages: 331-337

DOI Bookmark: 10.1109/CSCI51800.2020.00063