Marathwada Shikshan Prasarak Mandal’s

**Deogiri Institute of Engineering and Management Studies,**

**Aurangabad**

**Seminar Report**

**On**

**Classification Algorithms**

Submitted By

**Shriram Mogal (36053)**

**Dr. Babasaheb Ambedkar Technological University**

**Lonere (M.S.)**



Department of Computer Science and Engineering

**Deogiri Institute of Engineering and Management Studies,**

**Aurangabad**

(2019- 2020)

**Seminar Report**

**On**

**Classification Algorithms**

Submitted By

**Shriram Mogal (36053)**

**In partial fulfillment of**

**Bachelor of Technology**

**(Computer Science & Engineering)**

Guided By

**Prof. Amruta Joshi**

Department of Computer Science & Engineering

**Deogiri Institute of Engineering and Management Studies,**

**Aurangabad**

(2019- 2020)

**CERTIFICATE**

This is to certify that, the Seminar entitled “**Classification Algorithms**” submitted by

**Shriram Mogal (36053)** is a bonafide work completed under my supervision and guidance in partial fulfillment for award of Bachelor of Technology (Computer Science and Engineering) Degree of Dr. Babasaheb Ambedkar Technological University, Lonere.

Place: Aurangabad

Date:

**Prof. Amruta Joshi Mr. S.B. Kalyankar**

**Guide Head**

**Dr. Ulhas D. Shiurkar**

**Director,**

**Deogiri Institute of Engineering and Management Studies,**

**Aurangabad**

**Abstract**

Supervised machine learning is the construction of algorithms that are able to produce general patterns and hypotheses by using externally supplied instances to predict the fate of future instances. Supervised machine learning classification algorithms aim at categorizing data from prior information. Classification is carried out very frequently in data science problems. Various successful techniques have been proposed to solve such problems viz. Rule-based techniques, Logic-based techniques, Instance-based techniques, stochastic techniques.

Unlike regression where you predict a continuous number, you use classification to predict a category. There is a wide variety of classification applications from medicine to marketing. Classification models include linear models like Logistic Regression, SVM, and nonlinear ones like K-NN, Kernel SVM and Random Forests.

Following are some Methods of Classification:

1. Logistic Regression
2. K-Nearest Neighbors (K-NN)
3. Support Vector Machine (SVM)
4. Kernel SVM
5. Naive Bayes
6. Decision Tree Classification
7. Random Forest Classification

###### Contents

**List of Figures i**

**List of Graphs ii**

**List of Screens iii**

1. **INTRODUCTION** 
   1. Introduction 1
   2. Decision Tree 3
   3. Naive Bayes 4
   4. Artificial Neural Networks 5
   5. k-Nearest Neighbour 6

1. **LITERATURE SURVEY**

2.1 Paper 1: Research of Decision Tree Classification 7

2.2 Paper 2 :Short Survey on Naive Bayes Algorithm 8

2.3 Paper 3: k-Nearest Neighbour Classiﬁers 9

1. **IMPLEMENTATION**

3.1Visualization11

3.2 Evaluation of Algorithms 13

1. **CONCLUSIONS** 4.1 Conclusion 15

4.2 Application 15

**REFERENCES**

**ACKNOWLEDGEMENT**

**List of Figures**

**Figure Illustration Page**

1.1 General Picture of Classification 1

1.2 Decision Tree 3

1.3 Artificial Neural Networks 13

i

**List of Graphs**

**Figure Illustration Page**

1.4 k-Nearest Neighbor 14

3.2 Visualization of Logistic Regression 11

3.3 Visualization of K-NN 11

3.4 Visualization of SVM 11

3.5 Visualization of Kernal-SVM 11

3.6 Visualization of Naive Bayes 12

3.7 Visualization of Decision Tree 12

3.8 Visualization of Random Forest 12

**ii**

**List of Screens**

**Figure Illustration Page**

3.1 Dataset 10

3.10 Evaluation of Logistic Regression 11

3.11 Evaluation of K-NN 11

3.12 Evaluation of SVM 11

3.13 Evaluation of Kernal-SVM 11

3.14 Evaluation of Naive Bayes 12

3.15 Evaluation of Decision Tree 12

3.16 Evaluation of Random Forest 12

iii

1. **INTRODUCTION**

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).



Fig. 1.1

For example, spam detection in email service providers can be identified as a classification problem. This is s binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email.Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

There are two types of learners in classification as lazy learners and eager learners.

1. **Lazy Learner**

Lazy learners simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting.*Ex. k-nearest neighbor, Case-based reasoning*

**2. Eager learners**

Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict.*Ex. Decision Tree, Naive Bayes, Artificial Neural Networks*

# Classification algorithms

There is a lot of classification algorithms available now but it is not possible to conclude which one is superior to other. It depends on the application and nature of available data set. For example, if the classes are linearly separable, the linear classifiers like Logistic regression, Fisher’s linear discriminant can outperform sophisticated models and vice versa.

2

**1.2 Decision Tree**

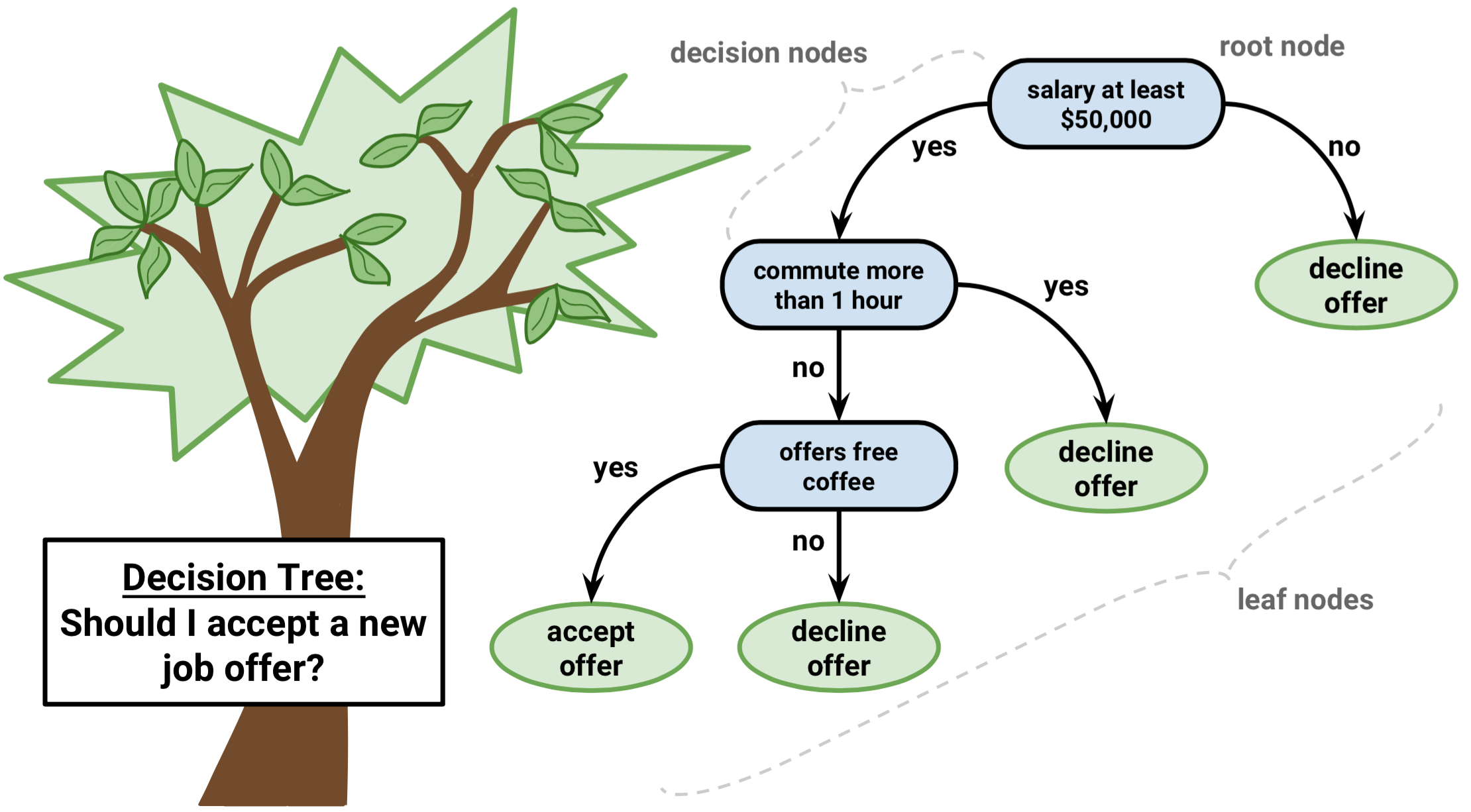


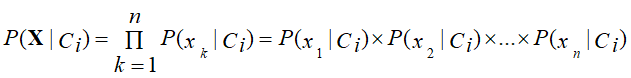
Fig. 1.2

Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition.The tree is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance..A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers. An over-fitted model has a very poor performance on the unseen data even though it gives an impressive performance on training data. This can be avoided by pre-pruning which halts tree construction early or post-pruning which removes branches from the fully grown tree.

3

## 1.3 Naive Bayes

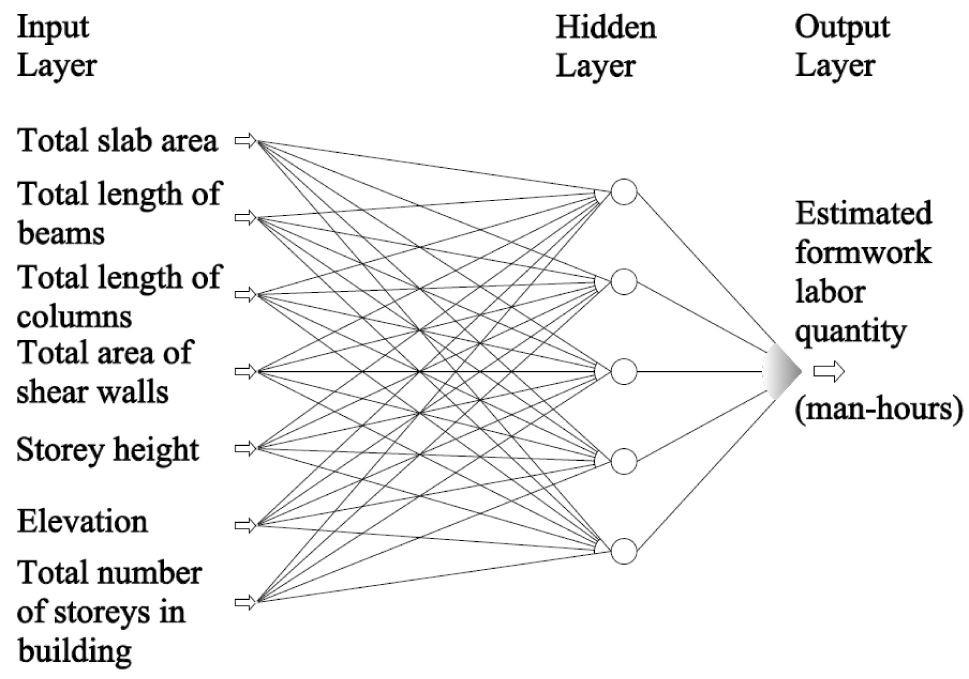
Naive Bayes is a probabilistic classifier inspired by the Bayes theorem under a simple assumption which is the attributes are conditionally independent.

The classification is conducted by deriving the maximum posterior which is the maximal P(Ci|**X**) with the above assumption applying to Bayes theorem. This assumption greatly reduces the computational cost by only counting the class distribution. Even though the assumption is not valid in most cases since the attributes are dependent, surprisingly Naive Bayes has able to perform impressively.

Naive Bayes is a very simple algorithm to implement and good results have obtained in most cases. It can be easily scalable to larger datasets since it takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.Naive Bayes can suffer from a problem called the zero probability problem. When the conditional probability is zero for a particular attribute, it fails to give a valid prediction. This needs to be fixed explicitly using a Laplacian estimator.

4

**1.4 Artificial Neural Networks**

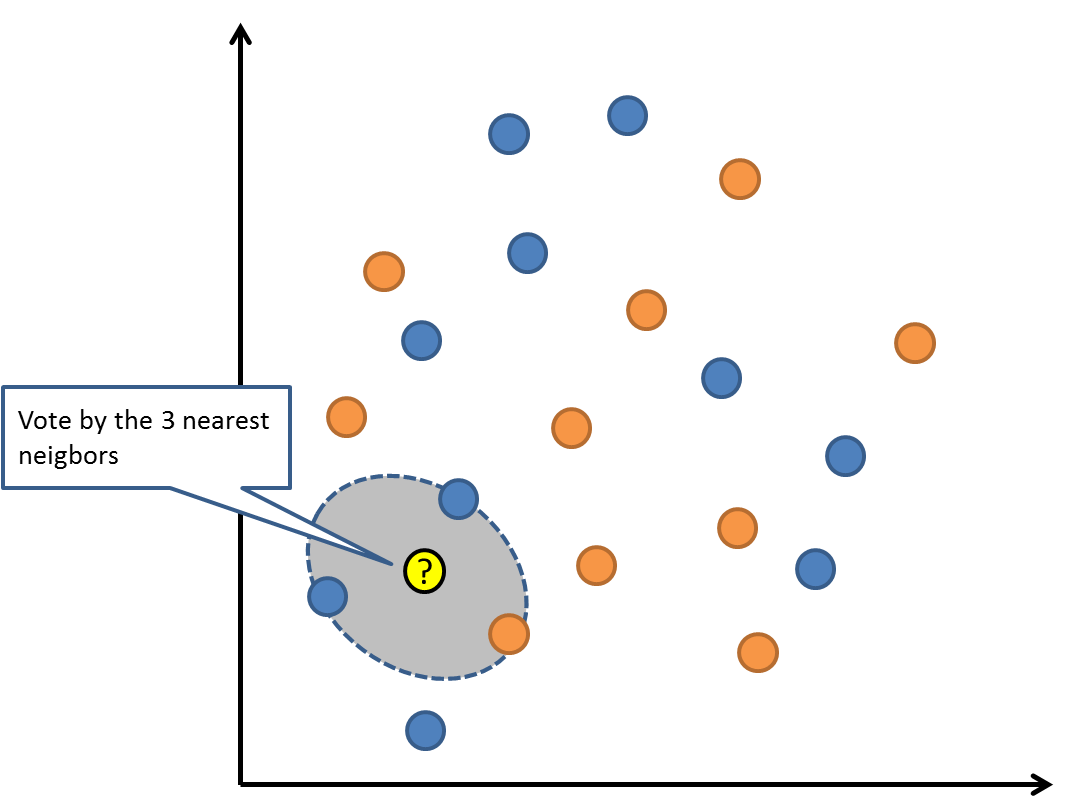
****

**Fig. 1.3**

Artificial Neural Network is a set of connected input/output units where each connection has a weight associated with it started by psychologists and neurobiologists to develop and test computational analogs of neurons. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples.There can be multiple hidden layers in the model depending on the complexity of the function which is going to be mapped by the model..However, when there are many hidden layers, it takes a lot of time to train and adjust wights. But Artificial Neural Networks have performed impressively in most of the real world applications. It is high tolerance to noisy data and able to classify untrained patterns. Usually, Artificial Neural Networks perform better with continuous-valued inputs and outputs.

5

**1.5 k-Nearest Neighbor (KNN)**

****

**Fig. 1.4**

*k*-Nearest Neighbor is a lazy learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors)and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors.In the distance-weighted nearest neighbor algorithm, it weights the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors.

6

1. **LITERATURE SURVEY**

I referred following research papers and studied them. This literature survey contains short information about them.

**2.1 Paper 1: Research of Decision Tree Classification Algorithm in Data Mining**

Decision tree is the main technology used for classification and prediction. Decision tree learning is a typical inductive algorithm based on instance, which focus on classification rules displaying as decision trees inferred from a group of disorder and irregular instance. In top-down recursive way, it compares attributes between internal nodes of decision tree, judges the downward branches according to different attributes of the node, and draws a conclusion from leaf nodes in the decision tree. So from a root to a leaf node corresponds to a conjunctive rule, and the entire tree corresponds to a group of disjunctive expression rules. Take the decision tree as a Boolean function. The input of the function is the object or all property of situation, and the output is the "yes" or "no" decision value. In the decision tree, each tree node corresponds to a property test, each leaf node corresponds to a Boolean value, and each branch represents one of the possible values of testing attribute.[4] ‘Short Survey on Naive Bayes Algorithm’ , International Journal of Advance Engineering and Research Development Volume 4, Issue 11, November -2017

7

**2.2 Paper 2 :Short Survey on Naive Bayes Algorithm**

Naïve Bayes is a subset of Bayesian decision theory. It’s called naive because the formulation makes some naïve assumptions. Python’s text-processing abilities which split up a document into a vector are used. This can be used to classify text. Classifies may put into human-readable form. It is a popular classification method in addition to conditional independence, overfitting, and Bayesian methods. In the face of the simplicity of Naive Bayes, it can classify documents surprisingly well. Instinctively a potential justification for the conditional independence assumption is that if the document is about politics, this is a good evidence of the kinds of other words found in the document. Naive Bayes is a reasonable classifier in this sense and has minimal storage and fast training, it is applied to time-storage critical applications, such as automatically classifying web pages into types and spam filtering. Considering a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, the aim is to create a rule which enables to allocate future objects to a class, given just the vectors of variables marking out the future objects. These problems are known as ―supervised classification problem‖, are worldwide, and most of the methods for constructing such rules have been developed. It is very easy to establish, and no need any complicated repetitive parameter estimation schemes. This means it should be applied to huge data sets. It is easy to interpret, so unskilled users in classifier technology can make out the reason for it is making the classification it makes. Finally, it often does surprisingly well: it should not be the best possible classifier in any particular application, but it can usually be relied on to be robust and to do well. [5] ‘Research of Decision Tree Classification Algorithm in Data Mining’,International Journal of Database Theory and Application Vol.9, No.5 (2016)

8

**2.3 Paper 3: k-Nearest Neighbour Classiﬁers**

The intuition underlying Nearest Neighbour Classiﬁcation is quite straightforward, examples are classiﬁed based on the class of their nearest neighbours. It is often useful to take more than one neighbour into account so the technique is more commonly referred to as k-Nearest Neighbour (k-NN) Classiﬁcation where k nearest neighbours are used in determining the class. Since the training examples are needed at run-time, i.e. they need to be in memory at run-time, it is sometimes also called Memory-Based Classiﬁcation. Because induction is delayed to run-time, it is considered a Lazy Learning technique. Because classiﬁcation is based directly on the training examples it is also called Example-Based Classiﬁcation or Case-Based Classiﬁcation.[6] ‘k-Nearest Neighbour Classiﬁers’, Technical Report UCD-CSI-2007-4 March 27, 2007

9

1. **IMPLEMENTATION**

The above mentioned classification methods are implemented here to show how they actually work.This all methods are implemented on the same dataset, so they can be compared. The data set consists of 400 records. The instance of data set shown below.

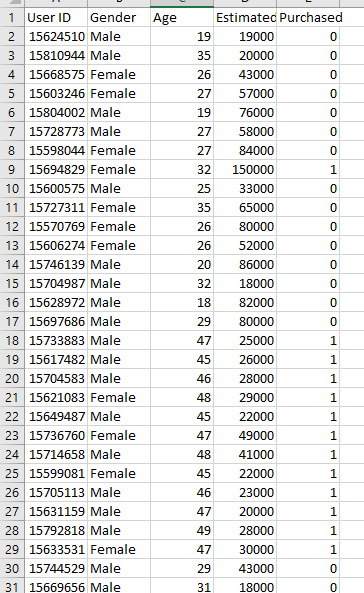


Fig. 3.1

10

**3.1 Visualization**

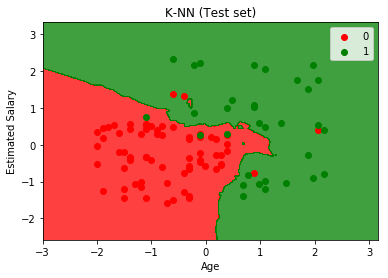
** **

Fig. 3.2 Fig. 3.4

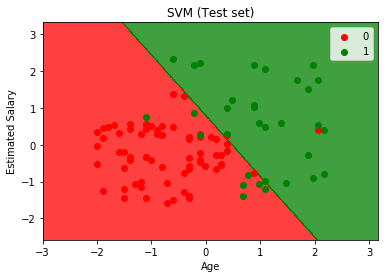
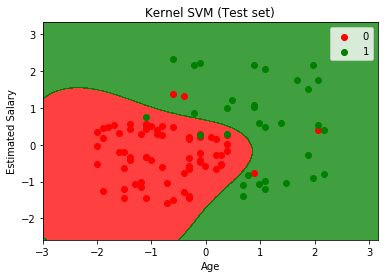
** **

Fig. 3.5 Fig. 3.6

11

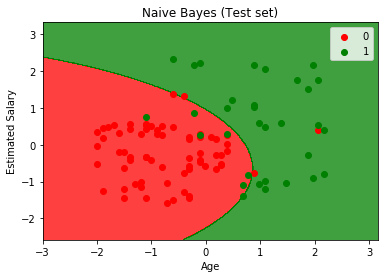
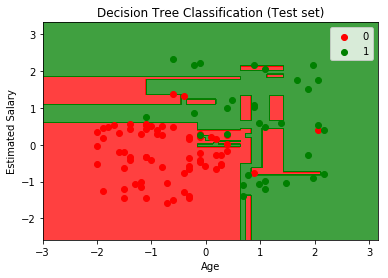
** **

Fig. 3.7 Fig. 3.8

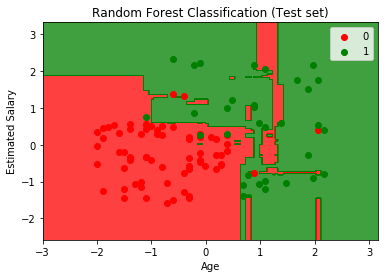
****

Fig. 3.9

12

**3.2 Evaluation of Algorithms**

I used confusion matrix method for evaluating the methods. In which principal diagonal shows the correct predicted tuples.

Logistic Regression K-NN

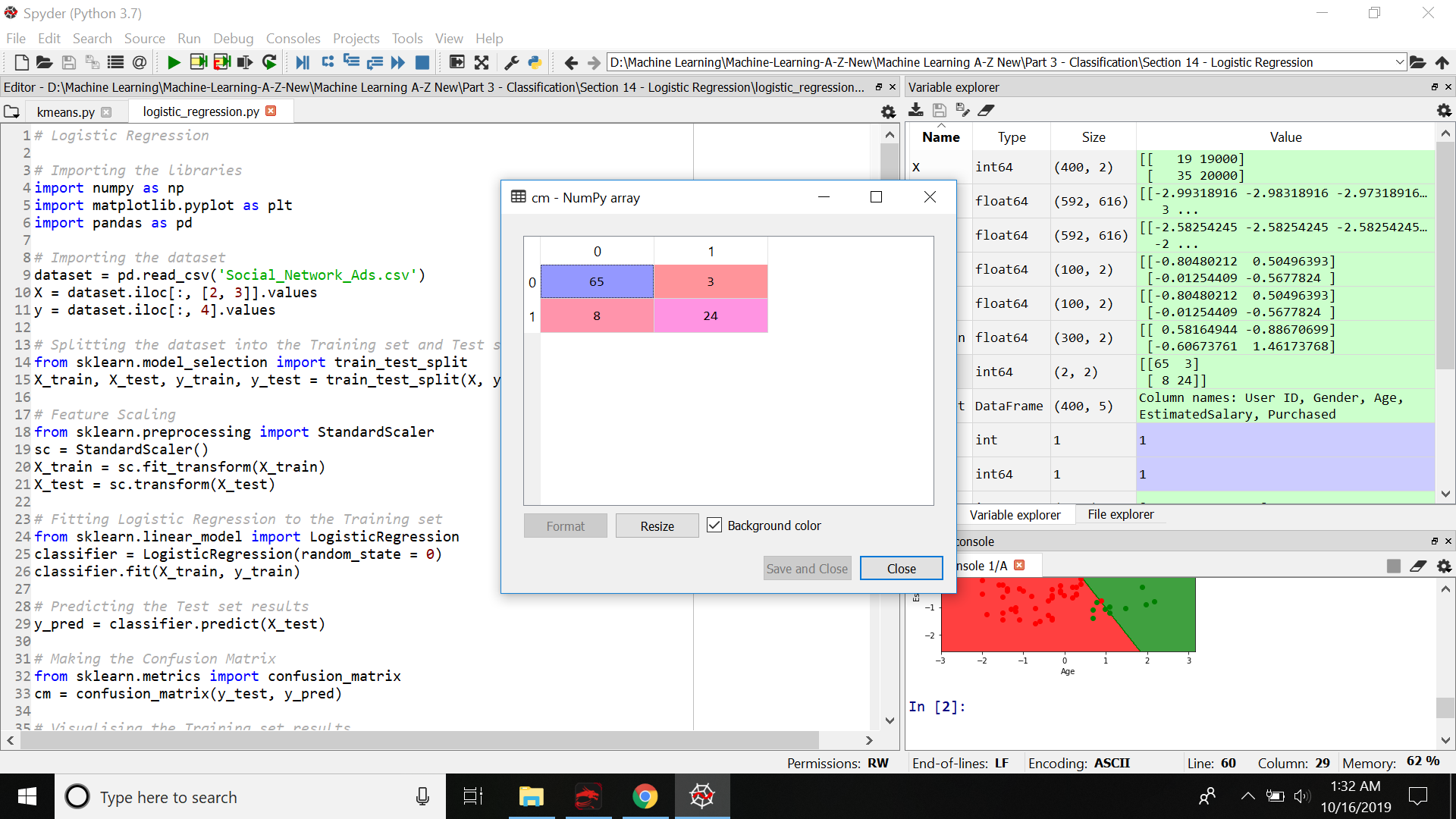
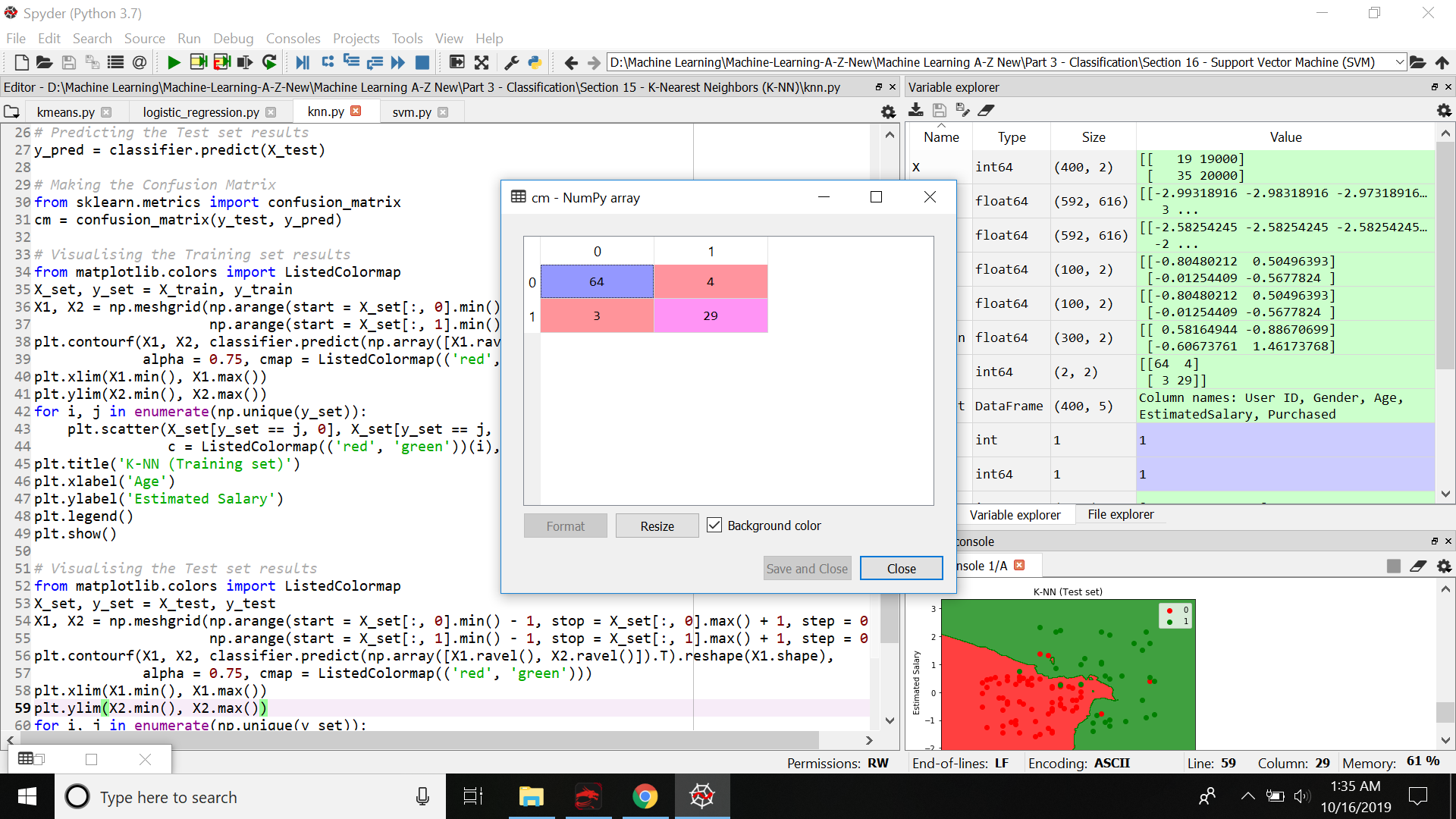
 

Fig. 3.10 Fig. 3.11

SVM Kernal SVM

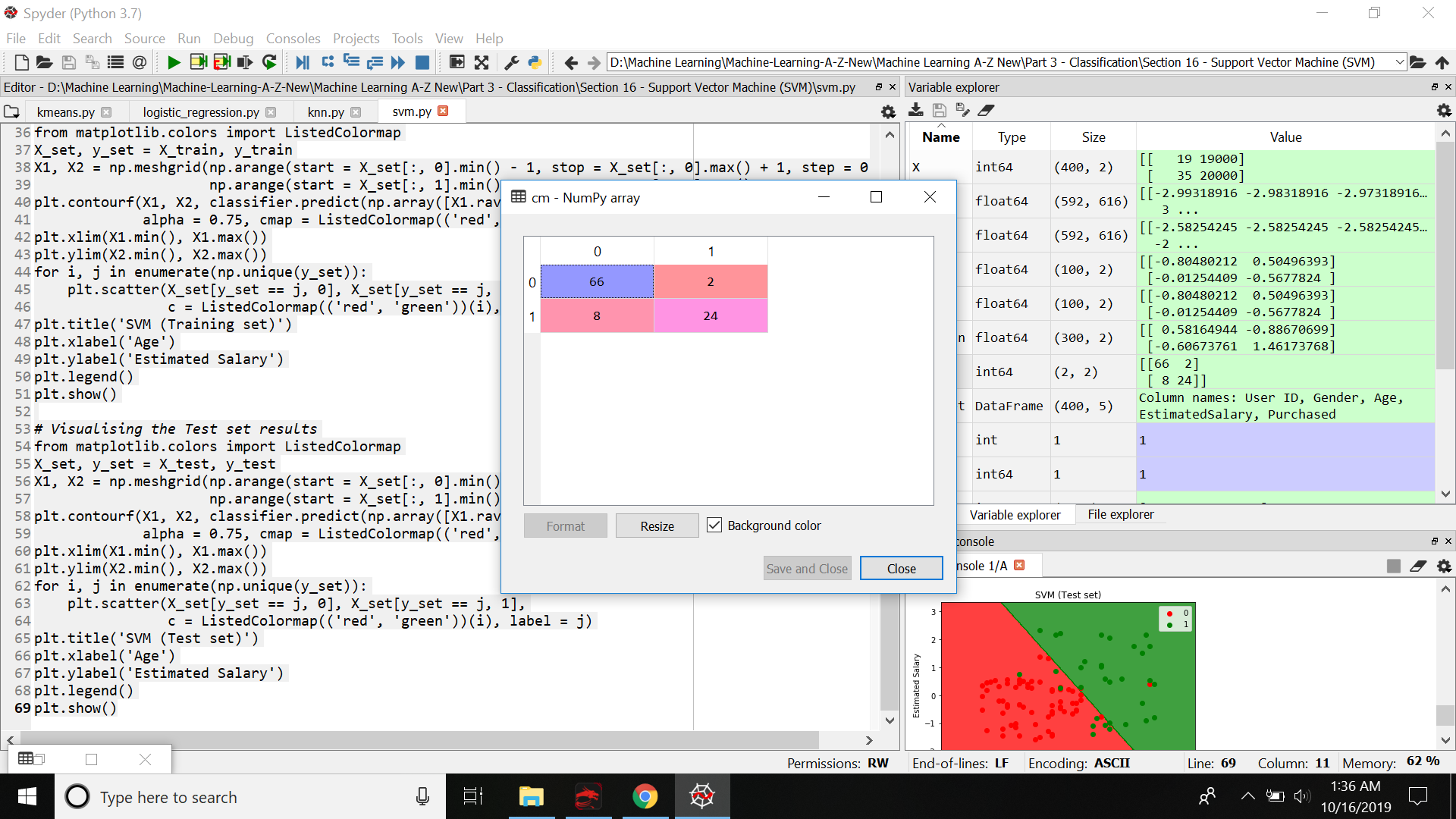
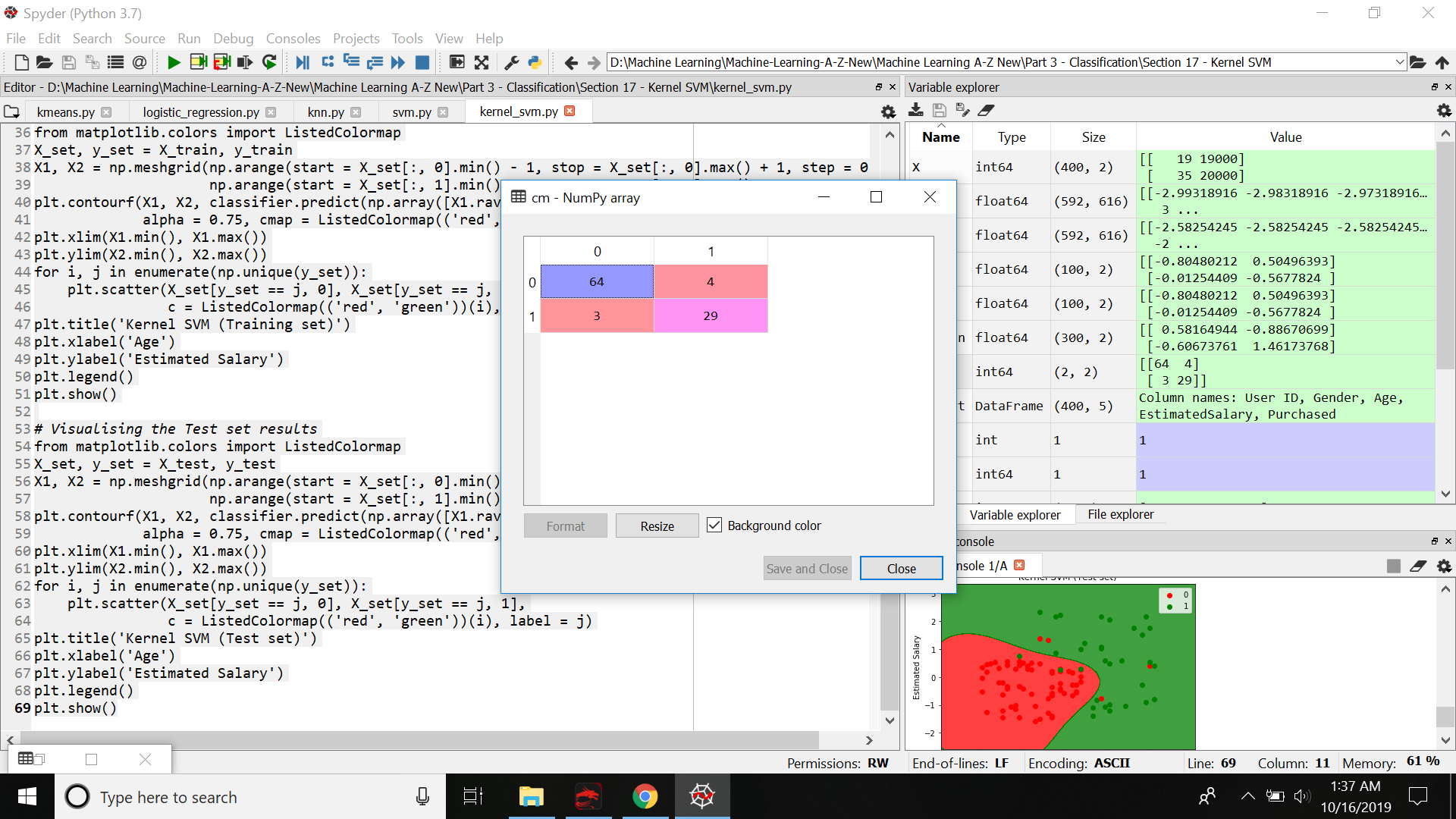
 

Fig. 3.12 Fig.3.13

13

Naive Bayes Decision Tree

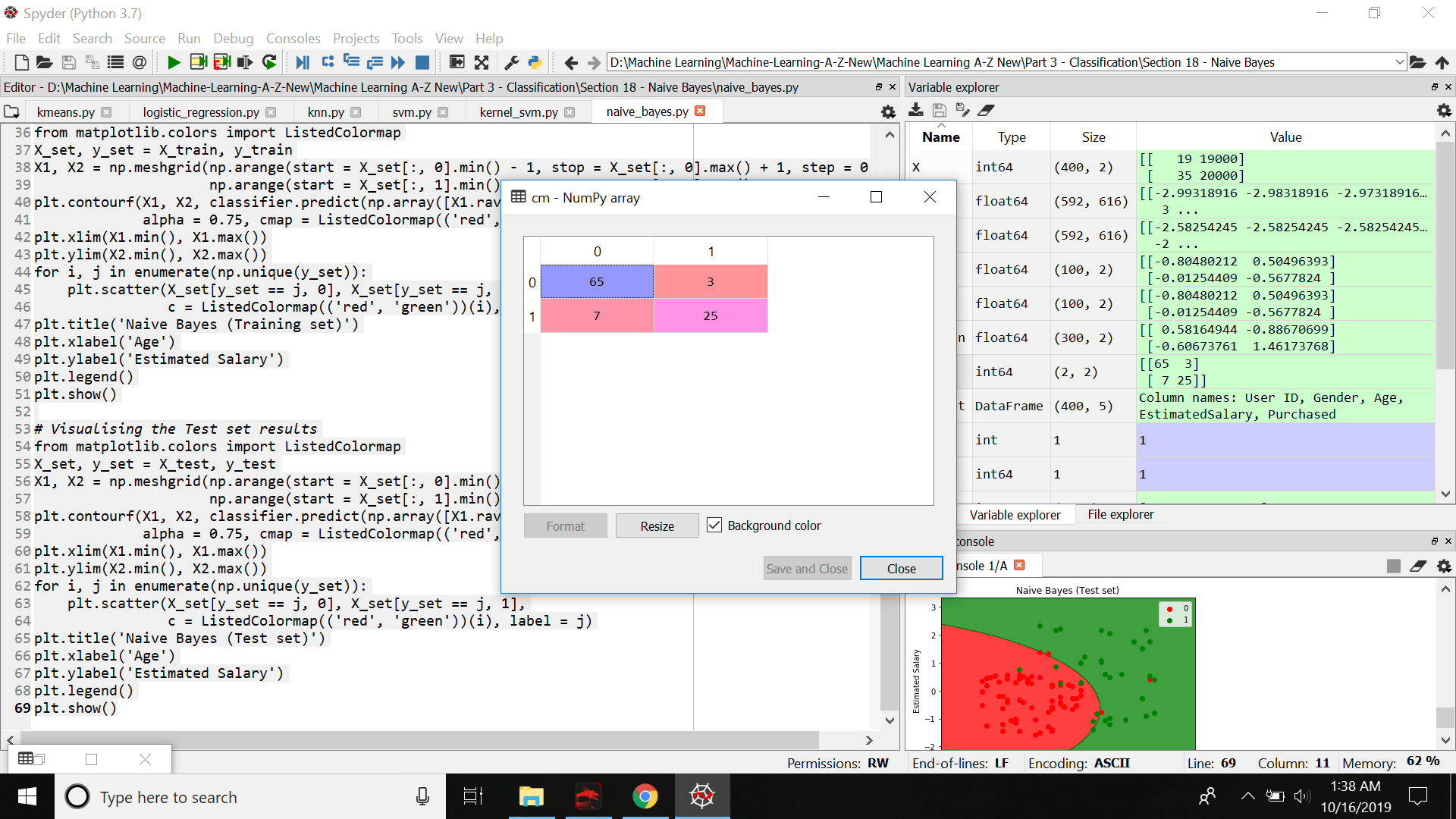
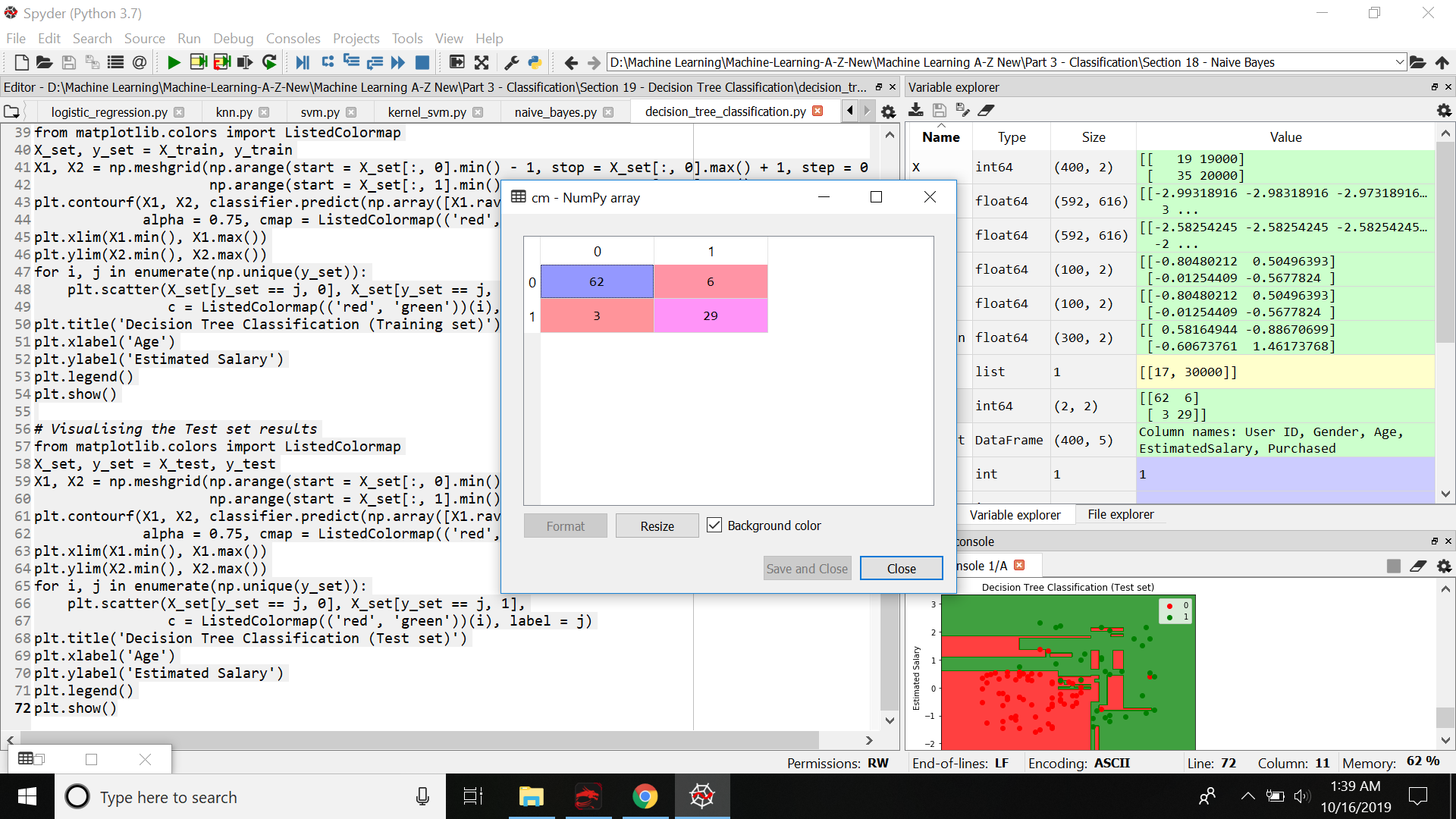
 

Fig. 3.14 Fig. 3.15

Random Forest

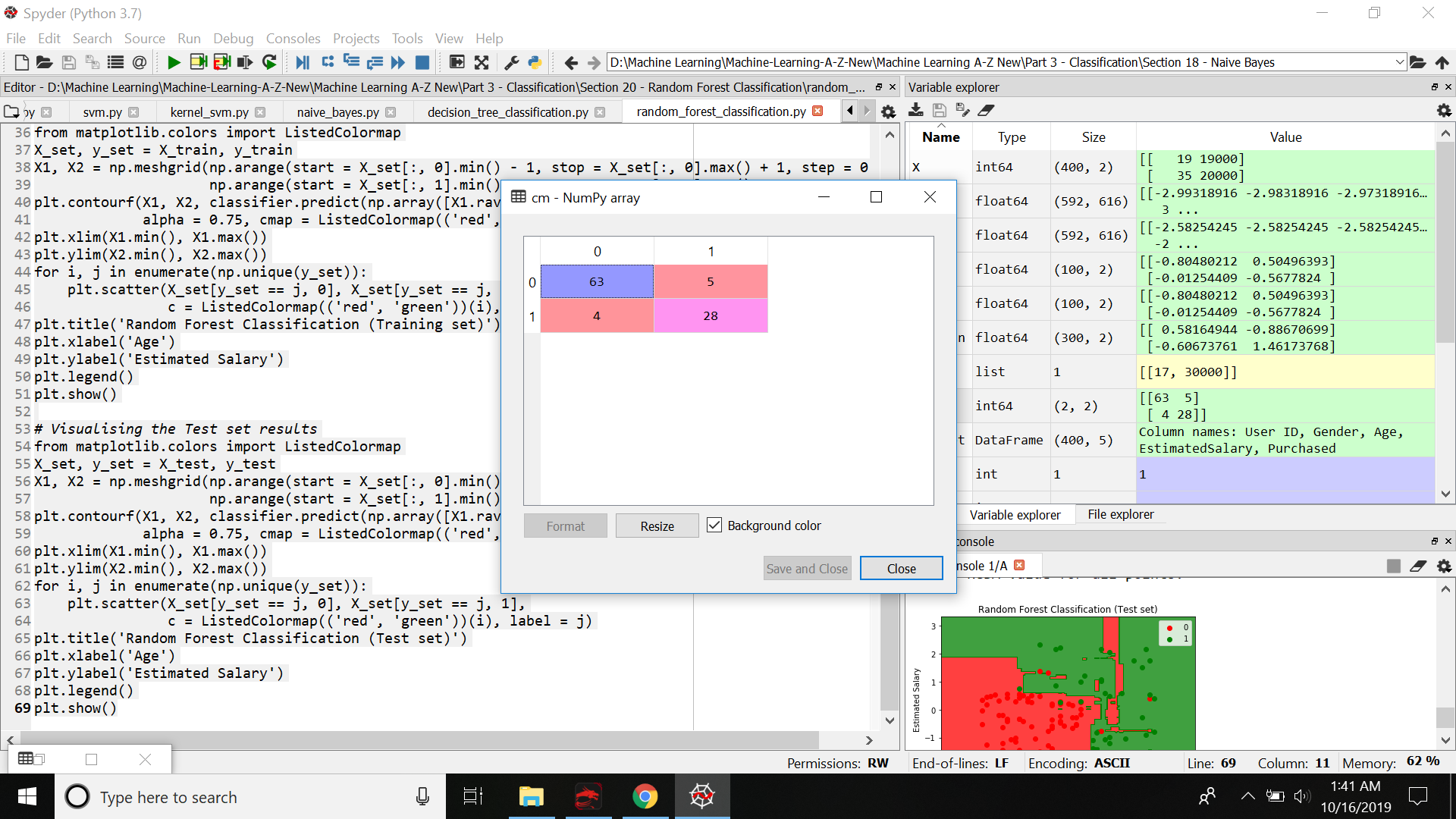


Fig. 3.16

14

**4. CONCLUSIONS**

**4.1 Conclusion**

This paper focuses on various classification techniques (statistical and machine learning based) used in data mining and a study on each of them. Data mining can be used in a wide area that integrates techniques from various fields including machine learning, Network intrusion detection, spam filtering, artificial intelligence, statistics and pattern recognition for analysis of large volumes of data. Classification methods are typically strong in modeling communications. Each of these methods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa. These classification algorithms can be implemented on different types of data sets like share market data, data of patients, financial data,etc. Hence these classification techniques show how a data can be determined and grouped when a new set of data is available. Each technique has got its own feature and limitations as given in the paper. Based on the Conditions, corresponding performance and feature each one as needed can be selected.

**4.2 Application**

1. Internet traffic interception - certain governments (possibly from the middle east) would like to restrict certain categories of web pages. For example, due to religious restrictions, certain movie pages may be restricted/censored. This is a clear application of classification. But a hard problem, given that this must be done in real-time and based on unstructured text content. We have developed a solution for one of our clients.

2. Video classification - as and when you upload a video on youtube, the video has to be classified into appropriate categories and meta-data added to it (annotations). Again a hard problem, given the state-of-the-art video processing and real-time nature of things.

3. Image classification - classic problem in image processing. Is this an image of a dog or a cat? Google recently has come up with a large scale training mechanism for neural networks: Large Scale Distributed Deep Networks. The same could also be used for face recognition.

4. Voice classification: classify and if possible, identify the voice. Very useful in Siri type applications.

15

**References**

[1] https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[2]http://www.computerscijournal.org/vol8no1/a-comparative-study-of-classification-techniques-in-data-mining-algorithms/

[3]https://www.quora.com/What-are-the-interesting-applications-of-classification-algorithms-like-neural-networks-and-random-forests

[4] ‘Short Survey on Naive Bayes Algorithm’ , International Journal of Advance Engineering and Research Development Volume 4, Issue 11, November -2017

[5] ‘Research of Decision Tree Classification Algorithm in Data Mining’,International Journal of Database Theory and Application Vol.9, No.5 (2016)

[6] ‘k-Nearest Neighbour Classiﬁers’, Technical Report UCD-CSI-2007-4 March 27, 2007

**ACKNOWLEDGEMENT**

I would like to place on record my deep sense of gratitude to Prof. S.B.Kalyankar , HOD-Dept. of Computer Science and Engineering, Deogiri Institute of Engineering and management Studies Aurangabad, for his generous guidance, help and useful suggestions.

I express my sincere gratitude to Prof. Amruta Joshi, Dept. of Computer Science and Engineering, Deogiri Institute of Engineering and management Studies Aurangabad, for her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I am extremely thankful to Dr. Ulhas Shi, Director, Deogiri Institute of Engineering and management Studies Aurangabad, for providing me infrastructural facilities to work in, without which this work would not have been possible.

**Signature of Student**

Shriram Mogal

Sign