Lakshmi sravanthi Naupada

**501076110**

Professor: Ceni Babaogula PhD

CIND820 Capstone project

04-04-2022

Table of Contents

[1. Abstract 5](#_Toc99992764)

[2. Objective and Research Question 5](#_Toc99992765)

[3. Previous Work and Results 5](#_Toc99992766)

[4. Our contribution and plan of work 6](#_Toc99992767)

[5. GITHUB repo 6](#_Toc99992768)

[6. Analysis conducted 6](#_Toc99992769)

[a. Initial Analysis 6](#_Toc99992770)

[b. Exploratory Analysis 7](#_Toc99992771)

[c. Experimental Design 7](#_Toc99992772)

[7. Key Findings 8](#_Toc99992773)

[a. Initial Analysis 8](#_Toc99992774)

[b. Results 15](#_Toc99992775)

[c. Comparative study of the three models 19](#_Toc99992776)

[d. Cross validation : K-fold cross validation 19](#_Toc99992777)

[8. Difference between the results with previous work 20](#_Toc99992778)

[9. Conclusion 20](#_Toc99992779)

[10. Shortcomings of the work and remarks 21](#_Toc99992780)

[11. Future work 21](#_Toc99992781)

[12. References: 21](#_Toc99992782)

### Abstract

Pregnancy is an important stage of woman’s life where she goes through a lot of physical and emotional changes in her physical body. It is very important to monitor the changes as these would affect the health of the newborn as well as the mother’s health. Maternity health and preventing maternal mortality are the topmost global targets of WHO for the year 2021, to meet the Sustainable developmental Goals.2 Although there has been a gradual decline in maternal mortality since 2000, an estimate of 810 women continue to die each day due to the complications of pregnancy and childbirth. This is more significant in the developing and underdeveloped nations due to poverty and less access to hospital care and regular monitoring of the health of pregnant women. Achieving this global target is essential for the well-being of future generations.1

### Objective and Research Question

The objective of this project is to develop a machine learning model that classifies pregnant women into high risk, mid risk and low risk based on the factors such as their age, blood pressure, blood glucose level and heart rate. The approach is based on answering the questions

* To determine the factors that would impact the most to pregnant woman placing them at high-risk. Using these models and findings, the maternity mortality rate could be reduced as these predictions give us an opportunity to monitor high risk group and thus take necessary precautions to reduce unwanted complications and mortality. 3
* To determine which model gives the higher accuracy in predicting the factors responsible

for maternal health risk.4 Will be a comparative study.

### Previous Work and Results

Marzia Ahmed et.al in the paper IoT Based Risk Level Prediction Model for Maternal Health Care in The Context of Bangladesh, worked on the maternity health dataset where the data was collected using IoT Internet of Things, web portals of the hospitals in Bangladesh. The work is based on looking at the factors responsible for high-risk pregnancy. For the analysis of risk factors categorizing and classification have been used. The proposed four continuous processes:

* Collection of patient health data using wearable sensors.
* Data stored in the local server and cloud server.
* Stored data has been classified and predicted using machine learning models.
* The predicted data was sent back to both hospitals and emergency care centers.

They used both Weka and Python for the machine learning models and compared the accuracy on both groups. The research work was applied on both uncategorized and preprocessed data. On applying different classification models on both groups Weka showed more accuracy on decision trees compared to Python.

The results of this work shows that Blood sugar level is the most important factor responsible for high risk pregnancy and also decision tree gives the highest accuracy of 97 % . The attribute selection was done by ranking system and chi square test it showed that blood sugar was the first factor responsible.

Rutvej Mehta et.al proposed a comparative study on different classification models and discussed the detailed information on the the different models and strategies on machine learning on maternal health7. This work will be one of my references for my work.

### Our contribution and plan of work

Considering these works as base I planned to determine the factors responsible for high risk pregnancy using R programming and worked on three classification models to compare which gives better accuracy and to determine which model performs well . A comparative approach on the performance and efficiency of these models were depicted in this paper .

### GITHUB repo

https://github.com/sravnaupada3?tab=repositories

### Analysis conducted

### Initial Analysis

**Data Description:** The dataset used for this research is the maternity health dataset from UCI machine learning repository. It has data collected from maternal health centres, hospitals, and community clinics from the rural areas of Bangladesh. This dataset has 7 attributes and 1014 instances. The attributes are Age, Systolic blood pressure, Diastolic blood pressure, Blood sugar, Body temperature, heart rate and risk level.

Among these the risk level is the categorical variable with levels low risk, mid risk, and high-risk levels. The remaining all variables are numeric5

The univariate analysis for each of the variable was performed to determine the data types , any missing values ,inconsistencies . At this stage the presence of outliers and determining the variance ,standard deviation gave us some information regarding the variable selection . And the column names and column datatypes were observed and deep understanding of the variables and selection of dependant variable was studied and the findings were reported in the key findings section.

Scatter plots and correlation matrix was used to determine the bivariate pairwise analysis and also to get some information on highly correlated variables . The results were included in the key findings.

### Exploratory Analysis

**Feature selection :** Rank feature importance is used to get the insights of the variable importance and varImp was used to determine the ranking of importance and among all variables the top three variables would be Blood sugar , Systolic Blood pressure and Age. The output is shown in the key findings.

### Experimental Design

1. **K-Fold Cross validation:** 10 fold cross validation - To evaluate the performance of models it is required to test on unseen data .Cross validation is one of the technique used to test the effectiveness of the models and involves resampling to evaluate the model .The most popular one is the 10 fold cross validation where the fitting procedure is performed ten times on ten folds . The results shows KNN has high accuracy compared to the other two models as shown in the key findings table for cross validation .
2. **Applied models and the study design :** Based on the experimental design and initial analysis the three models decision tree , Naïve Bayes and KNN algorithm are used to run the model and to analyze the performance and accuracy to compare the efficiency of models . The accuracy was initially calculated and based on confusion matrix the F1 score is considered to be the best measure for the unbalanced data and the precision , recall values calculated based on confusion matrix and the comparative study of the performance of models are depicted in the tables of key findings.

### Key Findings

### Initial Analysis

The dataset was imported using R programming and investigated for presence of any missing values. The result shows there are no missing values.

The imported dataset was investigated for the variables data types and found out to be the Risk Level as categorical and considered to be dependent variable as shown in the table 1.

**Table :1 Data type**

**Text, letter

Description automatically generated**

The minimum, maximum, mean values and the first, third quartiles for all the variables are given in the table 2.

**Table :2 Min, Max ,Median**

A screenshot of a computer

Description automatically generated with low confidence

By observing the column names there was a need for correction of one of the column name Age which was misspelt . So the column name for Age was modified . And the categorical variable was transformed from character into numeric into three levels 0 as low risk , 1 as mid risk and 2 as high risk table 3.

The mean median standard deviation of the variables and the variance were tabulated to see the distribution and to determine which variables have low variance as shown in the table 4

**Table :3**



**Table : 4 Standard deviation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Minimum** | **Maximum** | **Median** | **Mean** | **Standard Deviation** |
| **Age** | 10 | 70 | 26 | 29.87 | 13.47 |
| **Systolic BP** | 70 | 160 | 120 | 113.2 | 18.4 |
| **Diastolic BP** | 49 | 100 | 80 | 76.46 | 13.88 |
| **BS** | 6 | 19 | 7.5 | 8.7 | 3.29 |
| **Heart Rate** | 7 | 90 | 76 | 74.3 | 8.08 |
| **Body Temperature** | 98 | 103 | 98.6 | 98 | 1.37 |

The distribution of the dependant variable classes is analyzed using the histogram as shown below and shows that it is unbalanced .

**Figure:1 Histogram**



Since we observe the standard deviation values it is important to determine any outliers and we do this using boxplot. Boxplots for each of the variable shows the presence of outliers. The data points outside the whiskers show the presence of outliers as shown in table 5. The average age is 29.8 and the minimum and maximum age range is 10 and 70. The mean values of Systolic blood pressure and Diastolic blood pressure were 113. 2 and 76.46 respectively.

When we look at the BodyTemperatue the mean value is 98.67 and the mean of Blood sugar is 8.72. The average value of Heart Rate is 74.3

For all our quantitative variables the histograms show the distribution of variables in the figure 3.

The correlation matrix shown there was not much correlation between the variables as most of the values are below 0.7 except the variables Systolic and Diastolic Blood pressure where they have a high correlation of around 0.8. Shown in the figure 4 with dark blue circles.

**Figure:4 Correlation matrix**

**Chart, bubble chart

Description automatically generated**

Histogram is the most popular graphical representation of quantitative data. These shows the frequency distribution of the quantitative variable. So, in our dataset the variables Age, Systolic blood pressure, Diastolic Blood pressure, Heart rate, Blood glucose level and Body temperature are the quantitative variables and the following output table :3 shows the frequency distribution of these variables.

**Figure 3: Histograms of quantitative variables :**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated**

**Boxplots:** Boxplots are used to detect any outliers in the dataset and used R programming to analyze each variable in the dataset to generate boxplots. Based on the standard deviation values we can expect outliers in our variables. (Figure 2)

**Figure 2: Boxplots**

**Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated**

**Table

Description automatically generatedDiagram, box and whisker chart

Description automatically generated with medium confidence**

Based on all the above initial analysis the variable with high correlation that is systolic and diastolic blood pressure are considered and the diastolic blood pressure was removed from the analysis as it has less variance compared to systolic blood pressure .

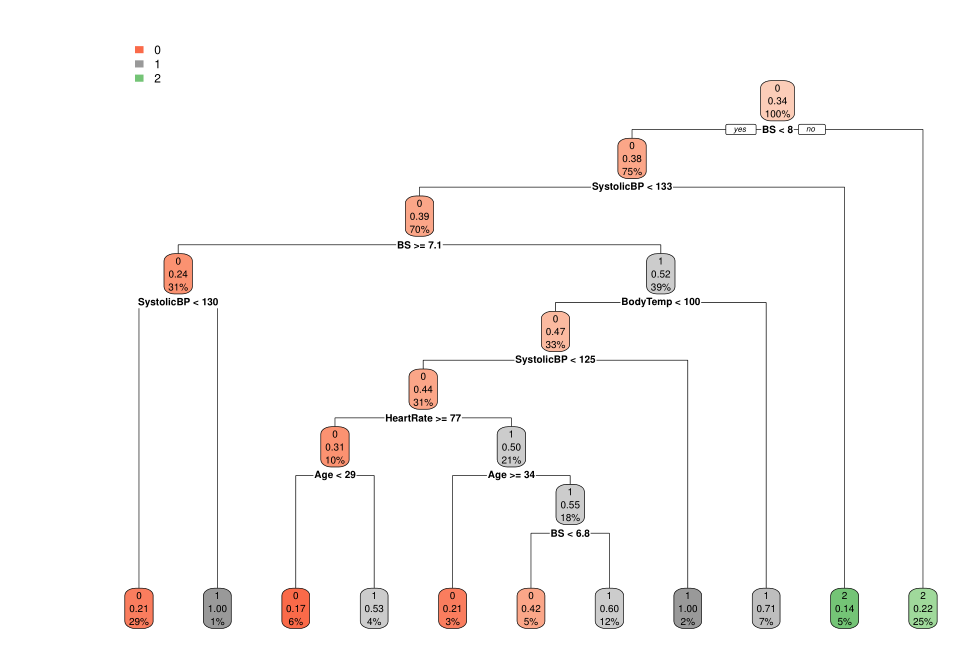
**Rank Feature Importance** : The feature selection shows that blood sugar has the highest rank followed by systolic blood pressure , diastolic blood pressure followed by Age ,heart rate and body temperature . So Blood sugar has the highest rank in the feature selection and least with body temperature.

**Figure 5:** **Rank Feature Importance** 

### Results

* + 1. **Decision Tree:** Decision tree is one of the commonly used classification algorithm for predicting the significant variables for the dependent class variable. In our project the Risk level will be the class variable. The figure shows the decision tree for Risk Level with classes 0,1 and 2 as low risk, mid risk, and high risk respectively.

**Figure 6: Decision tree for Risk Level Analysis**

****

According to the above tree BS, Systolic BP and BodyTemp are significant factors for the response variable. At the first level we see a threshold of BS (Blood sugar) at 8. And this indicated for the pregnant women with blood sugar greater than 8 leads to high-risk pregnancy indicating 2 in green color. And if the BS value greater than 9.5 is the upper limit where the person is high risk category. And the BS level between 8.5 and 9.5 indicate mid risk in grey color with number 1, but there is chance to get into high risk level. And for the pregnant individuals with BS less than 8 indicate low risk but depends on the other factors like Systolic Blood pressure and age.

And at the next level we see Systolic Blood pressure is the variable considered to be an important factor where for individuals with BS less than 8, if the Systolic blood pressure is higher than 133 it is considered as high-risk pregnancy with green 2. And the systolic Blood pressure less than 130 indicates low risk pregnancy with 0 and orange color.

Similarly, the Body temperature less than 100 indicated low risk in orange and body temperature higher than 100 indicate mid risk category.

Thus, using the decision tree, we can determine the factors responsible for classifying pregnant women into different ow risk md risk and high-risk categories.

**Table 5: Confusion matrix for prediction**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** | **Predicted 2** |
| **Actual 0** | 61 | 6 | 4 |
| **Actual 1** | 41 | 21 | 2 |
| **Actual 2** | 7 | 10 | 51 |

Apart from accuracy there is a need to analyze the other performance metrics such as

Precision and Recall . Since the confusion matrix we have for the decision tree is

unbalanced we consider F1 score as the main metrics of evaluation . The precision and re -call values for low risk mid risk and high risk are tabulated and the averages are used to calculate the F1 score as shown in the table 6 and 7. Since we consider F1 score as the main metric the value for this model would be **64%**

**Table 6: Precision and Recall Table**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Low risk 0** | **Mid risk 1** | **High risk2** |
| **Precision** | 0.56 | 0.56 | 0.89 |
| **Recall** | 0.85 | 0.32 | 0.75 |
|  |  |  |  |

**Table 7:**

|  |  |
| --- | --- |
| **Accuracy** | 66% |
| **F1 score** | 64% |

* + 1. **KNN algorithm:** This algorithm is the simplest and widely used algorithm, where the k nearest neighbors’ concept is used, it is a supervised model with no assumptions regarding the data. It works on the nearest neighbors’ model using Euclid distance between points.

**Table 8: Confusion matrix for KNN algorithm**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** | **Predicted 2** |
| **Actual 0** | 124 | 40 | 5 |
| **Actual 1** | 28 | 109 | 10 |
| **Actual 2** | 6 | 16 | 97 |

Based on the table 2 confusion matrix for knn algorithm 124 women are correctly classified as low risk , 109 are correctly classified as mid risk and 97 are correctly classified as high risk .

The accuracy for k = 1 is 77%

k= 5 is 65%

For this model we consider F1 score as best metric due to imbalance in the classes of the confusion matrix . So based on the precision and recall values in the table 9 ,the F score is calculated and found to be 75% shown in table 10

**Table 9: Precision and Recall**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Low Risk 0** | **Mid Risk 1** | **High Risk 2** |
| **Precision** | 0.78 | 0.66 | 0.86 |
| **Recall** | 0.73 | 0.74 | 0.81 |

**Table 10:**

|  |  |
| --- | --- |
| Accuracy | 77% |
| F1 score | 75% |

* + 1. **Naive Bayes Theorem:** Naïve Bayes theorem is a supervised learning model. This classifier is based on the Bayes theorem where the assumption is naïve where the occurrence of a feature is independent of the other.

**Table 11: Confusion matrix for Naïve Bayes theorem**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** | **Predicted 2** |
| **Actual 0** | 164 | 10 | 2 |
| **Actual 1** | 101 | 26 | 15 |
| **Actual 2** | 4 | 20 | 72 |

Based on the confusion matrix 164 cases were correctly classified as low risk and the remaining 105 are not correctly classified. And for the mid risk 26 individuals are correctly classified as mid risk and among 89 pregnant women 72 were correctly classified as high-risk pregnant women as shown in the table 11.

**Table 12: Overall statistics**

|  |  |
| --- | --- |
| **Overall Statistics:** |  |
| Accuracy | 0.6037 |
| 95% CI | 0.5559,0.65 |
| P value | 0.9972 |
| Mcnemar’s test P value | <2e-16 |
| Kappa | 0.3738 |
| No information Rate | 0.6659 |

This model has an accuracy of 60 % with p value less than 1.

The F1 score was calculated for this model based on the precision and recall values for the low risk , mid risk and high risk values and the score was found to be 61 % with accuracy of 60% as shown in the tables 13 and 14.

**Table 13: Precision and Accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Low Risk 0** | **Mid Risk 1** | **High Risk 2** |
| **Precision** | 0.60 | 0.46 | 0.80 |
| **Recall** | 0.93 | 0.18 | 0.75 |

**Table 14:**

|  |  |
| --- | --- |
| Accuracy | 60 % |
| F1 Score | 61% |

### Comparative study of the three models

To compare the performance of the models F 1score is considered as the main metric as the matrix is unbalanced . F 1 score is the harmonic mean between precision and recall the higher the score indicates the better the performance so among the above three models KNN algorithm gave the highest performance for the dataset as shown in the below table .

**Table 15: Comparative study of performance metrics**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** |
| **Decision Tree** | **66%** | **69%** |
| **KNN** | **77%** | **75%** |
| **Naive Bayes** | **60%** | **61%** |

### Cross validation : K-fold cross validation

The results for cross validation for the three models are shown below**:**

**Table :16 Accuracy table**

|  |  |
| --- | --- |
| **Models** | **Accuracy** |
| **Decision tree** | **0.62** |
| **KNN** | **0.68** |
| **Naïve Bayes** | **0.67** |

So even on unseen data the KNN algorithm has a good performance compared to other two models .

### Difference between the results with previous work

Past work was based on comparative model for the models decision tree , random forests , Naïve Bayes and the results show that decision tree has high accuracy and good performance compared to other models and they used python and weka for the analysis . And the cross validation 15 fold also showed the same results with decision tree highest accuracy with 98.51 % and 97% . And this research showed that the factors like blood pressure ,blood sugar and age are responsible for the high risk pregnancy .

Our research shows an F1 score of 75% for KNN algorithm for the comparative study that matches with the 10 fold cross validation with 68 % for KNN algorithm highest among the three models Decision tree , KNN and Naïve Bayes . And in our analysis blood sugar is the most important factor for high risk pregnancy followed by systolic blood pressure and then the Age . The least important is the body temperature .

### Conclusion

To conclude my work shows that the factors like Blood sugar, Systolic blood pressure and body temperature are the most important factors to be considered to classify the maternal women into low risk , mid risk and high risk . Among them blood sugar level is the top most important factor to consider as having high blood sugar results in high risk pregnancy as shown in our analysis .

And among the three models used for analysis KNN algorithm showed good accuracy and performance compared to other two models .

### Shortcomings of the work and remarks

As there is a considerable disparity compared to the original research findings there is further scope to improve the performance by using other machine learning models and selection of variables .

Normalising the numeric attributes might have given more accuracy although the numeric range is not significantly different from each other .

And the most important observation was that the Body temperature was found to be the least important on ranking in feature importance . But the classification models shows that it would be one of the top important factor . This difference would be due to the outliers in the variables so it is good to remove outliers in this case and reanalyze the models again.

### Future work

In future I would like to work on improving the model performance using other algorithms like random forests and improving the performance .

Would like to normalise the numeric data and check if it gives us a better performance and check the comparative study to determine the best model . And adding more variables and transforming the data to get more important variables to investigate if the performance will be improved. And I might also consider removing or replacing the outliers and perform the analysis to check if this improves performance

Reframing the research question to check if age and blood sugar are inter linked in putting the maternal health at high risk and decomposing into variables based on the high blood sugar levels by feature engineering .

### References:

* <https://www.who.int/news/item/05-10-2021-new-global-targets-to-prevent-maternal-deaths>
* <https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-020-03216-z>BMC Pregnancy Childbirth20, 518 (2020) Published 07 September 2020
* IoT Based Risk Level Prediction Model for Maternal Health Care In the Context of Bangladesh Marzia Ahmed1,2, and Mohammod Abdul Kashem2
* Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
* Comparative analysis of classification algorithms**,** Ravil Muhamedyev
* <https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set>
* A Survey on Data Mining Technologies for Decision Support System of Maternal Care Domain, Rutvej Mehta.
* Using Multinomial Logistic Regression to Examine the Relationship Between Children’s Work Status and Demographic Characteristics
* Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques I.A. Pasadana *et al* 2019 *J. Phys.: Conf. Ser.* **1255** 012024
* Diagnosis of Heart Disease Using Decision Tree Sabarinathan Vachiravel SRM Institute of Science and Technology
* Prediction of COVID-19 Possibilities using KNearest Neighbour Classification Algorithm Prasannavenkatesan Theerthagiri
* Breast Cancer Prediction using Naïve Bayes Classifier Megha Rathi Jaypee Institute of Information Technology