# Machine learning for the automated diagnosis of breast cancer

Joan Gomà, Pau Mateo

# ANNEX

## Descripció bàsica del dataset

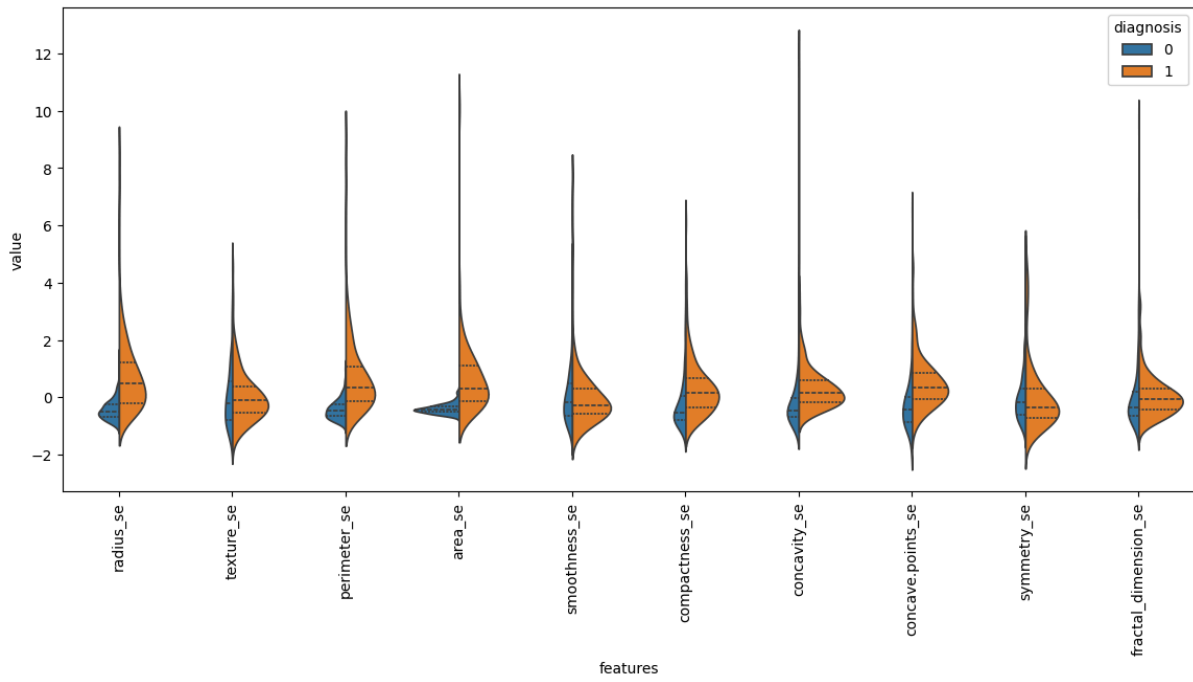| | radius_mean | texture_mean | perimeter_mean | area_mean | smoothness_mean | compactness_mean | concavity_mean | concave.points_mean | symmetry_mean | fractal_dimension_mean |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 |
| mean | 14.127 | 19.290 | 91.969 | 654.889 | 0.096 | 0.104 | 0.089 | 0.049 | 0.181 | -474.409 |
| std | 3.524 | 4.301 | 24.299 | 351.914 | 0.014 | 0.053 | 0.080 | 0.039 | 0.027 | 2127.701 |
| min | 6.981 | 9.710 | 43.790 | 143.500 | 0.053 | 0.019 | 0.000 | 0.000 | 0.106 | -9999.000 |
| 25% | 11.700 | 16.170 | 75.170 | 420.300 | 0.086 | 0.065 | 0.030 | 0.020 | 0.162 | 0.057 |
| 50% | 13.370 | 18.840 | 86.240 | 551.100 | 0.096 | 0.093 | 0.062 | 0.034 | 0.179 | 0.061 |
| 75% | 15.780 | 21.800 | 104.100 | 782.700 | 0.105 | 0.130 | 0.131 | 0.074 | 0.196 | 0.066 |
| max | 28.110 | 39.280 | 188.500 | 2501.000 | 0.163 | 0.345 | 0.427 | 0.201 | 0.304 | 0.097 |

| | radius_se | texture_se | perimeter_se | area_se | smoothness_se | compactness_se | concavity_se | concave.points_se | symmetry_se | fractal_dimension_se |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 |
| mean | 0.405 | 1.217 | 2.866 | 40.337 | 0.007 | 0.025 | 0.032 | 0.012 | 0.021 | -474.466 |
| std | 0.277 | 0.552 | 2.022 | 45.491 | 0.003 | 0.018 | 0.030 | 0.006 | 0.008 | 2127.688 |
| min | 0.112 | 0.360 | 0.757 | 6.802 | 0.002 | 0.002 | 0.000 | 0.000 | 0.008 | -9999.000 |
| 25% | 0.232 | 0.834 | 1.606 | 17.850 | 0.005 | 0.013 | 0.015 | 0.008 | 0.015 | 0.002 |
| 50% | 0.324 | 1.108 | 2.287 | 24.530 | 0.006 | 0.020 | 0.026 | 0.011 | 0.019 | 0.003 |
| 75% | 0.479 | 1.474 | 3.357 | 45.190 | 0.008 | 0.032 | 0.042 | 0.015 | 0.023 | 0.004 |
| max | 2.873 | 4.885 | 21.980 | 542.200 | 0.031 | 0.135 | 0.396 | 0.053 | 0.079 | 0.030 |

| | radius_worst | texture_worst | perimeter_worst | area_worst | smoothness_worst | compactness_worst | concavity_worst | concave.points_worst | symmetry_worst | fractal_dimension_worst |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 | 569.000 |
| mean | 16.269 | 25.677 | 107.261 | 880.583 | 0.132 | 0.254 | 0.272 | 0.115 | 0.290 | -474.389 |
| std | 4.833 | 6.146 | 33.603 | 569.357 | 0.023 | 0.157 | 0.209 | 0.066 | 0.062 | 2127.705 |
| min | 7.930 | 12.020 | 50.410 | 185.200 | 0.071 | 0.027 | 0.000 | 0.000 | 0.157 | -9999.000 |
| 25% | 13.010 | 21.080 | 84.110 | 515.300 | 0.117 | 0.147 | 0.115 | 0.065 | 0.250 | 0.070 |
| 50% | 14.970 | 25.410 | 97.660 | 686.500 | 0.131 | 0.212 | 0.227 | 0.100 | 0.282 | 0.079 |
| 75% | 18.790 | 29.720 | 125.400 | 1084.000 | 0.146 | 0.339 | 0.383 | 0.161 | 0.318 | 0.091 |
| max | 36.040 | 49.540 | 251.200 | 4254.000 | 0.223 | 1.058 | 1.252 | 0.291 | 0.664 | 0.207 |

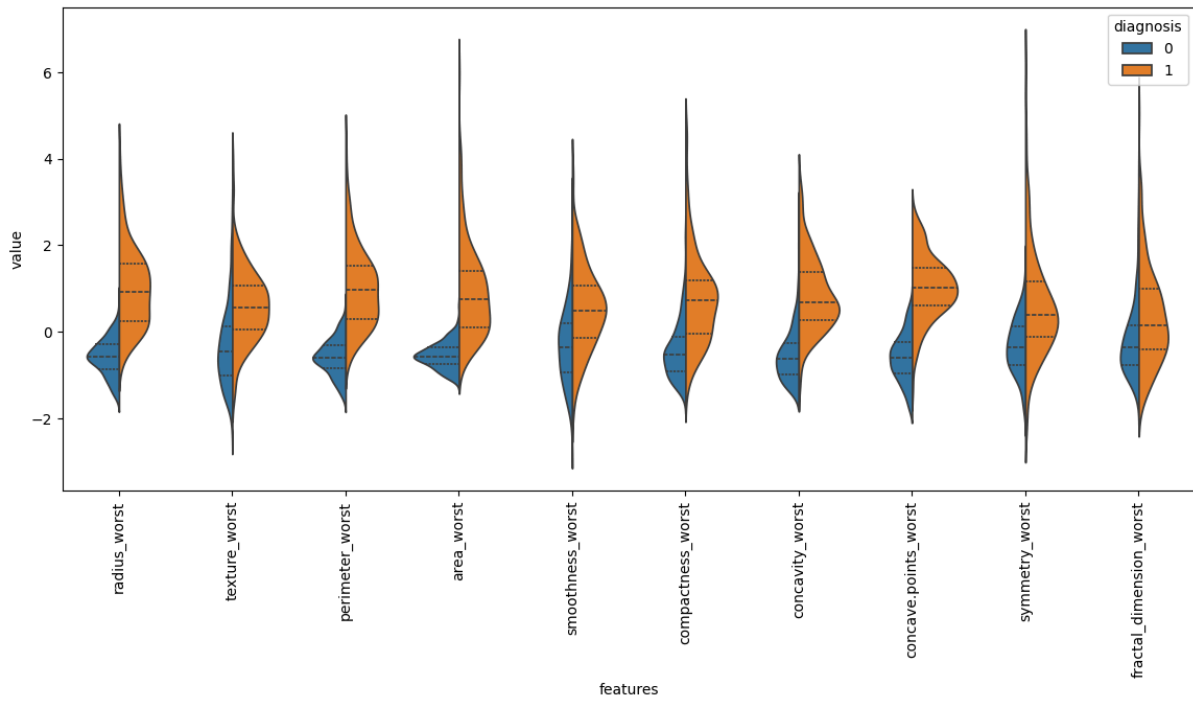## Descripció bàsica del les variables

```
        Data columns (total 33 columns):
  #    Column                  Non-Null Count  Dtype
 ---   ------                  --------------  -----
  0    Unnamed: 0              569 non-null    int64
  1    id                      569 non-null    int64
  2    diagnosis               569 non-null    object
  3    radius_mean             569 non-null    float64
  4    texture_mean            569 non-null    float64
  5    perimeter_mean          569 non-null    float64
  6    area_mean               569 non-null    float64
  7    smoothness_mean         569 non-null    float64
  8    compactness_mean        569 non-null    float64
  9    concavity_mean          569 non-null    float64
 10    concave.points_mean     569 non-null    float64
 11    symmetry_mean           569 non-null    float64
 12    fractal_dimension_mean  569 non-null    float64
 13    radius_se               569 non-null    float64
 14    texture_se              569 non-null    float64
 15    perimeter_se            569 non-null    float64
 16    area_se                 569 non-null    float64
 17    smoothness_se           569 non-null    float64
 18    compactness_se          569 non-null    float64
 19    concavity_se            569 non-null    float64
 20    concave.points_se       569 non-null    float64
 21    symmetry_se             569 non-null    float64
 22    fractal_dimension_se    569 non-null    float64
 23    radius_worst            569 non-null    float64
 24    texture_worst           569 non-null    float64
 25    perimeter_worst         569 non-null    float64
 26    area_worst              569 non-null    float64
 27    smoothness_worst        569 non-null    float64
 28    compactness_worst       569 non-null    float64
 29    concavity_worst         569 non-null    float64
 30    concave.points_worst    569 non-null    float64
 31    symmetry_worst          569 non-null    float64
 32    fractal_dimension_worst 569 non-null    float64
        dtypes: float64(30), int64(2), object(1)
              memory usage: 146.8+ KB
```
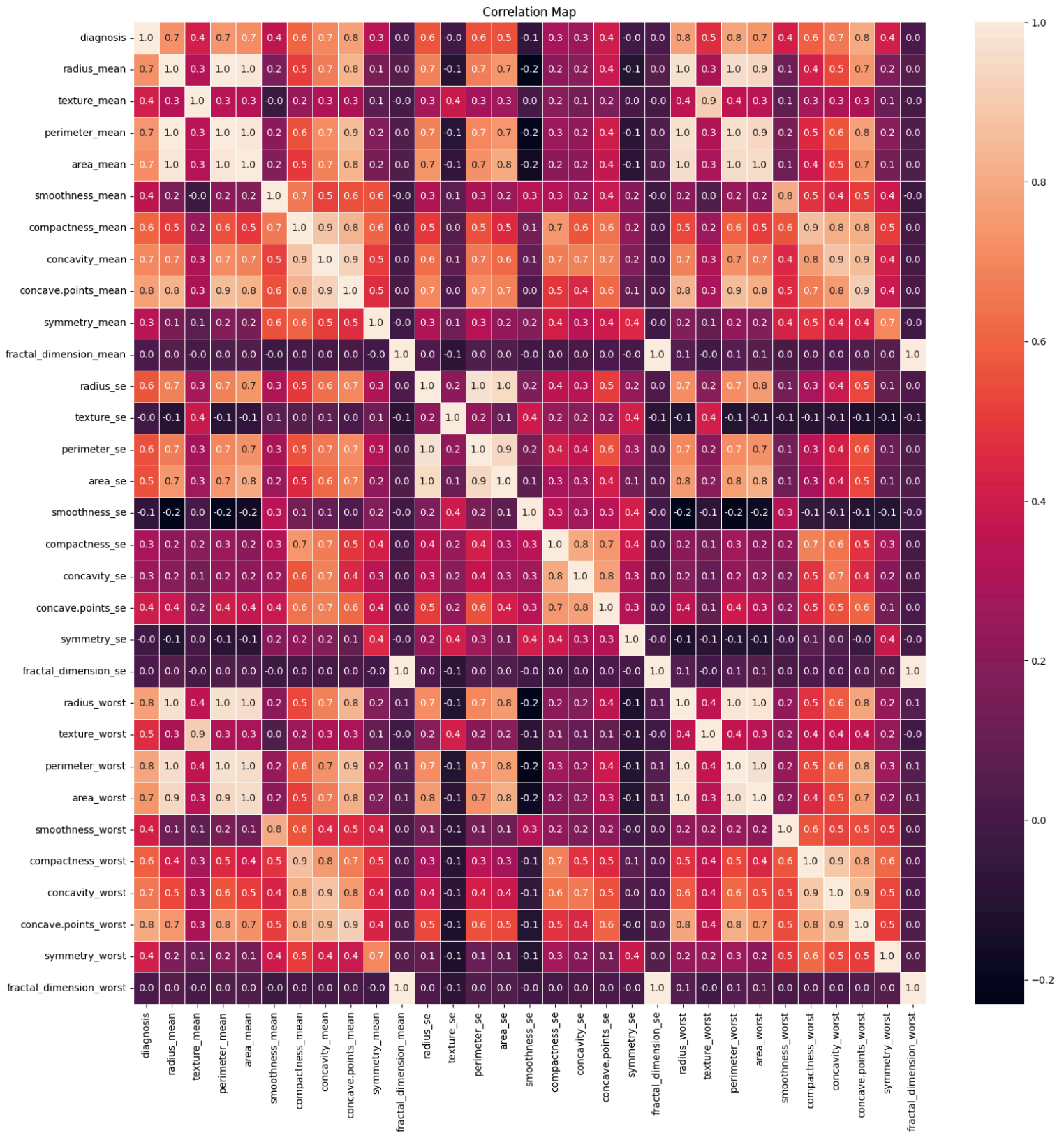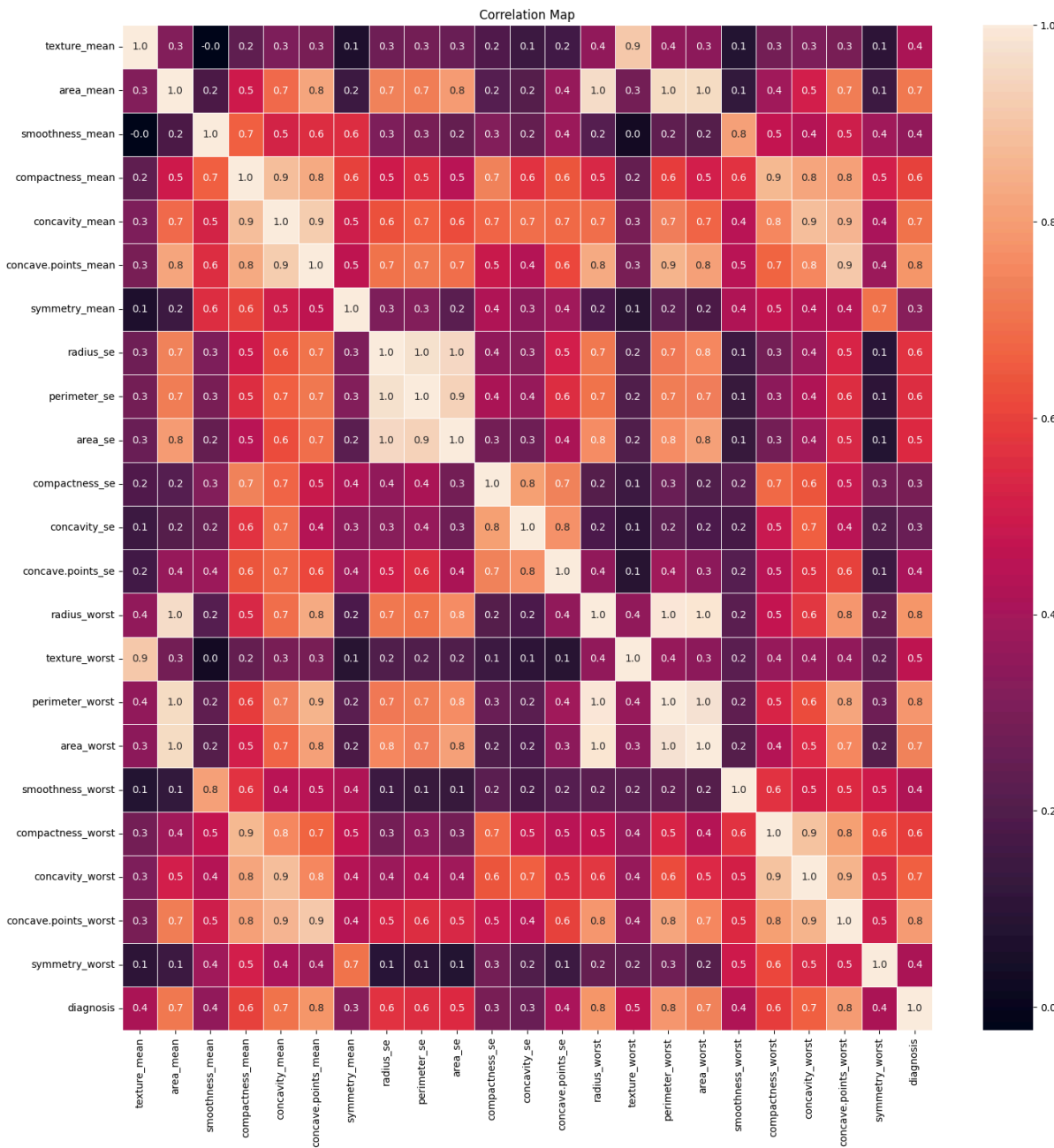
# Gràfica 1: violin plot de les variables del 10-20
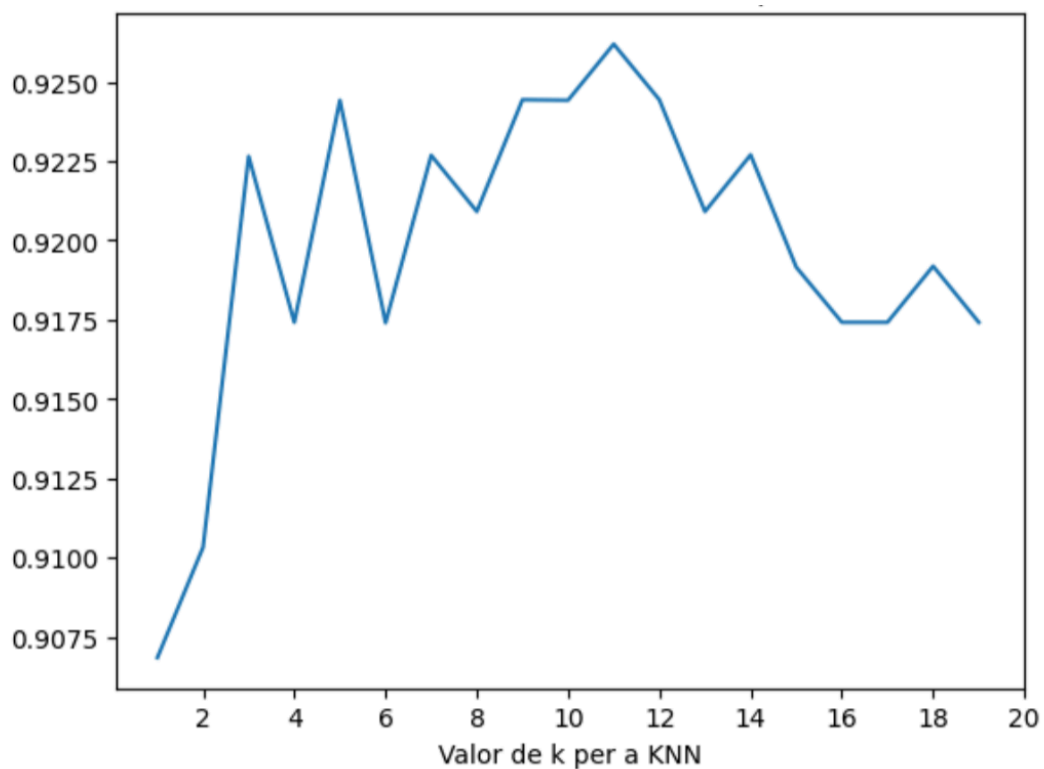


# Gràfica 2: violin plot de les variables del 20-30

# Gràfica 3: Correlation map



Correlation Map

# Gràfica 4: Correlation map variables reduïdes



Correlation Map

Gràfica 5: *accuracy* obtinguda amb validació creuada (10-fold) per
determinar el millor paràmetre k del k-NN



Gràfica 6: *recall* obtingut amb validació creuada (10-fold) per
determinar el millor paràmetre k del k-NN