FETAL HEALTH CLASSIFICATION

(Portfolio addition for module 3 & Bachelor thesis optimization)

BY: NOOR KHARISMAWAN AKBAR

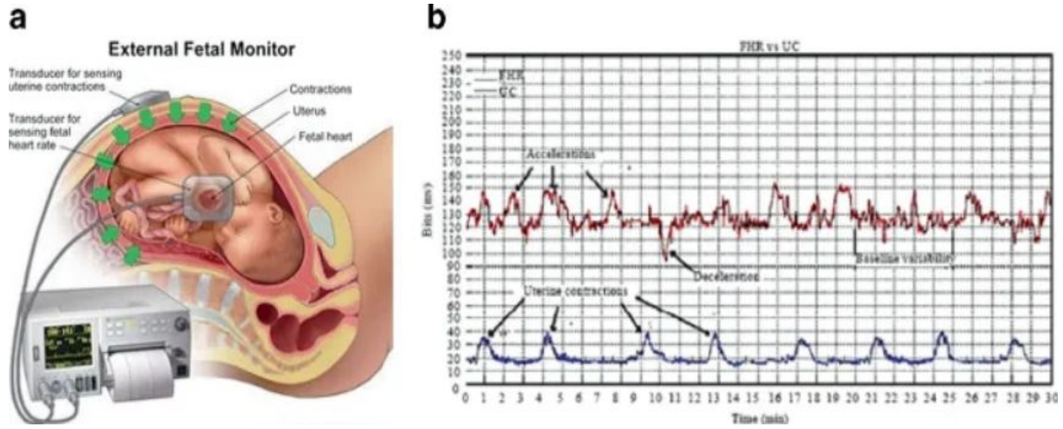
Purwadhika
Digital Technology School

PROBLEM UNDERSTANDING

BACKGROUND

The UN expects that by 2030, countries **end preventable deaths of newborns** and children under 5 years of age. Parallel to notion of child mortality is of course **maternal mortality**, which accounts for 295,000 deaths during and following pregnancy and childbirth (as of 2017).

Cardiotocograms (CTGs) are a simple and cost accessible option to assess fetal health, allowing healthcare professionals to take action in order to **prevent child and maternal mortality**.



(**a**) External Fetal Monitor, (**b**) CTGs Data

In this notebook we will try to **analyze the CTGs data to solve the above issues**.

PROBLEM UNDERSTANDING

PROBLEM STATEMENT

Using the available data, An appropriate model is needed to built to **determine fetal health** & knowing the **most important factor**.

GOAL & OBJECTIVE

The aim of this project is **to predict and classify the health of the fetus with the best accuracy possible**. Hopefully, with this model, it can prevent child and maternal deaths.

ANALYTICS APPROACH

Machine learning model will be built because we need to make predictions more than just using an inferential and/or descriptive analysis. The existing data has labels so we use a **supervised machine learning**. **Classification** model will be used because we need the model that can predict the correct label of a given input data.

MODEL EVALUATION

Accuracy R^2 (Coefficient of determination)

DATASET GENERAL INFO

Dataset Link:
https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

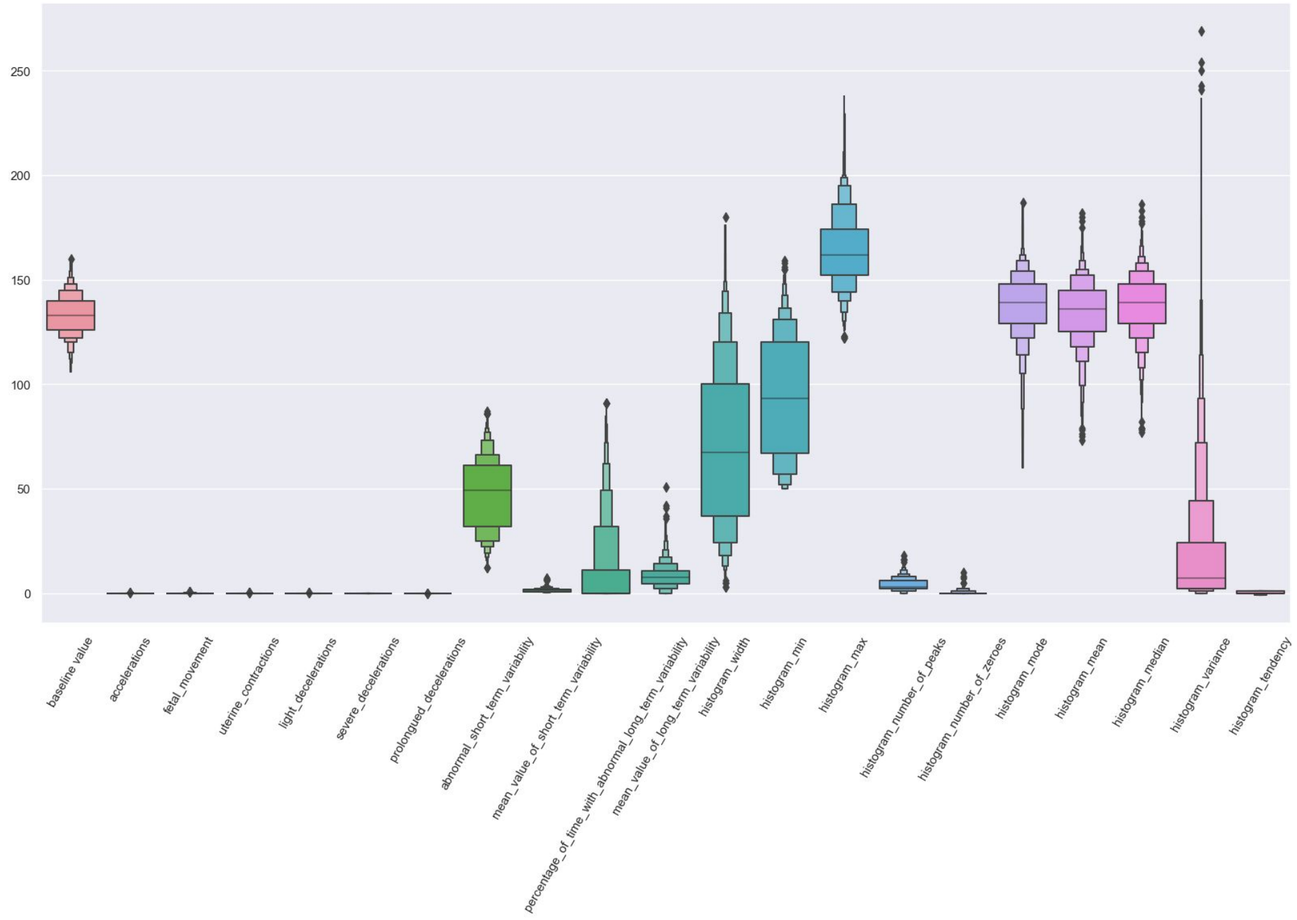
Rows	Columns	Data Type
2126	22	float64

Values	Sum
n.a.	0
null	0

Column Name	Description
baseline value	Baseline Fetal Heart Rate (FHR)
accelerations	Number of accelerations per second
fetal_movement	Number of fetal movements per second
uterine_contractions	Number of uterine contractions per second
light_decelerations	Number of LDs per second
severe_decelerations	Number of SDs per second
prolongued_decelerations	Number of PDs per second
abnormal_short_term_variability	Percentage of time with abnormal short term variability
mean_value_of_short_term_variability	Mean value of short term variability
percentage_of_time_with_abnormal_long_term_variability	Percentage of time with abnormal long term variability
mean_value_of_long_term_variability	Mean value of long term variability
fetal_health (TARGET)	Fetal health: 1 - Normal 2 - Suspect 3 - Pathological

*There are 10 columns again that explain the histogram of the dataset like width, min, max, num of peaks, num of 0, mode, mean, median, variance, and tendency.

DATASET INSIGHT



Column Histogram Plot

- This dataset has no missing value found.
- Some columns have right skewness data.

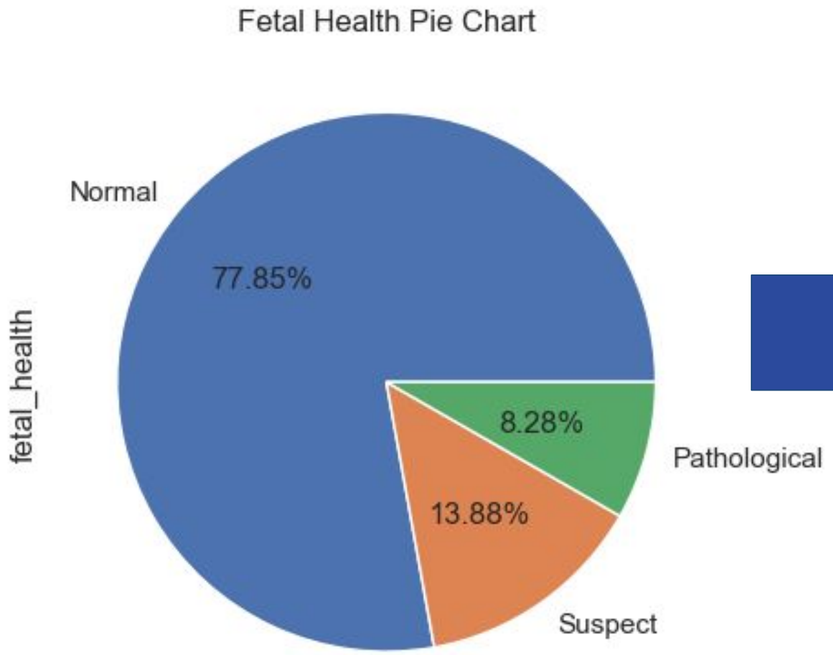
Column Box Plot

- Outlier Detection
- This dataset is quite clean with no extreme & unreasonable outlier.

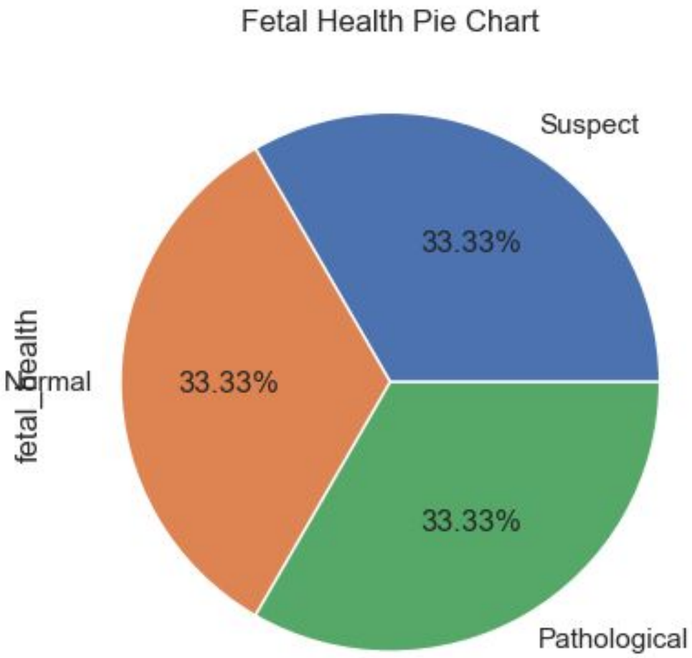
DATA CLEANING

Imbalance data handling: **Random Oversampler.**

Consideration: Data size is **not large enough (~2000)** & to maintain the dataset so that **no information is lost**.



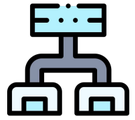
RandomOverSampler



fetal_health	Type	Qty
1.0	Normal	1,655
2.0	Suspect	295
3.0	Pathological	176

fetal_health	Type	Qty
1.0	Normal	1,655
2.0	Suspect	1,655
3.0	Pathological	1,655

MACHINE LEARNING MODELLING



TRAIN-TEST SPLIT

We will use a ratio of **70:30**. This refers to the reference from **baeldung.com** for a dataset of size **n<10,000**.

fetal_health	Train	Test
Ratio	70%	30%
Shape	(3,475; 21)	(1,490; 21)

NORMALIZATION

Variables that are measured at different scales do not contribute equally to the model fitting & model learned function and might end up creating a bias. So we make all data **standardized** using **StandardScaler** (μ=0, σ=1).

CROSS VALIDATION

Cross-Validation is a very useful technique to **assess the effectiveness of a machine learning model**, particularly in cases where you need to **mitigate overfitting**.

Parameter	Values
Sampling	StratifiedKFold
N_splits	5
N_jobs	2

HYPERPARAMETER TUNING RESULT

Models	Testing Score Before Tuning	Testing Score After Tuning	Diff
Support Vector Machine	**91.75%**	**97.25%**	**5.50%**
K-nearest neighbors	**94.56%**	**98.19%**	**3.62%**
Gradient Boosting classifier	**95.77%**	**97.58%**	**1.81%**
Random Forest	**98.05%**	**98.12%**	**0.07%**
Logistic Regression	85.97%	85.37%	-0.60%

- Hyperparameter tuning was successfully performed on 4 models, namely **Support Vector Machine, K-nearest neighbors, and Gradient Boosting Classifier, Random Forest**.
- Where that models experienced an increase in testing scores of **0-6%**.
- For the **Logistic Regression** model, the Testing Score has decreased. This can be caused because the default parameters still perform better than the parameters in the state for tuning.

MODELLING RESULT

Models	Training Score	Testing Score (After Tuning)	Error
K-nearest neighbors	**99.97%**	**98.19%**	**1.78%**
Random Forest	99.97%	98.12%	1.85%
Gradient Boosting classifier	99.97%	97.58%	2.39%
Support Vector Machine	99.97%	97.25%	2.72%
Logistic Regression	87.37%	85.37%	2.00%

- There are 4 models that have the largest training scores, namely: **K-nearest neighbors, Random Forest, and Gradient Boosting classifier, Support Vector Machine** with a value of **99.97%**.
- Of the three models, the K-nearest neighbors model has the largest testing score with a value of **98.19%** which produces the smallest error with a value of **1.78%**.
- This shows that the K-nearest neighbors is the **most suitable model** to predict this fetal_health.csv dataset.

BEST MODEL: KNN

1 MACHINE LEARNING MODEL

K-nearest Neighbors (KNN)

2 HYPERPARAMETER TUNING

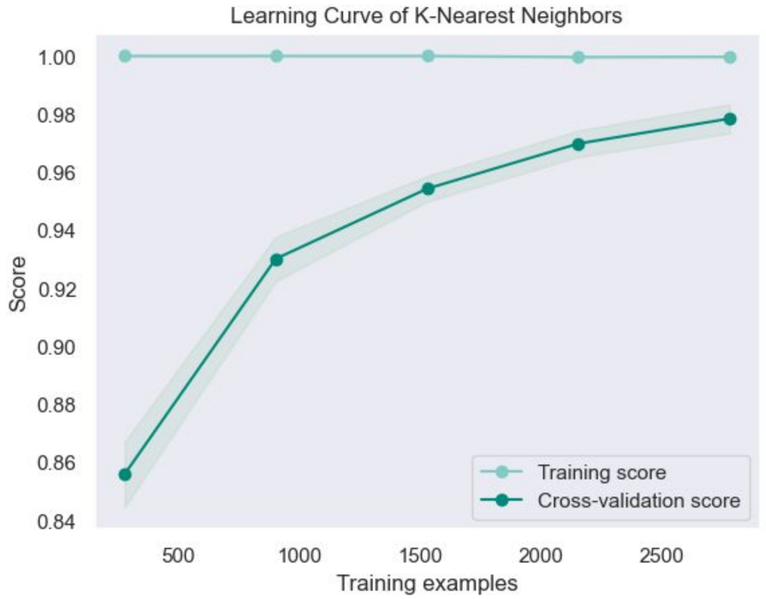
- Test data (before): 94.56%
- Test data (after): 98.19%
- Differences: 3.62%

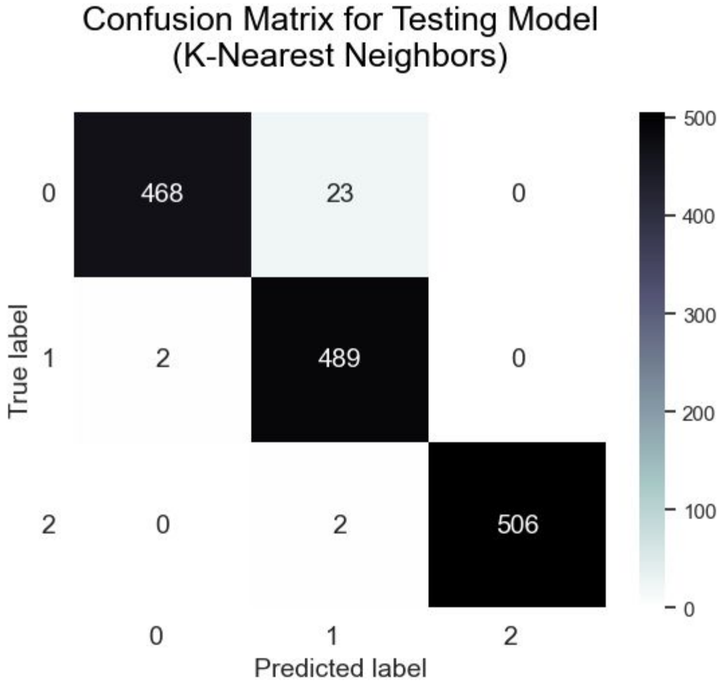
3 BEST PARAMETER TUNING

- **'leaf_size'**: [1, 2, 3, …, 27, 28, 29],
- **'n_neighbors'**: [1, 2, …, 10, 11, 12],
- **'p'**: [1, 2]

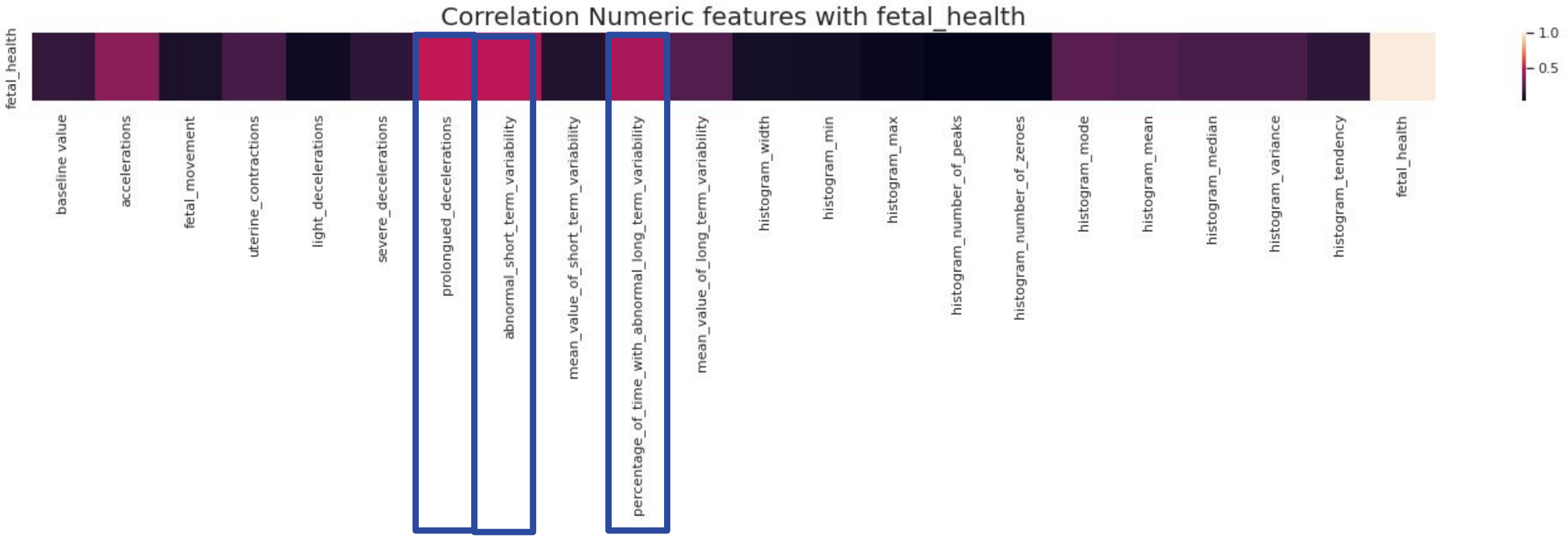
4 MODELLING RESULT

- Train data: 99.97%
- Test data: 98.19%
- Error margin: 1.78%





MOST CORRELATED FACTOR



	fetal_health
fetal_health	1.000000
prolongued_decelerations	0.484859
abnormal_short_term_variability	0.471191
percentage_of_time_with_abnormal_long_term_variability	0.426146
accelerations	0.364066
histogram_mode	0.250412
histogram_mean	0.226985
mean_value_of_long_term_variability	0.226797
histogram_variance	0.206630
histogram_median	0.205033
uterine_contractions	0.204894

Top 3 columns that has highest correlation with fetal_health:

- prolongued_decelerations
- abnormal_short_term_variability
- percentage_of_time_with_abnormal_long_term_variability

CONCLUSION

ABOUT DATASET

- This dataset is quite clean with no extreme & unreasonable outlier.
- This dataset also has no missing value found.
- But for the label (fetal_health), there found imbalance classification, where the dataset contain **77.85%** of **normal fetus label (1.0)**. This problem solved by do the oversampling.

BEST MODEL

K-nearest neighbors, with:

- Accuracy R^2 Train data: **99.97%**
- Accuracy R^2 Score Test data: **98.19%**
- Error margin: **1.78%**

MOST CORRELATED FACTOR

- Abnormal Fetal Heart Rate - FHR: **0.48**
- Abnormal Short Term Variability: **0.47**
- Abnormal Long Term Variability: **0.43**

RECOMMENDATION

- Building a model using **feature selection** to select column that only have high correlation with target (fetal_health).
- Hyperparameter tuning optimization can use **Optuna** to improve model performance

REFERENCES

- https://en.wikipedia.org/wiki/Cardiotocography
- https://www.datacamp.com/blog/classification-machine-learning
- https://onlinelibrary.wiley.com/doi/10.1002/1520-6661(200009/10)9:5%3C311::AID-MFM12%3E3.0.CO;2-9
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8442730/
- https://www.baeldung.com/cs/train-test-datasets-ratio
- https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html
- https://journal.universitasbumigora.ac.id/index.php/matrik/article/view/2515

THANK YOU

Contact me!

+6281227223150

akbar.noorkharismawan@gmail.com

http://www.linkedin.com/in/n-k-akbar

https://github.com/baramizzo58

https://public.tableau.com/app/profile/akbar2070

Purwadhika
Digital Technology School