FETAL HEALTH

A Machine Learning model to classify the outcome of Cardiotocogram test to detect potential health issues.

DATASET AND OBJECTIVE

2126 records of features extracted from cardiotocographic (**CTG**) examinations, providing important data on the condition of fetuses and mothers. Each of these records has been classified by expert obstetricians based on the health of the fetuses.

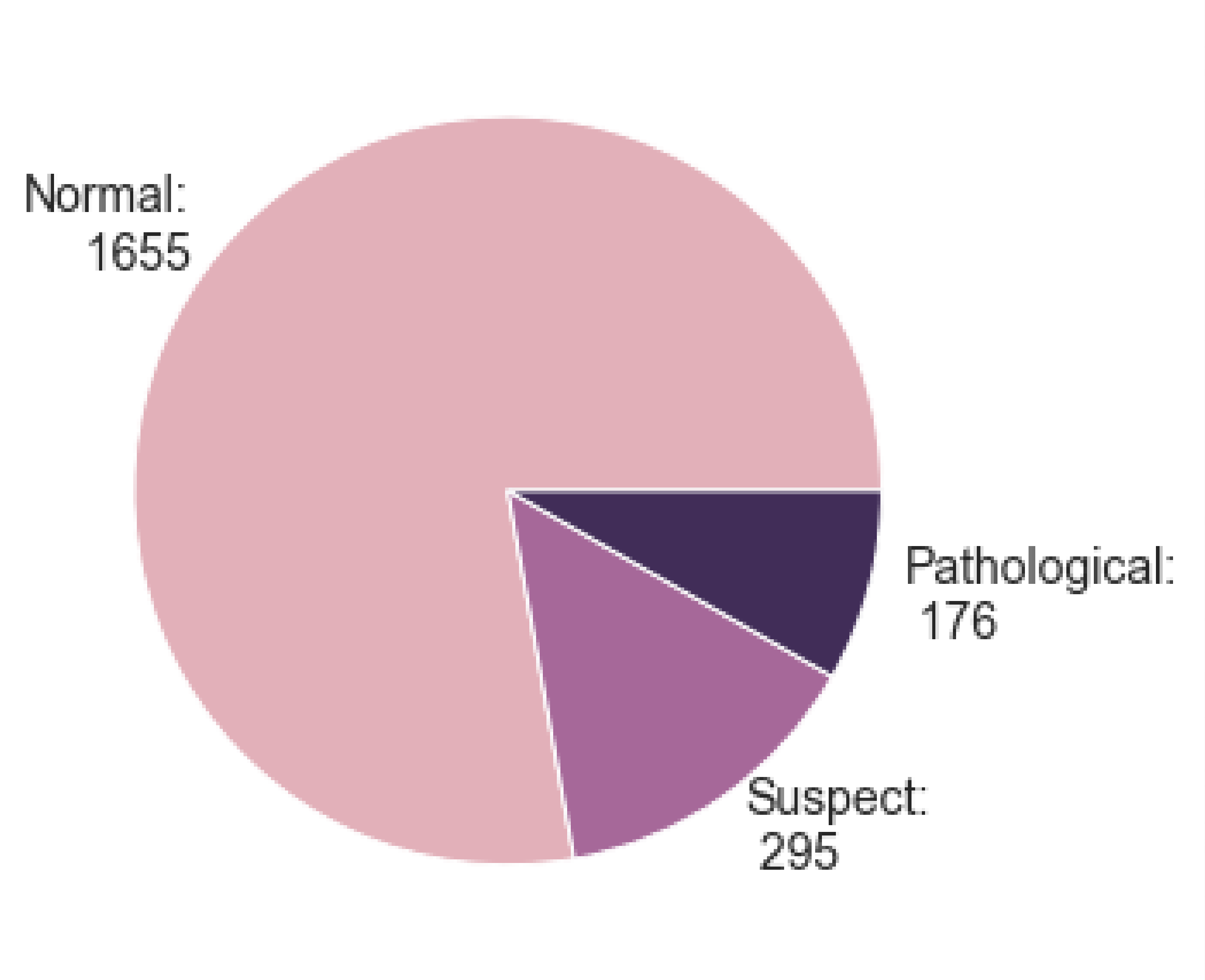
21 numerical features: provide information about heart rate, its short-term and long-term variability, fetal movements, and uterine contractions.
The last value is the **health status** of the fetus, classified into three ordered categories:

1. *'Normal'*
2. 'Suspect'
3. 'Pathological'

The goal is building a machine learning model able to predict the health status of a fetus by assigning it to one of three categories based on data provided by the CTG test.
This is a **classification problem** that would allow preventive action to reduce infant mortality based on a quick and non-invasive test.

DATA ANALYSIS
TARGET DISTRIBUTION

Fetal health class distribution



The classes are quite **unbalanced**.

- Approximately 78% of the dataset is populated by the *'Normal'* class.
- The remaining part is populated for nearly two-thirds by the *'Suspect'* class, which makes up 14% of the dataset.
- The *'Pathological'* class is a minority, its population is just over 1/10 of the 'Normal' class, around 8% of the total.

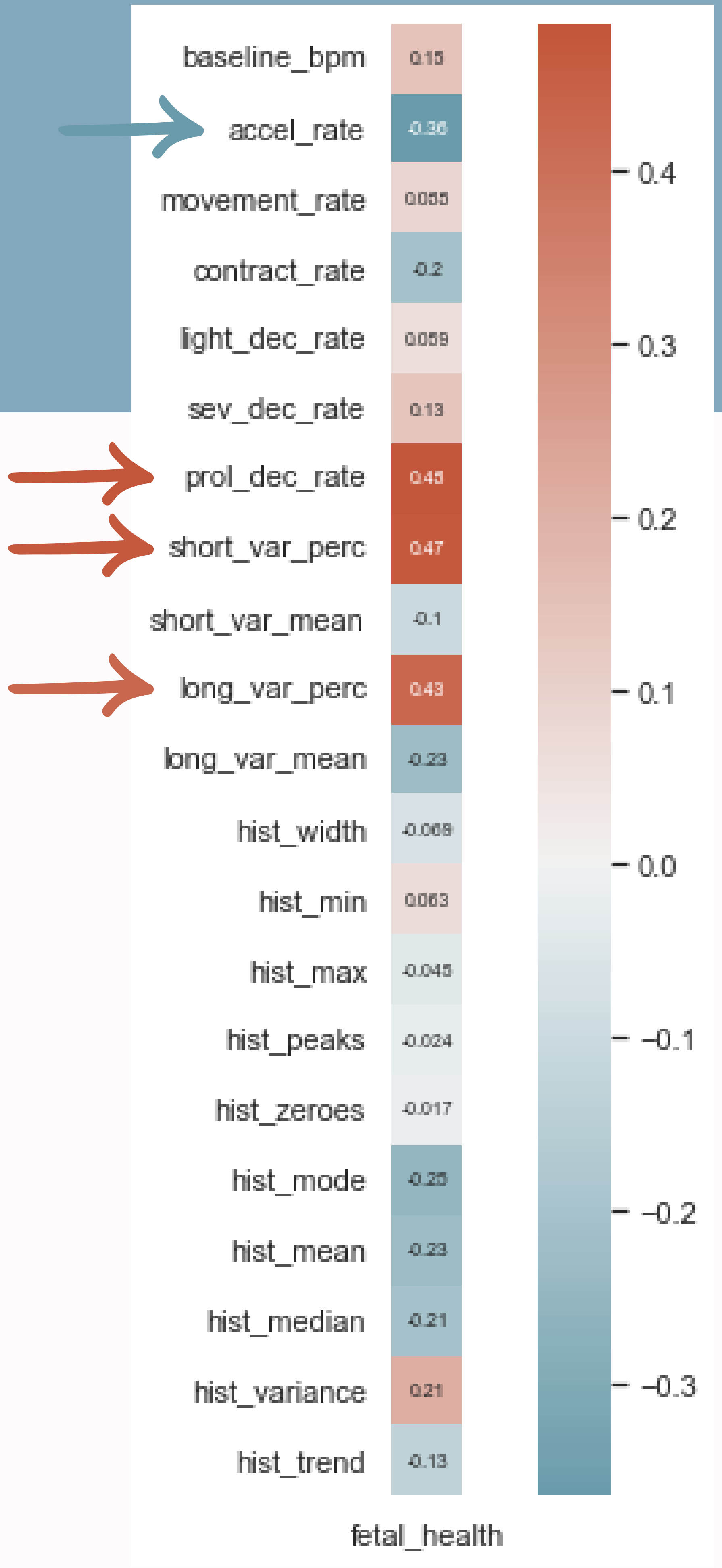
DATA ANALYSIS

CORRELATIONS

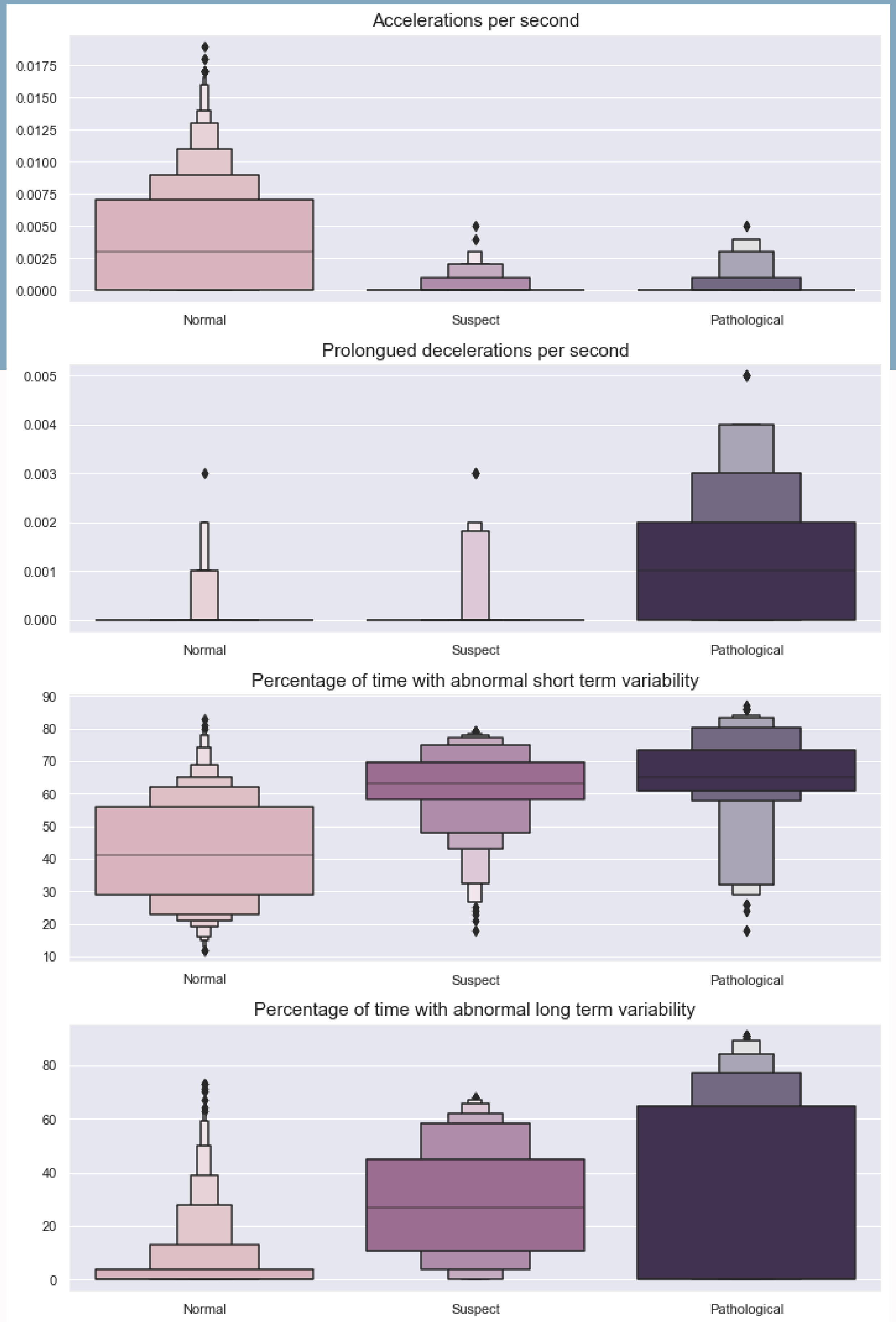
Correlations between *'fetal_health'* and other features.

There are no particularly strong correlations, but it's still worth to analyze in more detail the **most correlated features**.

- *'accel_rate'*: number of heaertbeat accelerations per second.
- *'prol_dec_rate'*: number of prolonged heartbeat decelerations per second.
- *short_var_perc'*: percentage of time with abnormal short-term heart rate variability.
- *'long_var_perc'*: percentage of time with abnormal long-term heart rate variability.



DATA ANALYSIS
FEATURE DISTRIBUTION



Boxplots of the distribution for the selected features.

- The distribution of accelerations per second for the suspected and pathological cases is very 'compressed' around 0.

- Pathological cases show greater variability in the rate of prolonged decelerations.

- The two percentages of time with short-term and long-term variability increase on average in a quite distinguishable way when moving from the 'Normal' class to the other ones, consistently with what is described by the correlation.

OUTLIERS AND METRICS

The charts reveal the presence of some **outliers**. These data **were retained** due to the nature of this analysis.
Our main purpose is to preemptively identify pregnancy cases that could compromise the health of the fetuses and mothers.

It is reasonable to assume, as a precaution, that anomalies in one or more of the features detected by the CTG exam may be indicative of such problems.
Consequently, the information provided by the records presenting these anomalies is potentially **valuable** for building a model aimed at correctly classifying as many at-risk cases as possible.

The same considerations guided the choice of the **model evaluation metric**: ***'recall'*** or *'sensitivity'*.

This metric allows for minimizing the number of cases belonging to a specific class but not classified as such: the **false negatives**.
As the number of false negatives decreases, recall increases, making it a suitable metric for our purposes.

MODEL SELECTION

I choose four classifiers to evaluate: **KNN**, **Decision Tree**, **Random Forest**, **XGBoost**.

The classifiers were trained on a training set with **80%** of the available data using **cross-validation**.

Classifier	***Recall***
KNN	0.7526941703273048
Decision Tree	0.880417505108254
Random Forest	0.8805101122662273
XGBoost	0.9084930261555423

The most performant classifier in terms of recall is **XGBoost**, a model that works by having decision trees operate in series, where each tree improves upon the result of the previous one. It will be the final classifier for the prediction.

PARAMETERS TUNING

Chosen a classifier, I set up a **Grid Search** on some parameters to improve its performance.
Grid Search works by setting a series of values for each parameter that you want to evaluate.
The model is trained and evaluated (with the chosen metric) for every combination of the values and the best combination is selected.

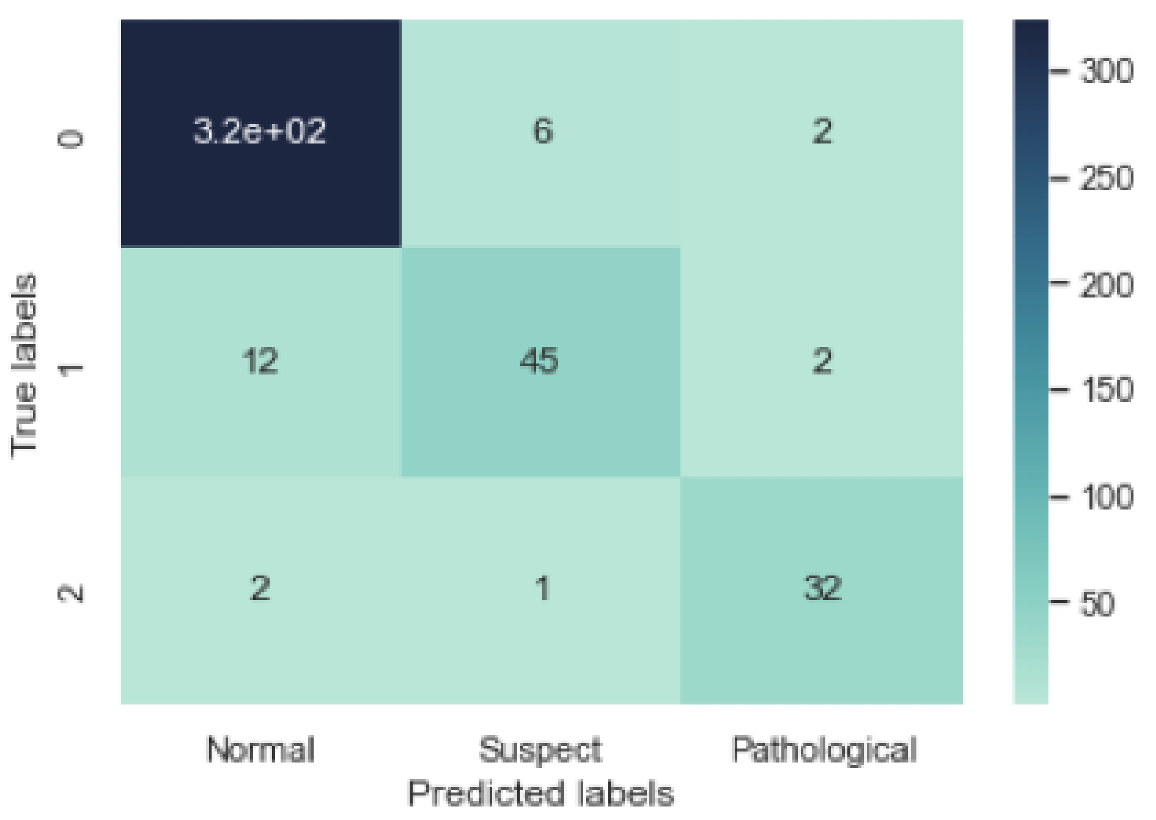
XGBoost parameters selected for the Grid Search:

- *'min_child_weight'*: minimum number of elements in a class, no new trees are generated below this number.
- *'max_depth'*: maximum depth of decision trees.
- *'subsample'*: fraction of the training set, randomly selected at each step, used to train the trees (useful to prevent overfitting).
- *'eta'*: magnitude of the 'correction' that each tree applies to the result of the previous one.

New *recall* score: 0.9130471352339388

FINAL PREDICTION

After building the final XGBoost classifier with the parameters obtained from the Grid Search and training it on the entire training set, it is finally possible to predict the classes of the records in the **test set** and compare the results with the actual classes.



	precision	recall	f1-score	support
0	0.9586	0.9759	0.9672	332
1	0.8654	0.7627	0.8108	59
2	0.8889	0.9143	0.9014	35
accuracy			0.9413	426
macro avg	0.9043	0.8843	0.8931	426
weighted avg	0.9399	0.9413	0.9401	426

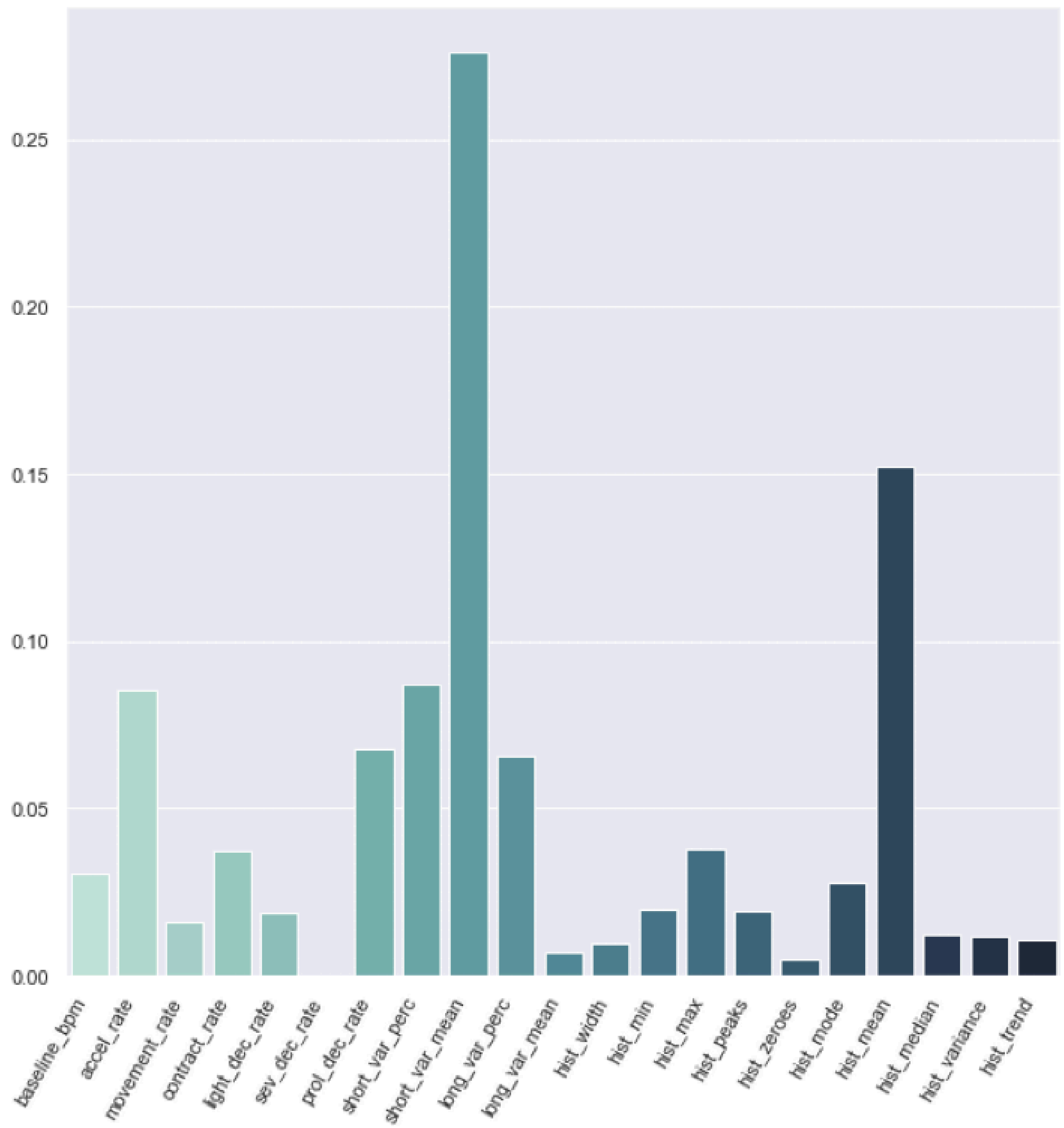
The final prediction *'recall'* seems sufficiently good.
However, analyzing the recall score for individual classes, it is evident that the recall for the second class (*'Suspect'*) is significantly lower than the other classes.
From the **confusion matrix**, it can be seen that a significant percentage of the elements from this class were classified as *'Normal'*, while only two of them were classified as *'Pathological'*.

FEATURE IMPORTANCE

The four features initially selected based on correlation are, as expected, among the most important in the prediction made by XGBoost.

However, the two features clearly most decisive are:

- *'short_var_mean'*, the average value of short-term irregularities.
- *'hist_mean'*, the average value of the histogram of the heart rate frequency distribution.



Kaggle dataset:
https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification?resource=download

Dataset authors:
Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318