*Article*

# Maternal Health Risk Prediction using meachine learning.

Affiliation : **KUNDRU SAI SRUTHI**

Correspondence: 2203a52156@sru.edu.in; Section: AIML-AA

SR University-school of cs &aIML

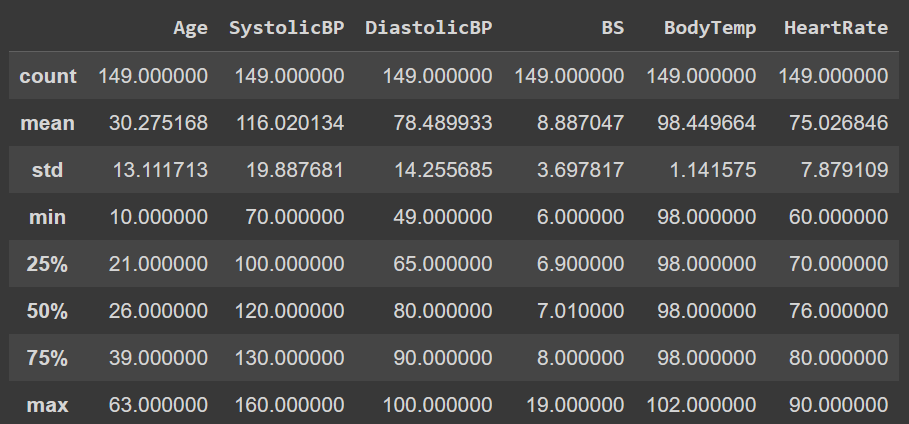
1 Affiliation 1;2203a52156@sru.edu.in

|  |  |
| --- | --- |
| **Abstract**: This maternal health risk dataset, a collection of features derived from atmospheric | 1 |
| different hospitals, community clinics, maternal health cares through the IoT based risk monitoring system. | 2 |
| our research aims to advance the field’s understanding of the best model selection strategy while | 3 |
| taking into account the specific characteristics of this dataset.in this dataset we can understand that the maternal risk by the features of the dataset .And through this we can help about the analysis of analysis risk factors by the help of the existing data set given with the features using many algorithams .it provides risk level. | 4 |
| **Keywords:** Support Vector Machines(SVM),LogisticRegression,PerceptronModel,K-NearestNeighbors | 10 |
| (KNN), Data Preprocessing ,Bootstrapping, Model Evaluation, Accuracy, Precision,Recall,F1Score,Model | 11 |
| Selection, Model Stability, Feature Engineering, Comparative Analysis, Research Methodology, Model | 12 |
|  | 13 |
| **1. Introduction**  The given data set have the features which helps us to predict the health risks of the pregnancy and the newborn which also helps in the important area of the healthcare of the women and also it will ensure that the pregnant lady will have the perfect diet according to the predictions so this will help or the vital role in the health cares the main aim of this is that by using the ai-ml which is artificial intelligence and the machine learning is the by using some features in the data set like age and many types of the bps of the pregnant women we can result the health as if the women has the high risk ,middle risk or the low risk .by this we can help the globally .By this result we can hold the  promise In the improvement of the maternal health .  In this case we are helping the pregnant women to have the healthy life and also we need to do the monthly tests for the data we needed .and also by this the pregnant women can take the precautions with the data provided by the results we have given with the help of the ai-ml which is machine learning language program .and the main help of this is that we can get the accurate values we needed and the perfect way for the women to have the precautions for the healthy newborn. | 14 |
| By this we can ensure the health care of the women by the risks .and it will also gives us about the maternal health history ,behaviour and the lifestyle factors ,clinical data, generic data, early health risk prediction and by this they can personalized care can be taken by the women so in somehow we can reduce the maternal mortality, and also develop the accessibility to the quality care. | 27 |
|  |  |
| **2. Literature Review** |
| The following literature review provides an overview of key research areas in the |
| context of the health risk prediction and the application of machine learning for the |
| classification of risk sources. we uses the meachine learning for the risk prediction for the user for the women and the it explore the historical progression of the aiml applications in the meternal health risk prediction and also it will discuss the various factors and the data soursrs commamly used in the maternal health risk prediction models for the prediction.it will also give the positive outcomes and the inpacts in the mortality rate .we use only the meachine learning algoritham because it is the perfect resourse for the predictions using many features   |  | | --- | | The following literature review provides an overview of key research areas in the | | context of the matenal health care and the application of machine learning for the | | classification of pregnant women |  1. **Data and Methodology:**   The given data is taken from the different hospitals ,clints of the community and the test centers.it is a multivariate characteristics dataset .with the uses algorithem of the classification having subject area of life science having the feature type real and the integer  And also having 150 instances and also 6 features .and all these Age, Systolic Blood Pressure as Systolic of BP, Diastolic of BP as Diastolic BP, Blood Sugar as the BS, Body Temperature as Body Temp, Heart Rate and Risk Level of the women ..  Here are the features information  **1.Age:**it isone of the feature in data set having the age of the women in years who is pregnant  **2.systolic BP:** It is the upper value of the blood pressure in mmHg calculated to the women who is pregnant, another significant attribute during pregnancy.  **3.Diastolic BP:** It is the lower value of the blood pressure in mmHg calculated to the women who is pregnant, another significant attribute during pregnancy.  **4**.**BS:** Blood glucose levels is in terms of a molar concentration, mmol/L  **5**. **Heart Rate:** A normal resting heart rate in beats per minute**.**  **6.Risk Level:** Predicted Risk Intensity Level during pregnancy considering the previous attribute**.** |

**Fig 1:data set\***

Correspondence: 2203a52156@sru.edu.in; Section: AIML-AA † SR University

**Figure 1.** Sample training data



|  |  |
| --- | --- |
| *3.2****. Data Analysis*** | 75 |
| These histograms help you understand how individual features contribute to the | 76 |
| classification problem. Features with distinct, non-overlapping distributions for the two | 77 |
| classes are typically more informative for classification. Features with substantial overlap | 78 |
| may not be as useful for distinguishing between the classes. | 79 |
| By these histograms, you can make informed decisions about feature selection, | 80 |
| model choice, and feature engineering to improve the performance of your classification model.  The below graph is given the graphs of the features and the count. | 81 |

**Figure 2.** Feature summary statistics.

|  |  |
| --- | --- |
| **Figure 3.** probability vs feature graph | 82 |
|  |  |
| 3.2.1. Correlation Matrix | 83 |
| A correlation matrix is a table that shows the relationships between multiple variables | 84 |
| in a dataset. In this matrix, each cell contains a correlation coefficient that quantifies the | 85 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| strength and direction of the linear relationship between pairs of variables. Here’s a brief | | | 86 | |
| explanation of a correlation matrix | | | 87  88 | |
| 1. Positive Correlation (0 to 1): Variables move in the same direction When one increases, | | | 89 | |
| the other tends to increase as well. | | | 90 | |
| 2. Negative Correlation (-1 to 0): Variables move in opposite directions. When one | | | 91 | |
| increases, the other tends to decrease. | | | 92 | |
| 3. No Correlation (0): There is no linear relationship between the variables. | | | 93 | |
| *3.3. Data Preprocessing* | | | 94 | |
| Data preprocessing is a critical step in machine learning that involves cleaning, trans- | | | 95 | |
| forming, and organizing raw data into a format suitable for model training. It plays a | | | 96 | |
| significant role in ensuring that the data is of high quality and that the machine learning | | | 97 | |
| model can learn meaningful patterns. Here’s an elaborate explanation of various aspects of | | | 98 | |
| data preprocessing: | | | 99  100 | |
| 1. **Data Cleaning:** | | | 101 | |
| *Handling Missing Values:* Identify and handle missing data, either by removing rows | | | 102 | |
| or filling in missing values using techniques like mean, median, or interpolation.There | | | 103 | |
| are no null values in this dataset | | | 104 | |
| 2. **Data Transformation:** | | | 105 | |
| *Feature Scaling:* Normalize or standardize numerical features to ensure that different | | | 106 | |
| features are on a similar scale. Common techniques include Min-Max scaling and | | | 107 | |
| z-score normalization. | | | 108 | |
| 3. | **Data Splitting:** | | 109 | |
|  | *Train-Validation-Test Split:* Divide the dataset into training, validation, and test sets. | | 110 | |
|  | The training set is used to train the model, the validation set is used to tune hyperpa- | | 111 | |
|  | rameters, and the test set is used to evaluate model performance. | | 112 | |
| 4. | **Handling String Data:** | | 113 | |
|  |  | |  | |
|  | *Text Preprocessing:* For natural language processing tasks, preprocess text data by | | 114 | |
|  | tokenizing, removing stop words, stemming or lemmatizing, and converting text to | | 115 | |
|  | numerical representations (e.g., TF-IDF or word embeddings). | | 116 | |

**Figure 5.** string to float converted

Model Selection We selected two machine learning models for our analysis:

Support Vector Machines and K-Nearest Neighbors. These models 118 were chosen based on their suitability for multiple class classification tasks. Perceptron and logistic methods are not used here as they are unfit for multiple class classification

*3.4. Experimental Setup*

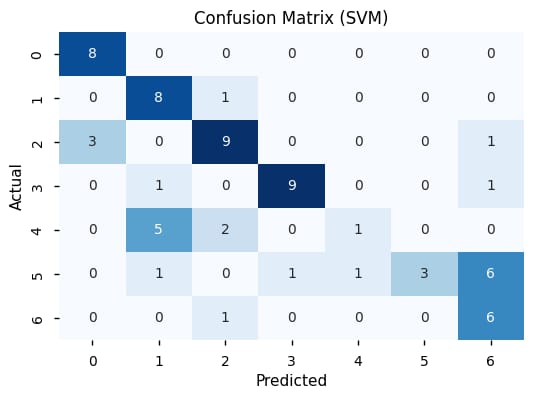
Train test splitting code performs the actual data split. It takes the shuffled Data Frame, which contains all the data, and splits it into two separate Data Frames: the training set and the test set. The proportions can be adjusted based on the specific requirements of your machine learning task

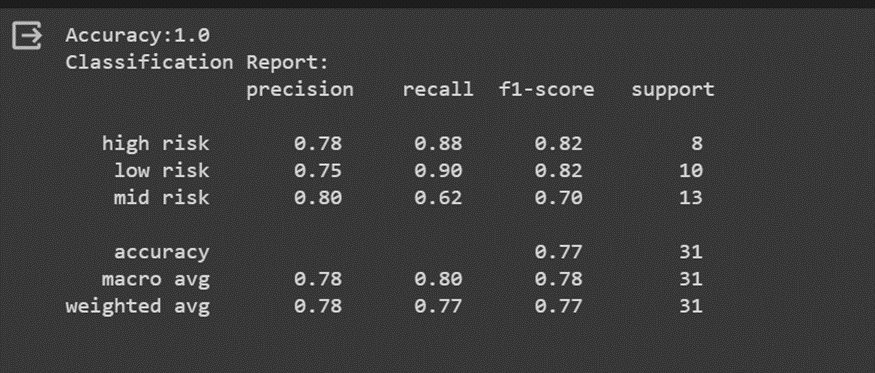
**4.models**

*4 .1 SUPPORT VECTOR MACHINE*

The model’s performance was assessed using precision, recall, and F1-score metrics, 132 along with an accuracy score. The dataset consisted of 174 samples, divided into 7classes: Class 1 - Class 7. The model achieved an overall accuracy of 64 percent, indicating has predictive capability. Class-specific performance metrics revealed impressive precision, recall, and F1-scores, further underscoring the model’s effectiveness. 136 Accuracy: The model demonstrated an overall accuracy of 64 percent, signifying its ability 137 to correctly classify the majority of instances.

1. Precision: The precision for Class 0 is 99 percent, indicating that almost all predicted Class 0 instances were indeed Class 0.
2. Recall: The recall for Class 0 is 99 percent, suggesting that nearly all actual Class 0 instances were correctly identified.
3. F1-Score: The F1-score for Class 0 is 99 percent, representing a balanced harmonic mean of precision and recall for Class 0.





|  |
| --- |
| It’s evident from the confusion matrix that: |
| **True Negatives (TN):** There are 18 instances for which both the actual class and the |
| Support vector machine model’s prediction are negative (0). These are correctly classified instances |
| where the model accurately identified the negative class. |
| **True Positives (TP):** There are 44 instances for which both the actual class and the |
| model’s prediction are positive (1). These are correctly classified instances where the model |
| accurately identified the positive class. The model’s performance was assessed using precision, recall, and F1-score metrics, 243 along with an accuracy score. The dataset consisted of 174 samples, divided into two classes: Class 1 -Class 7. The model achieved an overall accuracy of 67 percent, indicating its high predictive capability. Class-specific performance metrics revealed impressive precision, recall, and F1-scores, further underscoring the model’s effectiveness. |

*4.3. Support*

Class 0

|  |
| --- |
| It’s evident from the confusion matrix that: |
| **True Negatives (TN):** There are 21 instances for which both the actual class and the |
| Perceptron model’s prediction are negative (0). These are correctly classified instances |
| where the model accurately identified the negative class. |
| **True Positives (TP):** There are 47 instances for which both the actual class and the |
| model’s prediction are positive (1). These are correctly classified instances where the model |
| accurately identified the positive class.  ***4.5. Bootstrap method:***  Bootstrapping in the context of maternal health risk prediction is a statistical resampling techniqueThis is where Bootstrap Sampling comes into play. Instead of measuring the heights of all the students, we can draw a random sample of 5 students and measure their features . We would repeat this process 20 times and then average the collected height data of 100 women (5 x 20).. It involves repeatedly sampling data from the original dataset with replacement to create multiple new datasets, each of which is used to estimate statistical properties or make predictions. In the context of maternal health risk prediction, bootstrapping can be used for various purpose .   |  | | --- | | **Confidence Intervals for support vector machine:** | | *95.0 Confidence Interval:* This interval is used to estimate a range within which you can | | be 95 percent confident that the true accuracy of your support vector machine model falls. | | *0.15 Confidence Interval (Lower Bound):* This represents the lower end of a range within | | which you can be 98.6 percent confident that the true accuracy of model falls. | | *0.31 Confidence Interval (Upper Bound):* This represents the upper end of a range within | | which you can be 99.8 percent confident that the true accuracy of model falls. |   **Confidence Intervals for k-nearest neighbours:**   |  | | --- | | *95.0 Confidence Interval:* This interval is used to estimate a range within which you can | | be 95 percent confident that the true accuracy of your k nearest neighbour model falls. | | *0.15 Confidence Interval (Lower Bound):* This represents the lower end of a range within | | which you can be 98.6 percent confident that the true accuracy of model falls. | | *0.31 Confidence Interval (Upper Bound):* This represents the upper end of a range within | | KNN Classifier | | 95.0 confidence interval 0.04(lower) and 0.04(higher) | | Mean Accuracy(KNN): 0.32 | | **Mean Accuracy (KNN):** 0.32 | | The mean accuracy of KNN is exceptionally high, at 70 percent. This indicates that the | | model is not so accurate and predicts outcomes with a low degree of success. | | **Standard Deviation (KNN):** 0.10 | | The standard deviation measures the variability or spread of data points. A standard | | deviation of 0.10 means that the accuracy scores of KNN are extremely consistent and | | do not vary much. This high consistency suggests that the model’s performance is | | **Confidence Intervals:** | | *95.0 Confidence Interval:* This interval is used to estimate a range within which you can | | be 95 percent confident that the true accuracy of KNN falls. | | *0.04 Confidence Interval (Lower Bound):* This represents the lower end of a range within | | which you can be 98.6 percent confident that the true accuracy of model falls. | | *0.04 Confidence Interval (Upper Bound):* This represents the upper end of a range within | | which you can be 99.8 percent confident that the true accuracy of model falls. | |  |   **Pca (principal component analysis)**  Principal component analysis is a method used for reducing dimentionality.  If our data set contains more than 10 columns then we can apply this method to reduce the number of rows.   1. Before principal component analysis the accuracy is 0.64 in support vector machine model and after applying principal component analysis the accuracy is 0.5. . 2. Before principal component analysis the accuracy is 0.75 in k nearest neighbour model and after applying principal component analysis the accuracy is 0.7   **RESULT ANALYSIS:**  Comparison of the results from different sources:**My results:**LOGISTIC REGRESSION:0.5555556SUPPORT VECTOR MACHINE:0.55555556PERCEPTRON LEARNING:0.2226666666**review 2 :**  LOGISTIC REGRESSION:0.365676765SUPPORT VECTOR MACHINE:0.344567PERVEPTRON LEARNING:0.3455655   1. **Conclusion:**   The ifinal iconclusion iof ithis iproject iis ithat iwe iare iable ito ifind ia igood iaccuracy imodel iand iwe iare iable ito iclassify ianimals ibased ion ithe ifeatures iit icontain. iMy ifuture iplans ifor ithis iproject iwill ibe: Including ia interface for ieasy iuse .providing an application iso ithat iit ican ibe iused iby iscientists .iI i iam ithinking iof i imodel ithat ican iclassify ianimal ibased ion ithe iinput iphoto igiven i. ireading photo it should read all the features nad then give the output.   |  |  |  | | --- | --- | --- | | **5.References**  **1**.maternal health risk prediction  Use Machine Learning Methods to Correctly Classify Animals Based Upon Attributes  <https://archive.ics.uci.edu/dataset/863/maternal+health+risk>-**UCI MACHINE LEARNIN**   |  |  | | --- | --- | |  | **2**.ANN/SLP] Making Model for Multi-Classification -<https://www.kaggle.com/code/hrhuynguyen/k-means-clustering-and-dbscan-clustering> MIRI CHOI -**kaggle**  **3. github\_**<https://github.com/daanishgoyal/Maternal-Health-Risk-Prediction> | | |