FETAL HEALTH CLASSIFICATION

(Bachelor Thesis optimization)

BY: NOOR KHARISMAWAN AKBAR

PROJECT BACKGROUND

PROBLEM STATEMENT

In this work, we use machine learning to **predict fetal health** to prevent maternal and child deaths.

GOAL & OBJECTIVE

Classify the health of a fetus as Normal, Suspect or Pathological using CTG data

DATASET

Fetal_health.csv
https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification/download?datasetVersionNumber=1

MODEL EVALUATION

R^2(coefficient of determination)

WORKING FLOW

1 EDA

-Visualization every column

-Visualization group by fetal_health

2 DATA CLEANING

-Detecting n.a. values

-Detecting null values

Values	Sum
n.a.	0
Null	0

3 MODEL BUILDING

-Handling imbalanced data

fetal_health	1.0	2.0	3.0
Before	1,655	295	176
After	1,655	1,655	1,655

-Train-Test Split

fetal_health	Train	Test
Ratio	80%	20%
Shape	(3972, 21)	(993, 21)

-Normalization -> Using StandardScaler

-Cross-Validation

Values	Sum
Sampling	StratifiedKFold
N_splits	3
N_jobs	2

4 MACHINE LEARNING MODEL

- Gradient Boosting Machine (GBM)
- K-nearest neighbors (KNN)
- Logistic Regression (LR)
- Random Forest (RF)
- Support Vector Machine (SVM)

5 SAVING MODEL

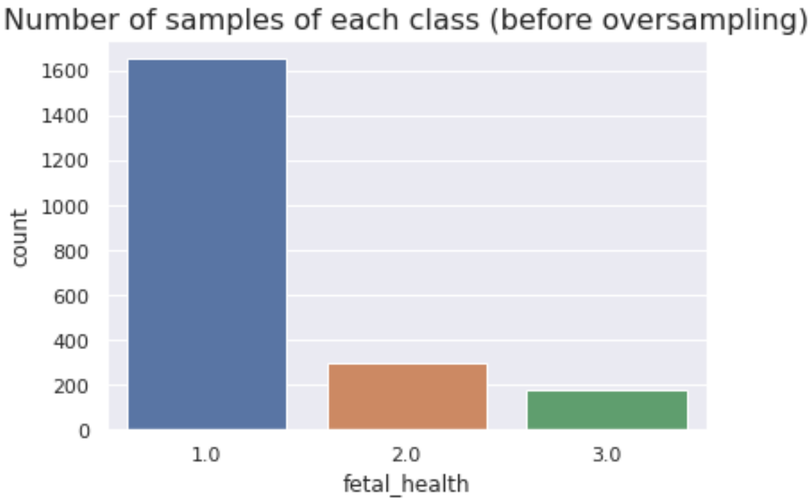
Using pickel

DATASET INSIGHT

Part 1

Rows	Columns	Data Type
2126	22	float64

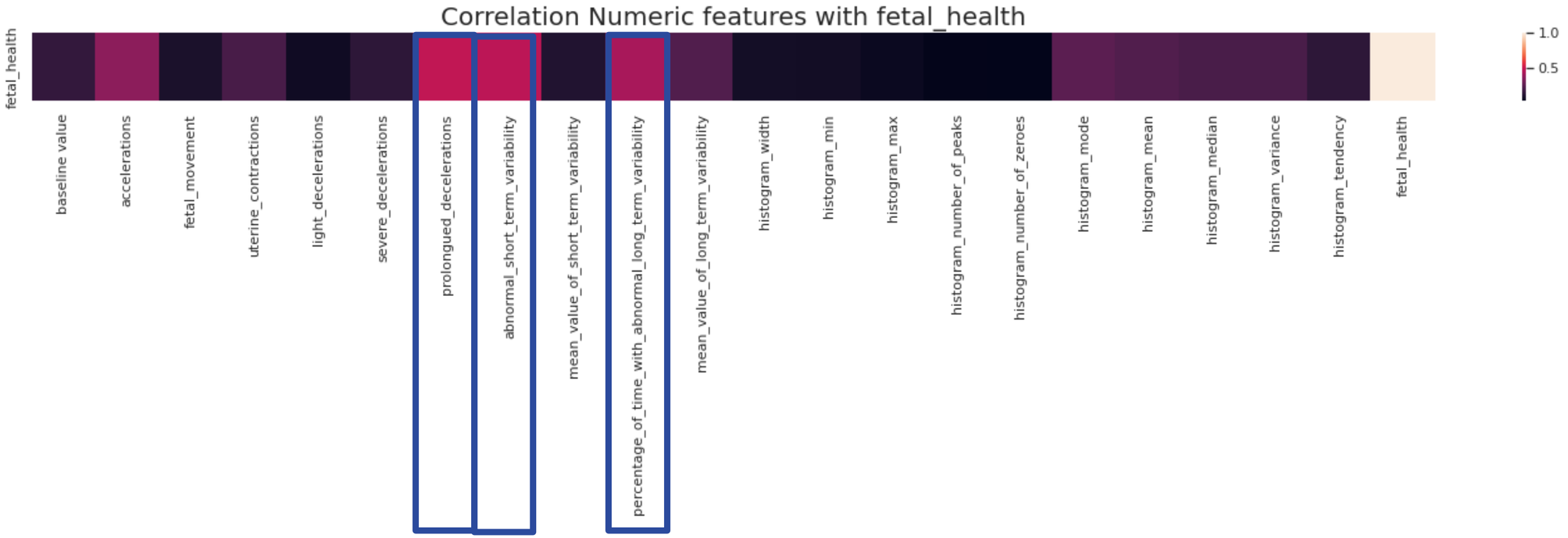
fetal_health	Sum
1.0	Normal
2.0	Suspect
3.0	Pathological



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

DATASET INSIGHT

Part 2



	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

MACHINE LEARNING MODELLING

Models	Training Score	Testing Score	Error
K-nearest neighbors	**99.95%**	**97.28%**	**2.67%**
Random Forest	99.95%	96.88%	3.07%
Gradient Boosting classifier	99.95%	95.97%	3.98%
Support Vector Machine	99.85%	98.59%	1.26%
Logistic Regression	85.88%	84.49%	1.38%

There are 3 models that have the largest training scores, namely: **K-nearest neighbors, Random Forest, and Gradient Boosting classifier** with a value of **99.95%**. Of the three models, the K-nearest neighbors model has the largest testing score with a value of **97.28%** which produces the smallest error with a value of **2.67%**.This shows that the K-nearest neighbors is the most suitable model to predict this fetal_health.csv dataset.

HYPERPARAMETER TUNING

Models	Testing Score Before Tuning	Testing Score Before Tuning	Diff
Support Vector Machine	**92.85%**	**98.59%**	**5.74%**
K-nearest neighbors	**94.36%**	**97.28%**	**2.92%**
Gradient Boosting classifier	**94.46%**	**95.97%**	**1.51%**
Random Forest	97.08%	96.88%	-0.02%
Logistic Regression	85.80%	84.49%	-1.30%

Hyperparameter tuning was successfully performed on 3 models, namely **Support Vector Machine, K-nearest neighbors, and Gradient Boosting Classifier**. where the three models experienced an increase in testing scores of **1-6%**. For the Random Forest & 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.

BEST MODEL: KNN

1 MACHINE LEARNING MODEL

K-nearest Neighbors (KNN)

2 PERFORMANCE ACCURACY

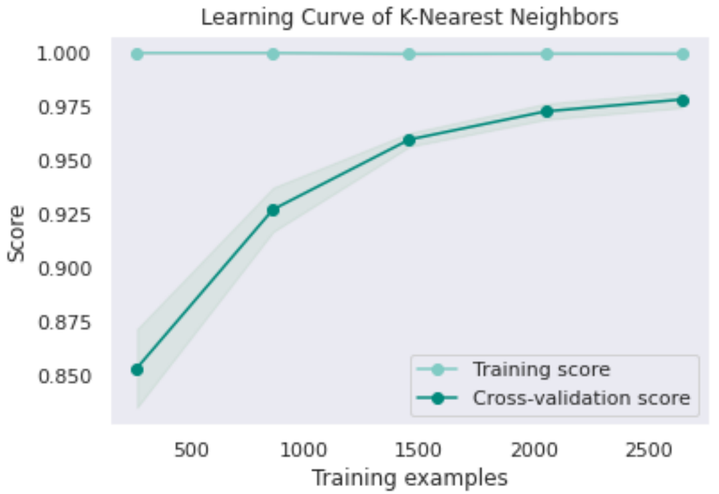
- Train data: 99.95%
- Test data: 97.28%
- Error margin: 2.67%

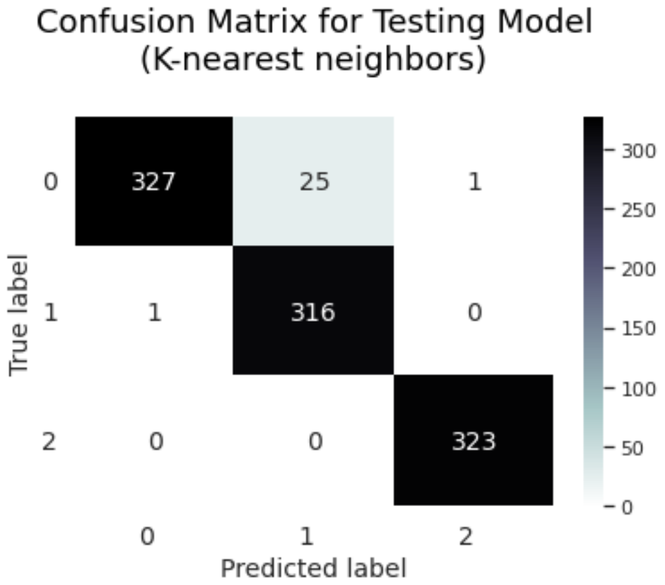
3 HYPERPARAMETER LIST

- **'leaf_size'**: [1, 2, 3, 4, 5, 6, 7, 8, 9],
- **'n_neighbors'**: [1, 2, 3, 4, 5, 6, 7, 8, 9],
- **'p'**: [1, 2]

4 TUNNING RESULT

- Test data (before): 94.36%
- Test data (after): 97.28%
- Differences: 2.92%





CONCLUSION

BEST MODEL

K-nearest neighbors, with:

- Train data: 99.95%
- Test data: 97.28%
- Error margin: 2.67%

HYPERPARAMETER TUNNING

Success tuning for **Support Vector Machine, K-nearest neighbors, and Gradient Boosting Classifier** with increase 1-6%

RECOMMENDATION

- Building a model using **feature selection** to select column that only have high correlation with target (fetal_health).
- Find the best parameter for **Random Forest** that can increase the accuracy of tuned score

THANK YOU