mmf175u4h

March 11, 2024

[1]:
```
import pandas as pd

# Define the file path
file_path = r"C:\Users\krestty\Downloads\Femor_img_dataset␣
 ↪(1)\Femor_img_dataset\dataset_bmd.csv"

# Load the dataset
data = pd.read_csv(file_path)

# Display the first few rows of the dataframe
print(data.head())
```

```
   Unnamed: 0  Participant ID  Age at recruitment  Sex  Genetic sex  \
0        2596         6022277                  58    0          0.0
1        4309         3361500                  42    1          1.0
2       12197         3056510                  69    0          0.0
3       12941         1182673                  62    1          1.0
4       15337         4606954                  58    0          0.0

   Standing height | Instance 0  Standing height | Instance 1  \
0                         149.5                          NaN
1                         185.0                        185.0
2                         173.0                        172.0
3                         172.0                          NaN
4                         158.0                        156.0

   Standing height | Instance 2  Standing height | Instance 3  \
0                         151.0                          NaN
1                         186.0                          NaN
2                         172.0                          NaN
3                         171.0                          NaN
4                         154.0                          NaN

   Weight | Instance 0(participant - p21002_i0)  …  \
0                                          63.6  …
1                                          83.6  …
2                                          81.8  …
3                                          91.7  …
```

```
4                                        58.9  …

   Treatment/medication code | Instance 1  \
0                                      NaN
1                    1140871310|1140865634
2                                      NaN
3                                      NaN
4                                    99999

   Treatment/medication code | Instance 2  \
0                                      NaN
1                    1140871310|1140865634
2                                1140865634
3                                      NaN
4                                      NaN

   Treatment/medication code | Instance 3  \
0                                      NaN
1                                      NaN
2                                      NaN
3                                      NaN
4                                      NaN

                                    Diagnoses - ICD10  \
0  E039|F171|F250|F329|F419|I10|R33|S7200|T814|W0…
1          M8725|N492|N508|S7210|T931|V134|Z470|Z874
2  D758|E538|I269|J90|K011|K029|K296|K30|K573|K57…
3  A099|B370|B955|B968|B972|D649|E559|E669|E780|F…
4             H001|I839|N950|R298|R42|Z866|Z871|Z888

                         Diagnoses - secondary ICD10  Diagnoses - ICD9  \
0              E039|F171|F329|F419|I10|R33|Z864|Z911               NaN
1                                   T931|Z470|Z874               NaN
2  D758|E538|I269|J90|K029|K296|K573|K590|M8199|N…              NaN
3  A099|B370|B955|B968|B972|D649|E559|E669|E780|F…              NaN
4                               R298|Z866|Z871|Z888               NaN

   Diagnoses - secondary ICD9  fracture  FRAX_bmd  FRAX_without_bmd
0                         NaN      True       2.6               0.6
1                         NaN      True       0.0               0.1
2                         NaN      True       0.5               2.5
3                         NaN      True       1.9               0.9
4                         NaN     False       4.3               0.9

[5 rows x 160 columns]
```

[2]:
```
# Display statistical summary of the dataframe
print(data.describe())
```

```
          Unnamed: 0  Participant ID  Age at recruitment         Sex  \
count     300.000000    3.000000e+02          300.000000  300.000000
mean   260278.200000    3.200962e+06           57.280000    0.473333
std    143370.259678    1.609800e+06            7.764614    0.500123
min      2596.000000    1.004467e+06           41.000000    0.000000
25%    140208.750000    1.674056e+06           51.000000    0.000000
50%    266014.500000    3.147242e+06           58.000000    0.000000
75%    380320.500000    4.574412e+06           64.000000    1.000000
max    501096.000000    6.022829e+06           69.000000    1.000000

       Genetic sex  Standing height | Instance 0  \
count   293.000000                    300.000000
mean      0.474403                    169.483667
std       0.500199                      8.975580
min       0.000000                    142.000000
25%       0.000000                    163.000000
50%       0.000000                    170.000000
75%       1.000000                    176.250000
max       1.000000                    191.000000

       Standing height | Instance 1  Standing height | Instance 2  \
count                     71.000000                    295.000000
mean                     169.339437                    168.721356
std                        8.118594                      9.065940
min                      151.000000                    142.000000
25%                      163.750000                    162.250000
50%                      170.000000                    170.000000
75%                      174.000000                    175.000000
max                      188.500000                    191.000000

       Standing height | Instance 3  \
count                     19.000000
mean                     171.736842
std                        9.780037
min                      151.000000
25%                      164.000000
50%                      174.000000
75%                      180.500000
max                      185.000000

       Weight | Instance 0(participant - p21002_i0)  …  \
count                                     300.000000  …
mean                                       76.977000  …
std                                        15.020066  …
```

```
min                                        50.200000  …
25%                                        65.675000  …
50%                                        75.200000  …
75%                                        85.500000  …
max                                       139.500000  …

       Calcium | Instance 1(participant - p100024_i1)  Calcium | Instance 2  \
count                                        98.000000             87.000000
mean                                        954.755816            968.173563
std                                         374.572966            431.965239
min                                         279.650000            196.900000
25%                                         680.927500            729.160000
50%                                         914.885000            924.110000
75%                                        1135.662500           1108.520000
max                                        2218.910000           3522.710000

       Calcium | Instance 3  Calcium | Instance 4  \
count            114.000000            109.000000
mean            1007.974035           1002.354679
std              370.459250            609.009466
min              156.030000            216.550000
25%              726.132500            737.010000
50%              979.935000            878.190000
75%             1219.772500           1136.510000
max             2511.780000           6095.940000

       Number of treatments/medications taken | Instance 0  \
count                                         300.000000
mean                                            2.033333
std                                             2.395416
min                                             0.000000
25%                                             0.000000
50%                                             1.000000
75%                                             3.000000
max                                            16.000000

       Number of treatments/medications taken | Instance 1  \
count                                          72.000000
mean                                            1.916667
std                                             2.180257
min                                             0.000000
25%                                             0.000000
50%                                             1.000000
75%                                             2.000000
max                                            10.000000

       Number of treatments/medications taken | Instance 2  \
count                                         273.000000
```

```
mean                                                 2.216117
std                                                  2.630596
min                                                  0.000000
25%                                                  0.000000
50%                                                  2.000000
75%                                                  3.000000
max                                                 18.000000

       Number of treatments/medications taken | Instance 3    FRAX_bmd  \
count                                          19.000000    300.000000
mean                                            1.315789      1.127333
std                                             2.161600      1.371540
min                                             0.000000      0.000000
25%                                             0.000000      0.200000
50%                                             0.000000      0.600000
75%                                             2.000000      1.425000
max                                             8.000000     11.000000

       FRAX_without_bmd
count        300.000000
mean           1.001667
std            0.810939
min            0.100000
25%            0.300000
50%            0.800000
75%            1.500000
max            4.300000

[8 rows x 129 columns]
```

[4]:
```
print(data.columns)
```

```
Index(['Unnamed: 0', 'Participant ID', 'Age at recruitment', 'Sex',
       'Genetic sex', 'Standing height | Instance 0',
       'Standing height | Instance 1', 'Standing height | Instance 2',
       'Standing height | Instance 3',
       'Weight | Instance 0(participant - p21002_i0)',
       …
       'Treatment/medication code | Instance 1',
       'Treatment/medication code | Instance 2',
       'Treatment/medication code | Instance 3', 'Diagnoses - ICD10',
       'Diagnoses - secondary ICD10', 'Diagnoses - ICD9',
       'Diagnoses - secondary ICD9', 'fracture', 'FRAX_bmd',
       'FRAX_without_bmd'],
      dtype='object', length=160)
```

[5]:
```
import pandas as pd

# Define the file path
file_path = r"C:\Users\krestty\Downloads\Femor_img_dataset␣
 ↪(1)\Femor_img_dataset\dataset_bmd.csv"

# Load the dataset into a DataFrame
df = pd.read_csv(file_path)

# Count the occurrences of each unique value in the 'fracture' column and print␣
 ↪the result
df_fracture_counts = df['fracture'].value_counts()
print(df_fracture_counts)
```

```
fracture
False    187
True     113
Name: count, dtype: int64
```

[6]:
```
# Extracting unique prefixes from column names
instances = set(col.split(' | ')[0] for col in df.columns)

# Initialize a list to store columns with maximum non-missing values for each␣
 ↪prefix
columns = []

# Iterate over unique prefixes
for prefix in instances:
    # Select columns starting with the current prefix
    columns_to_check = [col for col in df.columns if col.startswith(prefix)]

    # Calculate counts of non-missing values for each selected column
    non_missing_counts = df[columns_to_check].count()

    # Find column with maximum non-missing values
    column_with_max_non_missing = non_missing_counts.idxmax()

    # Append column with maximum non-missing values to the list
    columns.append(column_with_max_non_missing)

# Select columns with maximum non-missing values for each prefix and display␣
 ↪the resulting DataFrame
df_max_non_missing = df[columns]
print(df_max_non_missing.head())
```

```
   Femur neck BMC (bone mineral content) (right) | Instance 2  \
0                                                    3.40979
```

```
1                                                5.27111
2                                                5.45140
3                                                3.86018
4                                                3.29454

      Treatment/medication code | Instance 0  \
0           1140921600|1141191044|1140928274
1                                 1140884488
2                                        NaN
3                                        NaN
4  1189|1140909674|1141188442|1140876592|99999

   Femur neck BMD (bone mineral density) (right) | Instance 2  fracture  \
0                                                0.723206               True
1                                                0.868022               True
2                                                0.996041               True
3                                                0.687088               True
4                                                0.708290              False

   Femur wards BMD (bone mineral density) (right) | Instance 2  \
0                                                0.460578
1                                                0.717409
2                                                0.768620
3                                                0.436912
4                                                0.490887

  Illnesses of father | Instance 0  \
0                      8|6|1|-27
1                        8|2|-27
2                          2|-27
3                            3|2
4                          6|-27

   Femur troch BMD (bone mineral density) T-score (right) | Instance 2  \
0                                               -2.239100
1                                               -0.599261
2                                               -0.235811
3                                               -2.094850
4                                               -1.809550

   Tobacco smoking  \
0              NaN
1            114.0
2              NaN
3            113.0
4              NaN

   Femur upper neck BMD (bone mineral density) T-score (left) | Instance 2  \
```

```
0                                                  NaN
1                                                  NaN
2                                                 -0.9
3                                                  NaN
4                                                  NaN

  Type of special diet followed | Instance 4  …  \
0                                        NaN  …
1                                        NaN  …
2                                        NaN  …
3                                        NaN  …
4                                         13  …

   Country of Birth (non-UK origin) | Instance 0  \
0                                            NaN
1                                            NaN
2                                            NaN
3                                            NaN
4                                            NaN

  Femur upper neck BMD (bone mineral density) T-score (right) | Instance 2  \
0                                                    NaN
1                                                    NaN
2                                                   -0.9
3                                                    NaN
4                                                    NaN

   FRAX_bmd  Weight | Instance 0(participant - p21002_i0)  \
0       2.6                                          63.6
1       0.0                                          83.6
2       0.5                                          81.8
3       1.9                                          91.7
4       4.3                                          58.9

                       Diagnoses - secondary ICD10  \
0             E039|F171|F329|F419|I10|R33|Z864|Z911
1                                   T931|Z470|Z874
2  D758|E538|I269|J90|K029|K296|K573|K590|M8199|N…
3  A099|B370|B955|B968|B972|D649|E559|E669|E780|F…
4                               R298|Z866|Z871|Z888

   Femur total BMD (bone mineral density) (left) | Instance 2  \
0                                          0.760295
1                                          1.077890
2                                          0.929056
3                                          0.741237
4                                          0.619315
```

```
   Variation in diet | Instance 2  Diagnoses - ICD9  \
0                              3.0               NaN
1                              1.0               NaN
2                              2.0               NaN
3                              2.0               NaN
4                              2.0               NaN

  Ethnic background | Instance 0  Country of birth (UK/elsewhere) | Instance 0
0                          1001.0                                          1.0
1                          1001.0                                          1.0
2                          1001.0                                          1.0
3                          1001.0                                          1.0
4                          1001.0                                          1.0

[5 rows x 64 columns]
```

[7]:
```python
# Extracting unique prefixes from column names
instances = set(col.split(' | ')[0] for col in df.columns)

# Initialize a list to store columns with maximum non-missing values for each␣
 ↪prefix
columns = []

# Iterate over unique prefixes
for prefix in instances:
    # Select columns starting with the current prefix
    columns_to_check = [col for col in df.columns if col.startswith(prefix)]

    # Calculate counts of non-missing values for each selected column
    non_missing_counts = df[columns_to_check].count()

    # Find column with maximum non-missing values
    column_with_max_non_missing = non_missing_counts.idxmax()

    # Append column with maximum non-missing values to the list
    columns.append(column_with_max_non_missing)

# Print the length of the columns list
print("Number of selected columns:", len(columns))

# Print the contents of the columns list
print("Selected columns:", columns)
```

```
Number of selected columns: 64
Selected columns: ['Femur neck BMC (bone mineral content) (right) | Instance 2',
'Treatment/medication code | Instance 0', 'Femur neck BMD (bone mineral density)
(right) | Instance 2', 'fracture', 'Femur wards BMD (bone mineral density)
(right) | Instance 2', 'Illnesses of father | Instance 0', 'Femur troch BMD
```

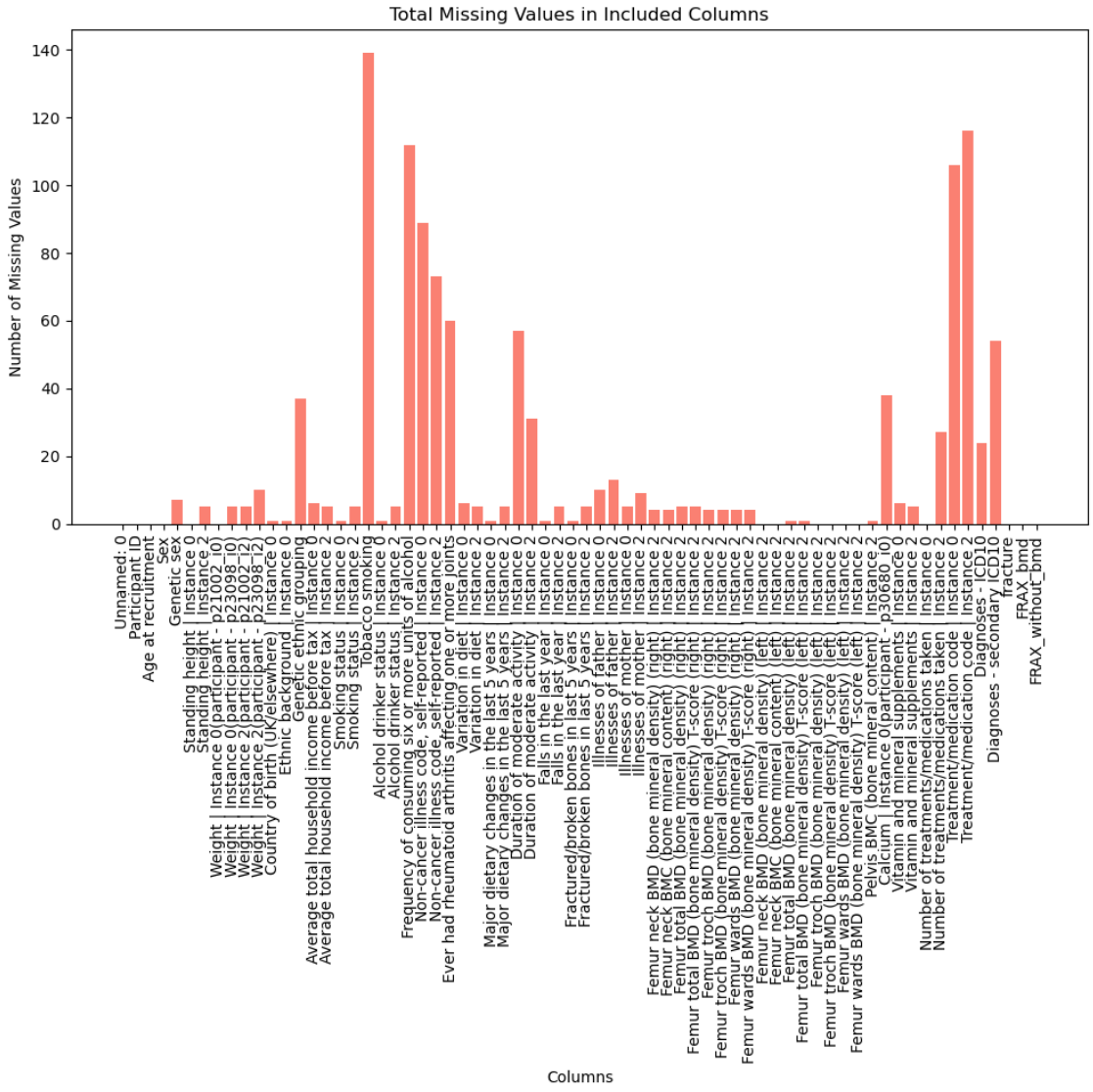
```
(bone mineral density) T-score (right) | Instance 2', 'Tobacco smoking', 'Femur
upper neck BMD (bone mineral density) T-score (left) | Instance 2', 'Type of
special diet followed | Instance 4', 'FRAX_without_bmd', 'Non-cancer illness
code, self-reported | Instance 2', 'Pelvis BMC (bone mineral content) | Instance
2', 'Frequency of consuming six or more units of alcohol', 'Number of
treatments/medications taken | Instance 0', 'Duration of moderate activity |
Instance 2', 'Femur shaft BMD (bone mineral density) T-score (left) | Instance
0', 'Femur upper neck BMD (bone mineral density) (right) | Instance 2', 'Vitamin
and mineral supplements | Instance 2', 'Genetic sex', 'Fractured/broken bones in
last 5 years | Instance 0', 'Age at recruitment', 'Femur neck BMC (bone mineral
content) (left) | Instance 2', 'Diagnoses - ICD10', 'Femur wards BMD (bone
mineral density) T-score (right) | Instance 2', 'Femur total BMD (bone mineral
density) (right) | Instance 2', 'Falls in the last year | Instance 0', 'Femur
troch BMD (bone mineral density) T-score (left) | Instance 2', 'Calcium |
Instance 0(participant - p30680_i0)', 'Standing height | Instance 0', 'Alcohol
drinker status | Instance 0', 'Femur shaft BMD (bone mineral density) T-score
(right) | Instance 0', 'Unnamed: 0', 'Participant ID', 'Fractured bone site(s) |
Instance 0', 'Femur lower neck BMD (bone mineral density) (right) | Instance 2',
'Smoking status | Instance 0', 'Femur upper neck BMD (bone mineral density)
(left) | Instance 2', 'Femur wards BMD (bone mineral density) T-score (left) |
Instance 2', 'Femur troch BMD (bone mineral density) (left) | Instance 2',
'Sex', 'Femur total BMD (bone mineral density) T-score (left) | Instance 2',
'Genetic ethnic grouping', 'Alcohol consumed | Instance 3', 'Femur lower neck
BMD (bone mineral density) (left) | Instance 2', 'Major dietary changes in the
last 5 years | Instance 0', 'Femur troch BMD (bone mineral density) (right) |
Instance 2', 'Femur total BMD (bone mineral density) T-score (right) | Instance
2', 'Illnesses of mother | Instance 0', 'Femur wards BMD (bone mineral density)
(left) | Instance 2', 'Diagnoses - secondary ICD9', 'Ever had rheumatoid
arthritis affecting one or more joints', 'Femur neck BMD (bone mineral density)
(left) | Instance 2', 'Average total household income before tax | Instance 2',
'Country of Birth (non-UK origin) | Instance 0', 'Femur upper neck BMD (bone
mineral density) T-score (right) | Instance 2', 'FRAX_bmd', 'Weight | Instance
0(participant - p21002_i0)', 'Diagnoses - secondary ICD10', 'Femur total BMD
(bone mineral density) (left) | Instance 2', 'Variation in diet | Instance 2',
'Diagnoses - ICD9', 'Ethnic background | Instance 0', 'Country of birth
(UK/elsewhere) | Instance 0']
```

[8]:
```python
# Calculate percentage of missing values in each column
missing_data = (df.isnull().sum() / len(df)) * 100

# Print the percentage of missing values for each column
print("Percentage of missing values in each column:")
print(missing_data)
```

```
Percentage of missing values in each column:
Unnamed: 0                      0.000000
Participant ID                  0.000000
Age at recruitment              0.000000
```

```
Sex                              0.000000
Genetic sex                      2.333333
                                      …
Diagnoses - ICD9                97.333333
Diagnoses - secondary ICD9      99.000000
fracture                         0.000000
FRAX_bmd                         0.000000
FRAX_without_bmd                 0.000000
Length: 160, dtype: float64
```

[10]:
```python
import matplotlib.pyplot as plt

X = []
y = []
excluded = []

# Iterate over the columns
for column in df.columns:
    # Check if the column has less than 150 missing values
    if df[column].isna().sum() < 150:
        X.append(df[column].isna().sum())
        y.append(column)
    else:
        # Add the column to the excluded list
        excluded.append(column)

# Remove excluded columns from the columns list
for column in excluded:
    df.drop(column, axis=1, inplace=True)

# Define a custom color for the bars
bar_color = 'salmon'

# Visualize the total missing values in the included columns with the specified␣
 ↪color
plt.figure(figsize=(10, 10))
plt.title("Total Missing Values in Included Columns")
plt.bar(y, X, color=bar_color)
plt.xticks(rotation=90)
plt.xlabel('Columns')
plt.ylabel('Number of Missing Values')
plt.tight_layout()
plt.show()
```



[11]:
```
df_selected_columns = df[columns].head()
print(df_selected_columns)
```

```
   Femur neck BMC (bone mineral content) (right) | Instance 2  \
0                                            3.40979
1                                            5.27111
2                                            5.45140
3                                            3.86018
4                                            3.29454

        Treatment/medication code | Instance 0  \
0             1140921600|1141191044|1140928274
1                                   1140884488
```

```
2                                          NaN
3                                          NaN
4  1189|1140909674|1141188442|1140876592|99999

   Femur neck BMD (bone mineral density) (right) | Instance 2  fracture  \
0                                          0.723206               True
1                                          0.868022               True
2                                          0.996041               True
3                                          0.687088               True
4                                          0.708290              False

   Femur wards BMD (bone mineral density) (right) | Instance 2  \
0                                          0.460578
1                                          0.717409
2                                          0.768620
3                                          0.436912
4                                          0.490887

  Illnesses of father | Instance 0  \
0                       8|6|1|-27
1                         8|2|-27
2                           2|-27
3                             3|2
4                           6|-27

   Femur troch BMD (bone mineral density) T-score (right) | Instance 2  \
0                                          -2.239100
1                                          -0.599261
2                                          -0.235811
3                                          -2.094850
4                                          -1.809550

   Tobacco smoking  FRAX_without_bmd  \
0              NaN               0.6
1            114.0               0.1
2              NaN               2.5
3            113.0               0.9
4              NaN               0.9

  Non-cancer illness code, self-reported | Instance 2  …  \
0                                          1226|1262   …
1                                               1465   …
2                                               1515   …
3                                    1459|1065|1526   …
4                                          1469|1538   …

   Ever had rheumatoid arthritis affecting one or more joints  \
0                                                  NaN
```

```
1                                                 0.0
2                                                 0.0
3                                                 NaN
4                                                 0.0

   Femur neck BMD (bone mineral density) (left) | Instance 2  \
0                                          0.710107
1                                          1.419990
2                                          1.015170
3                                          0.764439
4                                          0.655488

   Average total household income before tax | Instance 2  FRAX_bmd  \
0                                                 1.0           2.6
1                                                 4.0           0.0
2                                                 1.0           0.5
3                                                 2.0           1.9
4                                                 2.0           4.3

  Weight | Instance 0(participant - p21002_i0)  \
0                                        63.6
1                                        83.6
2                                        81.8
3                                        91.7
4                                        58.9

                       Diagnoses - secondary ICD10  \
0             E039|F171|F329|F419|I10|R33|Z864|Z911
1                                   T931|Z470|Z874
2  D758|E538|I269|J90|K029|K296|K573|K590|M8199|N…
3  A099|B370|B955|B968|B972|D649|E559|E669|E780|F…
4                              R298|Z866|Z871|Z888

   Femur total BMD (bone mineral density) (left) | Instance 2  \
0                                          0.760295
1                                          1.077890
2                                          0.929056
3                                          0.741237
4                                          0.619315

   Variation in diet | Instance 2  Ethnic background | Instance 0  \
0                             3.0                          1001.0
1                             1.0                          1001.0
2                             2.0                          1001.0
3                             2.0                          1001.0
4                             2.0                          1001.0

  Country of birth (UK/elsewhere) | Instance 0
```

```
0                                                 1.0
1                                                 1.0
2                                                 1.0
3                                                 1.0
4                                                 1.0

[5 rows x 50 columns]
```

[12]:
```
# Print the number of columns obtained after preprocessing
print("Number of columns after preprocessing:", len(columns))

# Print the list of column names
print("Columns obtained after preprocessing:")
print(columns)
```

```
Number of columns after preprocessing: 50
Columns obtained after preprocessing:
['Femur neck BMC (bone mineral content) (right) | Instance 2',
'Treatment/medication code | Instance 0', 'Femur neck BMD (bone mineral density)
(right) | Instance 2', 'fracture', 'Femur wards BMD (bone mineral density)
(right) | Instance 2', 'Illnesses of father | Instance 0', 'Femur troch BMD
(bone mineral density) T-score (right) | Instance 2', 'Tobacco smoking',
'FRAX_without_bmd', 'Non-cancer illness code, self-reported | Instance 2',
'Pelvis BMC (bone mineral content) | Instance 2', 'Frequency of consuming six or
more units of alcohol', 'Number of treatments/medications taken | Instance 0',
'Duration of moderate activity | Instance 2', 'Vitamin and mineral supplements |
Instance 2', 'Genetic sex', 'Fractured/broken bones in last 5 years | Instance
0', 'Age at recruitment', 'Femur neck BMC (bone mineral content) (left) |
Instance 2', 'Diagnoses - ICD10', 'Femur wards BMD (bone mineral density)
T-score (right) | Instance 2', 'Femur total BMD (bone mineral density) (right) |
Instance 2', 'Falls in the last year | Instance 0', 'Femur troch BMD (bone
mineral density) T-score (left) | Instance 2', 'Calcium | Instance 0(participant
- p30680_i0)', 'Standing height | Instance 0', 'Alcohol drinker status |
Instance 0', 'Unnamed: 0', 'Participant ID', 'Smoking status | Instance 0',
'Femur wards BMD (bone mineral density) T-score (left) | Instance 2', 'Femur
troch BMD (bone mineral density) (left) | Instance 2', 'Sex', 'Femur total BMD
(bone mineral density) T-score (left) | Instance 2', 'Genetic ethnic grouping',
'Major dietary changes in the last 5 years | Instance 0', 'Femur troch BMD (bone
mineral density) (right) | Instance 2', 'Femur total BMD (bone mineral density)
T-score (right) | Instance 2', 'Illnesses of mother | Instance 0', 'Femur wards
BMD (bone mineral density) (left) | Instance 2', 'Ever had rheumatoid arthritis
affecting one or more joints', 'Femur neck BMD (bone mineral density) (left) |
Instance 2', 'Average total household income before tax | Instance 2',
'FRAX_bmd', 'Weight | Instance 0(participant - p21002_i0)', 'Diagnoses -
secondary ICD10', 'Femur total BMD (bone mineral density) (left) | Instance 2',
'Variation in diet | Instance 2', 'Ethnic background | Instance 0', 'Country of
birth (UK/elsewhere) | Instance 0']
```

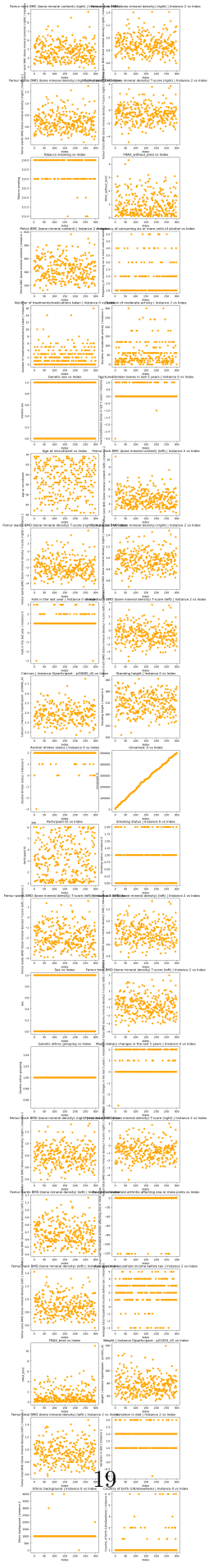
[13]:
```
df[columns].info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Data columns (total 50 columns):
 #   Column                                                                Non-
Null Count  Dtype
---  ------
--------------  -----
 0   Femur neck BMC (bone mineral content) (right) | Instance 2           296
non-null    float64
 1   Treatment/medication code | Instance 0                               194
non-null    object
 2   Femur neck BMD (bone mineral density) (right) | Instance 2           296
non-null    float64
 3   fracture                                                              300
non-null    bool
 4   Femur wards BMD (bone mineral density) (right) | Instance 2          296
non-null    float64
 5   Illnesses of father | Instance 0                                      290
non-null    object
 6   Femur troch BMD (bone mineral density) T-score (right) | Instance 2  296
non-null    float64
 7   Tobacco smoking                                                       161
non-null    float64
 8   FRAX_without_bmd                                                      300
non-null    float64
 9   Non-cancer illness code, self-reported | Instance 2                  227
non-null    object
 10  Pelvis BMC (bone mineral content) | Instance 2                       299
non-null    float64
 11  Frequency of consuming six or more units of alcohol                   188
non-null    float64
 12  Number of treatments/medications taken | Instance 0                 300
non-null    int64
 13  Duration of moderate activity | Instance 2                          269
non-null    float64
 14  Vitamin and mineral supplements | Instance 2                        295
non-null    object
 15  Genetic sex                                                           293
non-null    float64
 16  Fractured/broken bones in last 5 years | Instance 0                  299
non-null    float64
 17  Age at recruitment                                                    300
non-null    int64
 18  Femur neck BMC (bone mineral content) (left) | Instance 2            300
non-null    float64
```

```
 19  Diagnoses - ICD10                                                   276
non-null    object
 20  Femur wards BMD (bone mineral density) T-score (right) | Instance 2  296
non-null    float64
 21  Femur total BMD (bone mineral density) (right) | Instance 2          295
non-null    float64
 22  Falls in the last year | Instance 0                                  299
non-null    float64
 23  Femur troch BMD (bone mineral density) T-score (left) | Instance 2   300
non-null    float64
 24  Calcium | Instance 0(participant - p30680_i0)                        262
non-null    float64
 25  Standing height | Instance 0                                        300
non-null    float64
 26  Alcohol drinker status | Instance 0                                  299
non-null    float64
 27  Unnamed: 0                                                          300
non-null    int64
 28  Participant ID                                                      300
non-null    int64
 29  Smoking status | Instance 0                                         299
non-null    float64
 30  Femur wards BMD (bone mineral density) T-score (left) | Instance 2   300
non-null    float64
 31  Femur troch BMD (bone mineral density) (left) | Instance 2          300
non-null    float64
 32  Sex                                                                 300
non-null    int64
 33  Femur total BMD (bone mineral density) T-score (left) | Instance 2   299
non-null    float64
 34  Genetic ethnic grouping                                             263
non-null    float64
 35  Major dietary changes in the last 5 years | Instance 0              299
non-null    float64
 36  Femur troch BMD (bone mineral density) (right) | Instance 2         296
non-null    float64
 37  Femur total BMD (bone mineral density) T-score (right) | Instance 2  295
non-null    float64
 38  Illnesses of mother | Instance 0                                    295
non-null    object
 39  Femur wards BMD (bone mineral density) (left) | Instance 2          300
non-null    float64
 40  Ever had rheumatoid arthritis affecting one or more joints          240
non-null    float64
 41  Femur neck BMD (bone mineral density) (left) | Instance 2           300
non-null    float64
 42  Average total household income before tax | Instance 2             295
non-null    float64
```

```
 43  FRAX_bmd                                                    300
non-null    float64
 44  Weight | Instance 0(participant - p21002_i0)                300
non-null    float64
 45  Diagnoses - secondary ICD10                                 246
non-null    object
 46  Femur total BMD (bone mineral density) (left) | Instance 2  299
non-null    float64
 47  Variation in diet | Instance 2                              295
non-null    float64
 48  Ethnic background | Instance 0                              299
non-null    float64
 49  Country of birth (UK/elsewhere) | Instance 0                299
non-null    float64
dtypes: bool(1), float64(37), int64(5), object(7)
memory usage: 115.3+ KB
```

[14]:
```python
import itertools
import matplotlib.pyplot as plt

# Select only float64 and int64 columns
numeric_cols = df[columns].select_dtypes(include=['float64', 'int64']).columns

# Number of rows for the figure grid
n = len(numeric_cols)
ncols = 2
nrows = n // ncols
nrows += n % ncols

# Create a position index for the plots
position = list(itertools.product(range(nrows), range(ncols)))
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(10, 5*nrows))

# Define a custom color for the scatter plots
scatter_color = 'orange'

for col, pos in zip(numeric_cols, position):
    ax = axes[pos[0]][pos[1]]
    ax.scatter(df.index, df[col], color=scatter_color)
    ax.set_xlabel('Index')
    ax.set_ylabel(col)
    ax.set_title(f'{col} vs Index')

# Remove empty subplots
if len(numeric_cols) % 2 != 0:
    fig.delaxes(axes.flatten()[-1])
plt.show()
```


Femur neck BMC (bone mineral content) (right) | Instance 2 vs Index
Femur neck BMD (bone mineral density) (right) | Instance 2 vs Index
Femur neck BMD (bone mineral density) (right) | Instance 2 vs Index
Femur neck BMD (bone mineral density) T-score (right) | Instance 2 vs Index
Tobacco smoking vs Index
FRAX, without BMD vs Index
Pelvis BMC (bone mineral content) | Instance 2 vs Index
Number of consuming six or more units of alcohol vs Index
Number of treatments/medications taken | Instance 0 vs Index
Duration of moderate activity | Instance 2 vs Index
Genetic sex vs Index
Age at recruitment vs Index
Femur neck BMC (bone mineral content) (left) | Instance 2 vs Index
Femur neck BMD (bone mineral density) T-score (right) | Instance 2 vs Index
Falls in the last year | Instance 0 vs Index
Femur neck BMD (bone mineral density) T-score (left) | Instance 2 vs Index
Calcium | Instance 0|participant - p30680_i0 vs Index
Standing height | Instance 0 vs Index
Alcohol drinker status | Instance 0 vs Index
Townsend deprivation index at recruitment | Instance 0 vs Index
Participant ID vs Index
Smoking status | Instance 0 vs Index
Femur neck BMD (bone mineral density) T-score (left) | Instance 2 vs Index
Femur neck BMC (bone mineral content) (left) | Instance 2 vs Index
Sex vs Index
Femur total BMD (bone mineral density) T-score (left) | Instance 2 vs Index
Genetic ethnic grouping vs Index
Mouth dietary changes in the last 5 years | Instance 0 vs Index
Femur neck BMD (bone mineral density) (right) | Instance 2 vs Index
Femur total BMD (bone mineral density) T-score (right) | Instance 2 vs Index
Femur neck BMD (bone mineral density) (left) | Instance 2 vs Index
Diagnosed with arthritis affecting one or more joints vs Index
Femur neck BMD (bone mineral density) (left) | Instance 2 vs Index
Automatic income before tax | Instance 0 vs Index
FRAX_bmd vs Index
Weight | Instance 0|participant - p21002_i0 vs Index
Femur total BMD (bone mineral density) (left) | Instance 2 vs Index
Reaction time | Instance 0 vs Index
Ethnic background | Instance 0 vs Index
Country of birth (UK/elsewhere) | Instance 0 vs Index

[15]:
```python
# Define a function to handle outliers in a column using IQR method
def handle_outliers(column):
    Q1 = column.quantile(0.25)
    Q3 = column.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return column[(column > lower_bound) & (column < upper_bound)]

# Apply the function to handle outliers in each float column
float_cols = df.select_dtypes(include=['float64']).columns
for col in float_cols:
    df[col] = handle_outliers(df[col])

# Apply the function to handle outliers in each integer column
int_cols = df.select_dtypes(include=['int64']).columns
for col in int_cols:
    df[col] = handle_outliers(df[col])
```

[16]:
```python
# List of columns identified for removal due to little to no variation
to_drop = [
    ' Country of birth (UK/elsewhere) | Instance 0',
    ' Alcohol drinker status | Instance 0',
    'Ever had rheumatoid arthritis affecting one or more joints',
    'Ethnic background | Instance 0',
    'Genetic ethnic grouping',
    'Average total household income before tax | Instance 2',
    'Tobacco smoking'
]

# Extend the to_drop list with additional columns
to_drop.extend([
    'additional_column_1',
    'additional_column_2'
    # Add more columns as needed
])

# Filter out columns marked for removal from the original columns list
columns = [item for item in columns if item not in to_drop]
```

[18]:
```python
#Let's look at object columns
obj_cols = df[columns].select_dtypes(include=['object'])
obj_cols
```

[18]:

	Treatment/medication code | Instance 0
0	1140921600|1141191044|1140928274
1	1140884488
2	NaN
3	NaN
4	1189|1140909674|1141188442|1140876592|99999
..	…
295	NaN
296	1140879802|2038460150
297	1140883534
298	NaN
299	2038460150

	Illnesses of father | Instance 0
0	8|6|1|-27
1	8|2|-27
2	2|-27
3	3|2
4	6|-27
..	…
295	6|-27
296	NaN
297	-17|-27
298	2|-27
299	2|-27

	Non-cancer illness code, self-reported | Instance 2
0	1226|1262
1	1465
2	1515
3	1459|1065|1526
4	1469|1538
..	…
295	1111
296	1065|1286|1458|1078|1465
297	1473|1265|1081|1386|1066|1655|99999
298	1196
299	1111|1536|1312

	Vitamin and mineral supplements | Instance 2
0	-7
1	-7
2	-7
3	-7
4	7
..	…
295	-7

```
296                                                -7
297                                                -7
298                                                -7
299                                                 7

                                       Diagnoses - ICD10  \
0    E039|F171|F250|F329|F419|I10|R33|S7200|T814|W0…
1            M8725|N492|N508|S7210|T931|V134|Z470|Z874
2    D758|E538|I269|J90|K011|K029|K296|K30|K573|K57…
3    A099|B370|B955|B968|B972|D649|E559|E669|E780|F…
4               H001|I839|N950|R298|R42|Z866|Z871|Z888
..                                                    …
295                                                 NaN
296  F101|F314|F329|I10|K573|K579|M199|M819|R194|R9…
297   G439|H534|I632|I639|I694|Q211|R42|Z136|Z864|Z867
298  B968|F329|I10|N359|N390|S7200|W184|Z038|Z115|Z…
299  A099|D033|E780|G551|J459|M169|M179|M199|M2333|…

    Illnesses of mother | Instance 0  \
0                               12|10
1                             -17|-27
2                             8|1|-27
3                             -17|-27
4                             2|1|-27
..                                …
295                             9|-27
296                               NaN
297                           -17|-27
298                             2|-27
299                           10|2|-27

                          Diagnoses - secondary ICD10
0                E039|F171|F329|F419|I10|R33|Z864|Z911
1                                     T931|Z470|Z874
2    D758|E538|I269|J90|K029|K296|K573|K590|M8199|N…
3    A099|B370|B955|B968|B972|D649|E559|E669|E780|F…
4                                  R298|Z866|Z871|Z888
..                                                    …
295                                                 NaN
296  F101|F329|I10|K579|M199|M819|T404|T426|X619|X6…
297                  G439|H534|I694|Q211|R42|Z864|Z867
298   B968|F329|I10|N390|W184|Z115|Z501|Z507|Z864|Z874
299  E780|G551|J459|M169|M179|M199|M2578|M4806|M513…

[300 rows x 7 columns]
```

[19]:
```
# Print unique values for the specified columns
print(df['Vitamin and mineral supplements | Instance 2'].unique(), '\n')
print(df['Illnesses of mother | Instance 0'].unique(), '\n')
print(df['Illnesses of father | Instance 0'].unique(), '\n')
```

```
['-7' '7' '4' '4|7' nan '3|4' '2|3|7' '3|7' '3|4|6' '3' '2|3|4' '2' '1|4'
 '6' '3|4|7' '2|3|4|5|7' '2|7' '-3' '2|4' '4|5' '5' '2|6' '4|6' '1|2|5|6'
 '1|2|3|4|5|6|7' '6|7']

['12|10' '-17|-27' '8|1|-27' '2|1|-27' nan '10|-27' '-11|-27' '4|-17'
 '10|8|5' '8|-21' '1|-27' '11|2' '8|2|1|-27' '12|4|-17' '8|4' '9|8|-27'
 '8|-27' '2|-27' '10|9|-27' '11|10' '5|1' '11|5|-17' '8|2|-27' '11|-17'
 '12|8' '10|8|2|1|-27' '6|-27' '6|1|-27' '11|8' '8|5' '3|-17' '6|3'
 '9|-27' '9|8|2|1|-27' '12|-17' '5|-17' '9|8|5|2|1' '8|6|-27' '9|2|-27'
 '10|2|-21' '12|8|1' '-11|-21' '8|4|1' '9|8|3|1' '9|8|5|2' '8|2|1|-21'
 '12|9' '3|2' '4|1' '1|-21' '8|6|1|-27' '10|1|-27' '6|-21' '11|8|2'
 '9|8|4' '12|-11' '9|-21' '12|4|2|1' '10|8|-21' '8|5|2' '5|3|-17'
 '9|1|-27' '9|8|2|-27' '12|1' '9|8|-21' '10|2|-27']

['8|6|1|-27' '8|2|-27' '2|-27' '3|2' '6|-27' nan '3|-17' '8|1|-27' '4|-17'
 '2|1|-27' '-17|-27' '6|-21' '10|-27' '8|-27' '-11|-27' '8|2|-21' '6|3|1'
 '1|-27' '6|1|-27' '13|-17' '13|8|1' '10|8|1|-27' '9|1|-21' '12|2|1'
 '8|4|1' '12|-17' '8|6|2|-21' '-11|-21' '9|3|1' '9|-21' '3|-11' '4|1'
 '6|3' '9|8|-27' '10|9|2|-27' '9|-27' '12|9|1' '9|1|-27' '13|12|8|1'
 '9|2|1|-27' '9|2|-27' '13|9|8|6|1' '6|4|1' '10|6|-27' '-17|-21'
 '13|3|-17' '6|4' '1|-21' '12|4|-17' '9|8|1|-27' '9|8|2|-27' '12|9|8|1'
 '9|4' '8|2|1|-27' '12|9' '12|2' '10|9|-27' '13|1' '13|8' '8|6|-27'
 '10|-21' '11|-17' '13|9|8|2' '12|8' '8|3' '10|2|-27' '8|-21' '13|9|4|3|1'
 '12|8|1' '9|8|1|-21']
```

[20]:
```
# List of columns to drop
to_drop = ['Illnesses of father | Instance 0', 'Vitamin and mineral supplements␣
 ↪| Instance 2',
           'Illnesses of mother | Instance 0', 'Diagnoses - secondary ICD10',
           'Treatment/medication code | Instance 0']

# Remove specified columns from the list of columns
columns = [item for item in columns if item not in to_drop]

# Display the updated list of columns
columns
```

[20]:
```
['Femur neck BMC (bone mineral content) (right) | Instance 2',
 'Femur neck BMD (bone mineral density) (right) | Instance 2',
 'fracture',
 'Femur wards BMD (bone mineral density) (right) | Instance 2',
```

```
'Femur troch BMD (bone mineral density) T-score (right) | Instance 2',
'FRAX_without_bmd',
'Non-cancer illness code, self-reported | Instance 2',
'Pelvis BMC (bone mineral content) | Instance 2',
'Frequency of consuming six or more units of alcohol',
'Number of treatments/medications taken | Instance 0',
'Duration of moderate activity | Instance 2',
'Genetic sex',
'Fractured/broken bones in last 5 years | Instance 0',
'Age at recruitment',
'Femur neck BMC (bone mineral content) (left) | Instance 2',
'Diagnoses - ICD10',
'Femur wards BMD (bone mineral density) T-score (right) | Instance 2',
'Femur total BMD (bone mineral density) (right) | Instance 2',
'Falls in the last year | Instance 0',
'Femur troch BMD (bone mineral density) T-score (left) | Instance 2',
'Calcium | Instance 0(participant - p30680_i0)',
'Standing height | Instance 0',
'Alcohol drinker status | Instance 0',
'Unnamed: 0',
'Participant ID',
'Smoking status | Instance 0',
'Femur wards BMD (bone mineral density) T-score (left) | Instance 2',
'Femur troch BMD (bone mineral density) (left) | Instance 2',
'Sex',
'Femur total BMD (bone mineral density) T-score (left) | Instance 2',
'Major dietary changes in the last 5 years | Instance 0',
'Femur troch BMD (bone mineral density) (right) | Instance 2',
'Femur total BMD (bone mineral density) T-score (right) | Instance 2',
'Femur wards BMD (bone mineral density) (left) | Instance 2',
'Femur neck BMD (bone mineral density) (left) | Instance 2',
'FRAX_bmd',
'Weight | Instance 0(participant - p21002_i0)',
'Femur total BMD (bone mineral density) (left) | Instance 2',
'Variation in diet | Instance 2',
'Country of birth (UK/elsewhere) | Instance 0']
```

[21]:
```
# Copy the DataFrame with selected columns
df_copy = df[columns].copy()

# Select columns with data type 'object'
object_columns = df_copy.select_dtypes(include=["object"])

# Display the DataFrame containing only object columns
object_columns
```

[21]:
```
    Non-cancer illness code, self-reported | Instance 2  \
0                                                1226|1262
1                                                     1465
2                                                     1515
3                                          1459|1065|1526
4                                                1469|1538
..                                                     …
295                                                   1111
296                              1065|1286|1458|1078|1465
297                     1473|1265|1081|1386|1066|1655|99999
298                                                   1196
299                                          1111|1536|1312

                                       Diagnoses - ICD10
0    E039|F171|F250|F329|F419|I10|R33|S7200|T814|W0…
1            M8725|N492|N508|S7210|T931|V134|Z470|Z874
2    D758|E538|I269|J90|K011|K029|K296|K30|K573|K57…
3    A099|B370|B955|B968|B972|D649|E559|E669|E780|F…
4              H001|I839|N950|R298|R42|Z866|Z871|Z888
..                                                   …
295                                                NaN
296  F101|F314|F329|I10|K573|K579|M199|M819|R194|R9…
297   G439|H534|I632|I639|I694|Q211|R42|Z136|Z864|Z867
298  B968|F329|I10|N359|N390|S7200|W184|Z038|Z115|Z…
299  A099|D033|E780|G551|J459|M169|M179|M199|M2333|…

[300 rows x 2 columns]
```

[23]:
```
import pandas as pd
import numpy as np

# Expand the entries in the 'Diagnoses - ICD10' column into separate columns␣
 ↪and create dummy variables
df_copy['Diagnoses - ICD10'] = df_copy['Diagnoses - ICD10'].apply(lambda x: x.
 ↪split('|') if pd.notnull(x) else np.nan)
df1 = df_copy['Diagnoses - ICD10'].str.join('|').str.get_dummies()

# Do the same for the cancer illness column
df_copy['Non-cancer illness code, self-reported | Instance 2'] =␣
 ↪df_copy['Non-cancer illness code, self-reported | Instance 2'].apply(lambda␣
 ↪x: x.split('|')  if pd.notnull(x) else np.nan)
df2 = df_copy['Non-cancer illness code, self-reported | Instance 2'].str.
 ↪join('|').str.get_dummies()

df1.head()
```

[23]:
```
   A039  A048  A084  A099  A415  A418  A419  A498  B07  B181  …  Z954  Z955  \
0     0     0     0     0     0     0     0     0    0     0  …     0     0
1     0     0     0     0     0     0     0     0    0     0  …     0     0
2     0     0     0     0     0     0     0     0    0     0  …     0     0
3     0     0     0     1     0     0     0     0    0     0  …     0     0
4     0     0     0     0     0     0     0     0    0     0  …     0     0

   Z958  Z960  Z961  Z966  Z967  Z968  Z974  Z988
0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     0     0
2     0     0     0     0     0     0     0     0
3     0     0     0     1     0     0     0     0
4     0     0     0     0     0     0     0     0

[5 rows x 1157 columns]
```

[24]:
```
df2.head()
```

[24]:
```
   1065  1066  1072  1074  1075  1078  1081  1082  1093  1094  …  1653  \
0     0     0     0     0     0     0     0     0     0     0  …     0
1     0     0     0     0     0     0     0     0     0     0  …     0
2     0     0     0     0     0     0     0     0     0     0  …     0
3     1     0     0     0     0     0     0     0     0     0  …     0
4     0     0     0     0     0     0     0     0     0     0  …     0

   1655  1656  1657  1665  1666  1668  1670  1677  99999
0     0     0     0     0     0     0     0     0      0
1     0     0     0     0     0     0     0     0      0
2     0     0     0     0     0     0     0     0      0
3     0     0     0     0     0     0     0     0      0
4     0     0     0     0     0     0     0     0      0

[5 rows x 186 columns]
```

[25]:
```
# Concatenate all three dataframes together along the columns (axis=1)
df_final = pd.concat([df_copy, df1, df2], axis=1)

# Drop the original columns 'Diagnoses - ICD10' and 'Non-cancer illness code,␣
 ↪self-reported | Instance 2'
# axis=1 indicates that we're dropping columns, inplace=True modifies the␣
 ↪DataFrame in place
df_final.drop(['Diagnoses - ICD10', 'Non-cancer illness code, self-reported |␣
 ↪Instance 2'], axis=1, inplace=True)

# Display the information about the final DataFrame
df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Columns: 1381 entries, Femur neck BMC (bone mineral content) (right) | Instance
2 to 99999
dtypes: bool(1), float64(33), int64(1347)
memory usage: 3.2 MB
```

[26]:
```
# Create a cross-tabulation of 'fracture' by 'Sex'
cross_sex = pd.crosstab(df['Sex'], df['fracture'])

# Display the cross-tabulation
cross_sex
```

[26]:

fracture	False	True
Sex		
0	88	70
1	99	43

[27]:
```
from sklearn.feature_selection import VarianceThreshold

# Define the threshold for variance
threshold = 0.01

# Create a VarianceThreshold object
selector = VarianceThreshold(threshold=threshold)

# Fit the selector to the DataFrame
selector.fit(df_final)

# Get the boolean array indicating selected features
features = selector.get_support()

# Get the names of the selected features
kept_features = df_final.columns[features]

# Update the DataFrame to contain only selected features
df_final = df_final[kept_features]

# Display the DataFrame with selected features
df_final
```

```
C:\Users\krestty\anaconda3\Lib\site-
packages\sklearn\feature_selection\_variance_threshold.py:112: RuntimeWarning:
Degrees of freedom <= 0 for slice.
  self.variances_ = np.nanvar(X, axis=0)
```

[27]:

	Femur neck BMC (bone mineral content) (right) | Instance 2	Femur neck BMD (bone mineral density) (right) | Instance 2	fracture	Femur wards BMD (bone mineral density) (right) | Instance 2	Femur troch BMD (bone mineral density) T-score (right) | Instance 2
0	3.40979	0.723206	True	0.460578	-2.239100
1	5.27111	0.868022	True	0.717409	-0.599261
2	5.45140	0.996041	True	0.768620	-0.235811
3	3.86018	0.687088	True	0.436912	-2.094850
4	3.29454	0.708290	False	0.490887	-1.809550
..	…	…	…	…	…
295	4.15081	0.942686	False	0.767654	0.477970
296	2.88761	0.593618	True	0.433743	
297	5.59233	1.179270	False	0.871136	
298	NaN	NaN	True	NaN	
299	4.66290	0.801873	True	0.507284	

```
296                                        -1.331690
297                                         0.162941
298                                              NaN
299                                        -1.909560

     FRAX_without_bmd  Pelvis BMC (bone mineral content) | Instance 2  \
0                 0.6                                          196.965
1                 0.1                                          384.824
2                 2.5                                          322.331
3                 0.9                                          333.434
4                 0.9                                          206.776
..                …                                              …
295               0.4                                          264.125
296               1.3                                          198.665
297               0.1                                          377.965
298               1.6                                          311.616
299               2.6                                          345.933

     Frequency of consuming six or more units of alcohol  \
0                                                    NaN
1                                                    2.0
2                                                    1.0
3                                                    NaN
4                                                    1.0
..                                                   …
295                                                  NaN
296                                                  NaN
297                                                  3.0
298                                                  2.0
299                                                  1.0

     Number of treatments/medications taken | Instance 0  \
0                                                    3.0
1                                                    1.0
2                                                    0.0
3                                                    0.0
4                                                    5.0
..                                                   …
295                                                  0.0
296                                                  2.0
297                                                  1.0
298                                                  0.0
299                                                  1.0

     Duration of moderate activity | Instance 2  …  1567  1568  1569  1570  \
0                                          -1.0  …     0     0     0     0
1                                          10.0  …     0     0     0     0
```

```
2                                                  30.0  …     0     0     0     0
3                                                  20.0  …     0     0     0     0
4                                                  -1.0  …     0     0     0     0
..                                                  …    …     …     …     …     …
295                                                60.0  …     0     0     0     0
296                                                60.0  …     0     0     0     0
297                                                60.0  …     0     0     0     0
298                                                60.0  …     0     0     0     0
299                                                -1.0  …     0     0     0     0

     1571  1597  1598  1617  1637  99999
0       0     0     0     0     0      0
1       0     0     0     0     0      0
2       0     0     0     0     0      0
3       0     0     0     0     0      0
4       0     0     0     0     0      0
..    …     …     …     …     …      …
295     0     0     0     0     0      0
296     0     0     0     0     0      0
297     0     0     0     0     0      1
298     0     0     0     0     0      0
299     0     0     0     0     0      0

[300 rows x 336 columns]
```

[28]:
```python
# Calculate correlation matrix
correlation_matrix = df_final.corr()

# Extract correlation with the target variable 'fracture' and sort by absolute␣
 ↪values in descending order
fracture_correlation = correlation_matrix['fracture'].drop('fracture')  # Drop␣
 ↪'fracture' column for self-correlation
fracture_correlation = fracture_correlation.abs().sort_values(ascending=False)

# Print correlation with likelihood of hip fractures
print("Correlation with likelihood of hip fractures:")
print(fracture_correlation)

# Define threshold for correlation
threshold = 0.2

# Identify features with correlation below the threshold
unnecessary_features = fracture_correlation[fracture_correlation < threshold].
 ↪index

# Remove unnecessary features from the DataFrame
df_final.drop(columns=unnecessary_features, inplace=True)
```

```
# Print DataFrame after removing unnecessary features
print("\nDataFrame after removing unnecessary features:")
print(df_final)
```

```
Correlation with likelihood of hip fractures:
S7200                                                                 0.782352
Femur total BMD (bone mineral density) T-score (left) | Instance 2    0.500803
Femur total BMD (bone mineral density) T-score (right) | Instance 2   0.491086
Z501                                                                  0.490624
Femur total BMD (bone mineral density) (left) | Instance 2            0.484233
                                                                        ...
Unnamed: 0                                                            0.001387
R31                                                                   0.000816
1396                                                                  0.000816
1154                                                                  0.000569
M1399                                                                 0.000569
Name: fracture, Length: 335, dtype: float64

DataFrame after removing unnecessary features:
     Femur neck BMC (bone mineral content) (right) | Instance 2  \
0                                                      3.40979
1                                                      5.27111
2                                                      5.45140
3                                                      3.86018
4                                                      3.29454
..                                                         ...
295                                                    4.15081
296                                                    2.88761
297                                                    5.59233
298                                                        NaN
299                                                    4.66290

     Femur neck BMD (bone mineral density) (right) | Instance 2  fracture  \
0                                                     0.723206        True
1                                                     0.868022        True
2                                                     0.996041        True
3                                                     0.687088        True
4                                                     0.708290       False
..                                                         ...         ...
295                                                   0.942686       False
296                                                   0.593618        True
297                                                   1.179270       False
298                                                        NaN        True
299                                                   0.801873        True

     Femur wards BMD (bone mineral density) (right) | Instance 2  \
```

```
0                                                 0.460578
1                                                 0.717409
2                                                 0.768620
3                                                 0.436912
4                                                 0.490887
..                                                     …
295                                               0.767654
296                                               0.433743
297                                               0.871136
298                                                    NaN
299                                               0.507284

     Femur troch BMD (bone mineral density) T-score (right) | Instance 2  \
0                                                -2.239100
1                                                -0.599261
2                                                -0.235811
3                                                -2.094850
4                                                -1.809550
..                                                     …
295                                               0.477970
296                                              -1.331690
297                                               0.162941
298                                                    NaN
299                                              -1.909560

     FRAX_without_bmd  Pelvis BMC (bone mineral content) | Instance 2  \
0                 0.6                                          196.965
1                 0.1                                          384.824
2                 2.5                                          322.331
3                 0.9                                          333.434
4                 0.9                                          206.776
..                …                                              …
295               0.4                                          264.125
296               1.3                                          198.665
297               0.1                                          377.965
298               1.6                                          311.616
299               2.6                                          345.933

     Age at recruitment  \
0                    58
1                    42
2                    69
3                    62
4                    58
..                  …
295                  56
296                  65
297                  43
```

```
298                  62
299                  68

     Femur neck BMC (bone mineral content) (left) | Instance 2  \
0                                                    3.27879
1                                                        NaN
2                                                    5.62017
3                                                    4.33504
4                                                    3.56223
..                                                       …
295                                                  4.26806
296                                                  2.80757
297                                                  5.18932
298                                                  3.86859
299                                                  4.35525

     Femur wards BMD (bone mineral density) T-score (right) | Instance 2  …  \
0                                                  -3.457090                 …
1                                                  -1.866090                 …
2                                                  -1.087540                 …
3                                                  -4.023750                 …
4                                                  -3.223950                 …
..                                                       …                 …
295                                                -1.094970                 …
296                                                -3.663520                 …
297                                                -0.683568                 …
298                                                      NaN                 …
299                                                -3.482430                 …

     V184  W010  W019  W190  Z115  Z130  Z470  Z501  Z507  Z966
0       0     0     0     0     0     0     0     0     0     0
1       0     0     0     0     0     0     1     0     0     0
2       0     0     0     0     1     1     0     0     0     0
3       0     1     0     0     0     0     0     1     0     1
4       0     0     0     0     0     0     0     0     0     0
..    …   …   …   …   …   …   …   …   …   …
295     0     0     0     0     0     0     0     0     0     0
296     0     1     0     0     0     0     0     0     0     1
297     0     0     0     0     0     0     0     0     0     0
298     0     0     0     0     1     0     0     1     1     0
299     0     0     0     0     1     0     0     1     1     1

[300 rows x 35 columns]
```

[30]:
```
# Select numeric columns
numeric_cols = df_final.select_dtypes(include=['float64', 'int64']).columns
```

```
# Fill missing values with the mean of each numeric column
df_final[numeric_cols] = df_final[numeric_cols].fillna(df_final[numeric_cols].
  ↪mean())

# Display the first few rows of the updated DataFrame
print(df_final.head())
```

```
   Femur neck BMC (bone mineral content) (right) | Instance 2  \
0                                           3.40979
1                                           5.27111
2                                           5.45140
3                                           3.86018
4                                           3.29454

   Femur neck BMD (bone mineral density) (right) | Instance 2  fracture  \
0                                          0.723206               True
1                                          0.868022               True
2                                          0.996041               True
3                                          0.687088               True
4                                          0.708290              False

   Femur wards BMD (bone mineral density) (right) | Instance 2  \
0                                          0.460578
1                                          0.717409
2                                          0.768620
3                                          0.436912
4                                          0.490887

   Femur troch BMD (bone mineral density) T-score (right) | Instance 2  \
0                                         -2.239100
1                                         -0.599261
2                                         -0.235811
3                                         -2.094850
4                                         -1.809550

   FRAX_without_bmd  Pelvis BMC (bone mineral content) | Instance 2  \
0               0.6                                          196.965
1               0.1                                          384.824
2               2.5                                          322.331
3               0.9                                          333.434
4               0.9                                          206.776

   Age at recruitment  \
0                  58
1                  42
2                  69
3                  62
```

```
4                  58

   Femur neck BMC (bone mineral content) (left) | Instance 2  \
0                                                3.278790
1                                                4.708105
2                                                5.620170
3                                                4.335040
4                                                3.562230

   Femur wards BMD (bone mineral density) T-score (right) | Instance 2  …  \
0                                                -3.45709                  …
1                                                -1.86609                  …
2                                                -1.08754                  …
3                                                -4.02375                  …
4                                                -3.22395                  …

   V184  W010  W019  W190  Z115  Z130  Z470  Z501  Z507  Z966
0     0     0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     1     0     0     0
2     0     0     0     0     1     1     0     0     0     0
3     0     1     0     0     0     0     0     1     0     1
4     0     0     0     0     0     0     0     0     0     0

[5 rows x 35 columns]
```

[31]:
```
df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Data columns (total 35 columns):
 #   Column                                                                 Non-
Null Count  Dtype
---  ------
--------------  -----
 0   Femur neck BMC (bone mineral content) (right) | Instance 2           300
non-null    float64
 1   Femur neck BMD (bone mineral density) (right) | Instance 2           300
non-null    float64
 2   fracture                                                               300
non-null    bool
 3   Femur wards BMD (bone mineral density) (right) | Instance 2          300
non-null    float64
 4   Femur troch BMD (bone mineral density) T-score (right) | Instance 2  300
non-null    float64
 5   FRAX_without_bmd                                                       300
non-null    float64
 6   Pelvis BMC (bone mineral content) | Instance 2                        300
non-null    float64
```

```
 7   Age at recruitment                                                 300
non-null    int64
 8   Femur neck BMC (bone mineral content) (left) | Instance 2          300
non-null    float64
 9   Femur wards BMD (bone mineral density) T-score (right) | Instance 2  300
non-null    float64
 10  Femur total BMD (bone mineral density) (right) | Instance 2        300
non-null    float64
 11  Femur troch BMD (bone mineral density) T-score (left) | Instance 2   300
non-null    float64
 12  Femur wards BMD (bone mineral density) T-score (left) | Instance 2   300
non-null    float64
 13  Femur troch BMD (bone mineral density) (left) | Instance 2          300
non-null    float64
 14  Femur total BMD (bone mineral density) T-score (left) | Instance 2   300
non-null    float64
 15  Femur troch BMD (bone mineral density) (right) | Instance 2         300
non-null    float64
 16  Femur total BMD (bone mineral density) T-score (right) | Instance 2  300
non-null    float64
 17  Femur wards BMD (bone mineral density) (left) | Instance 2          300
non-null    float64
 18  Femur neck BMD (bone mineral density) (left) | Instance 2           300
non-null    float64
 19  FRAX_bmd                                                           300
non-null    float64
 20  Femur total BMD (bone mineral density) (left) | Instance 2          300
non-null    float64
 21  J189                                                               300
non-null    int64
 22  M819                                                               300
non-null    int64
 23  S7200                                                              300
non-null    int64
 24  S7210                                                              300
non-null    int64
 25  V184                                                               300
non-null    int64
 26  W010                                                               300
non-null    int64
 27  W019                                                               300
non-null    int64
 28  W190                                                               300
non-null    int64
 29  Z115                                                               300
non-null    int64
 30  Z130                                                               300
non-null    int64
```

```
 31  Z470                                                            300
non-null    int64
 32  Z501                                                            300
non-null    int64
 33  Z507                                                            300
non-null    int64
 34  Z966                                                            300
non-null    int64
dtypes: bool(1), float64(19), int64(15)
memory usage: 80.1 KB
```

[32]:
```
X = df_final.copy()
X.drop(['fracture','FRAX_bmd', 'FRAX_without_bmd'], axis=1, inplace=True)
y = df_final['fracture'].astype('int')
X, y
```

[32]:
```
(     Femur neck BMC (bone mineral content) (right) | Instance 2  \
 0                                                    3.409790
 1                                                    5.271110
 2                                                    5.451400
 3                                                    3.860180
 4                                                    3.294540
 ..                                                        …
 295                                                  4.150810
 296                                                  2.887610
 297                                                  5.592330
 298                                                  4.693209
 299                                                  4.662900

      Femur neck BMD (bone mineral density) (right) | Instance 2  \
 0                                                    0.723206
 1                                                    0.868022
 2                                                    0.996041
 3                                                    0.687088
 4                                                    0.708290
 ..                                                        …
 295                                                  0.942686
 296                                                  0.593618
 297                                                  1.179270
 298                                                  0.886000
 299                                                  0.801873

      Femur wards BMD (bone mineral density) (right) | Instance 2  \
 0                                                    0.460578
 1                                                    0.717409
 2                                                    0.768620
 3                                                    0.436912
```

```
4                                                 0.490887
..                                                     …
295                                               0.767654
296                                               0.433743
297                                               0.871136
298                                               0.672579
299                                               0.507284

     Femur troch BMD (bone mineral density) T-score (right) | Instance 2  \
0                                                -2.239100
1                                                -0.599261
2                                                -0.235811
3                                                -2.094850
4                                                -1.809550
..                                                     …
295                                               0.477970
296                                              -1.331690
297                                               0.162941
298                                              -0.616048
299                                              -1.909560

     Pelvis BMC (bone mineral content) | Instance 2  Age at recruitment  \
0                                            196.965                  58
1                                            384.824                  42
2                                            322.331                  69
3                                            333.434                  62
4                                            206.776                  58
..                                               …                   …
295                                          264.125                  56
296                                          198.665                  65
297                                          377.965                  43
298                                          311.616                  62
299                                          345.933                  68

     Femur neck BMC (bone mineral content) (left) | Instance 2  \
0                                                  3.278790
1                                                  4.708105
2                                                  5.620170
3                                                  4.335040
4                                                  3.562230
..                                                     …
295                                                4.268060
296                                                2.807570
297                                                5.189320
298                                                3.868590
299                                                4.355250
```

```
     Femur wards BMD (bone mineral density) T-score (right) | Instance 2  \
0                                                  -3.457090
1                                                  -1.866090
2                                                  -1.087540
3                                                  -4.023750
4                                                  -3.223950
..                                                       …
295                                                -1.094970
296                                                -3.663520
297                                                -0.683568
298                                                -2.013155
299                                                -3.482430

     Femur total BMD (bone mineral density) (right) | Instance 2  \
0                                                   0.723348
1                                                   0.970351
2                                                   0.973764
3                                                   0.763192
4                                                   0.757765
..                                                       …
295                                                 1.036840
296                                                 0.737125
297                                                 1.165070
298                                                 0.946033
299                                                 0.857143

     Femur troch BMD (bone mineral density) T-score (left) | Instance 2  …  \
0                                                  -1.658510                 …
1                                                   0.129092                 …
2                                                  -0.859427                 …
3                                                  -2.649160                 …
4                                                  -1.824990                 …
..                                                       …                 …
295                                                 0.432288                 …
296                                                -1.493160                 …
297                                                -0.466274                 …
298                                                -1.161010                 …
299                                                -2.272810                 …

     V184  W010  W019  W190  Z115  Z130  Z470  Z501  Z507  Z966
0       0     0     0     0     0     0     0     0     0     0
1       0     0     0     0     0     0     1     0     0     0
2       0     0     0     0     1     1     0     0     0     0
3       0     1     0     0     0     0     0     1     0     1
4       0     0     0     0     0     0     0     0     0     0
..    …   …   …   …   …   …   …   …   …   …
295     0     0     0     0     0     0     0     0     0     0
```

```
296     0     1     0     0     0     0     0     0     0     1
297     0     0     0     0     0     0     0     0     0     0
298     0     0     0     0     1     0     0     1     1     0
299     0     0     0     0     1     0     0     1     1     1

[300 rows x 32 columns],
0      1
1      1
2      1
3      1
4      0
      ..
295    0
296    1
297    0
298    1
299    1
Name: fracture, Length: 300, dtype: int32)
```

[33]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y,␣
 ↪test_size=0.2, train_size=0.8, random_state=42)
```

[35]:
```
#Cross Validation
from sklearn.model_selection import KFold
from sklearn.ensemble import RandomForestClassifier

kf = KFold(n_splits=5, shuffle=True)
kf
```

[35]:
```
KFold(n_splits=5, random_state=None, shuffle=True)
```

[37]:
```
from sklearn.ensemble import RandomForestClassifier

# Create an instance of RandomForestClassifier
model = RandomForestClassifier()

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test data
predictions = model.predict(X_test)
```

[38]:
```
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score,␣
 ↪recall_score, f1_score
```

```
def calculate_metrics(true, predictions):
    # Calculate confusion matrix
    grid_cm = confusion_matrix(true, predictions)

    # Extract TP, TN, FP, FN from confusion matrix
    tn, fp, fn, tp = grid_cm.ravel()

    # Calculate specificity, sensitivity, precision, f1 score, and accuracy
    grid_specificity = tn / (tn + fp)
    grid_sensitivity = tp / (tp + fn)
    grid_precision = precision_score(true, predictions)
    grid_f1 = f1_score(true, predictions)
    grid_accuracy = accuracy_score(true, predictions)

    # Print confusion matrix
    print("The Confusion Matrix is:\n")
    print(grid_cm)

    # Return a list containing the calculated metrics
    return [(grid_accuracy, grid_sensitivity, grid_specificity, grid_precision,␣
 ↪grid_f1)]

# Example usage:
# metrics = calculate_metrics(true_labels, predicted_labels)
```

[52]:
```
from sklearn.model_selection import KFold

# Define the number of splits for cross-validation
n_splits = 5

# Initialize KFold object
kf = KFold(n_splits=n_splits, shuffle=True)
```

[48]:
```
# Use Gridsearch to find the best hyperparameters
from sklearn.model_selection import GridSearchCV

# Define the parameters grid
params = {
    'n_estimators': [40, 50, 100],
    'max_features': ['sqrt', 'log2'],
    'max_depth': [4, 5, 6, 8],
    'min_samples_split': [10, 12, 13],
    'min_samples_leaf': [6, 8, 10]
}

# Instantiate GridSearchCV
grid_search = GridSearchCV(model, param_grid=params, cv=kf, scoring='recall')
```

```
# Fit the model
grid_search.fit(X_train, y_train)

# Get the best hyperparameters
best_params = grid_search.best_params_

# Get the best estimator
best_model = grid_search.best_estimator_

# Get the best score
best_score = grid_search.best_score_

# Get the CV results
cv_results = grid_search.cv_results_

# Print the best parameters and best score
print("Best Hyperparameters:", best_params)
print("Best Score:", best_score)
```

```
Best Hyperparameters: {'max_depth': 5, 'max_features': 'log2',
'min_samples_leaf': 6, 'min_samples_split': 12, 'n_estimators': 50}
Best Score: 0.8156625074272134
```

[50]:
```
#Use Gridsearch to find the best hyper parameters
from sklearn.model_selection import GridSearchCV,RandomizedSearchCV
from  random import randint
params = {
    'n_estimators' : [40,50,100],
    'max_features': [ 'sqrt', 'log2'],
    'max_depth': [4,5,6,8],
    'min_samples_split': [10,12,13],
    'min_samples_leaf': [6,8,10],

}
#Fit the model
grid_search = GridSearchCV(model, param_grid=params, cv = kf, scoring='recall').
 ↪fit(X_train, y_train)
grid_search
```

[50]:
```
GridSearchCV(cv=KFold(n_splits=5, random_state=None, shuffle=True),
             estimator=RandomForestClassifier(),
             param_grid={'max_depth': [4, 5, 6, 8],
                         'max_features': ['sqrt', 'log2'],
                         'min_samples_leaf': [6, 8, 10],
                         'min_samples_split': [10, 12, 13],
                         'n_estimators': [40, 50, 100]},
```

```
                scoring='recall')
```

[51]:
```
print('Best parameters:', grid_search.best_params_)
print('Best score:', grid_search.best_score_)
```

```
Best parameters: {'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 6,
'min_samples_split': 13, 'n_estimators': 40}
Best score: 0.8358434547908233
```

[53]:
```
# Predict target variable for the training set
y_pred_tr = grid_search.best_estimator_.predict(X_train)

# Predict target variable for the test set
y_pred = grid_search.best_estimator_.predict(X_test)
```

[55]:
```
train_metrics = calculate_metrics(true=y_train, predictions=y_pred_tr)
test_metrics = calculate_metrics(true=y_test, predictions=y_pred)

grid_scores = pd.DataFrame(data=train_metrics + test_metrics,␣
 ↪columns=['Accuracy', 'Sensitivity', 'Specificity', 'Precision', 'F1 Score'])
grid_scores['Model'] = ["Random Forest", 'Random Forest']
grid_scores['Data'] = ['Train Data', 'Test Data']

grid_scores
```

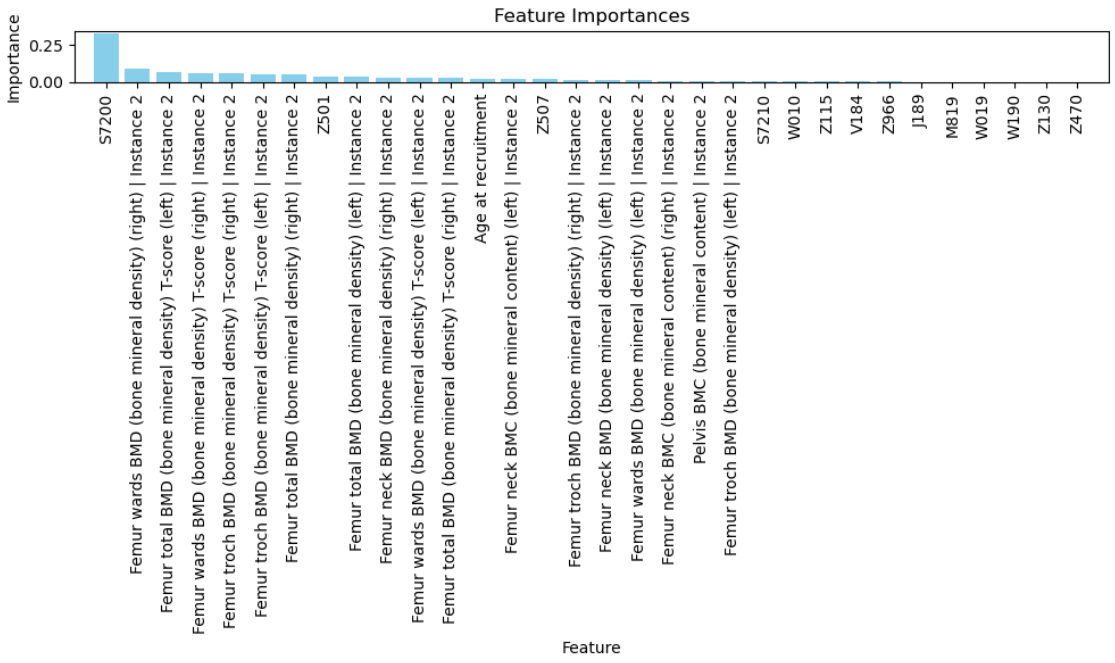
```
The Confusion Matrix is:

[[144   6]
 [ 15  75]]
The Confusion Matrix is:

[[34  3]
 [ 5 18]]
```

[55]:

	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Model	Data
0	0.912500	0.833333	0.960000	0.925926	0.877193	Random Forest	Train Data
1	0.866667	0.782609	0.918919	0.857143	0.818182	Random Forest	Test Data

[57]:
```
best_model = grid_search.best_estimator_

# Get feature importances
importances = best_model.feature_importances_
```

```
# Get feature names
feature_names = X_train.columns

# Sort feature importances in descending order
indices = np.argsort(importances)[::-1]

# Plot
plt.figure(figsize=(10, 6))
plt.title("Feature Importances")
plt.bar(range(X_train.shape[1]), importances[indices], align="center",␣
 ↪color='skyblue')  # Change color here
plt.xticks(range(X_train.shape[1]), [feature_names[i] for i in indices],␣
 ↪rotation=90)
plt.xlim([-1, X_train.shape[1]])
plt.xlabel("Feature")
plt.ylabel("Importance")
plt.tight_layout()
plt.show()
```



[ ]:
```
from sklearn.model_selection import cross_val_score

score2 = cross_val_score(grid_search, X_train, y_train, cv=kf, scoring='recall')
grid_cv_score = score2.mean()
grid_cv_stdev = stdev(score2)
print('Cross Validation Recall scores are: {}'.format(score2))
```

```
print('Average Cross Validation Recall score: ', grid_cv_score)
print('Cross Validation Recall standard deviation: ', grid_cv_stdev)
```

[ ]: