Machine Learning

Name: Dimitrios

Surname: Koutrintzes

AM: ---

Mail: dkoutrintzes@gmail.com

1. The dataset

The dataset contains data from women during pregnancy and it targets to classify the risk level that they have during that period. It contains the following 6 arguments for each patient.

- Age: Age in years when a woman is pregnant.
- SystolicBP: Upper value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- DiastolicBP: The lower value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- BS: Blood glucose levels are in terms of a molar concentration, mmol/L.
- HeartRate: A normal resting heart rate in beats per minute.
- Risk Level: Predicted Risk Intensity Level during pregnancy considering the previous attribute.

Data has been collected from different hospitals, community clinics, maternal health cares through the IoT-based risk monitoring system. The dataset includes 1014 registries.

For the experiments, we break the dataset in three categories. First, we have 811 registries for the train validation, 198 registries for the evaluation, and 5 registries for examples at the end.

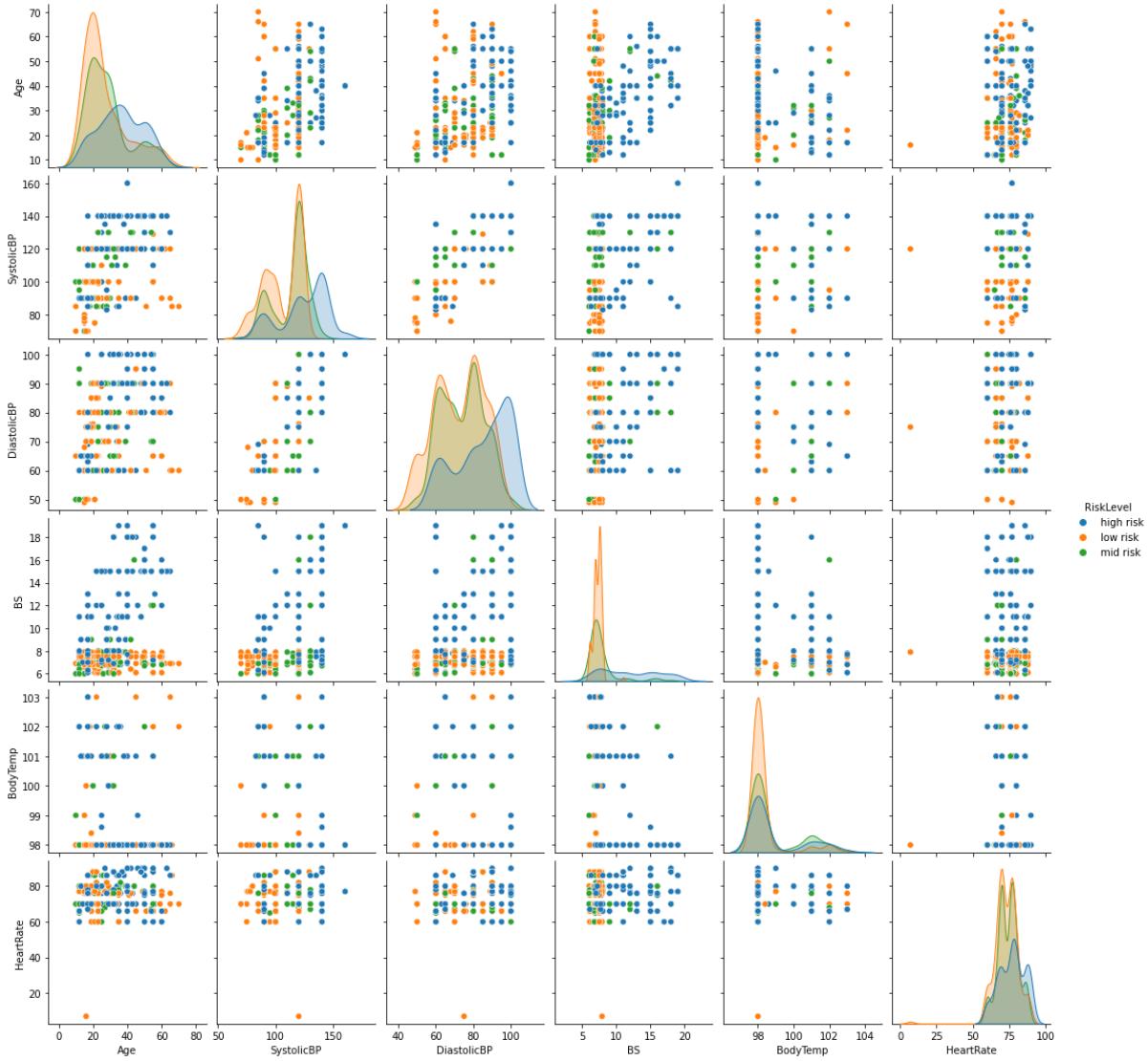
The kaggle link for the data is:
https://www.kaggle.com/bjoernjostein/predicting-health-risks-for-pregnant-patients

2. Classifier

For classifier will be using the RandomForest from the python library sklearn. We select this classifier because it returns the most persistent and most of the time better results.
In the training process, we test the classifier with tree sizes between 2 - 24. We select the best classifier at the end.
The random forest classifier operates by contrasting a multitude of decision trees at training time and at a classification task the prediction is the answer of the most trees.

3. Method

As we can see on the distributions of the data below its clear that the category high risk is more easily separated from the other two categories. In this method will review a two-step architecture, at first we will separate the high risk category from the low and mid risk and then a second classifier will separate the last to categories. Both Classifiers are based on the previous section.



* The image was taken from a submitted method in the original Kaggle page
https://www.kaggle.com/bjoernjostein/predicting-health-risks-for-pregnant-patients

The training process will be repeated 20 times to produce an average performance report and then we will be using the best model for the examples.

4. Evaluation

For the evaluation we will use the Test data that we separate earlier that the model haven't seen yet. Firstly as we see below we have an average accuracy score of 0.80 and a best of 0.84. The precision and recall are the same as the accuracy.

Average Accuracy	0.803
Average Precision	0.803
Average Recall	0.803

The Best Accuracy	0.843
The Best Precision	0.843
The Best Recall	0.843

Examining the classification report of the best model we can see that the high risk perform very well and second comes the low-risk category with a good precision but a lower recall, the mid category has a similarly recall but low precision. In general, we see good results on the high-risk category which is the most important.

	precision	recall	f1-score	support
high risk	0.92	0.91	0.91	53
low risk	0.90	0.83	0.86	87
mid risk	0.71	0.81	0.76	58
accuracy			0.84	198
macro avg	0.85	0.85	0.84	198
weighted avg	0.85	0.84	0.85	198

5. Conclusion

As we can see immediately if we run the 5 examples we kept in the beginning, one of them, which should be in the low-risk category failed and was predicted as mid-risk. This method provides a valid way to separate the classification process into two problems and leaves open the ability to even use different models and preprocessing methods for each. In future expansion, we could use a heavier architecture in the separation of low mid-risk by utilizing a neuron network to maximize the results between those categories.