Subarna Chowdhury Soma
SJSU ID: 014549587

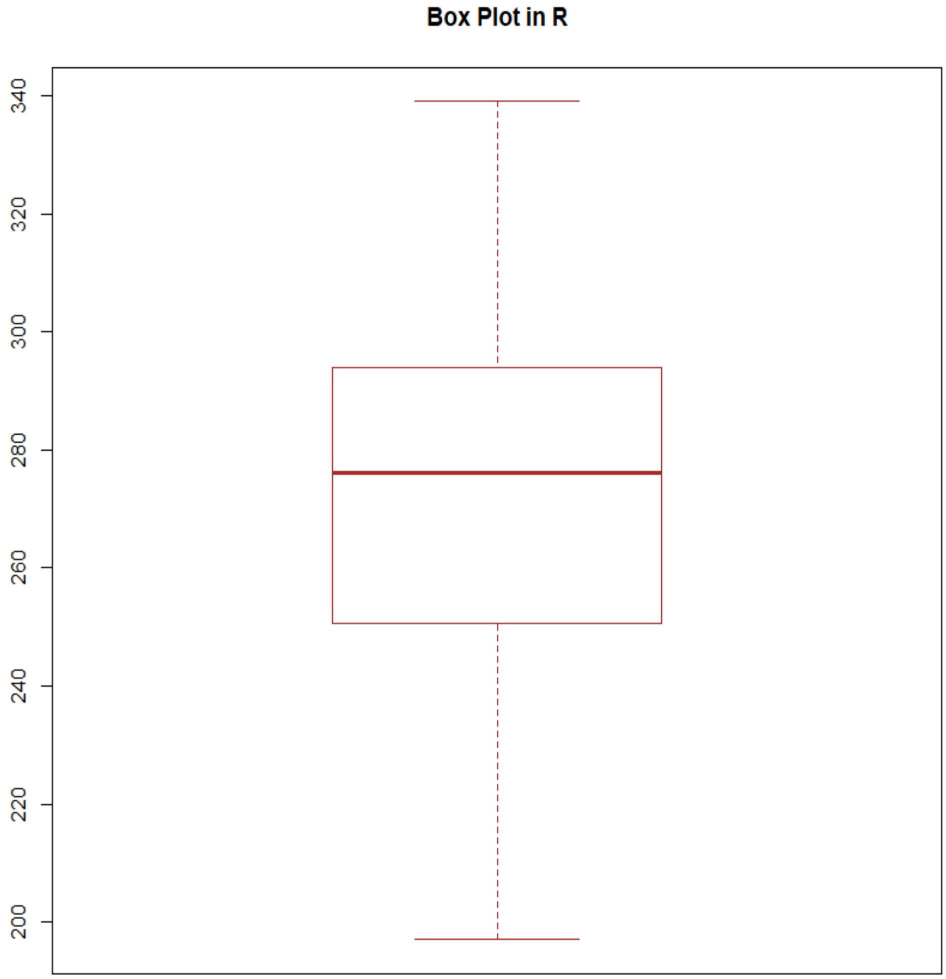
1. a. Use R to draw a box plot for the data, 197, 199, 234, 267,269,276,281,289, 299, 301, and 339 and determine any outliers. Implement the same in Python. Remember to submit the code, screenshot of the plot, and the printed list of outliers (if any).

Answer:

- ❑ Boxplot using R:

  File name: Box_Plot_and_Finding_Outliers.R

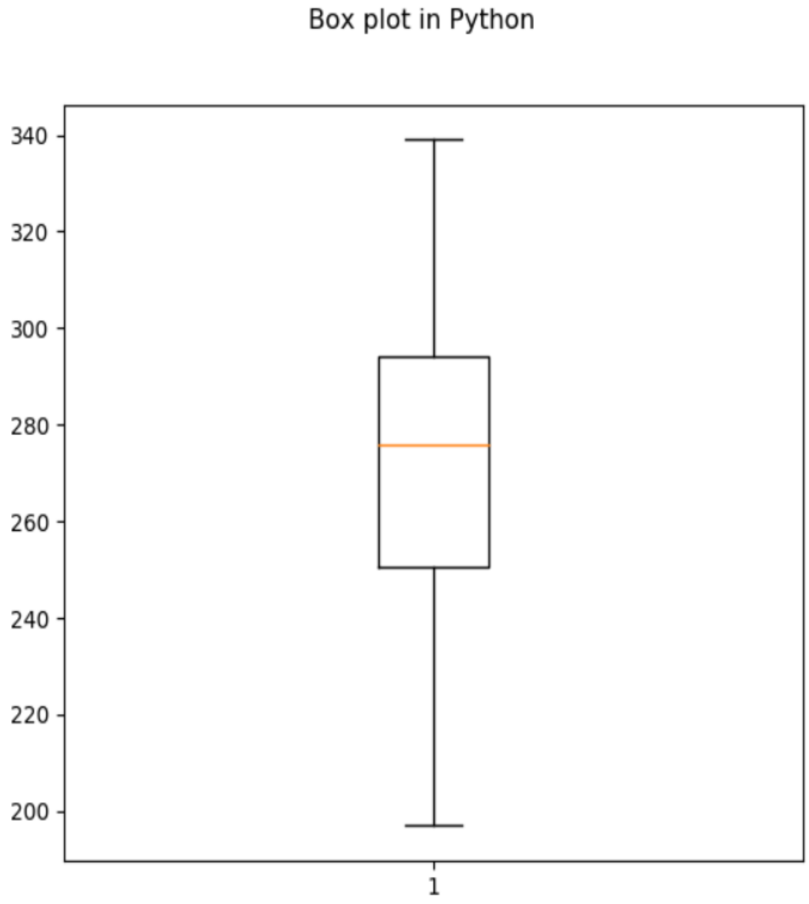
  Plot:



- Boxplot using Python:

  File name: Box_Plot_and_Finding_Outliers_Python.py

  Plot:



- Observation: Both the boxplots in R and Python show that there are no outliers in data.

b. ***Outliers are often discarded as noise. However, they can be of tremendous use in certain real-world scenarios. Briefly describe as many such scenarios as possible.***

Answer:

Outliers have numerous applications in a wide variety of domains. Real life scenarios where outliers can be very useful:

1) Medical applications:

- ❑ Outlier detection can be very useful for the diagnosis of critical diseases in various medical applications. For example, an abnormal MRI image may indicate the presence of malignant tumors, an unusual heart beat change can be an early sign of critical heart disease etc. Most hospitals often employ patient monitoring and alerting system. Detection of anomalies corresponding to unusual patient management actions will help to identify potential patient management errors. In public health data, outlier detection techniques are widely used to detect anomalous patterns in patient medical records which could be symptoms of a new disease.

2) Sensors (IoT):

- ❑ Internet of Things (IoT) devices are capable of generating massive amounts of data with the help of built in sensors and take decisions on the basis of this data. Event detection(Outlier) in sensor networks is very essential in IOT to detect unusual behaviours from data that can contain indication of vital hidden information. For example, lots of sensors are being used to measure the quality of air. Any abrupt change in values (outliers) are considered as a sign of incoming natural disaster.

3) Credit card fraud detection:

- ❑ Credit card fraud detection is an important application of outlier detection. Outliers in credit card transaction data could indicate credit card or identity theft. Due to a drastic increase in digital frauds in e-commerce marketplaces, there is a loss of billions of dollars and therefore different robust mechanisms are evolved for fraud detection and applied to diverse business fields. These techniques include various offline or online outlier detection methods such as statistical, neural networks, clustering, genetic algorithms, decision tree etc.

4) Intrusion Detection Systems:

❏ Outlier detection techniques are being widely used in Intrusion Detection System (IDS) to ensure network security. An Intrusion Detection System (IDS) is a software application or device that monitors policy violations or malicious activities of the system network and generates reports to the management system. The main focus of an intrusion detection system is to detect the attacks efficiently at a beginning stage in order to reduce their impacts.

5) Criminal Activity Detection in defence and surveillance:

❏ Outlier detection has a wide variety of use for security and crime prevention such as in military surveillance for enemy activities. It is being used in the battlefield to detect any incoming threat. Outlier detection technologies are capable of detecting suspicious behaviour that could reduce the occurrences or provide fast response in case of an incident. Finding suspicious behaviour (outliers) during tracking people in multiple city blocks can solve smaller crimes such as robberies and thefts that is helpful for police forces as well.

6) Earth and Environmental Science:

❏ Outliers detection turned out to be an important method in the field of earth and environmental science. Now a days, online environmental monitoring system is being set up as a part of any industrial plant operation to measure high concentrations of residuals products produced by the process and its negative impact on human-health and environment. These measurements are carried out by identifying outliers. Morovers outlier detection also helps to study pollution levels, weather, fire prevention, climate changes or incoming disaster such as earthquake etc.

7) Fault detection in critical systems:

❏ Outlier detection has been found to be useful and directly applicable to detect faults in critical systems. For example, abnormal readings from a spacecraft would signify a fault in any certain component of the craft. Different data driven tools are there to detect the presence of outliers in spacecraft system data.

8) Transportation:

❏ The technique of outlier mining has been introduced for detecting and analysing outliers in available urban traffic data. Transportation systems use outlier detection methods as a preventive mechanism to control accident and handle traffic congestion.

9) Spam detection:

- Detecting outliers has important applications in spam detection. Every fields, such as e-commerce websites are becoming extensively available for people to purchase different types of products online. Outlier detection plays an important role here to identify fake/spam reviews.

10) Social media applications:

- Social media outlier detection is of critical importance to prevent malicious activities such as bullying, terrorist attack planning, scam, and fraud information spreading. For example, in a popular social network such as Facebook, a new set of social network attacks may include unnecessary friend requests and can be considered as outliers.

2. a. Feature Engineering: List out as many features as possible, that can help determine the veracity of the following

Answer:

i) Web Page:

a. Authority (If the page contains information about author credentials and its domain etc.)
b. Accuracy (Spelling error free, well-written and grammatically correct etc.)
c. Objectivity (If there is no evidence of biased content etc.)
d. Currency (If the page is current and updated regularly etc.)
e. Coverage (If the topics are successfully presented with proper argument and source etc.)
f. Appearance (If the site is well organized and well maintained etc.)
g. Audience (Web page’s intended audience)
h. Chain link with other authenticated sites

ii) Student Homework

a. Old record of that student
b. Wrong citation method

c. No reference
d. Relevancy of the used material with homework
e. Repetition of the same mistake
f. Plagiarism check

iii) News Item in a daily

a. Currency (Whether the news article is recent or not)
b. Relevance (Relevancy of the content with headline)
c. Authority (Details of authority and credentials)
d. Accuracy (Unbiased and verifiable news source, spelling mistakes etc.)
e. Supporting sources/ Reliability
f. Purpose/Point of view (Intention of the news)
g. Type of the news
h. Title/Domain of the news paper

iv) Image on Instagram

a. Image quality
b. Any bent or distorted part of image
c. Reflection (Shadow etc.)
d. Identical patterns (multiple photoshopped copies of the same pattern)
e. Perspective, shadows, and proportions
f. Abnormalities in color and tones
g. Inappropriate body parts/measurements
h. Image metadata
i. Authenticity of the publisher profile

v) Video on YouTube

a. Authenticity of the publisher
b. Number of followers of the video uploader/channel
c. Video title (lots of videos with similar title)
d. Transcript check
e. Quality of the video
f. Content of the video

b. *The following are the scores of a 23-student class on Machine Learning, out of 100. As can be seen, based on the grading scale listed in your syllabus sheet, most of the students will fail. The professor therefore decides to grade on the*

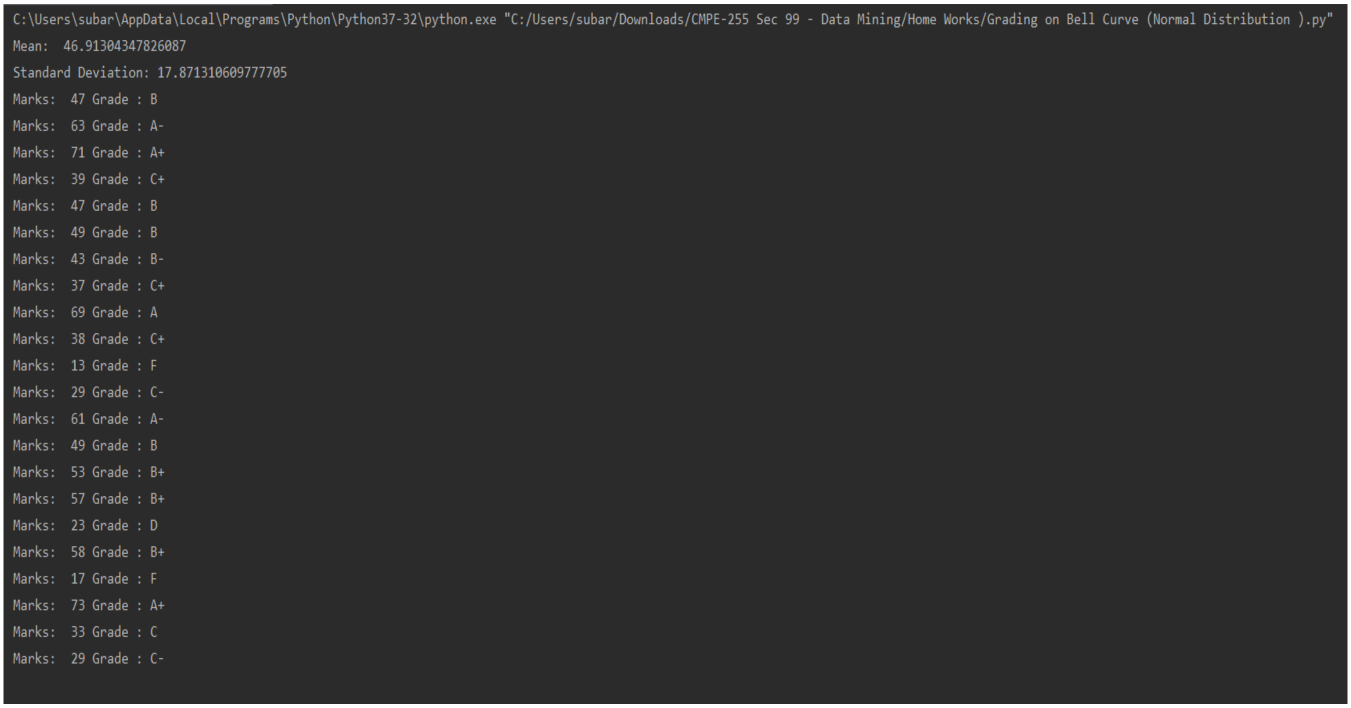
(normal distribution) curve with the mean score receiving a B- grade, the grade going up or down one step from there with every 1/3rd of the standard deviation change in the score. Assume SJSU letter grading [F to A+]. Write a program in python to determine each student's grade.

47, 63, 71, 39, 47, 49, 43, 37, 81, 69, 38, 13, 29, 61, 49, 53, 57, 23, 58, 17, 73, 33, 29

Answer:

File Name: Grading_on_Bell_Curve(NormalDistribution).py

Output:



```
C:\Users\subar\AppData\Local\Programs\Python\Python37-32\python.exe "C:/Users/subar/Downloads/CMPE-255 Sec 99 - Data Mining/Home Works/Grading on Bell Curve (Normal Distribution ).py"
Mean:  46.91304347826087
Standard Deviation: 17.871310609777705
Marks:  47 Grade : B
Marks:  63 Grade : A-
Marks:  71 Grade : A+
Marks:  39 Grade : C+
Marks:  47 Grade : B
Marks:  49 Grade : B
Marks:  43 Grade : B-
Marks:  37 Grade : C+
Marks:  69 Grade : A
Marks:  38 Grade : C+
Marks:  13 Grade : F
Marks:  29 Grade : C-
Marks:  61 Grade : A-
Marks:  49 Grade : B
Marks:  53 Grade : B+
Marks:  57 Grade : B+
Marks:  23 Grade : D
Marks:  58 Grade : B+
Marks:  17 Grade : F
Marks:  73 Grade : A+
Marks:  33 Grade : C
Marks:  29 Grade : C-
```

3a. A couple of years ago, the Indian Government awarded a $100 million contract for detecting tax evasion signs from data on Social Media. Search for more such examples of how Social Media and Big Data in general is helping with the problem of veracity. Describe them in as many technical terms as possible.

Answer:

❏ Billions of pieces of content are being posted every day on social media and fact-checkers cannot review every item to check the veracity of the news. Machine learning plays a big role here. Facebook uses machine learning to identify duplicates of debunked stories. For example, a fact-checker in France debunked the claim that one can save a person having a stroke by using a needle to prick their finger and draw blood. This allowed Facebook to identify over 20 domains and over 1,400 links spreading that same claim.

❏ Most of the fake news are intentionally written to mislead readers, which makes it not trivial to detect simply based on news content. The content of fake news is rather diverse in terms of topics, styles and media platforms. Fake news attempts to distort truth with diverse linguistic styles while simultaneously mocking true news and makes it difficult to measure the veracity of the news. To support a non-factual claim, false news also cites true evidence within the incorrect context. For example, some report shows that Russia created fake accounts and social bots to spread false stories to affect the result of the presidential election of USA in 2016.

❏ The first step of measuring veracity using machine learning is feature extraction from the news. Generally, there are three major aspects of the social media context features to extract: users based, generated posts based, and network based. User based features represents whether the news is created or spread by non-human accounts, such as social bots or cyborgs or human profile. Thus, capturing users' profiles and characteristics by user-based features can provide useful information about the veracity of the news. Post-based features focus on identifying useful information to infer the veracity of news from various aspects of relevant social media posts. These features can be categorized as post level, group level, and temporal level. Network-based features are extracted via constructing specific networks among the users who published related social media posts. Different types of network can be constructed such as the stance network (built with the nodes to indicate all the relevant tweet to the news and weighted edge), the friendship network (indicates the following/followee structure of users who post related tweets). Existing approaches for social context modeling can be classified into two categories: Stance-based and Propagation-based. Most stance classification methods mainly rely on hand-crafted linguistic or embedding features on individual posts to predict stances. Using these methods, the veracity of the news based on the stance values of relevant posts can be inferred. Propagation-based approaches for fake news detection deals with the interrelations of relevance and use predictive algorithms to predict news credibility. Another new area of veracity measurement of news is rumor classification using machine learning methods.

b. Briefly describe in your own words, five recent news items, most significant in your opinion, which demonstrate how Machine Learning is being used to advance the society.

Answer:

1. Detecting patients' pain levels via their brain signals:
   - ❏ Recently researchers from MIT and elsewhere have developed a system that can measure a patient's pain level by analyzing brain activity from a portable neuroimaging device. The researchers developed personalized machine-learning models using the measured brain signals to detect patterns of oxygenated hemoglobin levels associated with pain responses. The models can detect whether a patient is experiencing pain with around 87 percent accuracy. The system could help doctors to diagnose and treat pain in unconscious and noncommunicative patients and could reduce the risk of chronic pain that can occur after surgery.

2. Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model by Google
   - ❏ Google machine learning team are working to improve the quality of speech recognition for speakers of data-scarce languages. This led them to study multilingual speech recognition where a single model learns to transcribe multiple languages. They have presented an end-to-end system trained as a model, which allows for multilingual speech recognition. For this study they have chosen India which is inherently a multilingual society.

Using nine Indian languages, google team has demonstrated a dramatic improvement in the ASR quality on several data-scarce languages.

3. Machine learning for comprehensive forecasting of Alzheimer’s Disease progression:
    - ❑ A recent study shows that unsupervised machine learning model can be used to predict alzheimer disease progression among people. Alzheimer’s Disease (AD) and Mild Cognitive Impairment (MCI) are complex neurodegenerative diseases with multiple cognitive and behavioral symptoms. They have used synthetic patient data including the evolution of each sub-component of cognitive exams, laboratory tests, etc. to train a logistic regression model and an unsupervised model that predicts alzheimer disease progression.

4. Food detection mobile application:
    - ❑ A research team at McGill University in Canada has developed a mobile application using machine learning technologies that can recognize food items inside an overall meal in real-time, providing useful nutrition-related information just from the food image. Using this app patient can keep track of their daily consumption and nutrition. This allows the app to provide important nutrition-related information (e.g. calories, amounts, etc.) for each food component detected by the CNN-based model.

5. AI-powered weather forecasts for smart grids' energy outputs:
    - ❑ National Grid in the UK is using machine learning to help predict how much energy they will reap from turbines and solar panels when the wind is blowing or the sun is shining. They are expecting that improved solar forecasts will help them to run the system more efficiently with lower bills for consumers. Moreover they aim to achieve a zero-carbon electricity system by 2025 by with the help of these machine learning technologies.

4. Follow the simple tutorial at https://www.machinelearningplus.com/logistic-regression-tutorial-examples-r/

(Links to an external site.)

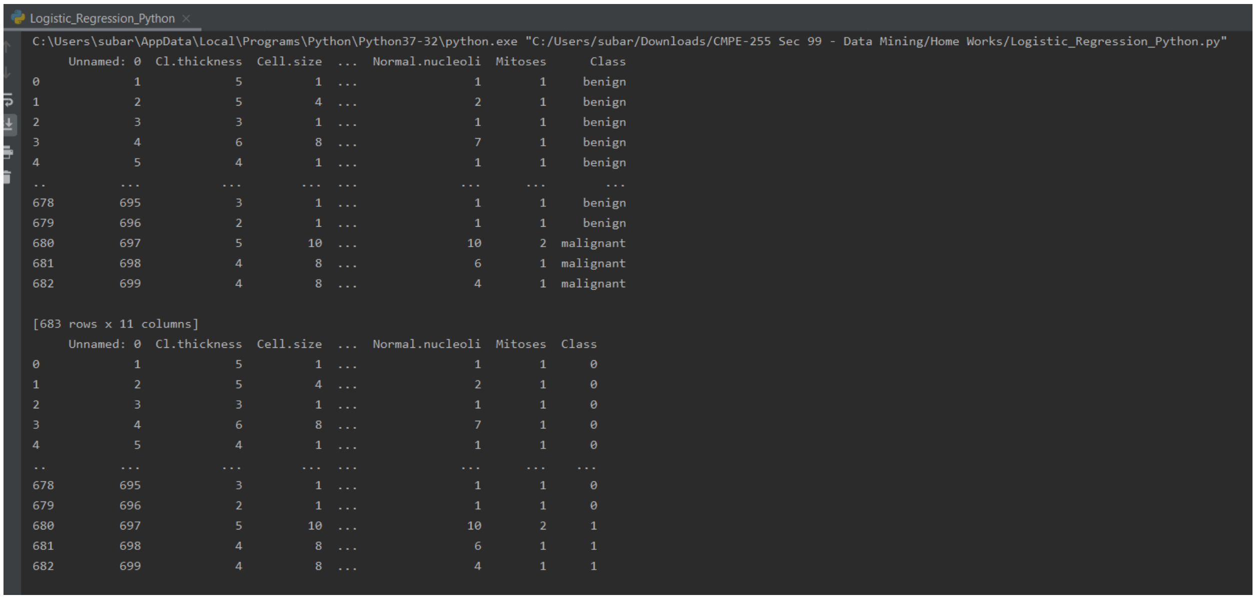
to see Logistic Regression in action. Implement the same functionality for the same dataset in Python. Do you achieve the same accuracy as with R?

Answer:

File Name:

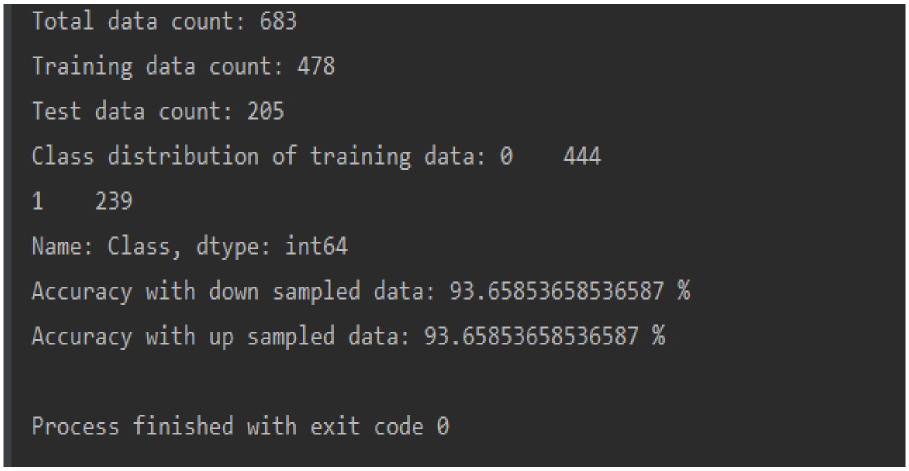
1. Logistic_Regression_Python.py
2. Logistic_Regression_Python.R
3. Mydata.csv [Breast Cancer dataset, Location: https://drive.google.com/open?id=1NMKxhBO7U2cS7hJpvU1uyH0txFwZmYQ6]

Output: To execute, put the csv file in the same folder



```
C:\Users\subar\AppData\Local\Programs\Python\Python37-32\python.exe "C:/Users/subar/Downloads/CMPE-255 Sec 99 - Data Mining/Home Works/Logistic_Regression_Python.py"
     Unnamed: 0  Cl.thickness  Cell.size  ...  Normal.nucleoli  Mitoses      Class
0             1             5          1  ...                1        1     benign
1             2             5          4  ...                2        1     benign
2             3             3          1  ...                1        1     benign
3             4             6          8  ...                7        1     benign
4             5             4          1  ...                1        1     benign
..          ...           ...        ...  ...              ...      ...        ...
678         695             3          1  ...                1        1     benign
679         696             2          1  ...                1        1     benign
680         697             5         10  ...               10        2  malignant
681         698             4          8  ...                6        1  malignant
682         699             4          8  ...                4        1  malignant

[683 rows x 11 columns]
     Unnamed: 0  Cl.thickness  Cell.size  ...  Normal.nucleoli  Mitoses  Class
0             1             5          1  ...                1        1      0
1             2             5          4  ...                2        1      0
2             3             3          1  ...                1        1      0
3             4             6          8  ...                7        1      0
4             5             4          1  ...                1        1      0
..          ...           ...        ...  ...              ...      ...    ...
678         695             3          1  ...                1        1      0
679         696             2          1  ...                1        1      0
680         697             5         10  ...               10        2      1
681         698             4          8  ...                6        1      1
682         699             4          8  ...                4        1      1
```



```
Total data count: 683
Training data count: 478
Test data count: 205
Class distribution of training data: 0    444
1    239
Name: Class, dtype: int64
Accuracy with down sampled data: 93.65853658536587 %
Accuracy with up sampled data: 93.65853658536587 %

Process finished with exit code 0
```

Observation: In Python, accuracy is almost the same as with R. Accuracy in Python is same for up and down sampling of data. It is: 93.6585 ≈ 94 %

5. a. ***Sigmoid functions such as the logistic function played a major role during the discussion on Machine Learning for Veracity of the tweets. What are some of the other mathematical functions that can possibly take the place of the Sigmoid function and help with Machine Learning? Comment on their effectiveness.***

Answer:

Some alternatives of sigmoid function are given below:

1. Tanh — Hyperbolic tangent:
    - ❑ It's mathematical formula is **f(x) = 1 — exp(-2x) / 1 + exp(-2x) = x / (1 + |x|)**
    - ❑ Range: (-1, 1)
    - ❑ Effectiveness: Optimization is easier in this method. Tanh function is also sigmoidal ("s"-shaped). Unlike logistic function, tanh can map strongly negative inputs to negative outputs. Additionally, only zero-valued inputs are mapped to near-zero outputs. These properties make the network less likely to get "stuck" during training.
2. Softsign:
    - ❑ It's mathematical formula is **f(x) = x / (1 + |x|)**
    - ❑ Range: (-1, 1)
    - ❑ Effectiveness: Softsign function converges polynomially. Tanh and softsign functions are closely related but in some cases soft sign activation function is smoother than tanh activation function. And this gentler non-linearity actually results in better faster learning. Researchers have found that soft sign has prevented neurons in machine learning neural network processes from being saturated early.
3. ReLu- Rectified Linear units:
    - ❑ It's mathematical formula is **f(x) = max(0,x) i.e if x < 0 , f(x) = 0 and if x >= 0 , f(x) = x**
    - ❑ Range: [0, ∞)
    - ❑ Effectiveness: This function is very simple and efficient and recently proved to be 6 times more convergent compared to Tanh function. It avoids and rectifies vanishing gradient problem . Almost all deep learning Models use ReLu nowadays. Some widely used variations of ReLu are Bipolar rectified linear unit (BReLU), Leaky rectified linear unit (Leaky ReLU), Randomized leaky rectified linear unit (RReLU), Exponential linear unit (ELU) etc.

4. Heaviside step function:
    - It’s mathematical formula is **$f(x) = 0$ if $x < 0$, $f(x) = 1$ if $x >= 0$**
    - Range: {0,1}
    - Effectiveness: The function produces 1 (or true) when input passes threshold limit whereas it produces 0 (or false) when input does not pass threshold. Therefore, it is also known as binary step function, which is very useful for binary classification studies.
5. Inverse Hyperbolic Sine function and its derivative function:
    - It’s mathematical formula is **$f(x) = \sinh^{-1}(x) = \log[x + \sqrt{(1 + x^2)}]$; $f'(x) = 1/\sqrt{(1 + x^2)}$**
    - Range: (-∞, ∞)
    - Effectiveness: The inverse hyperbolic sine is a multivalued function and hence requires a branch cut in the complex plane. The function produces outputs in scale of (-∞, ∞) .
6. Inverse Secant Function/ Inverse Hyperbolic Cosine Function :
    - It’s mathematical formula is **$f(x) = \cosh^{-1}(x)= 2 / (\exp(-x) + \exp(x))$**
    - Range: (0,1)
    - Effectiveness: The function produces outputs in scale of [0,1]. Output decreases and becomes close to 0 when x goes to infinite. However, it will never produce 0 as output even for very large inputs except ±∞. Mostly being used in neural network of machine learning.

b. ***Describe the use of Machine Learning, if any, in the following computer systems, clearly identifying and explaining whether it is supervised, unsupervised, semi-supervised, reinforcement learning or a combination of two or more of them:***

i) ***A coin classification system for a vending machine based on exact coin specifications from the U.S. Mint. The vending machine uses a statistical model of the size, weight, and denomination to classify coins.***

Answer: Size, weight and denomination are fixed for a coin. So supervised learning techniques can be used for coin classification system as model can be trained with an abundance of data about the dimensionalities of a coin.

ii) ***Detection of violence from surveillance camera feeds.***

Answer:
For violence detection, supervised deep learning methods can be used. From the camera feeds and other violent videos, features can be extracted to build a supervised

linear classifier model (e.g. SVM) to classify and detect violence from surveillance camera feeds.

iii) ***Detection of disease re-emergence based on past observations and present conditions.***

Answer: As we have past observation and present condition data of the disease to detect, combination of supervised and semi-supervised learning methods can be used here. We can use past data to train the model using supervised learning. After that, semi-supervised learning methods can be used with limited labeled data (present condition) to detect the re-emergence of disease.

iv) ***Identifying newer plant diseases based on leaf images.***

Answer: Unsupervised learning methods can be used here. Because we have to identify newer plant diseases, so training data will be unlabeled. Without having any prior knowledge, unsupervised techniques can be applied here to train the model to identify newer plant diseases.

v) ***Strategized Chess-playing by playing repeatedly and adjusting the strategy by penalizing moves that eventually lead to losing.***

Answer: For strategized chess-playing system reinforcement learning techniques can be used. Reinforcement learning system, called agent, can observe the environment, select and perform actions, and get reward or penalties in return. It can then learn by itself what is the best policy to get the most reward over time. In chess, similarly the reinforcement system will learn the strategy of winning the game by training with such rewards and penalties.