Data Witches

Name	Student ID
Claessen, VVHJAE	i6339543
Ovsiannikova, AM	i6365923
Pubben, J	i6276134
Roca Cugat, M	i6351071
Záboj, J	i6337952

Logbook

Changes Also see Git Commit History.

Version	Changes	Date
v0.0	Dataset loaded, EDA, cleaning	18-11-25
v1.1	***	XX-11-25
v0.2	***	XX-11-25

Methods

Let's ensure we all use the same names for all components.

Variable	Name
Raw ECG dataframe	df
Label dataframe	df_labels
HRV features (train)	hrv_train
HRV features (test)	hrv_test
HRV extraction type	FULL (nk.hrv — time + freq + nonlinear + RSA)
Clean HRV dataframe (train)	hrv_train_clean

Variable	Name
Clean HRV dataframe (test)	hrv_test_clean
HRV + labels (train)	hrv_train_with_labels
Winsorized HRV column	HRV_MedianNN_winsor
Model feature matrix (train)	X_train
Model feature matrix (test)	X_test
Model target vector (train)	y_train
Model target vector (test)	y_test

Function	Description	Arguments
corr_plot_hrv()	Correlation plot for HRV features	df, cols=None
distplots_hrv()	Distribution plots (hist + KDE)	df, cols=None
boxplots_hrv()	Boxplots for selected HRV variables	df, cols
check_missing_hrv()	Missingness summary	df
identify_outliers()	IQR-based outlier detection	df, column_name, threshold=1.5
model_evaluation()	Confusion matrix + classification report	model
model_desc()	Accuracy, CV, ROC-AUC, model performance	model

Preface

Packages imports

```
!wget https://raw.githubusercontent.com/MAI3003-Data-Witches/Data-Witches_Project2/re
```

```
--2025-11-23 15:14:25--  https://raw.githubusercontent.com/MAI3003-Data-Witches/Data-
Witches_Project2/refs/heads/main/requirements.txt
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 1
85.199.111.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.110.133|:
443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 95 [text/plain]
Saving to: 'requirements.txt'

requirements.txt    100%[====================>]      95  --.-KB/s    in 0s

2025-11-23 15:14:25 (6.33 MB/s) - 'requirements.txt' saved [95/95]


Requirement already satisfied: kaggle in /opt/conda/lib/python3.10/site-packages (fro
m -r requirements.txt (line 1)) (1.7.4.5)
Requirement already satisfied: plotly in /opt/conda/lib/python3.10/site-packages (fro
m -r requirements.txt (line 2)) (6.5.0)
Requirement already satisfied: matplotlib in /opt/conda/lib/python3.10/site-packages
(from -r requirements.txt (line 3)) (3.7.1)
Requirement already satisfied: numpy in /opt/conda/lib/python3.10/site-packages (from
-r requirements.txt (line 4)) (1.26.4)
Requirement already satisfied: seaborn in /opt/conda/lib/python3.10/site-packages (fr
om -r requirements.txt (line 5)) (0.12.2)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-packages (fro
m -r requirements.txt (line 6)) (2.0.2)
Requirement already satisfied: colorama in /opt/conda/lib/python3.10/site-packages (f
rom -r requirements.txt (line 7)) (0.4.6)
Requirement already satisfied: scikit-learn in /opt/conda/lib/python3.10/site-package
s (from -r requirements.txt (line 8)) (1.7.2)
Requirement already satisfied: shap in /opt/conda/lib/python3.10/site-packages (from
-r requirements.txt (line 9)) (0.49.1)
Requirement already satisfied: statsmodels in /opt/conda/lib/python3.10/site-packages
(from -r requirements.txt (line 10)) (0.14.0)
Requirement already satisfied: neurokit2 in /opt/conda/lib/python3.10/site-packages
(from -r requirements.txt (line 11)) (0.2.12)
Requirement already satisfied: bleach in /opt/conda/lib/python3.10/site-packages (fro
m kaggle->-r requirements.txt (line 1)) (6.0.0)
Requirement already satisfied: certifi>=14.05.14 in /opt/conda/lib/python3.10/site-pa
ckages (from kaggle->-r requirements.txt (line 1)) (2025.10.5)
Requirement already satisfied: charset-normalizer in /opt/conda/lib/python3.10/site-p
ackages (from kaggle->-r requirements.txt (line 1)) (3.1.0)
Requirement already satisfied: idna in /opt/conda/lib/python3.10/site-packages (from
kaggle->-r requirements.txt (line 1)) (3.4)
Requirement already satisfied: protobuf in /opt/conda/lib/python3.10/site-packages (f
rom kaggle->-r requirements.txt (line 1)) (6.31.1)
Requirement already satisfied: python-dateutil>=2.5.3 in /opt/conda/lib/python3.10/si
te-packages (from kaggle->-r requirements.txt (line 1)) (2.8.2)
Requirement already satisfied: python-slugify in /opt/conda/lib/python3.10/site-packa
ges (from kaggle->-r requirements.txt (line 1)) (8.0.4)
Requirement already satisfied: requests in /opt/conda/lib/python3.10/site-packages (f
rom kaggle->-r requirements.txt (line 1)) (2.31.0)
Requirement already satisfied: setuptools>=21.0.0 in /opt/conda/lib/python3.10/site-p
ackages (from kaggle->-r requirements.txt (line 1)) (67.7.2)
Requirement already satisfied: six>=1.10 in /opt/conda/lib/python3.10/site-packages
(from kaggle->-r requirements.txt (line 1)) (1.16.0)
```

```
Requirement already satisfied: text-unidecode in /opt/conda/lib/python3.10/site-packa
ges (from kaggle->-r requirements.txt (line 1)) (1.3)
Requirement already satisfied: tqdm in /opt/conda/lib/python3.10/site-packages (from
kaggle->-r requirements.txt (line 1)) (4.65.0)
Requirement already satisfied: urllib3>=1.15.1 in /opt/conda/lib/python3.10/site-pack
ages (from kaggle->-r requirements.txt (line 1)) (2.0.2)
Requirement already satisfied: webencodings in /opt/conda/lib/python3.10/site-package
s (from kaggle->-r requirements.txt (line 1)) (0.5.1)
Requirement already satisfied: narwhals>=1.15.1 in /opt/conda/lib/python3.10/site-pac
kages (from plotly->-r requirements.txt (line 2)) (2.12.0)
Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-packages
(from plotly->-r requirements.txt (line 2)) (25.0)
Requirement already satisfied: contourpy>=1.0.1 in /opt/conda/lib/python3.10/site-pac
kages (from matplotlib->-r requirements.txt (line 3)) (1.0.7)
Requirement already satisfied: cycler>=0.10 in /opt/conda/lib/python3.10/site-package
s (from matplotlib->-r requirements.txt (line 3)) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in /opt/conda/lib/python3.10/site-pa
ckages (from matplotlib->-r requirements.txt (line 3)) (4.39.4)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/conda/lib/python3.10/site-pa
ckages (from matplotlib->-r requirements.txt (line 3)) (1.4.4)
Requirement already satisfied: pillow>=6.2.0 in /opt/conda/lib/python3.10/site-packag
es (from matplotlib->-r requirements.txt (line 3)) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in /opt/conda/lib/python3.10/site-pac
kages (from matplotlib->-r requirements.txt (line 3)) (3.0.9)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-package
s (from pandas->-r requirements.txt (line 6)) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in /opt/conda/lib/python3.10/site-packa
ges (from pandas->-r requirements.txt (line 6)) (2023.3)
Requirement already satisfied: scipy>=1.8.0 in /opt/conda/lib/python3.10/site-package
s (from scikit-learn->-r requirements.txt (line 8)) (1.15.3)
Requirement already satisfied: joblib>=1.2.0 in /opt/conda/lib/python3.10/site-packag
es (from scikit-learn->-r requirements.txt (line 8)) (1.5.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /opt/conda/lib/python3.10/site
-packages (from scikit-learn->-r requirements.txt (line 8)) (3.6.0)
Requirement already satisfied: slicer==0.0.8 in /opt/conda/lib/python3.10/site-packag
es (from shap->-r requirements.txt (line 9)) (0.0.8)
Requirement already satisfied: numba>=0.54 in /opt/conda/lib/python3.10/site-packages
(from shap->-r requirements.txt (line 9)) (0.62.1)
Requirement already satisfied: cloudpickle in /opt/conda/lib/python3.10/site-packages
(from shap->-r requirements.txt (line 9)) (2.2.1)
Requirement already satisfied: typing-extensions in /opt/conda/lib/python3.10/site-pa
ckages (from shap->-r requirements.txt (line 9)) (4.15.0)
Requirement already satisfied: patsy>=0.5.2 in /opt/conda/lib/python3.10/site-package
s (from statsmodels->-r requirements.txt (line 10)) (0.5.3)
Requirement already satisfied: PyWavelets>=1.4.0 in /opt/conda/lib/python3.10/site-pa
ckages (from neurokit2->-r requirements.txt (line 11)) (1.4.1)
Requirement already satisfied: llvmlite<0.46,>=0.45.0dev0 in /opt/conda/lib/python3.1
0/site-packages (from numba>=0.54->shap->-r requirements.txt (line 9)) (0.45.1)

[notice] A new release of pip is available: 25.2 -> 25.3
[notice] To update, run: pip install --upgrade pip
```

```python
import sys
import random
import os.path
import warnings
```

```
from tqdm import tqdm

import numpy as np
import pandas as pd
import seaborn as sns
import neurokit2 as nk
from scipy import stats
from scipy.signal import welch
import matplotlib.pyplot as plt
from joblib.testing import xfail
import plotly.graph_objects as go
from colorama import Fore, Back, Style
from plotly.subplots import make_subplots

from sklearn import tree
from sklearn.pipeline import Pipeline
from sklearn.compose import make_column_transformer
from sklearn.preprocessing import MinMaxScaler, StandardScaler

from sklearn.neural_network import MLPClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaB
from sklearn.neighbors import KNeighborsClassifier, RadiusNeighborsClassifier

from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV, F
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix, (
from sympy import false
```

Options settings

```
pd.set_option('display.max_columns', None)
warnings.filterwarnings("ignore")
random.seed(3003)
IN_COLAB = 'google.colab' in sys.modules
DATA_PRESENT = os.path.isfile("data/Physionet2017Training.tar.xz")
```

Dataset download

```
if not DATA_PRESENT:
    !mkdir data
    !wget https://github.com/MAI3003-Data-Witches/Data-Witches_Project2/raw/refs/heads
    !tar -xf data/Physionet2017Training.tar.xz -C data
else: print("You already have the dataset downloaded, skipping")
```

```
 You already have the dataset downloaded, skipping
```

```
df = pd.read_csv('data/Physionet2017TrainingData.csv', header=None, index_col=False)
#TODO: is it actually needed to convert to mV?

df.head()
```

	0	1	2	3	4	5	6	7	8	9
0	-126000	-141000	-152000	-159000	-160000	-153000	-140000	-126000	-111000	-93000	-720
1	23000	25000	33000	61000	79000	85000	89000	93000	96000	98000	980
2	-116000	-146000	-174000	-190000	-199000	-207000	-211000	-211000	-207000	-201000	-1940
3	-500000	-597000	-695000	-792000	-873000	-917000	-918000	-873000	-822000	-767000	-7510