PUBLIC HEALTH FOUNDATION OF INDIA

The 5th Asian Population Association Virtual Conference

# Application of Machine Learning in Demographic Analysis through Python software: An Exploration through Anaconda Navigator

**Author: Mr. Vijay Kumar Mishra, Research Scientist, Public Health Foundation of India (PHFI)**

## INTRODUCTION & OBJECTIVES

Nowadays, Machine learning (ML) has become very popular in data science for resolving complex issues as well as for big-data analytics. This study tried to explore various aspects of machine learning to explore analysis of secondary data (Demographic and Health Surveys, DHS-VII, 2017-18), through specially designed python software libraries. The main aim of this study is to understand the analytical aspects of machine learning (KNN classification, Decision Tree, Logistic regression, Confusion matrix and ROC) in demography. Also, this study tried to establish a predictive model of outcome(anemia) and predictors.

## DATA SOURCE AND METHODS

This study used Nationally Representative Demographic and Health Survey data (DHS-VII, 2017-18) of Albania. The DHS-VII in Albania, was executed by the Institute of Statistics. The survey provided information on important indicators of maternal and child health, fertility and mortality. DHS-VII in Albania included 15,823 households, 15,000 women, 6,142 men, 2,696 couples, 2,762 children (<60 months), 16,128 all births and 54,019 household members. The data analyses (unweighted) were performed after removing the list-wise missing cases. This study included data of children (n=1,958) and related mothers for all analytical aspects.The outcome variable in this study was the presence of anemia in children (<60 months). The variable, 'anemia level' was dichotomized into the variable 'anemia' (no = 0, yes = 1) for analysis. The predictor variables used were 'age(child)', 'maternal education', 'wealth index(household)', 'exposure to media(maternal)', 'number of ANC visits(mother)', and 'birth weight(child)'.

The main python libraries used in this study are 'pandas', 'numpy', 'matplotlib', 'pyreadstat', 'itertools', 'statmodels, 'skicit-learn', 'seaborn'. The descriptive analyses including exploratory graphs were done to understand the distribution of data. The cross-tab was shown with the help of heat map. The correlation heat map was created to show the relationship between outcome and predictors. The predictive model for outcome variable (Anemia) was revealed through 'confusion matrix' and 'ROC'. All the statistical analyses including data mining, were performed using python software (version 3.7), through Jupyter inbuilt in Anaconda Navigator.

**Compliance with Ethical Standards:** This study does not contain any studies with human participants performed by the author. The secondary data used in this study is publicly available on the website of Demographic and Health Surveys (DHS). Author has received access of data on request from DHS.

## RESULTS

This study found that around 28 percent of children are anemic in Albania. The prevalence of anemia varied from poorest (highest, 11.03%) to richest (lowest, 1.84%). Most of the anemic children belonged to poorest section of the society. The overall accuracy of the logistic regression model was 73%.

*Fig.1 Anaconda Navigator: An Integrated Development Environment (IDE) for data analysis*

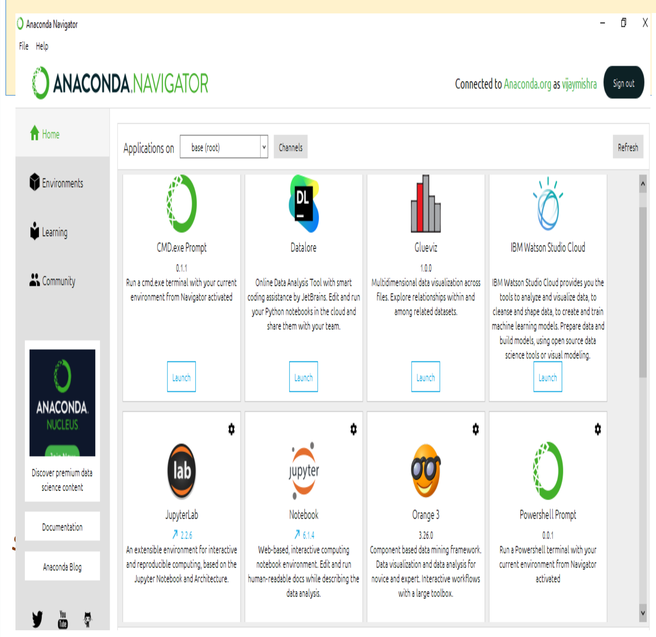



*Fig.2 Prevalence of anemia among children (<60 months)*

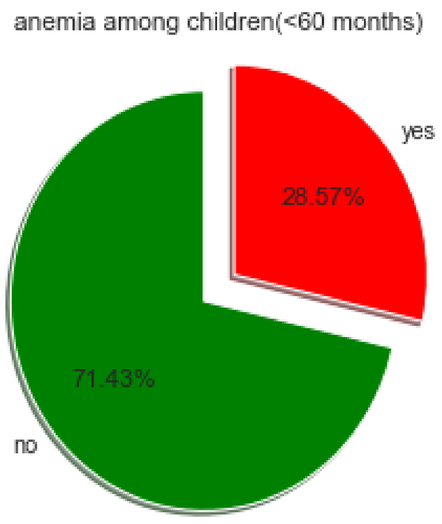



*Fig.3 Prevalence of anemia among children by households' wealth index*

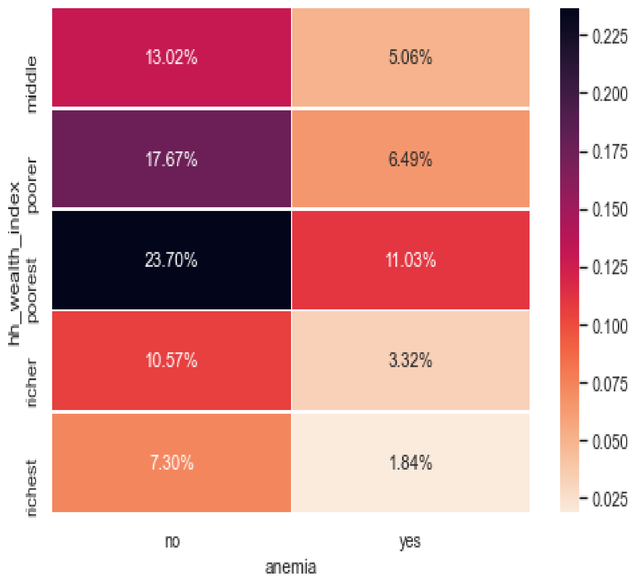



*Fig.4 Heatmap showing relationship between hemoglobin level and selected predictors*

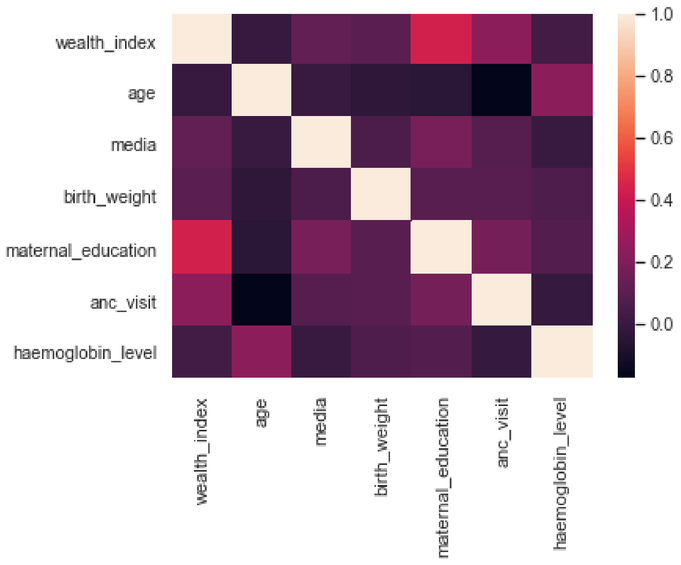



*Fig.5 Pair-plots showing distribution of anemia and predictors*

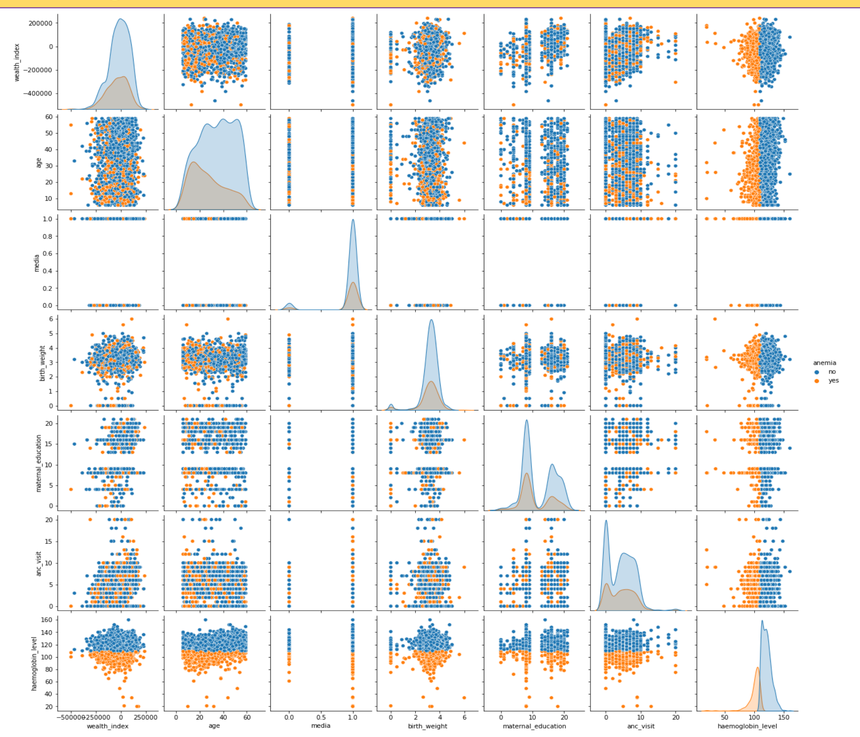
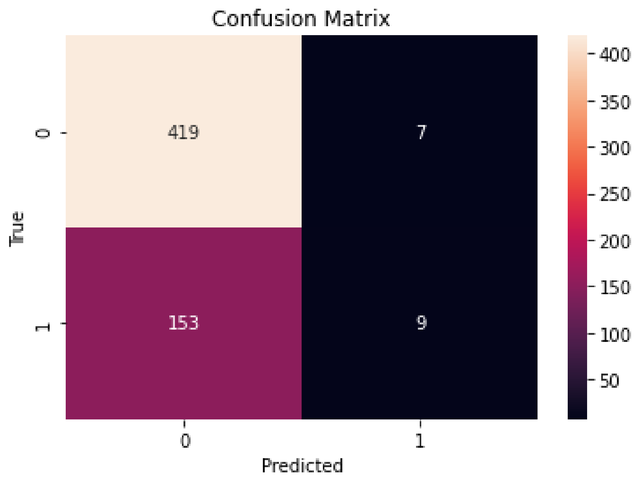



*Fig.6 Confusion matrix (heatmap) showing true and predicted values with accuracy and precision*




| | precision | recall | f1-score | |
|---|---|---|---|---|
| accuracy | | | | 0.73 |
| macro avg | 0.65 | 0.52 | 0.47 | |
| wighted avg | 0.69 | 0.73 | 0.64 | |

*Fig.7 Receiver operating characteristics curve (ROC) showing a plot between True ve+ and False +ve for regression model*

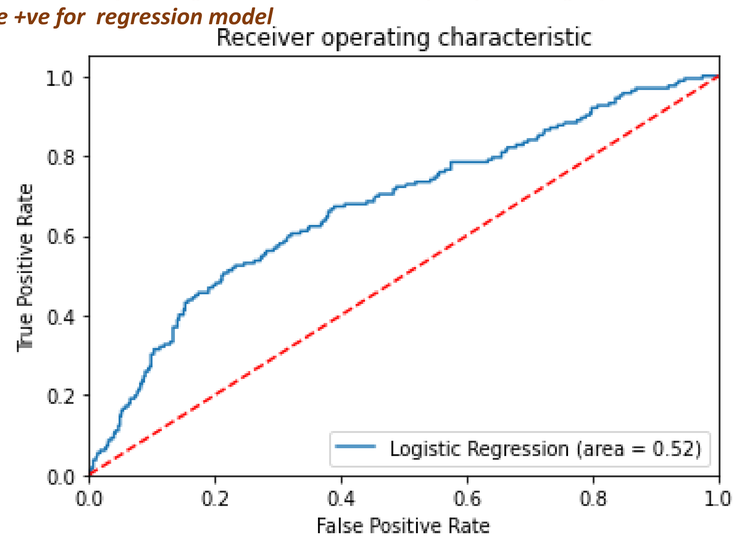
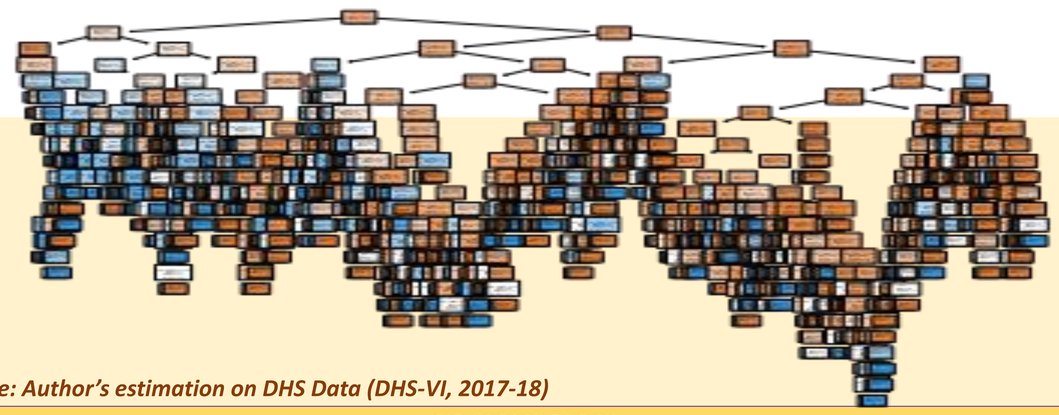



*Fig.8 Decision tree classification model in Logistic regression*



*Source: Author's estimation on DHS Data (DHS-VI, 2017-18)*

## CONCLUSION

- **Anemia (outcome) among children can be controlled by increasing awareness among mothers on risk-factors, through mass media. The wealth index, maternal education and birth weight of the child were found as highly significant predictors of anemia, through predictive modeling.**
- **This study found that most of the analyses used in demography; such as descriptive statistics including cross-tabulation, and predictive modeling (like logistic regression), can be performed easily in open source python software. The vital python packages used for various analyses in demography are 'numpy', 'pandas', 'statmodels' and 'skicit-learn'.**
- **The study also suggests on skipping the step of test-train data spilt for predictive modeling, if number of cases in data<2000. Since, most of the machine learning models provide better precision and accuracy, only if we deal with big data.**
- **The analytical aspects like heat-map based cross tabs, confusion matrix, decision tree and ROC are very useful in Demography.**