**ABSTRACT**

Prostate cancer has been on the increase within the past years and direful cases being found in men in their 20’s. The matter is that the majority of the cases are diagnosed at late stages so the death rates are high. In recent years data, analytic studies have become a common complement with novel and new research where various tools and algorithms are taking a center stage in cancer research. In this research, the main goal is to use machine learning algorithm (in this case KNN-algorithm) to derive an outcome or make a decision, which may be used in building a prognostic tool that helps in treatment. In this project, we used KNN - algorithm of Machine Learning (using python) to predict if the tumor is “B” benign or “M” malignant, given the description of the tumor.

"The data set consists of 100 observations and 10 variables (out of which 8 numeric variables and one categorical variable (label) and one is ID) which are as follows:"

"1. Radius

2. Texture

3. Perimeter

4. Area

5. Smoothness

6. Compactness

7. Symmetry

8. Fractal dimension"

First, we load the data (records or features of the tumor), then divide or group the data into train (or learn) and test. After this grouping, we train the algorithm on the training data, after which it is ready to predict the label of testing data.

This project contributes to academics, society, and cancer research which ultimately assist in the reduction of mortality rates by use of pattern recognitions, which leads to better decision making.

**INTRODUCTION**

**Background of Study**

Prostate cancer is among the highest killing cancers. It's been ranked sixth globally and third in the African nation as a reason for death after infectious and cardiovascular diseases. Family history, age, diet, weight (obesity), speculative habits (smoking, significant drinking), and exposure to environmental pollutants facilitate to predict an individual’s risk of developing cancer. Prostate cancer is a noncommunicable disease that solely affects men. It's common among men, who are above fifty years of age, however, a number of cases reported in men below fifty years. Most of the cases are diagnosed in men over sixty-five years of age.

The medical specialty of prostate cancer is complicated and has few established risk factors. However, the most established risk is related to family history, age, race, country and androgenic hormone deficiency. American Cancer Society shows additional modifiable risk factors, which might be modified. These modifiable risks are smoking, obesity, unhygienic workplace exposure, sexually transmitted diseases, and diet.

According to (KNCCS, 2011) the illness can't be eradicated however its effects may be considerably reduced if effective measures are in place to manage risk factors, observe cases early and supply excellent care to those with the illness.

Machine learning algorithms have been used successfully recently to learn the cancer pattern. In this project, KNN algorithm of machine learning is used with Python - which contains various useful open source packages like numpy, scipy, matplotlib, scikit-learn and pandas and several state-of-the-art learning algorithms- to predict prostate cancer patterns. Policy makers, medical practitioners (physicians), and patients can use this prediction to provide essential input into the rational planning of cancer control programmes.

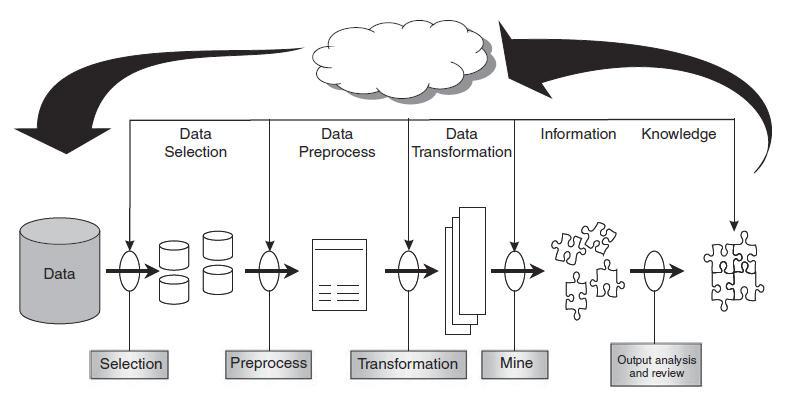


Figure 1: Data Mining Process

The use of data mining techniques and knowledge management technologies in disease prediction and prognosis is part of a growing trend towards personalized, predictive medicine. This shifting towards predictive medicine is essential, not only for the patients (in terms of life-style and quality-of-life decisions) but also for the physicians (in making treatment decisions) and health economists and policy planners (in implementing large-scale cancer prevention or cancer treatment policies)

**Statement of the Problem**

An ideal scenario would be a cancer-free continent, however, the truth is that this illness is turning into common and resulting in many deaths annually. The vast medical data on prostate cancer patients should be carefully studied so that good decision can be applied to reduce this epidemic.

This project carried out a comparative analysis of machine learning techniques in prostate cancer, which were able to detect patterns. These patterns were used to check prostate cancer prevalence by incorporating major variables or feature of the tumor such as radius, texture, perimeter, area, smoothness, compactness, symmetry, fractal dimension.

This project carried out a comparative analysis of machine learning techniques in prostate cancer, which were able to detect patterns. These patterns were used to check prostate cancer prevalence by incorporating major variables or feature of the tumor such as radius, texture, perimeter, area, smoothness, compactness, symmetry, fractal dimension.

**Project Objective**

The main goal of this project was to use KNN algorithm to classify and derive patterns from the eight variables or features of tumor of 100 records of people, which will be used in predicting if the tumor is "B" benign or "M" malignant, that helps in the identification of the problem in advance and deciding the treatment technique. This prognostic tool can assist in early detection of the prostate cancer, therefore, reduce on high rates of reportable cases of advanced prostate cancer and deaths.

* Recognize data-mining techniques suitable for prostate cancer analysis.
* Suggest a model for identifying patterns of prostate cancer incidences.
* Build a prototype for community awareness (prognostic tool).

**Literature Review**

**>Purpose of machine learning**

* Push the work-load to self-sufficient machines
* Pattern recognition which might not be recognized manually
* Better analysis of data

There are mainly four types of machine learning system: Supervised, Unsupervised learning, Statics, and Model evaluation.

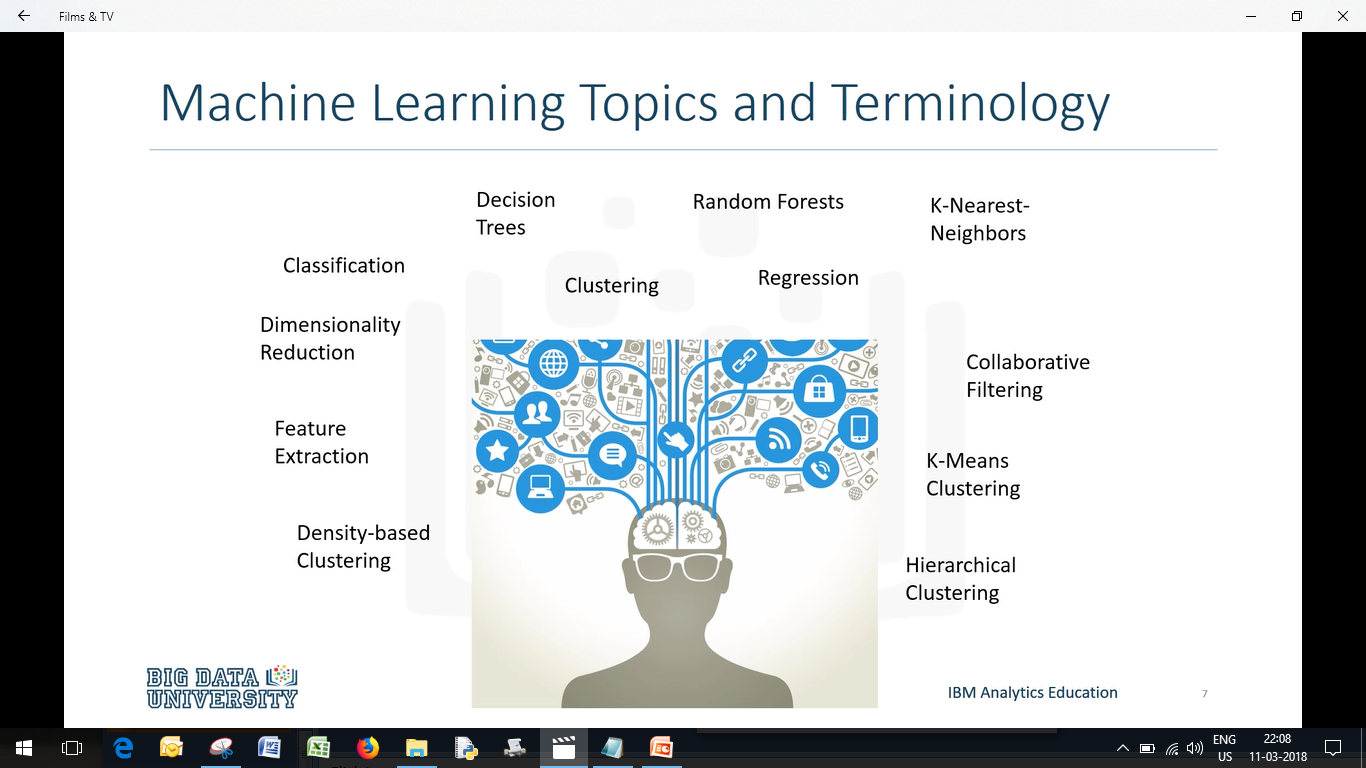
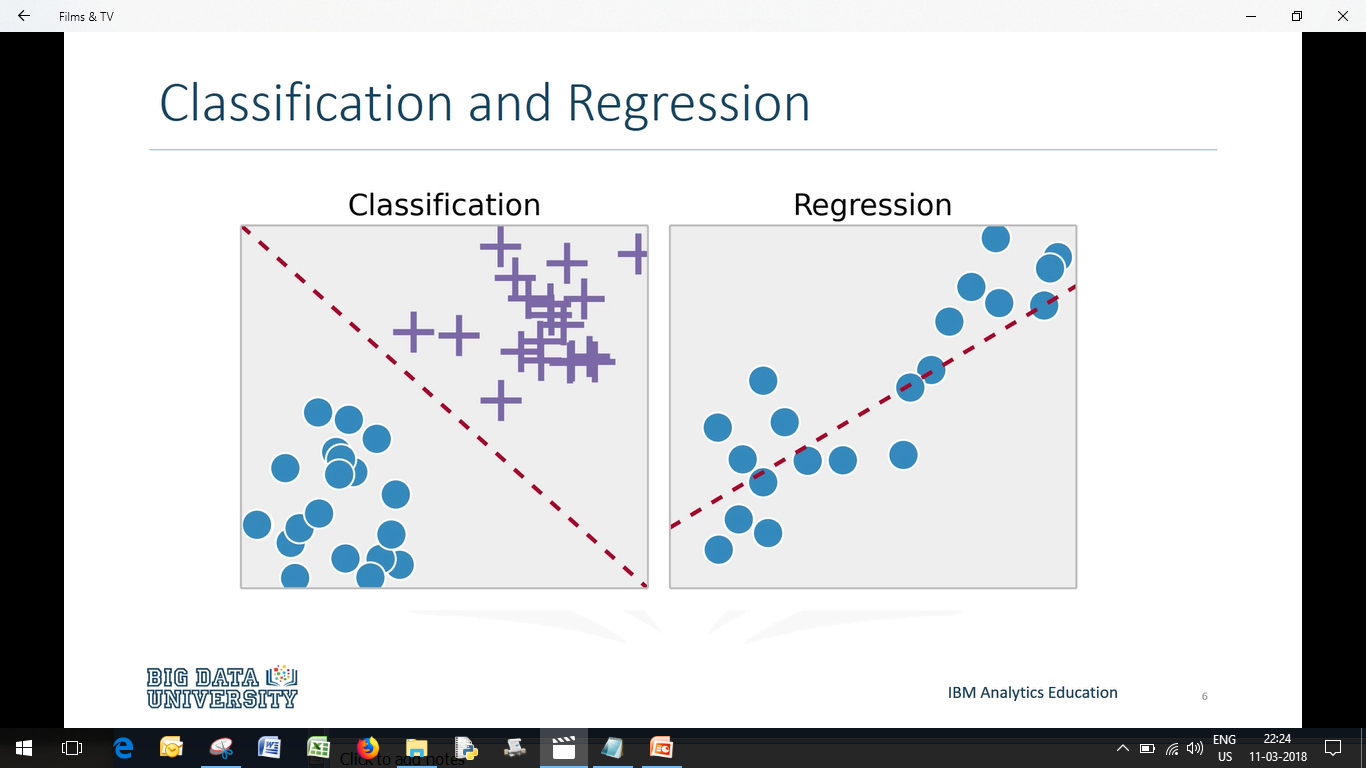


Figure 2: Machine Learning Algorithms

**Supervised Learning**

* “Teach the model”, then with that knowledge, have it predict future instances.
* Generally speaking, the model is trained on a labeled dataset, so it can predict the outcome of out-of-sample data.

**Classification and Regression**



Classification

The concept of categorizing data is based on training with a set of data so that the machine learning can essentially learn boundaries that separate categories of data. Therefore, new data inputted into the model can be categorized based on where the point exists.

**K-Nearest Neighbours – A Classification Model**

Know- Nearest Neighbor simply works on supervised learning, where the label is given for the data.

In KNN, data points are categorized and when determining the category of a new data points, the K nearest points are used in this process. The K-NN classification rule is to assign to a test sample the majority class label of its K nearest training samples. In practice, K is typically chosen to be odd, therefore as to avoid ties. The K = 1. rule is usually referred to as the nearest-neighbor classification rule.

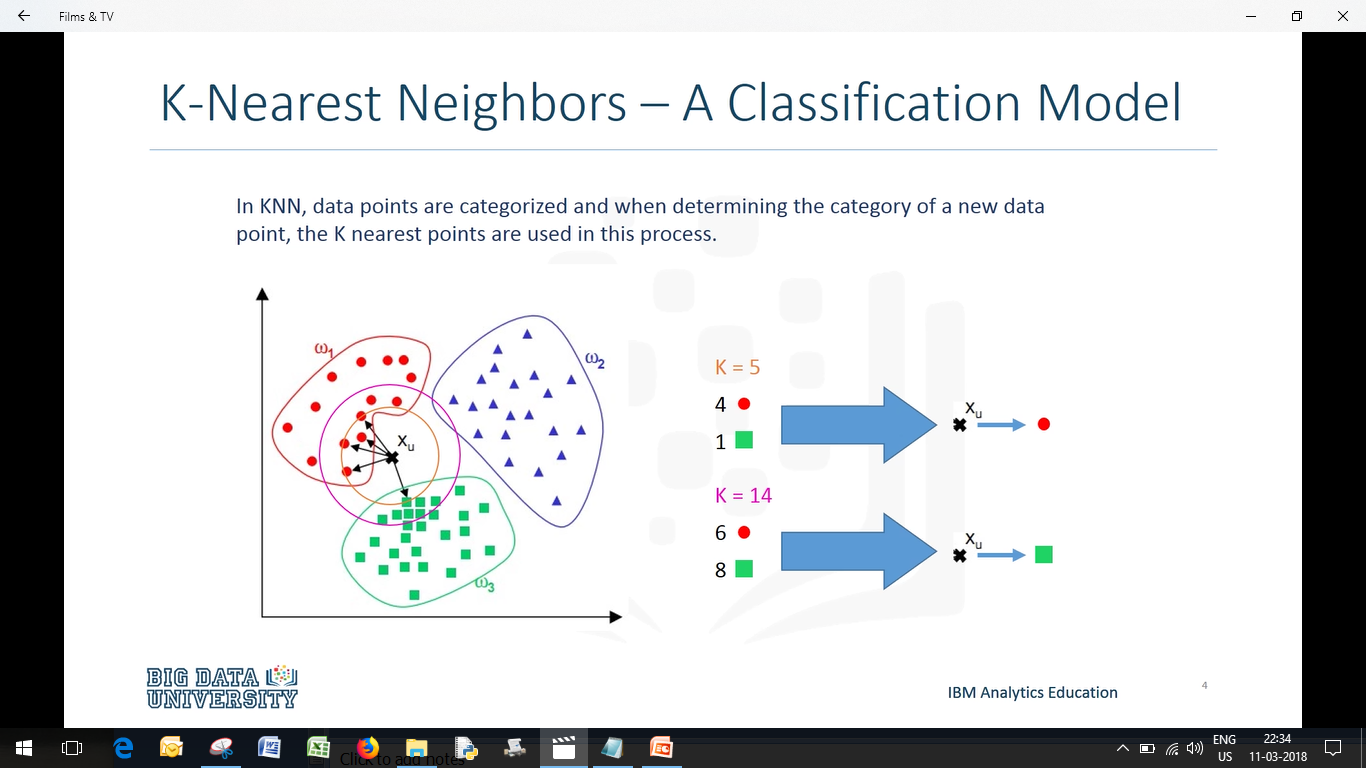


Figure 3: K-NN

How K-NN algorithm work

According to the Euclidean formula (for distance), the dist(d) between 2 points in a plane with coordinates (x, y) and (a, b) is given by

dist(d) = sqrt[sq(x-a) + sq(y-b)]

**Preprocessing your dataset**

**"**

* Dataset may need to be preprocessed to ensure more reliable data mining results
* Conversion of non-numeric data to numeric data
* Calibration of numeric data to reduce effects of disparate ranges
* Particularly when using the Euclidean distance metric"

**Data Cleaning**

"The data available to us was raw information, therefore having heaps of impurities that required to be removed. Using Microsoft Excel, we did data cleansing wherever needed. We had a dataset of a 145, however, after cleaning were reduced to a 100. Most of the deleted information had no elaborated location. Any information that was blank in any of the attributes was additionally replaced."

**Modeling Techniques, Tools and Modeling Infrastructure Used**

"The software tools used, software- Anaconda, which is open-source and

downloadable from the World Wide Web. It comes with around 256 packages/modules pre-installed, which makes it very efficient. It also has many python IDEs(Jupiter, spyder etc) present, you just have to install it once. Data can be stored in My-SQL Server database."

**k-NN Time Complexity**

"

* Suppose there are m instances and n features in the dataset
* The nearest neighbor algorithm requires computing m distances
* Each distance computation involves scanning through each feature value
* Running time complexity is proportional to m X n "

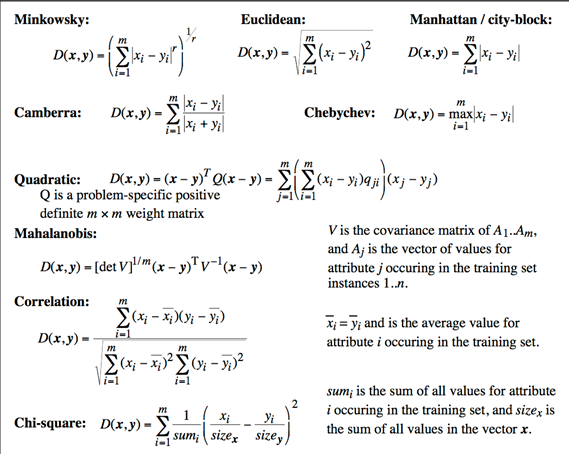


Figure 4: Distance Formulas used in ML

WORKFLOW

Figure 5: Workflow Sequence

Dataset Description

We are going to use Prostate\_Cancer dataset. The data set consists of 100 observations and 10 variables (out of which 8 numeric variables and one categorical variable and one are ID) which are as follows: 1. Radius 2. Texture 3. Perimeter 4. Area 5. Smoothness 6. Compactness 7. Symmetry 8. Fractal dimension

Given a description of the tumor, we are going to predict if the tumor is “B” benign or “M” malignant.

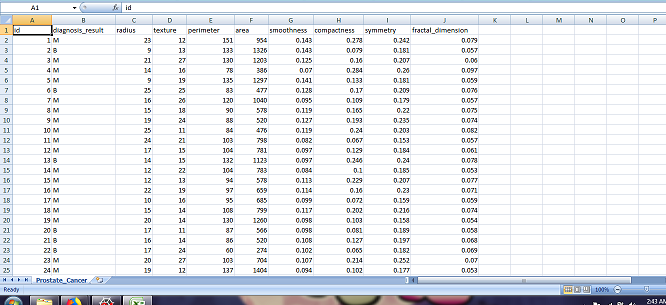


Figure 6: Tumor Data (organized)

**Requirements:**

Operating System: Win XP, 7 (or higher)/ Mac OS

Python 2.7 (or higher), Anaconda can also be used, which is very convenient for beginners

Python IDE used: Jupyter and Spyder

Hardware: Laptop with minimum 4 gigabytes of RAM

Approaches

After loading the data externally, of the observations, the label (benign or malignant) to which the observation belongs is displayed, after prediction.

The algo of the k-in, that we apply in our project is as follows:

1. Calculate "distance" from a test record to the training records.
2. To find the "k- nearest" training records
3. Check the majority class from the k - nearest training records
4. The diagnosis result label for the training record is predicted as the class with the majority votes/weight among the k - nearest training.

The objective of the project using k-NN

* Finally, at the end, we will be classifying the prostate cancer based on their details into two labels "B" or "M".
* Test the algorithm using a range of neighbors (in this case column 2-9), to see which no of neighbor yields the best result in terms of production with the help of graph
* Finding the accuracy of the approach
* Getting confusion matrix and accuracy score of the problem

WORKPLAN

DATA

KNN BOOT (Library)

KNN FIT (Learning)

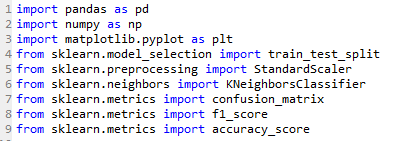
RESULT

KNN PREDICT (Testing)

**Packages / Libraries / Headers**

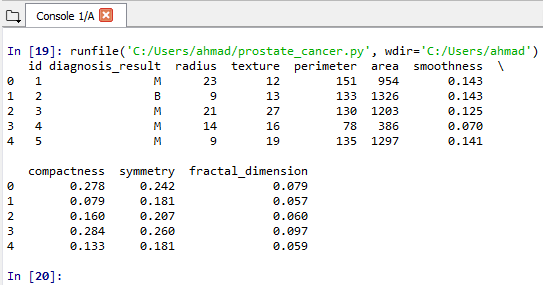
Python packages used: SciPy, Matplotlib, NumPy, Scikit-Learn, and Pandas.

K-NN specific modules : train\_test\_split, KNeighborsClassifier, accuracy\_score, confusion\_matrix.



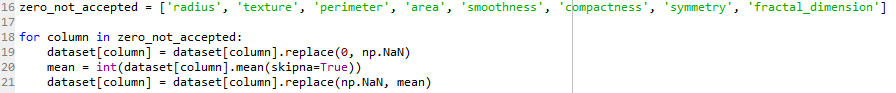
**Loading the dataset into the program from an external source**

C:\Users\ahmad\Desktop\header.png



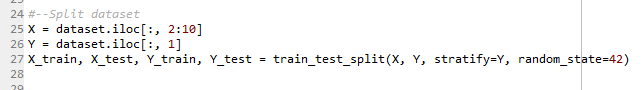
**Replace zero values**

Values of columns like 'radius', 'smoothness', 'compactness', 'fractal compactness' cannot be accepted as zeroes because it will affect the outcome. We can replace such values with the mean of the respective column.

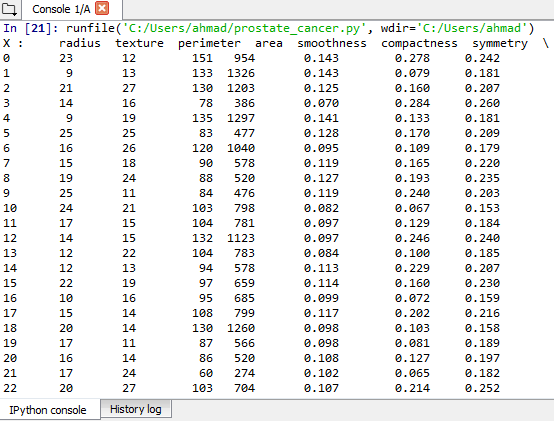
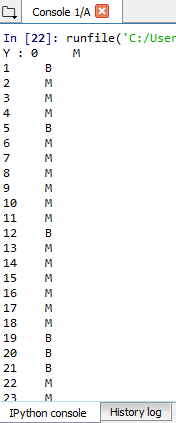


**Splitting the dataset**

Before proceeding further, lets split the dataset into train and test:

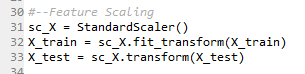


X has all the features or variables (total of 8 columns) whereas Y has the label column ()

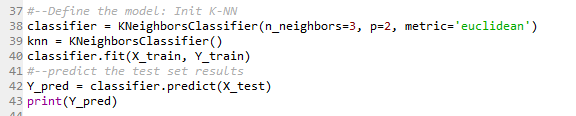
**Feature: Scaling**

It is a thumb rule, if any algorithm, which computes distance or assumes normality, scales your features.



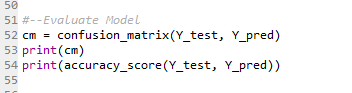
**K-NeighborsClassifier**

Define the model using KNeighborsClassifier and fit the train data in the model



**Evaluation Model**

It is important to evaluate the model, we used a confusion matrix to do that.



**RESULTS AND DISCUSSION**

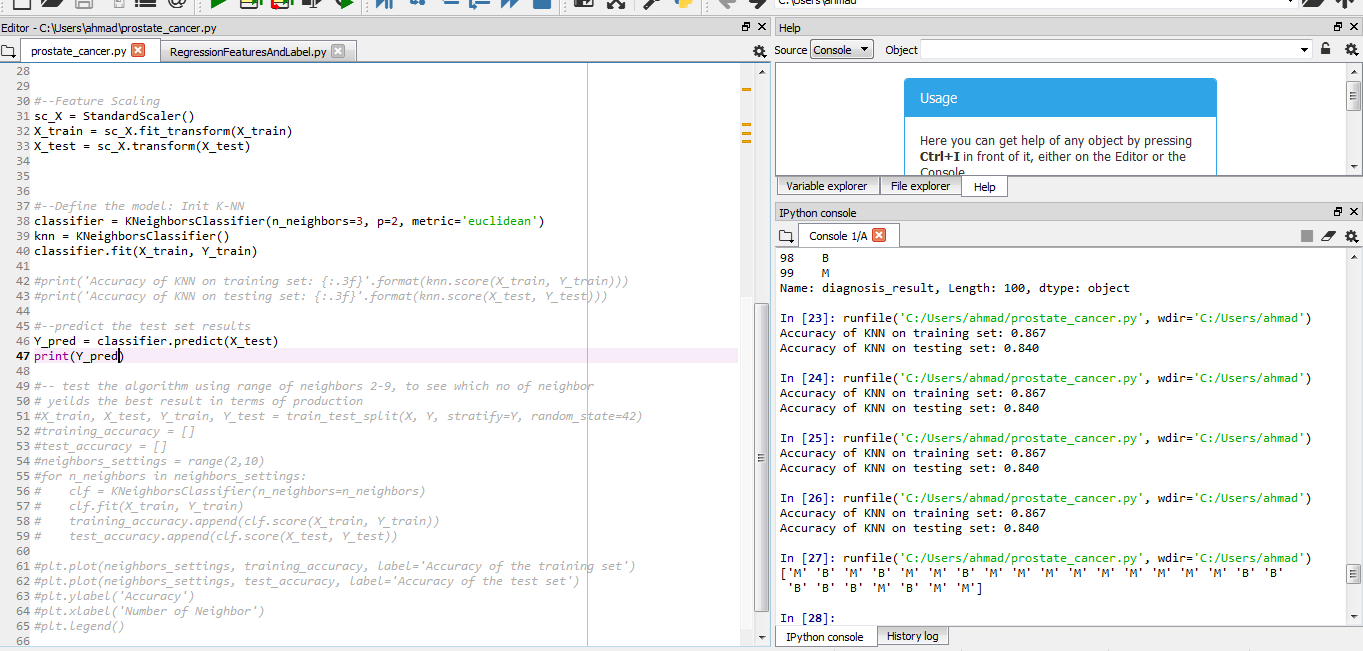


Figure 7.1: Prediction output and accuracy

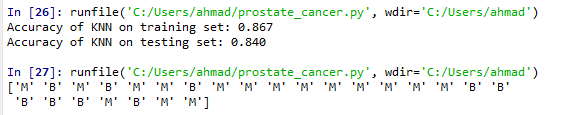


Figure 7.2: Prediction and Training and Testing Accuracy

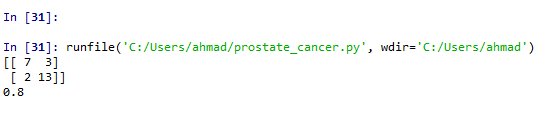


Figure 8: Output of the evaluation model and accuracy score

**->** Testing the algorithm using a range of all neighbors, in this case, we are testing for columns 2-9 (for that we have to mention range(2-10) in the program). This is done to see which no of neighbor yields the best result in terms of production. Below is the program used for it, a for loop will be used to consider all the cases and n\_neighbors is used to consider for all the cases (default is 5 for the k-nn algorithm).

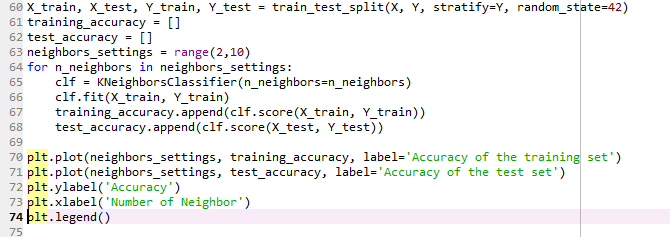


Figure 9: a Testing algorithm for a range of all neighbors

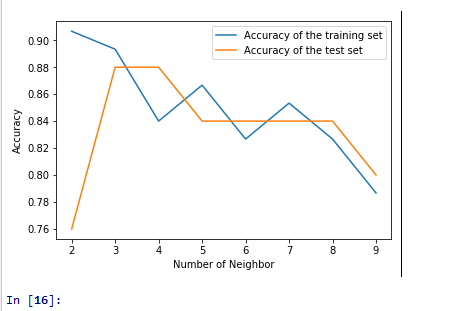


Figure 10: Output of figure 9

-> As you can see the yield is highest for the test set at 3 and 4, hence for the value of

n\_neighbor = 3 or 4.

**Applications and Future Scope**

"

Machine learning allows building model to swiftly analyze data and bring outcome, leveraging the chronological and real-time data. With ML, health-care service suppliers will construct good decision on patient’s diagnoses and treatment choices, that ultimately direct to the overall advancement of health-care services."

"

Earlier, it was difficult for health-care professionals to collect, organize and analyze the large quantity of information for effective prediction and treatment as there have been no technologies or tools existing. Now with ML, it is easy, as big-data technologies like Hadoop are sufficient enough for wide and large-scale adoption."

"In fact, 54% of organizations are using or considering Hadoop as big data processing tool to get important insights on healthcare according to the [Ventana Research Survey](http://news.sys-con.com/node/1920943). 94% of Hadoop users out of existing users perform analytics on voluminous data which they believe was not possible before."

The data-mining technique would be helpful in providing essential statistics, facts and advance analytics in terms of disease, test results such as blood pressure, clinical trial data, etc., to doctors.

"

**Drug Discovery**

Development and discovery of drugs are very expensive and takes a lot of time and energy. Generally, a new drug takes more than 10 years to develop and get into the market. And the whole process may costs roughly approx 2.6 billion dollars

A drug discovery program is meant at identifying a compound that reacts with affected molecules of the body, leading the disease to cure. However there is the possibility of the drug compound may react to non-affected molecules inside the body adversely, which may be harmful and cause dangerous side effects

Machine learning approach based on large information of approved clinic data and failed results in identifying a toxic compound in advance, that may later cause side effects, this can ultimately save a lot of time and resources before going into clinical trials.

"

"

### Robotics In Surgery

Today robotics is already being used and spearheading in the field of surgery. Robotics high-powered by AI and mil algorithms improve accuracy and exactness of surgical tools by incorporating time period surgery metrics, information from productive surgical experiences, and information from pre-op medical records among the surgery."

### Personalized Treatment And Behavioral Modification

### From,"2012 to 2017, the penetration rate of Electronic Health Records in health-care increased from 40% to 67%. This naturally suggests that a lot of access to individual patient health knowledge. By collection, this personal medical knowledge of individual patients with ml applications and algorithms, health care suppliers (or providers) will observe and assess health problems better. supported supervised learning, medical professionals will predict the risks and threats to a patient’s health in line with the symptoms and genetic data in his medical history."

### "

### Predicting Epidemic Outbreaks

Healthcare organizations are applying ml and AI algorithms to observe and predict the potential epidemic outbreaks which may take over numerous components of the globe. By collection information from satellites, time period updates on social media, and different important data from the world wide web, these digital tools will predict epidemic outbreaks. this could be a boon notably for the third-world countries that lack correct health care infrastructure and aids.

For example, support-vector machines and new technologies such as artificial neural networks have helped a lot to predict the out-break of malaria by taking into consideration factors such as temperature, climatic condition, average monthly rainfall, etc. "