CIND820

Capstone Project: Initial Results

Lakshmi sravanthi Naupada

Student number : 501076110

Supervisor : Ceni Babaoglu, Ph. D

**Initial Results**

In this section the initial results for the project questions are included.

The objective of this project is to develop a machine learning model to classify the pregant women into low risk, mid risk, and high-risk categories. And, to determine which factor is most responsible for the high-risk pregnancy and finally to compare the machine learning models to determine the best model with high accuracy.

The following machine learning models are develpoed for the dataset health i.e maternity health dataset and the initial results are incuded in this section.

**Preprocessing of data:**

The dataset has one categorical variable RiskLevel with categories low risk, mid risk and high risk. The remaining variables are all numeric variables. But for the convenience of applying the machine learning models we are transforming the response variable RiskLevel into three numeric categories lowrisk as 0, mid risk as 1 and high risk as 2.

RiskLevel = (Lowrisk = 0, Midrisk=1, Highrisk = 2)

**1.Decision Tree:**

Decision tree is one of the commoly used classification algorithm for predicting the significant variables for the dependant class variable. In our project the Risklevel will be the class variable. The figure shows the decision tree for RiskLevel with classes 0,1 and 2 as low risk, mid risk, and high risk respectively.

**Figure :1 Decision tree for RiskLevel Analysis**

Timeline

Description automatically generated

According to the above tree BS, Age, Systolic BP, Diastolic 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 Bloodpressure and Diastolic Bloodpressure 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 Bloodpressure 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 preganancy with 0 and orange color.

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

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 :1 Confusion matrix for prediction**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted 0** | **Predicted 1** | **Predicted 2** |
| **Actual 0** | 53 | 23 | 2 |
| **Actual 1** | 20 | 30 | 16 |
| **Actual 2** | 1 | 7 | 51 |

Accuracy for model is 0.66 that is 66%

**2.KNN algorithm:** This algorithm is the simplest and widely used alogorith, 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 euclaid distance between points.

**Table 2: 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= 15 is 65%

**3.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 occurance of a feature is independent of the other.

**Table :3 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 3.

|  |  |
| --- | --- |
| **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 inforamtion Rate | 0.6659 |

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