diabetes

July 26, 2024

[176]:
```
import sqlite3
```

[177]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

[178]:
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,␣
 ↪roc_auc_score
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn import datasets
from sklearn.model_selection import train_test_split
```

0.1 Load the data

[181]:
```
# Load the dataset
diabetes_df = pd.read_csv("C:/Users/deeps/OneDrive/Documents/WEBSTER/DATASET/
 ↪Excel/diabetes_data.csv")

# Display the first few rows of the dataset
diabetes_df.head(10)
```

[181]:

	PatientID	Age	Gender	Ethnicity	SocioeconomicStatus	EducationLevel	\
0	6000	44	0	1	2	1	
1	6001	51	1	0	1	2	
2	6002	89	1	0	1	3	
3	6003	21	1	1	1	2	
4	6004	27	1	0	1	3	
5	6005	65	0	0	0	0	
6	6006	61	1	2	1	3	
7	6007	74	1	3	0	3	
8	6008	54	0	0	1	2	
9	6009	82	1	0	1	1	

```
         BMI  Smoking  AlcoholConsumption  PhysicalActivity  …  \
0  32.985284        1            4.499365          2.443385  …
1  39.916764        0            1.578919          8.301264  …
2  19.782251        0            1.177301          6.103395  …
3  32.376881        1            1.714621          8.645465  …
4  16.808600        0           15.462549          4.629383  …
5  15.820815        1           17.781024          9.252522  …
6  20.075147        0            1.086479          8.745650  …
7  29.438938        0            6.187378          9.114535  …
8  15.027557        0           19.505734          0.590771  …
9  34.300044        1           15.943844          6.056621  …

   TinglingHandsFeet  QualityOfLifeScore  HeavyMetalsExposure  \
0                  1           73.765109                    0
1                  0           91.445753                    0
2                  0           54.485744                    0
3                  0           77.866758                    0
4                  0           37.731808                    0
5                  0           86.378969                    0
6                  0           86.036931                    0
7                  0           47.315820                    0
8                  0           88.638130                    0
9                  0           96.636541                    0

   OccupationalExposureChemicals  WaterQuality  MedicalCheckupsFrequency  \
0                              0             0                  1.782724
1                              0             1                  3.381070
2                              0             0                  2.701019
3                              0             1                  1.409056
4                              0             0                  1.218452
5                              0             0                  1.535161
6                              0             0                  0.578208
7                              0             0                  1.659424
8                              0             0                  3.675916
9                              0             0                  2.567315

   MedicationAdherence  HealthLiteracy  Diagnosis  DoctorInCharge
0             4.486980        7.211349          1    Confidential
1             5.961705        5.024612          1    Confidential
2             8.950821        7.034944          0    Confidential
3             3.124769        4.717774          0    Confidential
4             6.977741        7.887940          0    Confidential
5             9.682226        2.744281          0    Confidential
6             1.175504        1.229453          0    Confidential
7             2.258377        9.035877          0    Confidential
8             2.006186        3.452805          1    Confidential
```

```
9             4.031643        2.633287          0    Confidential

[10 rows x 46 columns]
```

[182]:
```
diabetes_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1879 entries, 0 to 1878
Data columns (total 46 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   PatientID                       1879 non-null   int64
 1   Age                             1879 non-null   int64
 2   Gender                          1879 non-null   int64
 3   Ethnicity                       1879 non-null   int64
 4   SocioeconomicStatus             1879 non-null   int64
 5   EducationLevel                  1879 non-null   int64
 6   BMI                             1879 non-null   float64
 7   Smoking                         1879 non-null   int64
 8   AlcoholConsumption              1879 non-null   float64
 9   PhysicalActivity                1879 non-null   float64
 10  DietQuality                     1879 non-null   float64
 11  SleepQuality                    1879 non-null   float64
 12  FamilyHistoryDiabetes           1879 non-null   int64
 13  GestationalDiabetes             1879 non-null   int64
 14  PolycysticOvarySyndrome         1879 non-null   int64
 15  PreviousPreDiabetes             1879 non-null   int64
 16  Hypertension                    1879 non-null   int64
 17  SystolicBP                      1879 non-null   int64
 18  DiastolicBP                     1879 non-null   int64
 19  FastingBloodSugar               1879 non-null   float64
 20  HbA1c                           1879 non-null   float64
 21  SerumCreatinine                 1879 non-null   float64
 22  BUNLevels                       1879 non-null   float64
 23  CholesterolTotal                1879 non-null   float64
 24  CholesterolLDL                  1879 non-null   float64
 25  CholesterolHDL                  1879 non-null   float64
 26  CholesterolTriglycerides        1879 non-null   float64
 27  AntihypertensiveMedications     1879 non-null   int64
 28  Statins                         1879 non-null   int64
 29  AntidiabeticMedications         1879 non-null   int64
 30  FrequentUrination               1879 non-null   int64
 31  ExcessiveThirst                 1879 non-null   int64
 32  UnexplainedWeightLoss           1879 non-null   int64
 33  FatigueLevels                   1879 non-null   float64
 34  BlurredVision                   1879 non-null   int64
 35  SlowHealingSores                1879 non-null   int64
 36  TinglingHandsFeet               1879 non-null   int64
```

```
 37  QualityOfLifeScore              1879 non-null   float64
 38  HeavyMetalsExposure             1879 non-null   int64
 39  OccupationalExposureChemicals   1879 non-null   int64
 40  WaterQuality                    1879 non-null   int64
 41  MedicalCheckupsFrequency        1879 non-null   float64
 42  MedicationAdherence             1879 non-null   float64
 43  HealthLiteracy                  1879 non-null   float64
 44  Diagnosis                       1879 non-null   int64
 45  DoctorInCharge                  1879 non-null   object
dtypes: float64(18), int64(27), object(1)
memory usage: 675.4+ KB
```

[183]:
```
diabetes_df.T
```

[183]:

	0	1	2	\
PatientID	6000	6001	6002	
Age	44	51	89	
Gender	0	1	1	
Ethnicity	1	0	0	
SocioeconomicStatus	2	1	1	
EducationLevel	1	2	3	
BMI	32.985284	39.916764	19.782251	
Smoking	1	0	0	
AlcoholConsumption	4.499365	1.578919	1.177301	
PhysicalActivity	2.443385	8.301264	6.103395	
DietQuality	4.898831	8.941093	7.722543	
SleepQuality	4.049885	7.50815	7.708387	
FamilyHistoryDiabetes	1	0	1	
GestationalDiabetes	1	0	0	
PolycysticOvarySyndrome	0	0	0	
PreviousPreDiabetes	0	0	0	
Hypertension	0	0	0	
SystolicBP	93	165	119	
DiastolicBP	73	99	91	
FastingBloodSugar	163.687162	188.34707	127.703653	
HbA1c	9.283631	7.32687	4.083426	
SerumCreatinine	2.665607	4.172177	1.973168	
BUNLevels	28.190147	32.149491	10.018375	
CholesterolTotal	254.27067	155.358831	231.608922	
CholesterolLDL	86.993627	110.056105	62.035793	
CholesterolHDL	70.801469	39.900112	62.480666	
CholesterolTriglycerides	190.335834	81.172469	279.809069	
AntihypertensiveMedications	0	0	1	
Statins	0	0	1	
AntidiabeticMedications	1	0	0	
FrequentUrination	0	0	0	
ExcessiveThirst	0	0	0	

UnexplainedWeightLoss	0	0	0
FatigueLevels	9.534169	0.123214	9.64332
BlurredVision	0	0	0
SlowHealingSores	0	0	0
TinglingHandsFeet	1	0	0
QualityOfLifeScore	73.765109	91.445753	54.485744
HeavyMetalsExposure	0	0	0
OccupationalExposureChemicals	0	0	0
WaterQuality	0	1	0
MedicalCheckupsFrequency	1.782724	3.38107	2.701019
MedicationAdherence	4.48698	5.961705	8.950821
HealthLiteracy	7.211349	5.024612	7.034944
Diagnosis	1	1	0
DoctorInCharge	Confidential	Confidential	Confidential

	3	4	5	\
PatientID	6003	6004	6005	
Age	21	27	65	
Gender	1	1	0	
Ethnicity	1	0	0	
SocioeconomicStatus	1	1	0	
EducationLevel	2	3	0	
BMI	32.376881	16.8086	15.820815	
Smoking	1	0	1	
AlcoholConsumption	1.714621	15.462549	17.781024	
PhysicalActivity	8.645465	4.629383	9.252522	
DietQuality	4.804044	2.532756	2.309158	
SleepQuality	6.286548	9.771125	9.869401	
FamilyHistoryDiabetes	1	0	0	
GestationalDiabetes	1	0	0	
PolycysticOvarySyndrome	0	0	0	
PreviousPreDiabetes	1	0	0	
Hypertension	0	0	0	
SystolicBP	169	165	144	
DiastolicBP	87	69	64	
FastingBloodSugar	82.688415	90.743395	119.593839	
HbA1c	6.516645	5.607222	8.523665	
SerumCreatinine	3.057797	4.150353	0.733091	
BUNLevels	44.123281	7.757117	35.797135	
CholesterolTotal	176.592374	157.344121	250.001898	
CholesterolLDL	68.23841	66.476215	65.202003	
CholesterolHDL	46.977819	40.059755	24.705041	
CholesterolTriglycerides	112.751396	381.528785	395.494809	
AntihypertensiveMedications	0	1	0	
Statins	0	1	1	
AntidiabeticMedications	1	0	0	
FrequentUrination	0	0	0	

ExcessiveThirst	0	0	0
UnexplainedWeightLoss	0	0	0
FatigueLevels	3.403557	2.924687	1.973642
BlurredVision	0	0	0
SlowHealingSores	0	0	0
TinglingHandsFeet	0	0	0
QualityOfLifeScore	77.866758	37.731808	86.378969
HeavyMetalsExposure	0	0	0
OccupationalExposureChemicals	0	0	0
WaterQuality	1	0	0
MedicalCheckupsFrequency	1.409056	1.218452	1.535161
MedicationAdherence	3.124769	6.977741	9.682226
HealthLiteracy	4.717774	7.88794	2.744281
Diagnosis	0	0	0
DoctorInCharge	Confidential	Confidential	Confidential

	6	7	8	\
PatientID	6006	6007	6008	
Age	61	74	54	
Gender	1	1	0	
Ethnicity	2	3	0	
SocioeconomicStatus	1	0	1	
EducationLevel	3	3	2	
BMI	20.075147	29.438938	15.027557	
Smoking	0	0	0	
AlcoholConsumption	1.086479	6.187378	19.505734	
PhysicalActivity	8.74565	9.114535	0.590771	
DietQuality	4.70548	0.180463	7.95831	
SleepQuality	4.317813	5.365338	8.425994	
FamilyHistoryDiabetes	0	0	0	
GestationalDiabetes	0	1	0	
PolycysticOvarySyndrome	0	0	0	
PreviousPreDiabetes	0	0	0	
Hypertension	0	0	0	
SystolicBP	109	128	172	
DiastolicBP	96	98	66	
FastingBloodSugar	157.002741	81.507888	130.839112	
HbA1c	4.525074	7.426382	8.601526	
SerumCreatinine	3.624364	0.979222	0.962348	
BUNLevels	10.787199	36.844189	12.368956	
CholesterolTotal	258.393159	159.338689	168.14778	
CholesterolLDL	104.852442	108.548713	59.202961	
CholesterolHDL	24.457787	39.448009	85.458363	
CholesterolTriglycerides	83.546356	121.674277	257.74949	
AntihypertensiveMedications	0	0	0	
Statins	0	0	0	
AntidiabeticMedications	0	0	0	

FrequentUrination	1	0	0		
ExcessiveThirst	0	0	0		
UnexplainedWeightLoss	0	1	0		
FatigueLevels	6.519587	2.314413	4.965486		
BlurredVision	0	0	1		
SlowHealingSores	0	0	0		
TinglingHandsFeet	0	0	0		
QualityOfLifeScore	86.036931	47.31582	88.63813		
HeavyMetalsExposure	0	0	0		
OccupationalExposureChemicals	0	0	0		
WaterQuality	0	0	0		
MedicalCheckupsFrequency	0.578208	1.659424	3.675916		
MedicationAdherence	1.175504	2.258377	2.006186		
HealthLiteracy	1.229453	9.035877	3.452805		
Diagnosis	0	0	1		
DoctorInCharge	Confidential	Confidential	Confidential		

	9	…	1869	1870 \
PatientID	6009	…	7869	7870
Age	82	…	53	29
Gender	1	…	0	1
Ethnicity	0	…	3	0
SocioeconomicStatus	1	…	0	1
EducationLevel	1	…	1	3
BMI	34.300044	…	27.051115	38.813517
Smoking	1	…	1	0
AlcoholConsumption	15.943844	…	15.936018	2.806481
PhysicalActivity	6.056621	…	0.958362	0.864336
DietQuality	1.339302	…	6.634445	8.310398
SleepQuality	6.718609	…	8.921075	8.203492
FamilyHistoryDiabetes	0	…	1	0
GestationalDiabetes	0	…	1	0
PolycysticOvarySyndrome	0	…	0	0
PreviousPreDiabetes	0	…	0	1
Hypertension	0	…	0	0
SystolicBP	95	…	174	128
DiastolicBP	85	…	118	102
FastingBloodSugar	81.72107	…	88.070236	83.672246
HbA1c	7.424233	…	4.165395	7.159699
SerumCreatinine	3.347684	…	4.554104	2.432542
BUNLevels	7.351596	…	23.663565	24.671534
CholesterolTotal	164.024326	…	213.31769	229.036028
CholesterolLDL	77.850453	…	180.021644	194.18011
CholesterolHDL	55.621487	…	26.837201	33.817421
CholesterolTriglycerides	391.399137	…	132.71111	84.776976
AntihypertensiveMedications	0	…	1	0
Statins	0	…	0	0

		…		
AntidiabeticMedications	0	…	0	1
FrequentUrination	0	…	1	0
ExcessiveThirst	1	…	1	0
UnexplainedWeightLoss	0	…	0	0
FatigueLevels	4.60009	…	6.474088	7.19882
BlurredVision	1	…	0	0
SlowHealingSores	0	…	0	0
TinglingHandsFeet	0	…	0	0
QualityOfLifeScore	96.636541	…	40.088243	47.139988
HeavyMetalsExposure	0	…	1	0
OccupationalExposureChemicals	0	…	0	0
WaterQuality	0	…	1	0
MedicalCheckupsFrequency	2.567315	…	1.574857	3.046683
MedicationAdherence	4.031643	…	1.226518	1.124973
HealthLiteracy	2.633287	…	5.068996	2.814329
Diagnosis	0	…	1	1
DoctorInCharge	Confidential	…	Confidential	Confidential

	1871	1872	1873	\
PatientID	7871	7872	7873	
Age	24	59	75	
Gender	1	0	1	
Ethnicity	0	0	0	
SocioeconomicStatus	1	2	1	
EducationLevel	3	2	3	
BMI	15.342747	33.668673	30.378877	
Smoking	1	0	1	
AlcoholConsumption	17.267981	9.996765	9.235769	
PhysicalActivity	8.530101	0.267199	9.868789	
DietQuality	1.819866	4.79017	6.643934	
SleepQuality	7.212605	9.192673	8.085365	
FamilyHistoryDiabetes	1	0	0	
GestationalDiabetes	0	0	0	
PolycysticOvarySyndrome	0	0	0	
PreviousPreDiabetes	0	0	1	
Hypertension	1	0	0	
SystolicBP	142	110	168	
DiastolicBP	109	68	75	
FastingBloodSugar	141.423251	197.472046	101.371009	
HbA1c	5.379436	5.242416	4.371377	
SerumCreatinine	1.282775	4.056453	0.810429	
BUNLevels	23.155524	7.204079	36.124005	
CholesterolTotal	190.995858	285.227582	294.173133	
CholesterolLDL	66.474175	176.683341	141.918258	
CholesterolHDL	29.03003	40.618424	24.200696	
CholesterolTriglycerides	150.066888	213.461798	323.778594	
AntihypertensiveMedications	0	0	1	

Statins	1	0	0
AntidiabeticMedications	0	0	0
FrequentUrination	0	0	0
ExcessiveThirst	0	0	1
UnexplainedWeightLoss	0	0	0
FatigueLevels	9.355321	6.37512	5.758964
BlurredVision	0	0	1
SlowHealingSores	0	0	0
TinglingHandsFeet	1	0	0
QualityOfLifeScore	94.149617	22.218334	33.156263
HeavyMetalsExposure	0	0	0
OccupationalExposureChemicals	0	0	0
WaterQuality	0	1	1
MedicalCheckupsFrequency	1.872678	1.269005	2.863466
MedicationAdherence	6.849746	4.090977	5.486838
HealthLiteracy	8.692255	9.854819	2.516808
Diagnosis	0	0	0
DoctorInCharge	Confidential	Confidential	Confidential

	1874	1875	1876 \
PatientID	7874	7875	7876
Age	37	80	38
Gender	0	1	1
Ethnicity	0	0	0
SocioeconomicStatus	2	2	0
EducationLevel	2	2	2
BMI	20.811137	27.694312	35.640824
Smoking	0	0	0
AlcoholConsumption	10.946207	16.067905	4.865124
PhysicalActivity	3.217636	7.107335	9.881212
DietQuality	8.338196	3.034771	2.657002
SleepQuality	8.70343	4.472689	4.81261
FamilyHistoryDiabetes	0	1	0
GestationalDiabetes	0	1	0
PolycysticOvarySyndrome	0	0	0
PreviousPreDiabetes	0	0	0
Hypertension	1	0	0
SystolicBP	104	166	128
DiastolicBP	74	115	70
FastingBloodSugar	109.832032	90.729361	149.366801
HbA1c	5.920723	7.332397	4.907208
SerumCreatinine	3.984707	2.132178	2.195365
BUNLevels	21.645433	7.433835	26.225481
CholesterolTotal	260.342336	273.728852	293.513379
CholesterolLDL	99.720234	179.858432	113.915759
CholesterolHDL	40.296248	48.873298	62.217083
CholesterolTriglycerides	198.613903	271.239061	374.429055

AntihypertensiveMedications	0	0	0
Statins	0	1	0
AntidiabeticMedications	0	0	0
FrequentUrination	0	0	0
ExcessiveThirst	1	0	0
UnexplainedWeightLoss	0	0	0
FatigueLevels	3.693506	4.225031	1.174257
BlurredVision	1	0	0
SlowHealingSores	0	1	0
TinglingHandsFeet	1	0	0
QualityOfLifeScore	88.122729	77.128599	13.148221
HeavyMetalsExposure	0	0	0
OccupationalExposureChemicals	0	0	0
WaterQuality	1	1	0
MedicalCheckupsFrequency	3.154225	0.424893	0.553757
MedicationAdherence	3.849584	5.217465	3.377744
HealthLiteracy	8.805087	0.915878	3.017481
Diagnosis	0	1	1
DoctorInCharge	Confidential	Confidential	Confidential

	1877	1878
PatientID	7877	7878
Age	43	85
Gender	0	1
Ethnicity	1	0
SocioeconomicStatus	2	2
EducationLevel	0	2
BMI	32.423016	33.145119
Smoking	0	0
AlcoholConsumption	6.362936	13.854861
PhysicalActivity	4.750079	5.434137
DietQuality	8.736024	5.127496
SleepQuality	7.01739	4.924963
FamilyHistoryDiabetes	1	1
GestationalDiabetes	0	0
PolycysticOvarySyndrome	1	0
PreviousPreDiabetes	0	0
Hypertension	0	0
SystolicBP	124	134
DiastolicBP	91	86
FastingBloodSugar	162.027044	175.011749
HbA1c	8.820613	7.814477
SerumCreatinine	0.893745	4.607711
BUNLevels	41.555665	28.471762
CholesterolTotal	178.55955	268.635952
CholesterolLDL	141.601955	57.431715
CholesterolHDL	74.116118	73.728242

```
CholesterolTriglycerides         171.298228    174.869266
AntihypertensiveMedications               1             0
Statins                                   1             1
AntidiabeticMedications                   1             0
FrequentUrination                         0             1
ExcessiveThirst                           0             0
UnexplainedWeightLoss                     0             0
FatigueLevels                      9.732583      4.360088
BlurredVision                             0             0
SlowHealingSores                          0             0
TinglingHandsFeet                         0             1
QualityOfLifeScore                54.37098      43.72086
HeavyMetalsExposure                       0             0
OccupationalExposureChemicals             0             0
WaterQuality                              0             1
MedicalCheckupsFrequency            1.13247      3.070583
MedicationAdherence                 0.00925      8.483128
HealthLiteracy                     4.914556      7.790921
Diagnosis                                 1             1
DoctorInCharge                Confidential  Confidential

[46 rows x 1879 columns]
```

[184]:
```
diabetes_df.dtypes
```

[184]:
```
PatientID                    int64
Age                          int64
Gender                       int64
Ethnicity                    int64
SocioeconomicStatus          int64
EducationLevel               int64
BMI                        float64
Smoking                      int64
AlcoholConsumption         float64
PhysicalActivity           float64
DietQuality                float64
SleepQuality               float64
FamilyHistoryDiabetes        int64
GestationalDiabetes          int64
PolycysticOvarySyndrome      int64
PreviousPreDiabetes          int64
Hypertension                 int64
SystolicBP                   int64
DiastolicBP                  int64
FastingBloodSugar          float64
HbA1c                      float64
SerumCreatinine            float64
```

```
BUNLevels                        float64
CholesterolTotal                 float64
CholesterolLDL                   float64
CholesterolHDL                   float64
CholesterolTriglycerides         float64
AntihypertensiveMedications        int64
Statins                            int64
AntidiabeticMedications            int64
FrequentUrination                  int64
ExcessiveThirst                    int64
UnexplainedWeightLoss              int64
FatigueLevels                    float64
BlurredVision                      int64
SlowHealingSores                   int64
TinglingHandsFeet                  int64
QualityOfLifeScore               float64
HeavyMetalsExposure                int64
OccupationalExposureChemicals      int64
WaterQuality                       int64
MedicalCheckupsFrequency         float64
MedicationAdherence              float64
HealthLiteracy                   float64
Diagnosis                          int64
DoctorInCharge                    object
dtype: object
```

[185]:
```
diabetes_df.describe()
```

[185]:

	PatientID	Age	Gender	Ethnicity	\
count	1879.000000	1879.000000	1879.000000	1879.000000	
mean	6939.000000	55.043108	0.487493	0.755721	
std	542.564896	20.515839	0.499977	1.047558	
min	6000.000000	20.000000	0.000000	0.000000	
25%	6469.500000	38.000000	0.000000	0.000000	
50%	6939.000000	55.000000	0.000000	0.000000	
75%	7408.500000	73.000000	1.000000	1.000000	
max	7878.000000	90.000000	1.000000	3.000000	

	SocioeconomicStatus	EducationLevel	BMI	Smoking	\
count	1879.000000	1879.000000	1879.000000	1879.000000	
mean	0.992017	1.699308	27.687601	0.281533	
std	0.764940	0.885665	7.190975	0.449866	
min	0.000000	0.000000	15.025898	0.000000	
25%	0.000000	1.000000	21.469981	0.000000	
50%	1.000000	2.000000	27.722988	0.000000	
75%	2.000000	2.000000	33.856460	1.000000	
max	2.000000	3.000000	39.998811	1.000000	

```
       AlcoholConsumption  PhysicalActivity  …  SlowHealingSores  \
count         1879.000000       1879.000000  …       1879.000000
mean            10.096587          5.200790  …          0.102714
std              5.914216          2.857012  …          0.303666
min              0.000928          0.004089  …          0.000000
25%              4.789725          2.751022  …          0.000000
50%             10.173865          5.249002  …          0.000000
75%             15.285359          7.671402  …          0.000000
max             19.996231          9.993893  …          1.000000

       TinglingHandsFeet  QualityOfLifeScore  HeavyMetalsExposure  \
count        1879.000000         1879.000000          1879.000000
mean            0.111229           48.508643             0.052155
std             0.314500           28.758488             0.222400
min             0.000000            0.002390             0.000000
25%             0.000000           23.974098             0.000000
50%             0.000000           47.519693             0.000000
75%             0.000000           72.883179             0.000000
max             1.000000           99.788530             1.000000

       OccupationalExposureChemicals  WaterQuality  MedicalCheckupsFrequency  \
count                    1879.000000   1879.000000               1879.000000
mean                        0.103246      0.200639                  1.997101
std                         0.304361      0.400585                  1.122632
min                         0.000000      0.000000                  0.004013
25%                         0.000000      0.000000                  1.057801
50%                         0.000000      0.000000                  1.987170
75%                         0.000000      0.000000                  2.946019
max                         1.000000      1.000000                  3.999715

       MedicationAdherence  HealthLiteracy    Diagnosis
count          1879.000000     1879.000000  1879.000000
mean              4.957539        5.011736     0.400213
std               2.910934        2.920908     0.490072
min               0.005384        0.000362     0.000000
25%               2.420024        2.410113     0.000000
50%               4.843886        5.035208     0.000000
75%               7.513933        7.586865     1.000000
max               9.997165        9.993029     1.000000

[8 rows x 45 columns]
```

[186]:
```
diabetes_df.isnull().sum()
```

[186]:
```
PatientID                    0
Age                          0
```

```
Gender                           0
Ethnicity                        0
SocioeconomicStatus              0
EducationLevel                   0
BMI                              0
Smoking                          0
AlcoholConsumption               0
PhysicalActivity                 0
DietQuality                      0
SleepQuality                     0
FamilyHistoryDiabetes            0
GestationalDiabetes              0
PolycysticOvarySyndrome          0
PreviousPreDiabetes              0
Hypertension                     0
SystolicBP                       0
DiastolicBP                      0
FastingBloodSugar                0
HbA1c                            0
SerumCreatinine                  0
BUNLevels                        0
CholesterolTotal                 0
CholesterolLDL                   0
CholesterolHDL                   0
CholesterolTriglycerides         0
AntihypertensiveMedications      0
Statins                          0
AntidiabeticMedications          0
FrequentUrination                0
ExcessiveThirst                  0
UnexplainedWeightLoss            0
FatigueLevels                    0
BlurredVision                    0
SlowHealingSores                 0
TinglingHandsFeet                0
QualityOfLifeScore               0
HeavyMetalsExposure              0
OccupationalExposureChemicals    0
WaterQuality                     0
MedicalCheckupsFrequency         0
MedicationAdherence              0
HealthLiteracy                   0
Diagnosis                        0
DoctorInCharge                   0
dtype: int64
```

[187]:
```
diabetes_df[diabetes_df['DietQuality'] >1]
```

[187]:

	PatientID	Age	Gender	Ethnicity	SocioeconomicStatus	EducationLevel	\
0	6000	44	0	1	2	1	
1	6001	51	1	0	1	2	
2	6002	89	1	0	1	3	
3	6003	21	1	1	1	2	
4	6004	27	1	0	1	3	
…	…	…	…	…	…	…	
1874	7874	37	0	0	2	2	
1875	7875	80	1	0	2	2	
1876	7876	38	1	0	0	2	
1877	7877	43	0	1	2	0	
1878	7878	85	1	0	2	2	

	BMI	Smoking	AlcoholConsumption	PhysicalActivity	…	\
0	32.985284	1	4.499365	2.443385	…	
1	39.916764	0	1.578919	8.301264	…	
2	19.782251	0	1.177301	6.103395	…	
3	32.376881	1	1.714621	8.645465	…	
4	16.808600	0	15.462549	4.629383	…	
…	…	…	…	…	…	
1874	20.811137	0	10.946207	3.217636	…	
1875	27.694312	0	16.067905	7.107335	…	
1876	35.640824	0	4.865124	9.881212	…	
1877	32.423016	0	6.362936	4.750079	…	
1878	33.145119	0	13.854861	5.434137	…	

	TinglingHandsFeet	QualityOfLifeScore	HeavyMetalsExposure	\
0	1	73.765109	0	
1	0	91.445753	0	
2	0	54.485744	0	
3	0	77.866758	0	
4	0	37.731808	0	
…	…	…	…	
1874	1	88.122729	0	
1875	0	77.128599	0	
1876	0	13.148221	0	
1877	0	54.370980	0	
1878	1	43.720860	0	

	OccupationalExposureChemicals	WaterQuality	MedicalCheckupsFrequency	\
0	0	0	1.782724	
1	0	1	3.381070	
2	0	0	2.701019	
3	0	1	1.409056	
4	0	0	1.218452	
…	…	…	…	
1874	0	1	3.154225	

```
1875                            0             1                  0.424893
1876                            0             0                  0.553757
1877                            0             0                  1.132470
1878                            0             1                  3.070583

      MedicationAdherence  HealthLiteracy  Diagnosis  DoctorInCharge
0                4.486980        7.211349          1    Confidential
1                5.961705        5.024612          1    Confidential
2                8.950821        7.034944          0    Confidential
3                3.124769        4.717774          0    Confidential
4                6.977741        7.887940          0    Confidential
…                     …               …          …               …
1874             3.849584        8.805087          0    Confidential
1875             5.217465        0.915878          1    Confidential
1876             3.377744        3.017481          1    Confidential
1877             0.009250        4.914556          1    Confidential
1878             8.483128        7.790921          1    Confidential

[1692 rows x 46 columns]
```

[188]:
```
diabetes_df['BMI'].mean()
```

[188]:
```
27.687601425288864
```

[189]:
```
diabetes_df['Smoking'].mean()
```

[189]:
```
0.28153273017562536
```

[190]:
```
diabetes_df.groupby("DoctorInCharge").mean()
```

[190]:
```
                PatientID        Age    Gender  Ethnicity  \
DoctorInCharge
Confidential       6939.0  55.043108  0.487493   0.755721

                SocioeconomicStatus  EducationLevel        BMI   Smoking  \
DoctorInCharge
Confidential               0.992017        1.699308  27.687601  0.281533

                AlcoholConsumption  PhysicalActivity  …  SlowHealingSores  \
DoctorInCharge                                        …
Confidential             10.096587           5.20079  …          0.102714

                TinglingHandsFeet  QualityOfLifeScore  HeavyMetalsExposure  \
DoctorInCharge
Confidential             0.111229           48.508643             0.052155

                OccupationalExposureChemicals  WaterQuality  \
```

```
DoctorInCharge
Confidential                          0.103246       0.200639

                MedicalCheckupsFrequency  MedicationAdherence  HealthLiteracy  \
DoctorInCharge
Confidential                    1.997101             4.957539        5.011736

                Diagnosis
DoctorInCharge
Confidential     0.400213

[1 rows x 45 columns]
```

[191]:
```
diabetes_df.Smoking.value_counts()
```

[191]:
```
Smoking
0    1350
1     529
Name: count, dtype: int64
```

[192]:
```
diabetes_df.Gender.value_counts()
```

[192]:
```
Gender
0    963
1    916
Name: count, dtype: int64
```

[193]:
```
diabetes_df.EducationLevel.value_counts()
```

[193]:
```
EducationLevel
2    725
1    615
3    376
0    163
Name: count, dtype: int64
```

[194]:
```
diabetes_df.groupby('Diagnosis')[['Ethnicity' , 'Smoking' ,␣
↪'PhysicalActivity']].agg(['mean', 'sum', 'min', 'max'])
```

[194]:

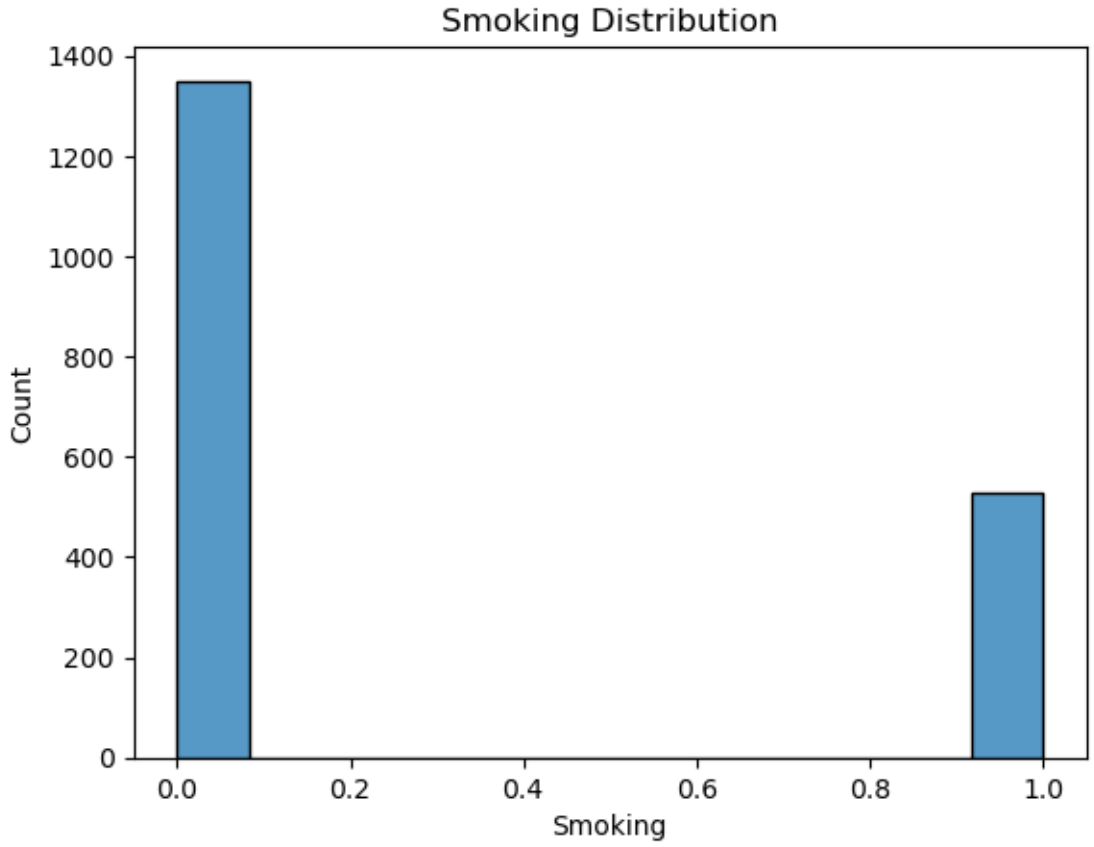
	Ethnicity				Smoking				PhysicalActivity
	mean	sum	min	max	mean	sum	min	max	mean
Diagnosis									
0	0.775510	874	0	3	0.261757	295	0	1	5.215752
1	0.726064	546	0	3	0.311170	234	0	1	5.178368

	sum	min	max
Diagnosis			

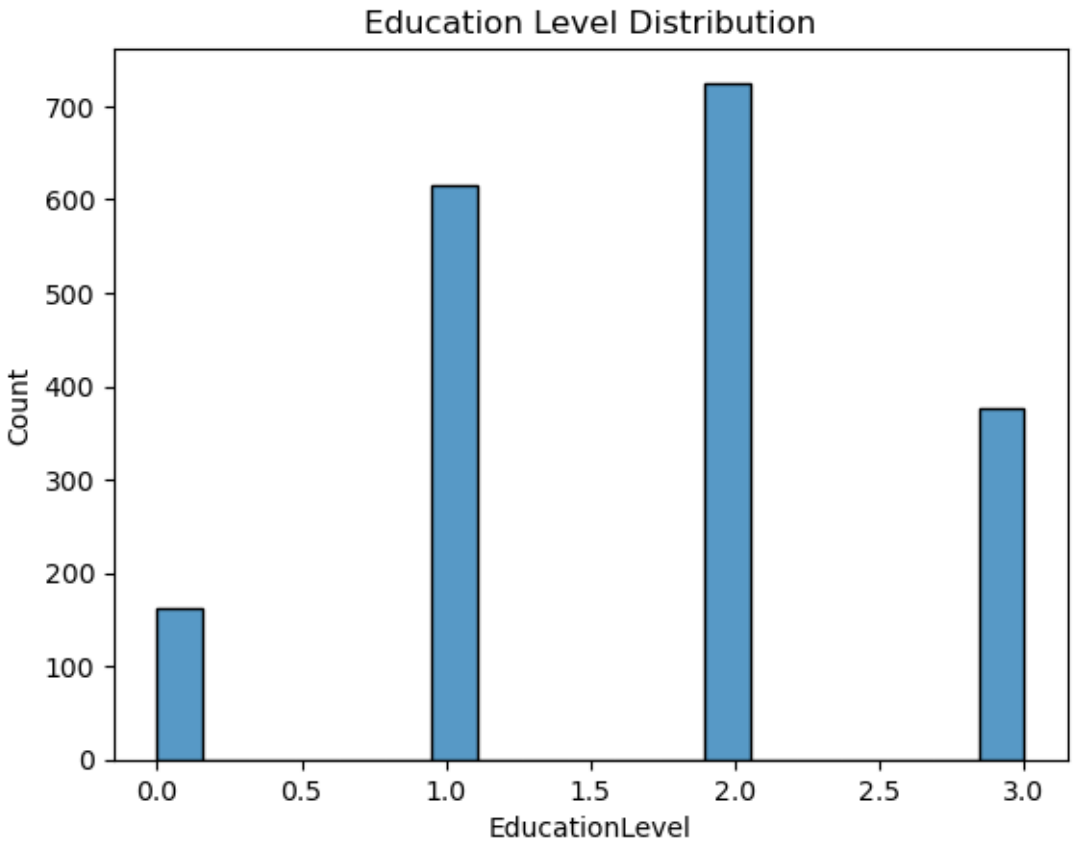
```
0          5878.152840  0.004089  9.993893
1          3894.132367  0.028360  9.969572
```

0.2 Data Analysis

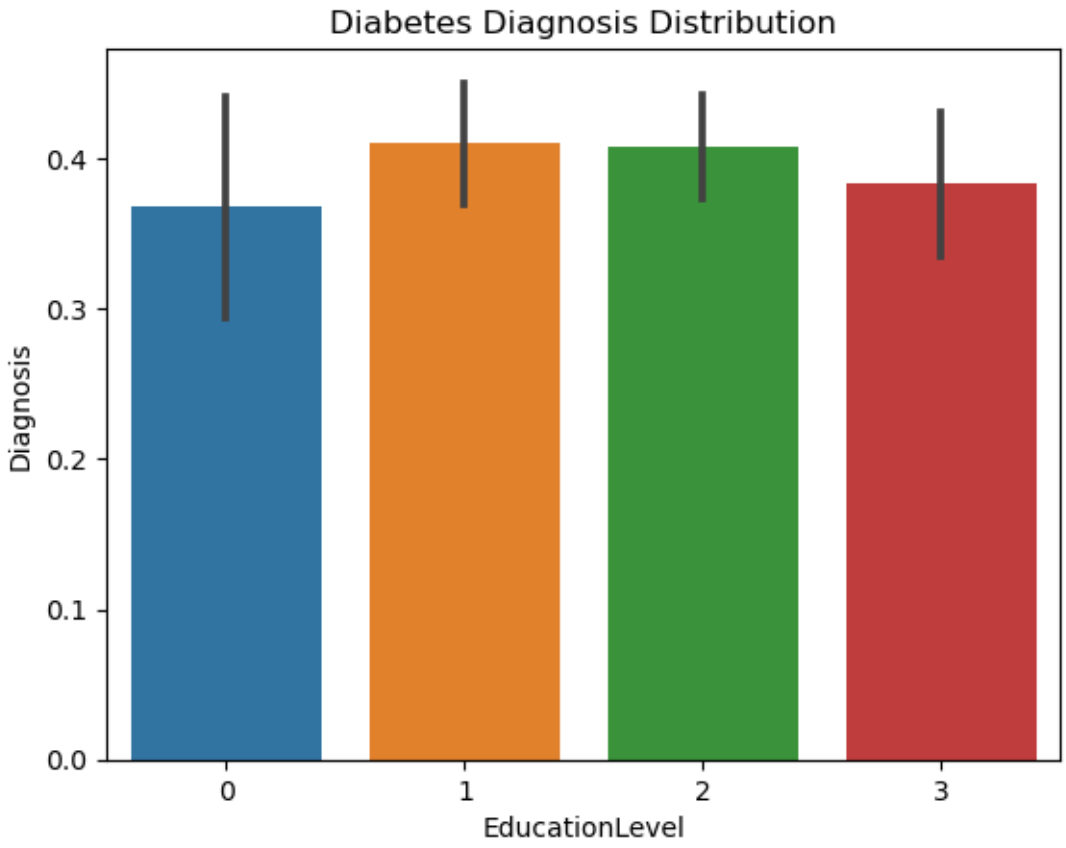
[195]:
```
sns.histplot(diabetes_df.Smoking)
plt.title('Smoking Distribution')
plt.show()
```



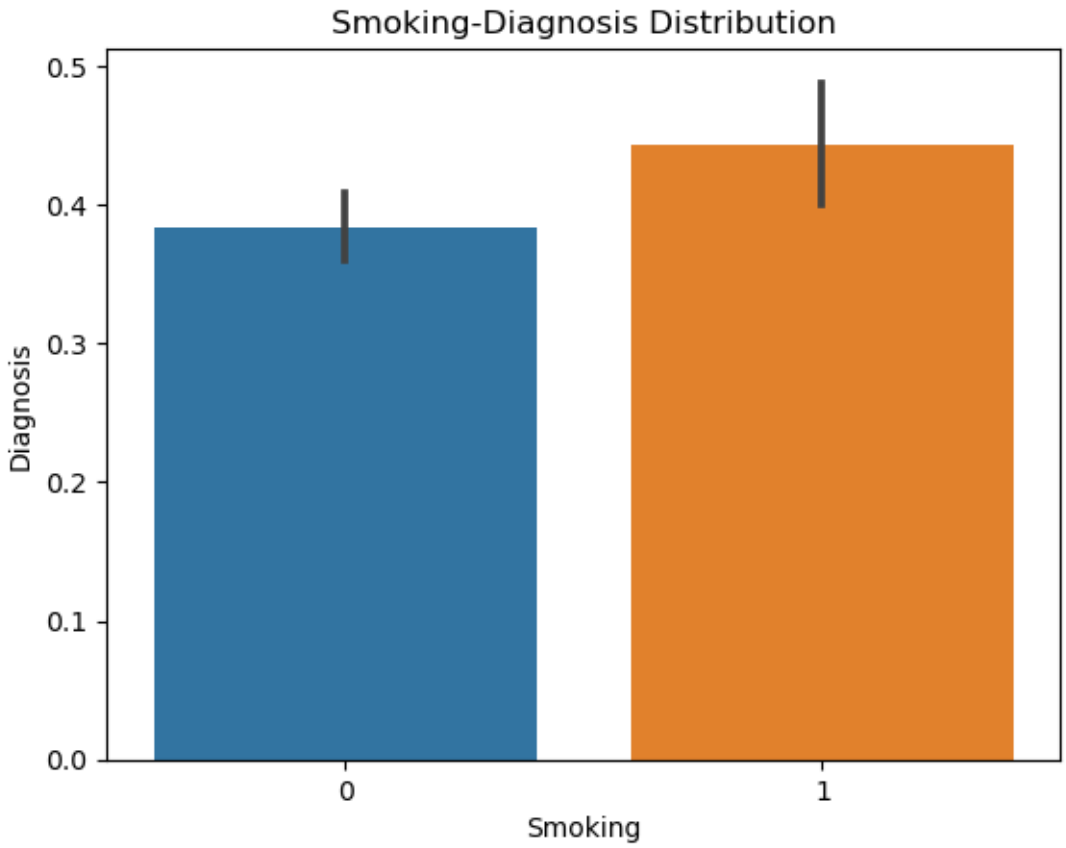
[196]:
```
sns.histplot(diabetes_df.EducationLevel)
plt.title('Education Level Distribution')
plt.show()
```



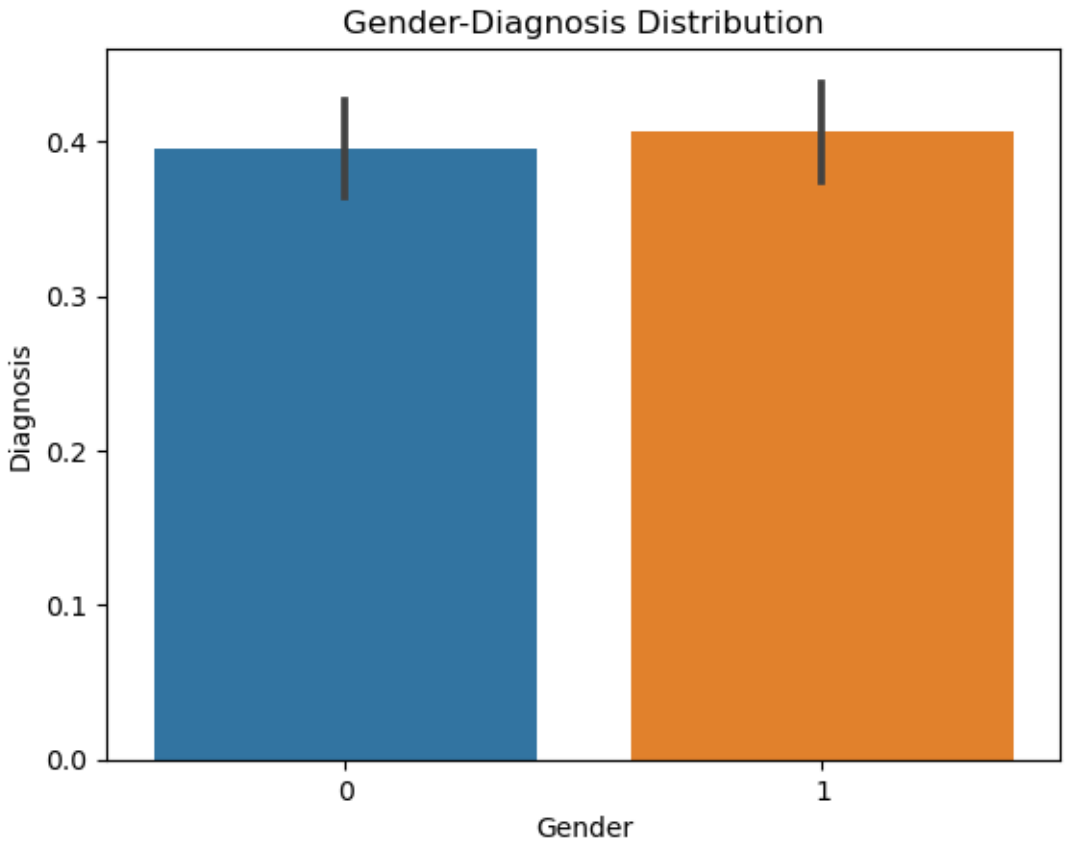
[197]:
```
sns.barplot(y= diabetes_df.Diagnosis, x= diabetes_df.EducationLevel)
plt.title('Diabetes Diagnosis Distribution')
plt.show()
```



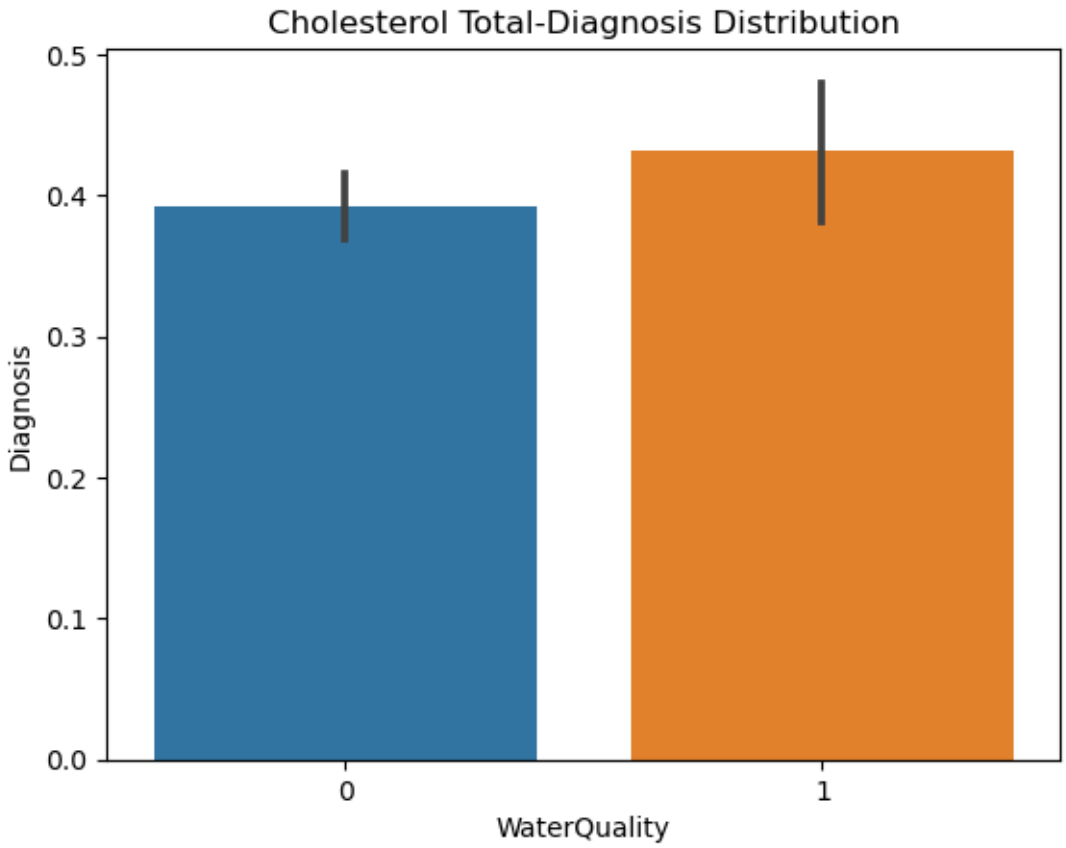
[198]:
```
sns.barplot(y= diabetes_df.Diagnosis, x= diabetes_df.Smoking)
plt.title('Smoking-Diagnosis Distribution')
plt.show()
```



[199]:
```
sns.barplot(y= diabetes_df.Diagnosis, x= diabetes_df.Gender)
plt.title('Gender-Diagnosis Distribution')
plt.show()
```



[200]:
```
sns.barplot(y= diabetes_df.Diagnosis, x= diabetes_df.WaterQuality)
plt.title('Cholesterol Total-Diagnosis Distribution')
plt.show()
```



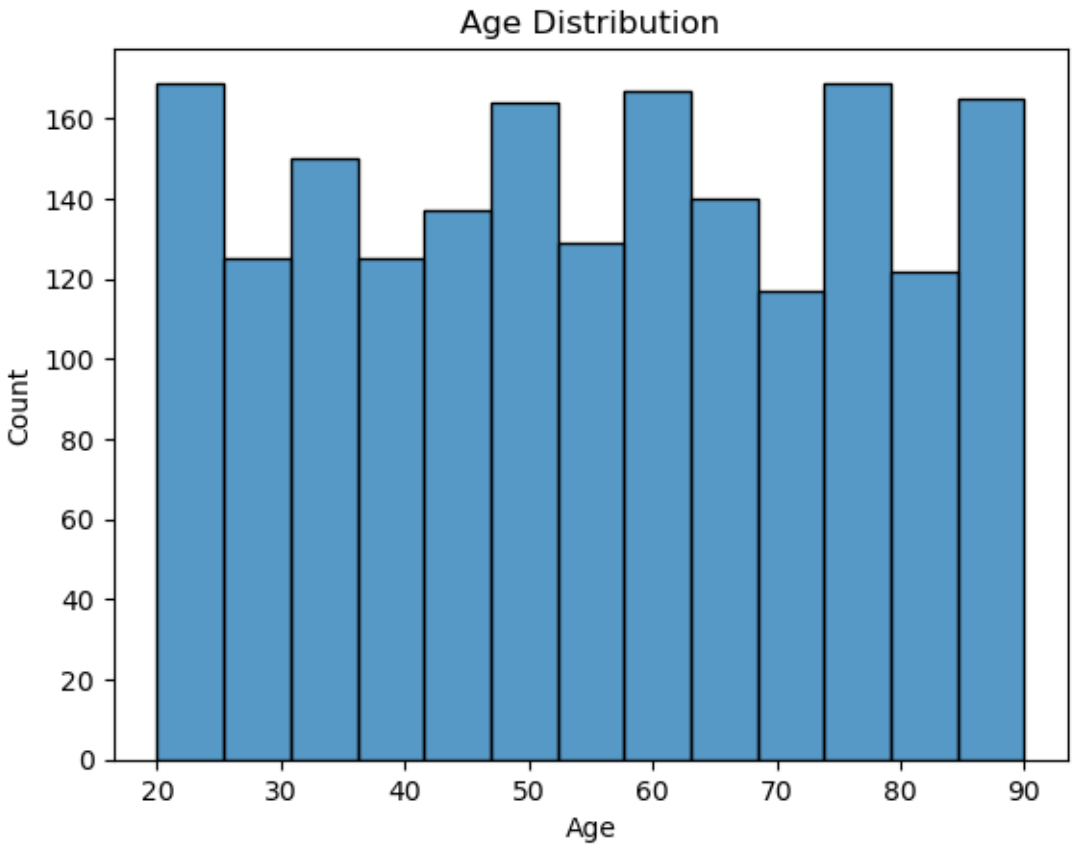
[201]:
```
diabetes_df.groupby('EducationLevel')[['Diagnosis' , 'Smoking' ,␣
↪'PhysicalActivity']].agg(['median', 'mean', 'min', 'max'])
```

[201]:

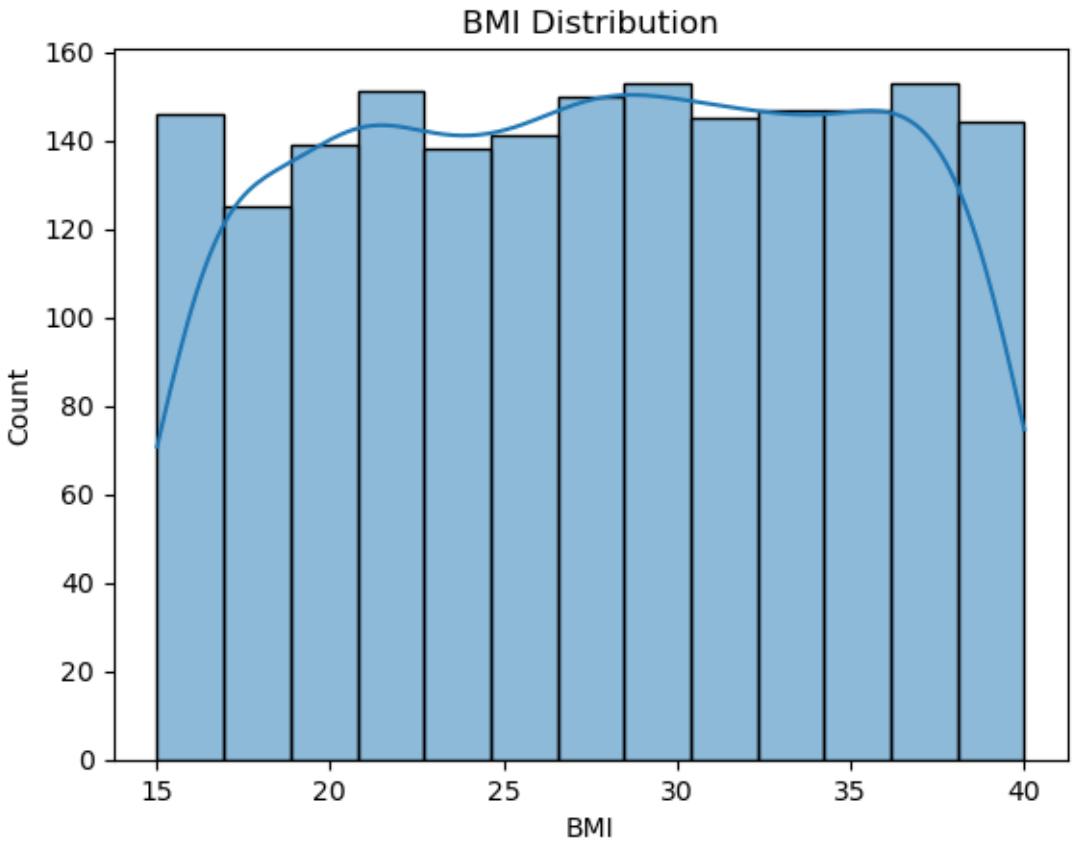
	Diagnosis				Smoking			
	median	mean	min	max	median	mean	min	max
EducationLevel								
0	0.0	0.368098	0	1	0.0	0.220859	0	1
1	0.0	0.409756	0	1	0.0	0.300813	0	1
2	0.0	0.408276	0	1	0.0	0.260690	0	1
3	0.0	0.382979	0	1	0.0	0.316489	0	1

	PhysicalActivity			
	median	mean	min	max
EducationLevel				
0	5.109473	5.288231	0.038327	9.935757
1	5.591697	5.327184	0.053623	9.993893
2	5.064152	5.125811	0.004089	9.980646
3	5.032038	5.100725	0.045043	9.943985

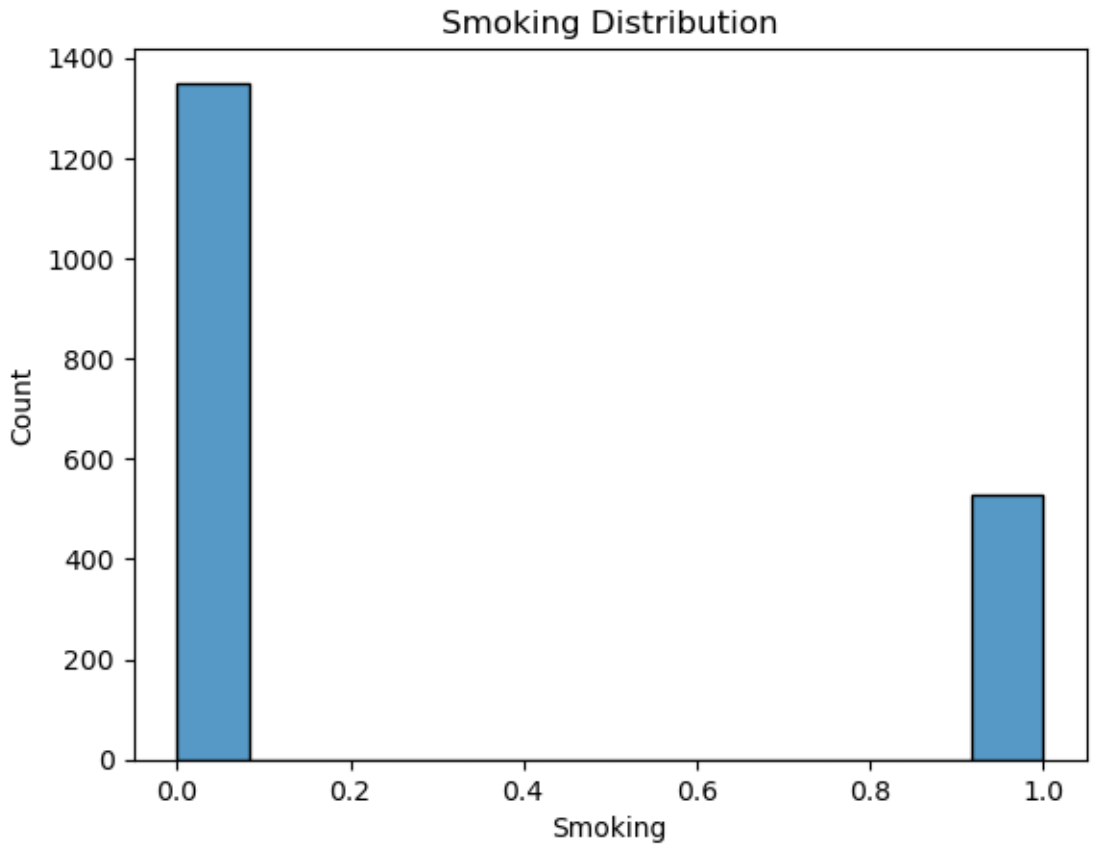
[202]:
```
# EDA: Visualize distributions
sns.histplot(diabetes_df['Age'], kde=False)
plt.title('Age Distribution')
plt.show()
```



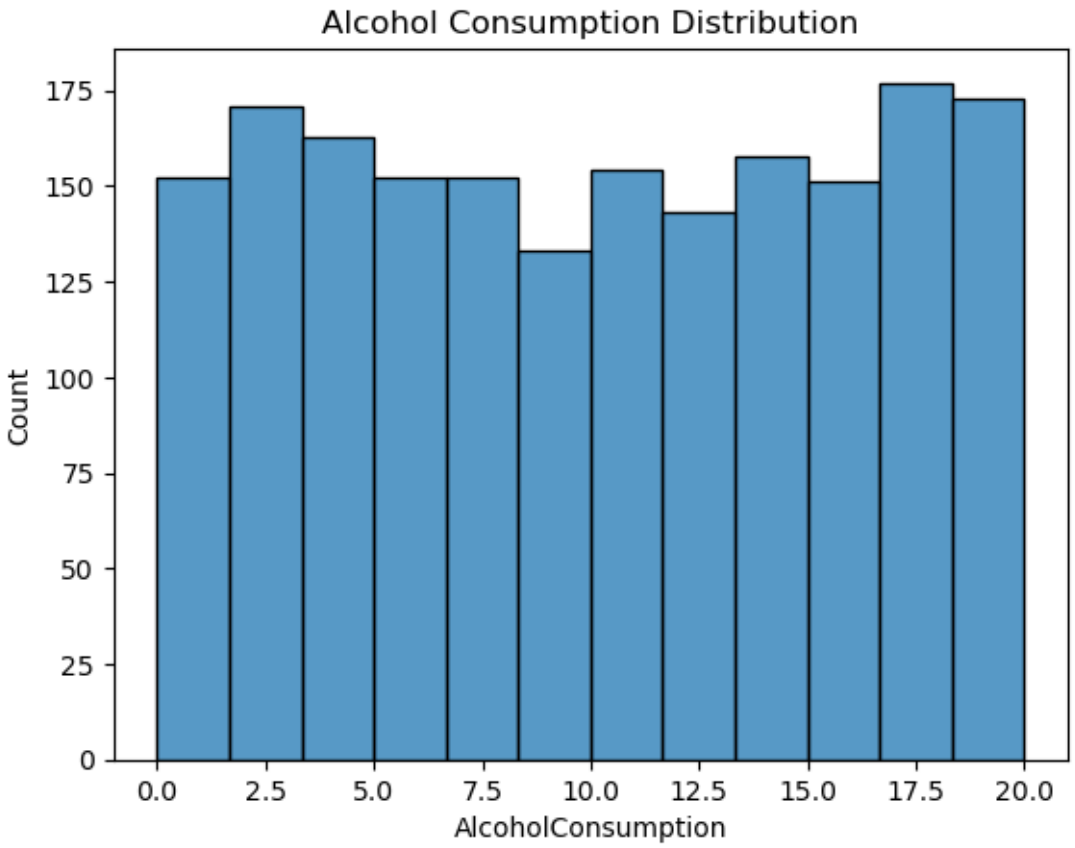
[203]:
```
sns.histplot(diabetes_df['BMI'], kde=True)
plt.title('BMI Distribution')
plt.show()
```



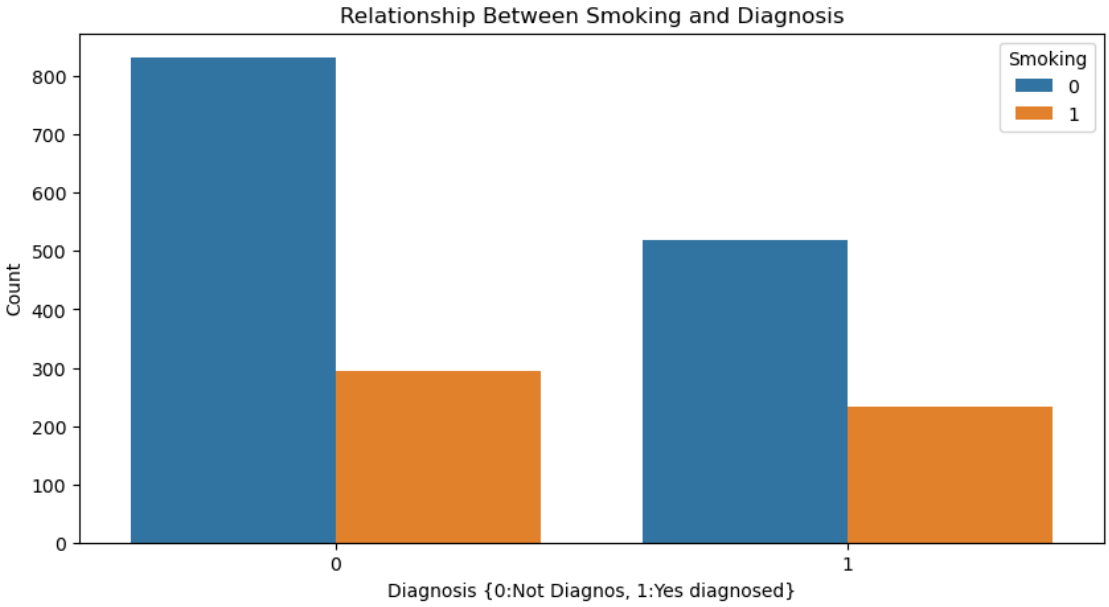
[204]:
```
sns.histplot(diabetes_df['Smoking'], kde=False)
plt.title('Smoking Distribution')
plt.show()
```



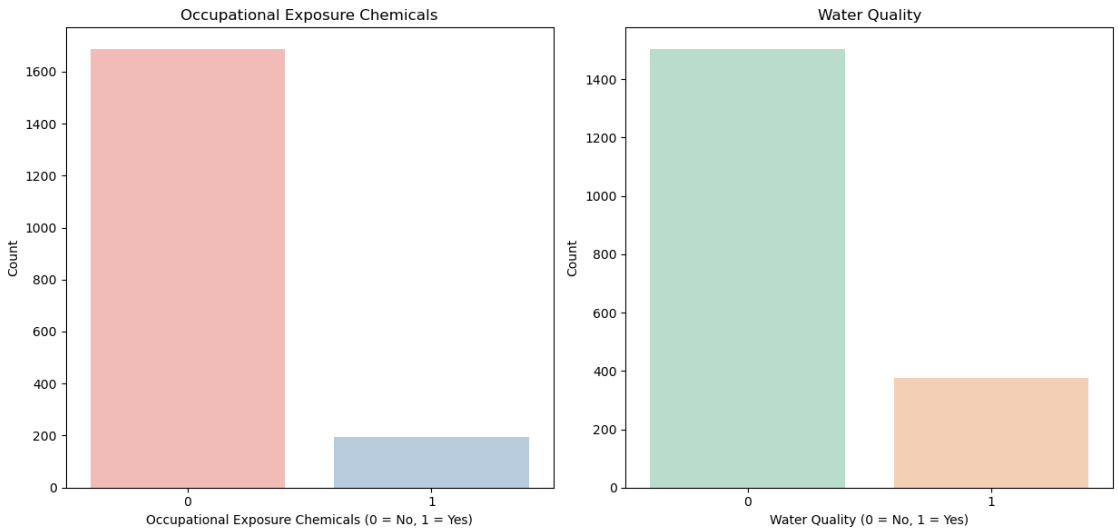
[205]:
```
sns.histplot(diabetes_df['AlcoholConsumption'], kde=False)
plt.title('Alcohol Consumption Distribution')
plt.show()
```



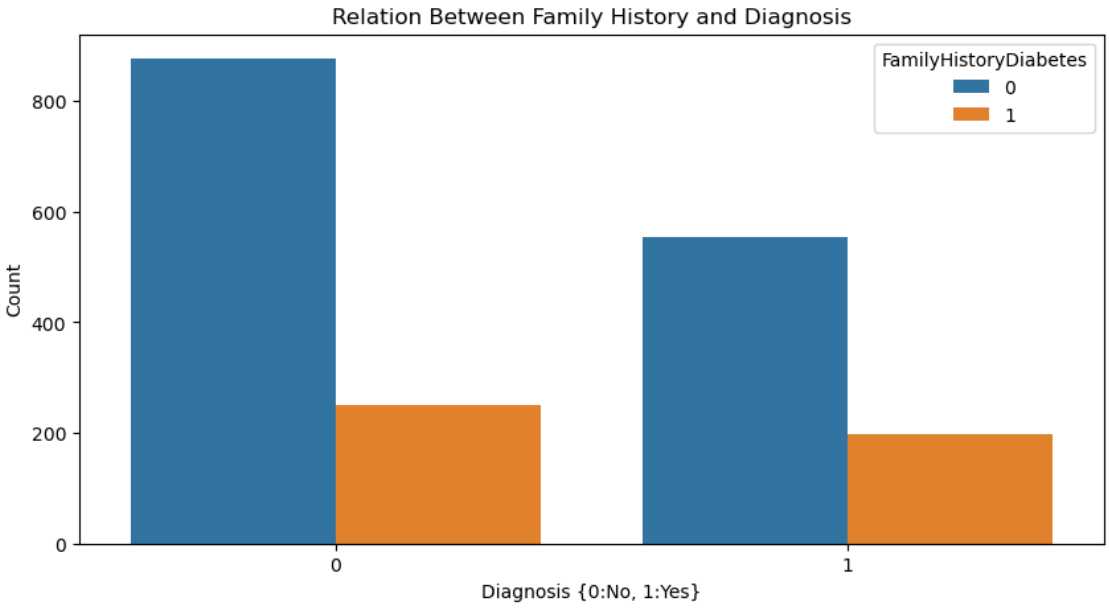
[206]:
```
# Relationship Between Smoking and Diagnosis
plt.figure(figsize = (10,5))
sns.countplot(data=diabetes_df, x='Diagnosis', y=None, hue='Smoking')
plt.title('Relationship Between Smoking and Diagnosis')
plt.xlabel('Diagnosis {0:Not Diagnos, 1:Yes diagnosed}')
plt.ylabel('Count')
plt.show()
```



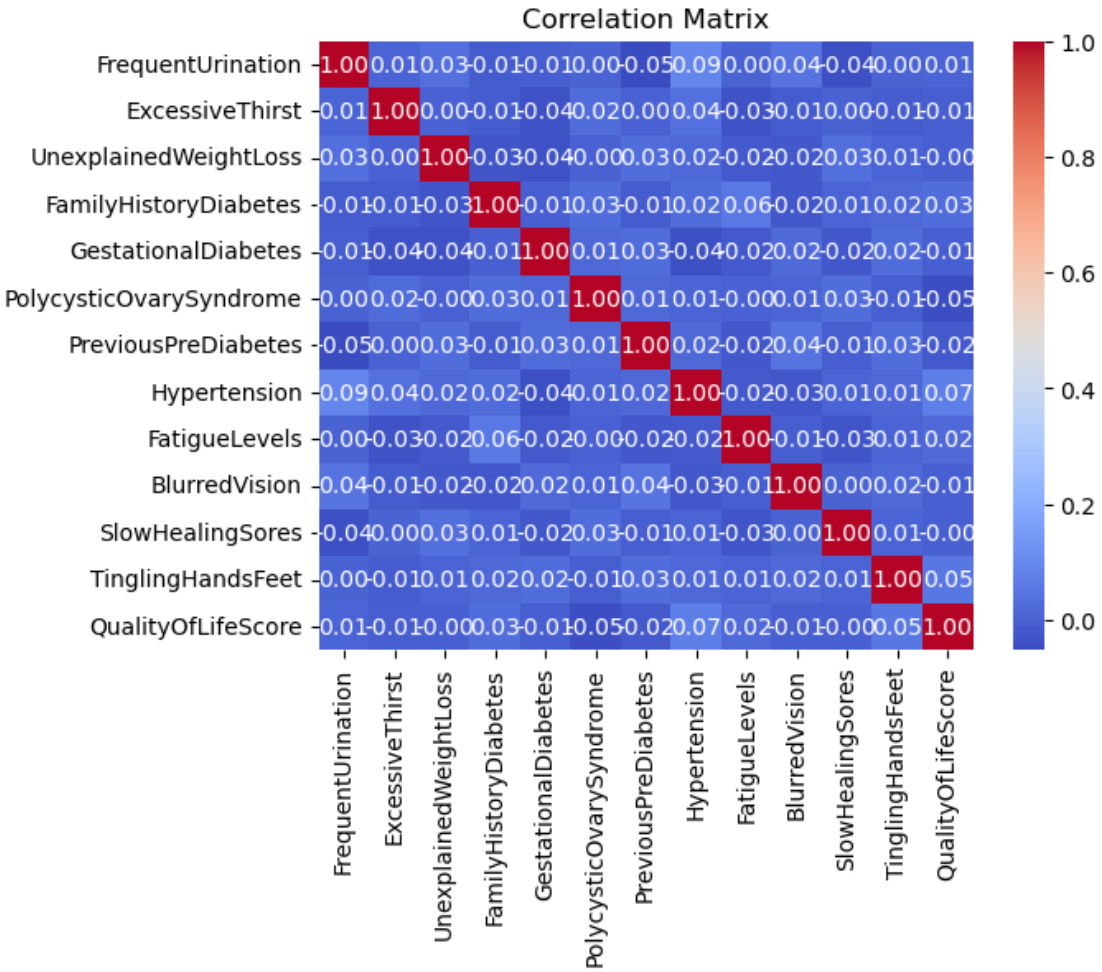
[207]:
```
plt.figure(figsize=(18, 6))

# Occupational Exposure Chemicals
plt.subplot(1, 3, 2)
sns.countplot(data=diabetes_df, x='OccupationalExposureChemicals',␣
 ↪palette='Pastel1')
plt.title('Occupational Exposure Chemicals')
plt.xlabel('Occupational Exposure Chemicals (0 = No, 1 = Yes)')
plt.ylabel('Count')

# Water Quality
plt.subplot(1, 3, 3)
sns.countplot(data=diabetes_df, x='WaterQuality', palette='Pastel2')
plt.title('Water Quality')
plt.xlabel('Water Quality (0 = No, 1 = Yes)')
plt.ylabel('Count')

plt.tight_layout()
plt.show()
```



[208]:
```
# Relation Between Family Diabetes related History and Diagnosis
plt.figure(figsize = (10,5))
sns.countplot(data=diabetes_df, x='Diagnosis', y=None,␣
 ↪hue='FamilyHistoryDiabetes')
plt.title('Relation Between Family History and Diagnosis')
plt.xlabel('Diagnosis {0:No, 1:Yes}')
plt.ylabel('Count')
plt.show()
```



[209]:
```
# Correlation Matrix
columns_of_interest = ['FrequentUrination', 'ExcessiveThirst',␣
 ↪'UnexplainedWeightLoss', 'FamilyHistoryDiabetes', 'GestationalDiabetes',
       'PolycysticOvarySyndrome', 'PreviousPreDiabetes', 'Hypertension',␣
 ↪'FatigueLevels', 'BlurredVision',␣
 ↪'SlowHealingSores','TinglingHandsFeet','QualityOfLifeScore']
cor_diabetes = diabetes_df[columns_of_interest]

sns.heatmap(cor_diabetes.corr(), annot=True, fmt='.2f', cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



0.2.1 Data Preparation & Data Partitioning

[210]:
```
X = diabetes_df.drop(['Diagnosis','DoctorInCharge'], axis=1)
y = diabetes_df['Diagnosis']

#DATA PARTITIONING
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=242)
```

0.2.2 Model Fitting

[211]:
```
## Model Fitting - Random Forest Classifier
model = RandomForestClassifier(n_estimators=100, random_state=242)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

[212]:
```
# Evaluate the model
print(confusion_matrix(y_test, y_pred))
```

```
[[224   6]
 [ 24 122]]
```

[213]:
```
print(classification_report(y_test, y_pred))
print("ROC-AUC Score:", roc_auc_score(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.90      0.97      0.94       230
           1       0.95      0.84      0.89       146

    accuracy                           0.92       376
   macro avg       0.93      0.90      0.91       376
weighted avg       0.92      0.92      0.92       376

ROC-AUC Score: 0.9047647409172126
```

[214]:
```
#Standardized data to same scale
standardizer = StandardScaler()
x_standardized =standardizer.fit_transform(X_train)
x_test_std = standardizer.fit_transform(X_test)
```

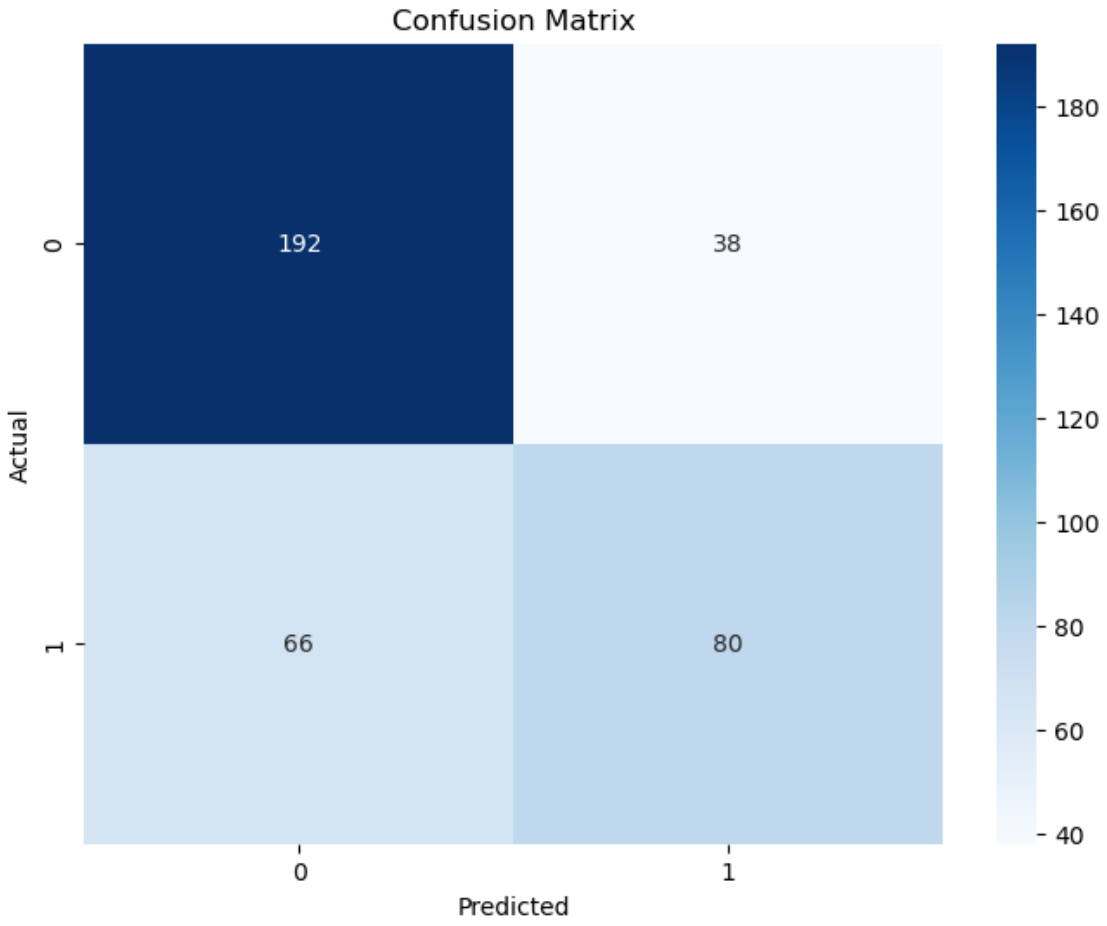
[215]:
```
# Train the Logistic Regression model
logistic_model = LogisticRegression()
logistic_model.fit(X_train, y_train)
```

[215]:
```
LogisticRegression()
```

[216]:
```
# Make predictions on the test set
logit_y_pred = logistic_model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, logit_y_pred)
conf_matrix = confusion_matrix(y_test, logit_y_pred)
class_report = classification_report(y_test, logit_y_pred)

print(f'Accuracy: {accuracy}')
print('Confusion Matrix:')
print(conf_matrix)
print('Classification Report:')
print(class_report)
```

```
Accuracy: 0.723404255319149
Confusion Matrix:
[[192  38]
 [ 66  80]]
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.83      0.79       230
           1       0.68      0.55      0.61       146

    accuracy                           0.72       376
   macro avg       0.71      0.69      0.70       376
weighted avg       0.72      0.72      0.72       376
```

[217]:
```
# Plotting the confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', cbar=True)
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```


Confusion Matrix
192
38
66
80
0
1
Actual
0
1
Predicted
180
160
140
120
100
80
60
40

[ ]:

[ ]: