Data Analysis on the Risks of Obesity and Overweight in Women-A Study

Dr. Manjula Sanjay Koti[1], Alamma B.H.[2]
[1]*Professor & Head, Dept. of MCA, Sir M. Visvesvaraya Institute of Technology, Bangalore-562 157, India*
[2]*Assistant Professor, Dayananda Sagar College of Engineering, Bangalore-560072, India*

Abstract **- Health problems in women are increasing globally and obesity being the main one which is higher in females than males. Along with under nutrition, obesity as epidemic is continued as a problem in some countries, including India, as double burden This is not just affecting adults but also children and adolescents. Other risks are Polycystic ovary syndrome, blood Pressure, blood Sugar, Thyroid and others. It has emerged and reached epidemic proportions seen in industrialized countries. These factors affect more on women of reproductive age. Compared with normal-weight women, obese women are prone to develop diseases like PCOS, Diabetes, CVD, Hypertension etc. This paper analyzes health risks of obesity and also prediction of other health risks like PCOS, Diabetes, Hypertension and Thyroid using the data containing health records by exploratory data analysis and machine learning techniques, respectively. We use Random Forest (RF) and Decision Tree (DT) classifiers for analysis of risk factors in women. The performance based on accuracy rate is measured by comparing with the two different classifiers.**

Index Terms **- Random Forest, Decision Tree, PCOS, diabetes, Hypertension.**

1.INTRODUCTION

Obesity has become an epidemic over last half centuries in developed countries which has been estimated to be over 1.5 billion adults who are obese, it is seen more in female and nearly 300 billion women are considered to be clinically obese. It is required to have public aware-ness to fight against this epidemic and prevent it from growing which is one of vital social responsibility. As per the world health organization a major portion of the total health budget of families and communities have to bear the cost of managing the associated medical condition for individuals being too fat (adiposity) causing significant health problems. Adiposity has particular consequences for women health like PCOS, Diabetes, CVD, Hypertension, Thyroid disease etc. [4][8]. Diabetes is a hyperglycemic state of the body in which there is high glucose level in the blood, It is another health problem in women which is risk factor for developing hypertension, PCOS, CVD, Thyroid disease etc. Diabetes can occur during pregnancy or menopause in women [5]. Polycystic ovary syndrome (PCOS) is another risk factor in women which is caused due to hormonal imbalance of women at reproductive age. It may have infrequent or prolonged menstrual periods or excess male hormone (androgen) levels. The ovaries may develop numerous small collections of fluid (follicles) and fail to regularly release eggs [6]. Blood pressure is determined both by the amount of blood heart pumps and the amount of resistance to blood flow in arteries. Uncontrolled high blood pressure increases the risk of serious health problems, including heart attack, stroke, renal failure, etc. Thyroid disease effects women causing hyperthyroidism, hypothyroidism, goiter, thyroid cancer, and thyroid nodules.

2 RESEARCH OBJECTIVES

2.1 To analyze the risks of Obesity in women using exploratory data analysis.
2.2 To predict common risk factors in women using machine learning techniques.

3. LITERATURE SURVEY

In [11] the paper authors have discussed about the obesity risk factor of activities associating with obese and normal people. The algorithm Random tree, and

Logistic of SPSS and WEKA was used for data analysis which resulted in class level accuracy evaluation technique. Accordingly, 58% of people were found to be obese. They also suggested that to prevent this disease Government should take the step to cumulative cope and vegetables.

In [12] the paper authors have described about the Analysis of the relationships of overweight and obesity with place of residence, education and wealth index carried out using logistic regression. They also ad-dressed that maintaining socio-cultural barriers for healthy body size can contribute to fight the over-weight and obesity epidemic.

In [13] the authors have explained about the infertility in women are hard to detect or diagnose and can be diagnosed with greater precision with the help of predictive modeling. They showed that the best prediction can be done with the Random Forest algorithm. Another interesting observation was that the key variable selection would improve the performance of predictive models and help for the timely detection and treatment of infertility problem.

In [14] the authors have analysed adolescent obesity using General Bayesian Network (GBN) with What-If analysis and have explored how it can be utilized in other areas of public health. They have done performance comparisons with other algorithms like Support Vector Machine (SVM) which showed very poor results in accuracy (45.431%), together with the Naïve Bayes Network (NBN) (45.627%) compared to the other models implemented. Finally, GBN-MB showed the best performance in all the tests.

4.METHODOLOGY

4.1 Data Set

In this research, data was collected from Kaggle related to polycystic ovary syndrome (PCOS) from the open source. The data set contains all physical and clinical parameters to determine PCOS and infertility related issue. The data set was collected from 10 different hospital across Kerala, India. Primarily the data set contains 42 columns linked with PCOS of 541 patients.

4.2 Data Preprocessing

Using Python, the data was prepared with few more variables i.e. classifying if the patient has thyroid, BP, Blood sugar and Obesity using the following medical standards:

TSH (normal range):
Non-Pregnant - 0.4-4.0MIU/L
Pregnant – 0.1 – 3.0 MIU/L
Blood Sugar (Normal range)
At fasting - 80-100,3hrs
After eating-120-140
BP (normal range)-120/80

Table 1 Classification Category of BMI

Classification	BMI(Kg/m^2)	Risk of comorbidities
Underweight	<18.5	Low (other health risk)
Healthy weight	18.5-24.9	Average
Overweight (pre-obesity)	25-29.9	Increased
Obesity. class I	30-34.9	Moderate
Obesity, class II	35-39.9	Severe
Obesity,class III	40	Very severe

Based on these classified variables i.e Thyroid, BP, PCOS (available in the data) and Blood sugar, a dependent variable is created by applying few rules, for example, if a patient has Thyroid value as “Yes” and other classified variables as “No”, it means that she has thyroid issue. Same logic was used for other variables, respectively. In case a patient has “No” for all the variables, she is considered as normal. On the other hand, there are patient who had two issues at a time, in such cases we have assigned the primary issue because of which the secondary risk will be caused based on medical information.

Other data cleaning includes the missing values treatment through mean imputation.

4.3 Data Analysis

4.3.1Exploratory data analysis

In statistics, exploratory data analysis is an approach to get an overview of the data set to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task [3]. In this paper, Pandas package in python was used for importing the data and created a dependent variable by classifying the data into Thyroid, BP, PCOS, Sugar patients using a unique feature of Pandas i.e., data. Loc and value_counts. The former considers index labels and returns row/ data frame if the index/ label exists in

the data based on a condition in our case which are mentioned in the data pre-processing details. The latter gives a count of such patients in the data after making the classification using data.loc feature.

4.3.2Descision Tree

Decision Tree is most popularly and quite often used supervised learning algorithm for classification problems. It works for both categorical and continuous dependent variables. In this algorithm the population are split into two or more homogeneous sets which is done on most significant attributes/independent variables to make as a distinct group as possible. Each interior node relates to the input variable and its divided in children nodes based on the input node variable values. Each leaf node or terminal node represents a particular value of output variable. On execution of decision tree, samples on each interior node are divided in subsets based on variable, and this process is looped in each subset of a recursive partition. During growth of decision tree at each step, one of the input variable is selected for division samples. The new position is determined for the selected variable through a value test of the division sample, the most common test are entropy and impurity [7]. This technique performs well on large datasets and are extremely fast but on the other hand, it has disadvantages like Requires algorithms capable of determining an optimal choice at each node. Promne to overfitting, especially when a tree is particularly deep. Ideally, both these error due to bias and variance need to be minimized. One such powerful model in this area would be random forest [1][9].

4.3.3Random forest

Random Forest is an ensemble of decision trees, it is a collection of decision trees known as forest. To classify a new object each tree gives a classification, and it votes for that class. If the samples are N then in Random forest, N cases are taken but with replacement. Each tree is grown to the largest extent possible for selected variables at random and the best split on these is used to split the node [10].

One-way Random Forests reduce variance is by training on different samples of the data and by using a random subset of features. For example, if we have 30 features, random forests will only use a few features in each model, say five. Unfortunately, we have omitted 25 features that could be useful.Thus, in each tree we can utilize five random features. If we use many trees in our forest, eventually many or all of our features will have been included. This inclusion of many features will help limit our error due to bias and error due to variance. Therefore, random forests are a strong modeling technique and much more robust than a single decision tree [2].

4.3.4 Data Modelling Procedure

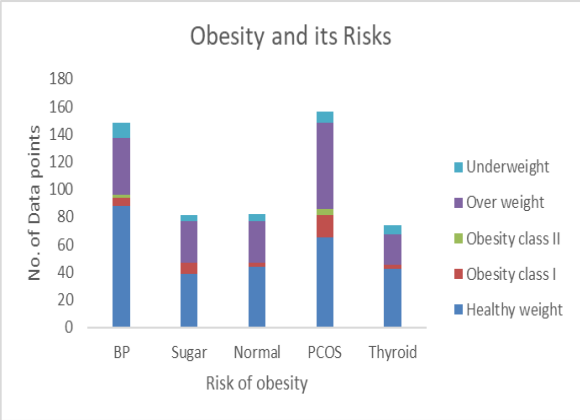
The data along with the calculated dependent variable was split into train (80%) and test sets (20%) using the sklearn.model_selecton package in Python. These sets were modelled using Decision tree and Random forest using sklearn.tree and sklear.ensemble packages with the functions DecisionTreeClassifier and RandomForestClassifier functions respectively. The evaluation metrics like accuracy, recall, f1 score and confusion matrix of the above mentioned models were obtained using the sklearn.metrics.

5. RESULTS AND DISCUSSION

The perusal of the Table 2 presents BMI classes v/s the number of people who are normal and also affected with the risks like BP, Sugar, PCOS and Thyroid. Majority of the section is affected with PCOS (156) followed by BP (148) and the least affected risk across the BMI classes is Thyroid. On the other hand, there are higher number of Healthy people (278) followed by overweight (185) and the section with least number are Obesity class II (7).

Table 2

BMI Classes/ Medical Condition	BP	Sugar	Normal	PCOS	Thyroid	Total
Healthy weight	88	39	44	65	42	278
Obesity class I	6	8	3	16	3	36
Obesity class II	2	0	0	5	0	7
Overweight	41	30	30	62	22	185
Underweight	11	4	5	8	7	35
Total	148	81	82	156	74	541



Explains that Random forest gives a better prediction of risks in women with an accuracy of 77% when compared to decision tree. Accuracy is the most intuitive performance measure, and it is simply a ratio of correctly predicted observation to the total observations. Accuracy is a great measure but only when you have symmetric datasets where values of correctly predicted values are almost same. Therefore, we look at F1 score as this parameter is more useful when the output has an uneven class of correctly predicted values. F1 score is also higher for the Random forest model with 75% compared to decision tree which has 72% score.

Table 3 Confusion Matrix of Decision Tree and Random Forest

Parameters	Decision Tree	Random Forest
Training accuracy	0.72	0.77
F1 score (weighted Avg.)	0.72	0.75

From the table 4 and 5, it is seen that in the 20% of test population, random forest gives higher number of correct predictions for PCOS (25), BP (28) and Blood sugar (16). Whereas decision tree output is higher for Normal (15) and thyroid (4) categories but again it is not with huge difference.

Table 4 Confusion Matrix of decision Tree

Actual/ Predicted	1-PCOS	2-Bp	3-Normal	4-Blood Sugar	5-Thyroıd
1-PCOS	24	5	1	1	0
2-Bp	2	23	1	0	4
3-Normal	1	0	15	1	0
4-Blood Sugar	0	0	3	13	0
5-Thyroıd	1	3	7	0	4

Table 5 Confusion Matrix of Random Forest

ual/Predicted	1-PCOS	2-Bp	3-Normal	4-Blood Sugar	5-Thyroıd
COS	25	4	1	1	0
p	1	28	1	0	0
ormal	3	0	12	1	1
lood Sugar	0	0	0	16	0
hyroıd	1	6	7	0	3

6.CONCLUSION

This study proposed a method for classification processes of patients that suffer from Thyroid, BP, Diabetes, PCOS and Obesity using the source" infertility Data set" by Exploratory data analysis and two machine learning technique like decision tree and Random forest. They were compared with their parametric values revealing that the best results obtained from random forest method, with accuracy of 77% and f1score of 75%. These results show that the propose method has high percentage in evaluated metrics which its evidence to indicate that its efficient and accurate method. Data Analysis of 541 medical records of women were taken out of which 228 women were found to be Overweight and Obese between age 20-48Among these overweight and obese women the risk of having Diabetes = 38, BP=49, thyroid= 25and PCOS= 83were found

ACKNOWLEDGEMENT

I thank for the great contribution of the repository machine learning uci, for providing dataset of information that allowed to submit a study on different methods or techniques.

REFERENCES

[1] (https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991);
[2] (https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991)
[3] 3.https://en.wikipedia.org/wiki/Exploratory_data_analysis
[4] Dr Saravanakumar, Eswari, Sampath, Lavanya "Predictive Methodology for Diabetic Data Analysis in Big Data," ELSEVIER, ISBCC (2015)

[5] Stephanie Revels, Sathish A.P Kumar and Ofir Ben-Assuli, Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytis, Health Policy and Technology, 2017, http://dx.doi.org/10.1016/j.hlpt.(2017)
[6] Vijayalakshmi N, UmaMaheswari M, data mining to elicit predominant factors causing infertility in women, IJCSMC, Vol. 5, Issue. 8, August (2016)
[7] Min Chen, YixueHao, Kai Hwang, Lu Wang and LigWang,b Disease prediction by machine learning over big data from Healthcare communities", IEEE Access,(2017)
[8] Yu, C.Y.; Li, X.X.; Yang, H.; Li, Y.H.; Xue, W.W.; Chen, Y.Z.; Tao, L.; Zhu, F. Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate. Int. J. Mol. Sci. (2018)
[9] Menzies, N.A.; Wolf, E.; Connors, D. Progression from latent infection to active disease in dynamic tuberculosis transmission models: A systematic review of the validity of modelling assumptions. LancetInfect. Dis. (2018).
[10] D.Sindhuja, R. JeminaPriyadarsini, "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder", International Journal of Computer Science and Mobile Computing, Vol.5, Issue.5, ISSN 2320–088X, May (2016).
[11] RifatHossaina, etal" PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques" Procedia Computer Science 132 1068–1076(2018).
[12] SubasNeupaneet al "Overweight and obesity among women: analysis of demographic and health survey data from 32 Sub-Saharan African Countries" DOI 10.1186/s12889-016-2698-5, (2016).
[13] Simi M S et al." Exploring Female Infertility Using PredictiveAnalytic"978-1-5090-6046-7/17/$31.00, IEEE, (2017).
[14] Cheong Kim et al." Predicting Factors Affecting Adolescent Obesity Using General Bayesian Network and What-If Analysis", Int. J. Environ. Res. Public Health, 16, 4684, (2019).