Logistic Regression Analysis of breast cancer tumor using Python IDE

Yashaswini B M
Dept of CSE
DBIT
Bangalore,India

Manjula K
Dept of CSE
DBIT
Bangalore,India

***Abstract*- In this paper we have used Logistic regression to the data set of size around 1200 patient data and achieved an accuracy of 89% to the problem of identifying whether the breast cancer tumor is cancerous or not using the logistic regression model in data analytics using python scripting language.**

Keywords: Data Science;Data Analytics;Logistic Regression;Python

I. INTRODUCTION

Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

Data analytics (DA) is the process of examining datasets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Linear and Logistic regressions are usually the first algorithms people learn in predictive modeling.[5] Regression analysis is an important tool for modelling and analyzing data. In our paper we have used Logistic regression to the data set of size around 1200 patient data and achieved an accuracy of 89% to the problem of identifying whether the breast cancer tumor is cancerous or not.

II DATA ANALYSIS IDE

Python in Data Analytics : Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable and it has fewer syntactical constructions than other languages. It has list of data structures which are used

Widely.[6]

Following are some data structures:

- Lists – Lists are one of the most versatile data structure in Python. A list can simply be defined by writing a list of comma separated values in square brackets. Python lists are mutable and individual elements of a list can be changed.
- Strings – Strings can simply be defined by use of single (‘), double (”) or triple (”’) inverted commas. Strings enclosed in tripe quotes (”’) can span over multiple lines and are used frequently in doc strings
- Tuples – A tuple is represented by a number of values separated by commas. Tuples are immutable and the output is surrounded by parentheses so that nested tuples are processed correctly.Since Tuples are immutable and cannot change, they are faster in processing as compared to lists. Hence, if your list is unlikely to change, you should use tuples, instead of lists.
- Dictionary – Dictionary is an unordered set of key: value pairs, with the requirement that the keys are unique (within one dictionary). A pair of braces creates an empty dictionary: {}.

Following are a list of libraries, used for data analysis:

- NumPy stands for Numerical Python. It contains basic linear algebra functions, Fourier transforms, advanced random number capabilities and tools for integration with other low level languages like Fortran, C and C++
- SciPy stands for Scientific Python. It is one of the most useful library for variety of high level science and engineering modules like discrete Fourier transform, Linear Algebra, Optimization and Sparse matrices.
- Matplotlib for plotting vast variety of graphs, starting from histograms to line plots to heat plots..
- Pandas for structured data operations and manipulations. It is extensively used for data munging and preparation. Pandas were added relatively recently to Python and have been instrumental in boosting Python’s usage in data scientist community.
- Scikit Learn for machine learning. Built on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.[6]

Analysis in Python using Anaconda and Pandas:

In this paper we are using Python Programming language and Anaconda as package manager and Pandas for data analysis. Anaconda is an easy-to-install, free package manager, environment manager, Python distribution, and collection of over 150 open source packages with free community support.Pandas is one of the most useful data analysis library in Python. They have been instrumental in increasing the use of Python in data science community.

III DATA ANALYTICS MODELS USED FOR ANALYSIS

- In this paper we have used Logistic Regression Model. Logistic regression is used to find the probability of event=Success and event=Failure. It is used for the classification problems. In this case we will use it for binary (1,0) classification.[5]
- Based on the observations in the histogram plots, we can reasonably hypothesize that the cancer diagnosis depends on the mean cell radius, mean perimeter, mean area, mean compactness, mean concavity and mean concave points. We can then perform a logistic regression analysis using those features as follows:

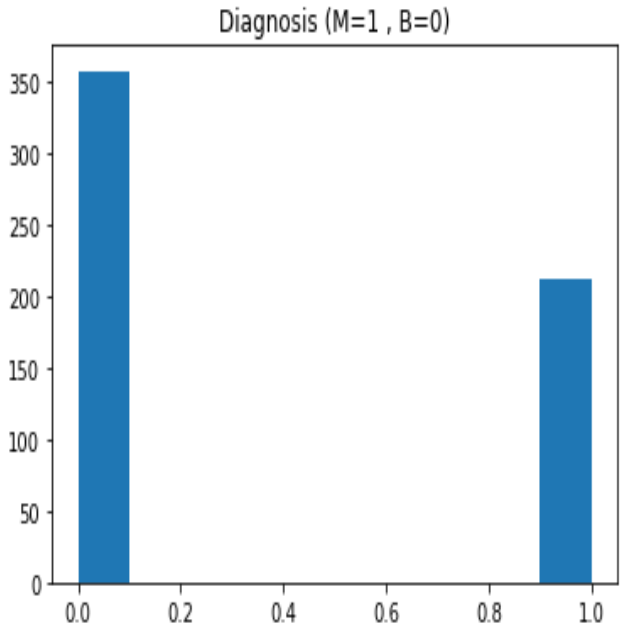
predictor_var= ['radius_mean','perimeter_mean','area_mean','compactness_mean','concave points_mean']

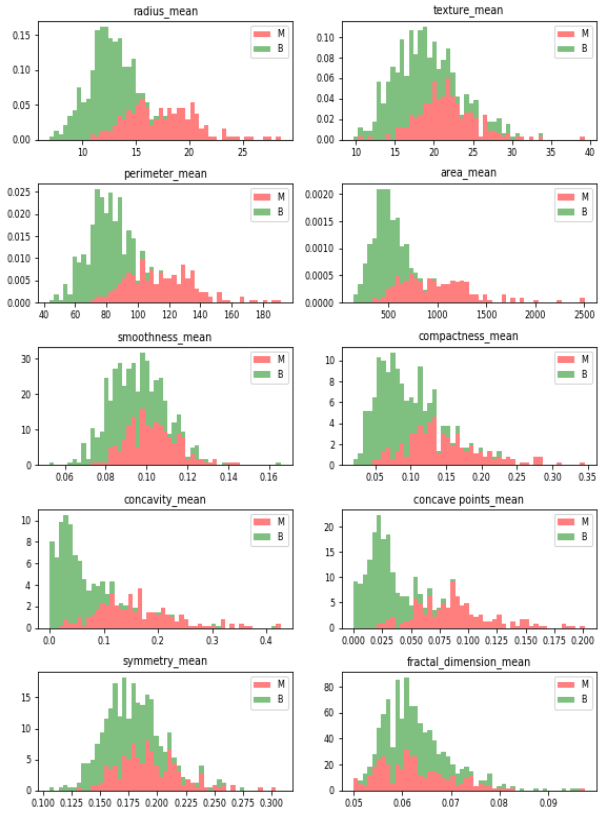
outcome_var='diagnosis'

model=LogisticRegression()

classification_model(model,traindf,predictor_var,outcome_var).

IV RESULTS





Observations from the output

- Mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters tends to show a correlation with malignant tumors.
- Mean values of texture, smoothness, symmetry or fractual dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no noticeable large outliers that warrants further cleanup.

Accuracy : 89.196%

Cross-Validation Score : 93.750%

Cross-Validation Score : 91.250%

Cross-Validation Score : 89.167%

Cross-Validation Score : 90.293%

Cross-Validation Score : 89.449%

Accuracy : 88.945%

Cross-Validation Score : 92.500%
Cross-Validation Score : 90.625%
Cross-Validation Score : 89.167%

Cross-Validation Score : 89.660%
Cross-Validation Score : 88.943%
V Conclusion

Based on the observations in the histogram plots, we can reasonably hypothesize that the cancer diagnosis depends on the mean cell radius, mean perimeter, mean area, mean compactness, mean concavity and mean concave points. Larger values of these parameters tends to show a correlation with malignant tumors.

In this paper we have used Logistic Regression del which is widely used for the classification of discrete data.It gives an accuracy of 89.94% by considering the respective parameters.

REFERENCES

[1] O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.
[2] W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.
[3] W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pages 77-87, April
[4] W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computerized breast cancer diagnosis and prognosis from fine needle aspirates. Archives of Surgery 1995;130:511-516.
[5] Data to establish inspection Applying logistic regression to maintenance intervals K.E. Spezzaferro.
[6] Introduction to Computation and Programming Using Python with Application to Understanding Data by GUTTAG JOHN V