Feature Selection with Filtering Method I Constant, Quasi Constant and Duplicate Feature Removal

Unnecessary and redundant features not only slow down the training time of an algorithm, but they also affect the performance of the algorithm.

There are several advantages of performing feature selection before training machine learning models:

- Models with less number of features have higher explainability
- It is easier to implement machine learning models with reduced features
- Fewer features lead to enhanced generalization which in turn reduces overfitting
- Feature selection removes data redundancy
- Training time of models with fewer features is significantly lower
- Models with fewer features are less prone to errors

What is filter method?

Features selected using filter methods can be used as an input to any machine learning models.

- Univariate -> Fisher Score, Mutual Information Gain, Variance etc
- Multi-variate -> Pearson Correlation

The univariate filter methods are the type of methods where individual features are ranked according to specific criteria. The top N features are then selected. Different types of ranking criteria are used for univariate filter methods, for example fisher score, mutual information, and variance of the feature.

Multivariate filter methods are capable of removing redundant features from the data since they take the mutual relationship between the features into account.

File failed to load: /extensions/MathZoom.js

- Constant Removal
- Quasi Constant Removal
- Duplicate Feature Removal

Constant Feature Removal

In [30]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [31]:
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.feature_selection import VarianceThreshold
```

In [32]:
```
data = pd.read_csv('Subject01_dataset_202002071_50ms.csv', nrows = 2861)
data.dropna(inplace=True)
data.head()
```

Out[32]:

	Time	DEMG1_AR1	DEMG1_AR2	DEMG1_AR3	DEMG1_AR4	DEMG1_AR5	DEMG1_AR6	DEM
0	0.00	0.622484	-0.260461	0.356585	-0.234134	0.211762	-0.032847	
1	0.05	0.104438	0.112400	0.282052	-0.097667	0.134183	-0.088160	
2	0.10	0.032687	0.044736	0.055276	0.095904	0.082075	-0.239689	
3	0.15	0.148638	0.044486	0.156377	-0.141641	-0.132405	0.049945	
4	0.20	0.298830	-0.126001	0.336519	-0.123507	-0.132748	-0.037417	

5 rows × 103 columns

File failed to load: /extensions/MathZoom.js

In [139]:
```
columns_to_drop = ["Right Knee Angle", "Left Knee Angle", "Right Hip Angle ",
X = data.drop(columns_to_drop, axis=1)
y = data[columns_to_drop]
print(y)

X.shape, y.shape
```

```
      Right Knee Angle  Left Knee Angle  Right Hip Angle    Left Hip Angle  \
0            -12.014975        -7.330916          8.248002         5.384330
1            -11.924520        -7.321892          8.347285         5.560538
2            -11.843123        -7.211070          8.544523         5.762084
3            -11.753622        -7.071632          8.778896         5.958890
4            -11.751801        -6.931188          8.842804         5.944573
...                 ...              ...               ...              ...
2855         -72.057638       -13.742484         30.679507         8.905448
2856         -65.581122       -12.133072         30.987638         7.561556
2857         -58.013923       -11.103924         29.903660         6.550607
2858         -50.129967       -10.827592         27.485598         5.810535
2859         -41.375815       -10.732799         23.995448         5.296846

      Right Knee Torque  Left Knee Torque  Right Hip Torque  Left Hip Torque
0             -9.804979        -16.450071        -17.707678       -19.771144
1            -10.773364        -17.615612        -18.388583       -21.755446
2            -11.342771        -18.882028        -18.200220       -23.449065
3            -11.675735        -20.124378        -17.944835       -25.288482
4            -11.339695        -20.413309        -17.190672       -25.727190
...                 ...               ...               ...              ...
2855          -7.743681        -29.369524         -3.139781       -40.041396
2856         -10.716236        -37.305377         -4.483967       -41.587591
2857         -11.643175        -41.695356         -5.863842       -43.933461
2858          -9.887130        -41.048253         -4.411930       -43.088404
2859          -8.166552        -36.671271         -0.390141       -40.891515

[2860 rows x 8 columns]
```

Out[139]: ((2860, 95), (2860, 8))

In [140]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, rando
```

File failed to load: /extensions/MathZoom.js

Constant Features Removal

In [38]:
```
constant_filter = VarianceThreshold(threshold=0)
constant_filter.fit(X_train)
```

Out[38]:
```
VarianceThreshold
VarianceThreshold(threshold=0)
```

In [39]:
```
constant_filter.get_support().sum()
```

Out[39]:
```
95
```

File failed to load: /extensions/MathZoom.js

In [40]:
```
constant_list = [not temp for temp in constant_filter.get_support()]
constant_list
```

File failed to load: /extensions/MathZoom.js

Out[40]:
```
[False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
```

File failed to load: /extensions/MathZoom.js

```
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False,
 False]
```

In [41]:
```
X.columns[constant_list]
```

Out[41]:
```
Index([], dtype='object')
```

In [42]:
```
X_train_filter = constant_filter.transform(X_train)
X_test_filter = constant_filter.transform(X_test)
```

In [43]:
```
X_train_filter.shape, X_test_filter.shape, X_train.shape
```

Out[43]:
```
((1716, 95), (1144, 95), (1716, 95))
```

Quasi constant feature removal

File failed to load: /extensions/MathZoom.js

In [44]:
```
quasi_constant_filter = VarianceThreshold(threshold=0.01)
```

In [45]:
```
quasi_constant_filter.fit(X_train_filter)
```

Out[45]:
```
VarianceThreshold
VarianceThreshold(threshold=0.01)
```

In [46]:
```
quasi_constant_filter.get_support().sum()
```

Out[46]: 72

File failed to load: /extensions/MathZoom.js

In [96]:
```
quasi_constant_list = [not temp for temp in quasi_constant_filter.get_support(
quasi_constant_list
```

File failed to load: /extensions/MathZoom.js

Out[96]:
```
[False,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
```

```
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False,
 True,
 True,
 False,
 False,
 False,
 False,
 False,
 False]
```

In [47]:
```
95-72
```

Out[47]:
```
23
```

In [97]:
```
X.columns[quasi_constant_list]
```

Out[97]:
```
Index(['DEMG1_RMS', 'DEMG1_MAV', 'DEMG2_RMS', 'DEMG2_MAV', 'DEMG3_RMS',
       'DEMG3_MAV', 'DEMG4_RMS', 'DEMG4_MAV', 'DEMG5_RMS', 'DEMG5_MAV',
       'DEMG6_AR6', 'DEMG6_RMS', 'DEMG6_MAV', 'DEMG7_RMS', 'DEMG7_MAV',
       'DEMG8_RMS', 'DEMG8_MAV', 'DEMG9_RMS', 'DEMG9_MAV', 'DEMG10_RMS',
       'DEMG10_MAV', 'DEMG11_RMS', 'DEMG11_MAV'],
      dtype='object')
```

File failed to load: /extensions/MathZoom.js

In [48]:
```
X_train_quasi_filter = quasi_constant_filter.transform(X_train_filter)
X_test_quasi_filter = quasi_constant_filter.transform(X_test_filter)
```

In [49]:
```
X_train_quasi_filter.shape, X_test_quasi_filter.shape
```

Out[49]:
```
((1716, 72), (1144, 72))
```

In [ ]:

Remove Duplicate Features

In [98]:
```
X_train_T = X_train_quasi_filter.T
X_test_T = X_test_quasi_filter.T
```

In [99]:
```
type(X_train_T)
```

Out[99]:
```
numpy.ndarray
```

In [100]:
```
X_train_T = pd.DataFrame(X_train_T)
X_test_T = pd.DataFrame(X_test_T)
```

In [101]:
```
X_train_T.shape, X_test_T.shape
```

Out[101]:
```
((72, 1716), (72, 1144))
```

In [102]:
```
X_train_T.duplicated().sum()
```

Out[102]:
```
0
```

In [103]:
```
duplicated_features = X_train_T.duplicated()
duplicated_features
```

Out[103]:
```
0     False
1     False
2     False
3     False
4     False
      ...
67    False
68    False
69    False
70    False
71    False
Length: 72, dtype: bool
```

File failed to load: /extensions/MathZoom.js

In [104]:
```
features_to_keep = [not index for index in duplicated_features]
```

In [105]:
```
X_train_unique = X_train_T[features_to_keep].T
X_test_unique = X_test_T[features_to_keep].T
```

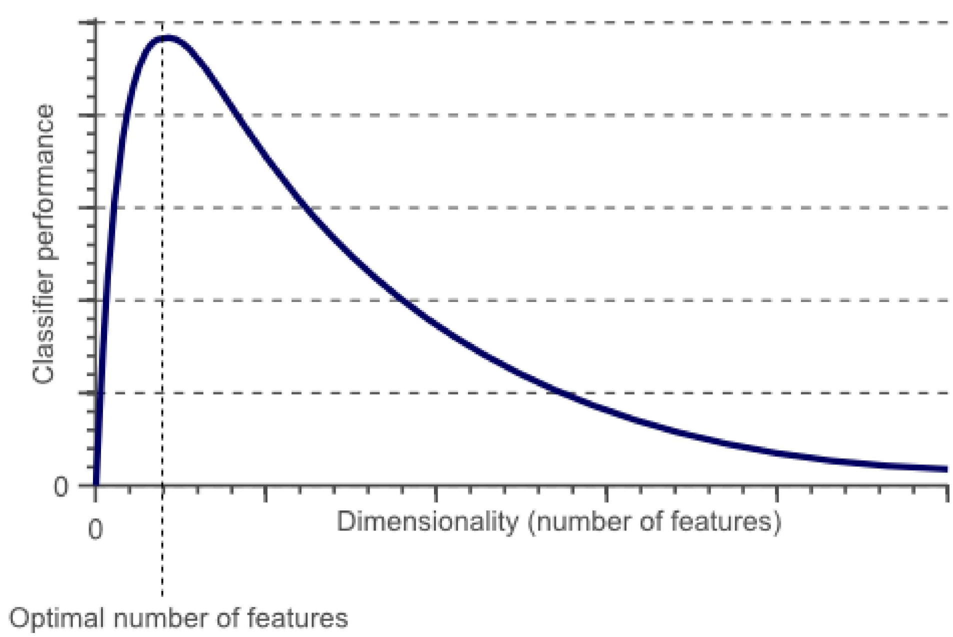
In [106]:
```
X_train_unique.shape, X_train.shape
```

Out[106]:
```
((1716, 72), (1716, 95))
```

In [ ]:

In [ ]:

Feature Selection with Filtering Method- Correlated Feature Removal



A dataset can also contain correlated features. Two or more than two features are correlated if they are close to each other in the linear space.

Correlation between the output observations and the input features is very important and such features should be retained

File failed to load: /extensions/MathZoom.js

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

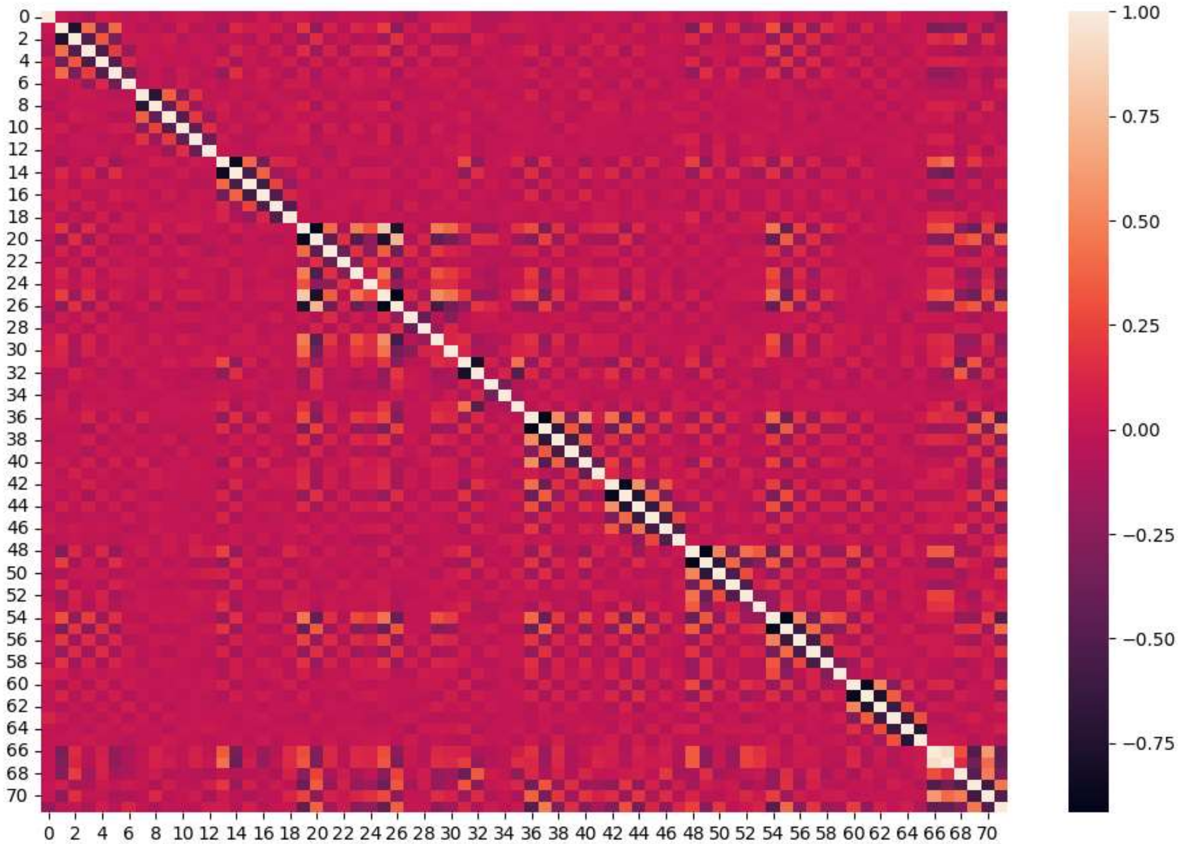
Summary

- Feature Space to target correlation is desired
- Feature to feature correlation is not desired
- If 2 features are highly correlated then either feature is redundant
- Correlation in feature space increases model complexity
- Removing correlated features improves model performance
- Different model shows different performance over the correlated features

In [107]:
```
corrmat = X_train_unique.corr()
```

In [108]:
```
plt.figure(figsize=(12,8))
sns.heatmap(corrmat)
```

Out[108]: <Axes: >



In [109]:
```
def get_correlation(data, threshold):
    corr_col = set()
    corrmat = data.corr()
    for i in range(len(corrmat.columns)):
        for j in range(i):
            if abs(corrmat.iloc[i, j])> threshold:
                colname = corrmat.columns[i]
                corr_col.add(colname)
    return corr_col
```

In [110]:
```
corr_features = get_correlation(X_train_unique, 0.85)
corr_features
```

Out[110]: {14, 20, 26, 37, 43, 49, 55, 61, 67}

In [111]:
```
len(corr_features)
```

Out[111]: 9

In [112]:
```
X_train_uncorr = X_train_unique.drop(labels=corr_features, axis = 1)
X_test_uncorr = X_test_unique.drop(labels = corr_features, axis = 1)
```

File failed to load: /extensions/MathZoom.js

In [113]:
```
X_train_uncorr.shape, X_test_uncorr.shape
```

Out[113]:
```
((1716, 63), (1144, 63))
```

Feature Grouping and Feature Importance

In [120]:
```
corrmat
```

Out[120]:

	0	1	2	3	4	5	6	7	
0	1.000000	0.018754	-0.026221	-0.104476	0.067727	0.026402	-0.004455	0.007820	-0.0588(
1	0.018754	1.000000	-0.759803	0.416815	-0.355066	0.404926	0.009202	-0.035641	0.0037'
2	-0.026221	-0.759803	1.000000	-0.500346	0.326546	-0.300512	0.029074	0.029226	0.0193‹
3	-0.104476	0.416815	-0.500346	1.000000	-0.541121	0.221033	-0.187547	0.028860	-0.0081(
4	0.067727	-0.355066	0.326546	-0.541121	1.000000	-0.485040	0.122426	-0.009438	0.0055!
...	...	...	...	...	...	...	...	...	
67	0.006357	-0.285234	0.103232	-0.035960	0.097295	-0.211737	-0.082617	-0.000141	0.0873‹
68	-0.010246	-0.263891	0.217593	-0.144773	0.093685	-0.101928	-0.081742	-0.004716	0.0432(
69	0.008946	0.071047	-0.101836	-0.079640	0.036567	-0.010714	0.046785	0.011727	-0.0917(
70	0.006684	-0.217061	0.160789	-0.001973	0.007566	-0.107230	-0.079670	-0.032523	0.1205
71	-0.183139	-0.014527	-0.009442	-0.085388	0.071454	-0.054630	0.108286	-0.001158	-0.0841'

72 rows × 72 columns

In [121]:
```
corrdata = corrmat.abs().stack()
corrdata
```

Out[121]:
```
0   0     1.000000
    1     0.018754
    2     0.026221
    3     0.104476
    4     0.067727
            ...
71  67    0.414080
    68    0.308669
    69    0.458294
    70    0.495667
    71    1.000000
Length: 5184, dtype: float64
```

File failed to load: /extensions/MathZoom.js

In [122]:
```
corrdata = corrdata.sort_values(ascending=False)
corrdata
```

Out[122]:
```
0   0     1.000000
27  27    1.000000
21  21    1.000000
22  22    1.000000
23  23    1.000000
            ...
49  10    0.000037
36  12    0.000004
12  36    0.000004
10  59    0.000003
59  10    0.000003
Length: 5184, dtype: float64
```

In [123]:
```
corrdata = corrdata[corrdata>0.85]
corrdata = corrdata[corrdata<1]
corrdata
```

Out[123]:
```
67  66    0.928561
66  67    0.928561
36  37    0.916683
37  36    0.916683
49  48    0.903687
48  49    0.903687
26  25    0.896532
25  26    0.896532
19  20    0.885856
20  19    0.885856
54  55    0.883545
55  54    0.883545
42  43    0.872013
43  42    0.872013
61  60    0.863946
60  61    0.863946
13  14    0.856168
14  13    0.856168
dtype: float64
```

In [ ]:

File failed to load: /extensions/MathZoom.js

In [124]:
```
corrdata = pd.DataFrame(corrdata).reset_index()
corrdata.columns = ['features1', 'features2', 'corr_value']
corrdata
```

Out[124]:

	features1	features2	corr_value
0	67	66	0.928561
1	66	67	0.928561
2	36	37	0.916683
3	37	36	0.916683
4	49	48	0.903687
5	48	49	0.903687
6	26	25	0.896532
7	25	26	0.896532
8	19	20	0.885856
9	20	19	0.885856
10	54	55	0.883545
11	55	54	0.883545
12	42	43	0.872013
13	43	42	0.872013
14	61	60	0.863946
15	60	61	0.863946
16	13	14	0.856168
17	14	13	0.856168

In [125]:
```
grouped_feature_list = []
correlated_groups_list = []
for feature in corrdata.features1.unique():
    if feature not in grouped_feature_list:
        correlated_block = corrdata[corrdata.features1 == feature]
        grouped_feature_list = grouped_feature_list + list(correlated_block.fe
        correlated_groups_list.append(correlated_block)
```

In [126]:
```
len(correlated_groups_list)
```

Out[126]: 9

In [127]:
```
X_train.shape, X_train_uncorr.shape
```

Out[127]: ((1716, 95), (1716, 63))

File failed to load: /extensions/MathZoom.js

In [128]:
```
for group in correlated_groups_list:
    print(group)
```

```
   features1  features2  corr_value
0         67         66    0.928561
   features1  features2  corr_value
2         36         37    0.916683
   features1  features2  corr_value
4         49         48    0.903687
   features1  features2  corr_value
6         26         25    0.896532
   features1  features2  corr_value
8         19         20    0.885856
    features1  features2  corr_value
10         54         55    0.883545
    features1  features2  corr_value
12         42         43    0.872013
    features1  features2  corr_value
14         61         60    0.863946
    features1  features2  corr_value
16         13         14    0.856168
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

File failed to load: /extensions/MathZoom.js

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

File failed to load: /extensions/MathZoom.js

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

File failed to load: /extensions/MathZoom.js

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

File failed to load: /extensions/MathZoom.js