knn1-1

November 21, 2023

1 ASSIGNMENT 4

NAME: RISHAV KUMAR

ROLL NO. 2301560042 My github account link : Github

2 Question 1

Dataset link: LINK

[180]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split

from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
```

[181]:
```
df= pd.read_csv(r'C:
↪\Users\risha\Documents\KRMU\AIML_assigment\datasets\children_anemia.csv')
```

[182]:
```
df.head()
```

[182]:

	Age in 5-year groups	Type of place of residence	Highest educational level	\
0	40-44	Urban	Higher	
1	35-39	Urban	Higher	
2	25-29	Urban	Higher	

```
3                25-29                      Urban                  Secondary
4                20-24                      Urban                  Secondary

  Wealth index combined  Births in last five years  \
0               Richest                          1
1               Richest                          1
2               Richest                          1
3               Richest                          1
4               Richest                          1

   Age of respondent at 1st birth  \
0                              22
1                              28
2                              26
3                              25
4                              21

   Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)  \
0                                                 NaN
1                                                 NaN
2                                                 NaN
3                                                95.0
4                                                 NaN

  Anemia level  \
0          NaN
1          NaN
2          NaN
3     Moderate
4          NaN

  Have mosquito bed net for sleeping (from household questionnaire)  \
0                                                 Yes
1                                                 Yes
2                                                  No
3                                                 Yes
4                                                 Yes

  Smokes cigarettes              Current marital status  \
0                No                 Living with partner
1                No                             Married
2                No                             Married
3                No                             Married
4                No  No longer living together/separated

  Currently residing with husband/partner When child put to breast  \
0                      Staying elsewhere              Immediately
```

```
1                          Living with her                 Hours: 1
2                          Living with her               Immediately
3                          Living with her                    105.0
4                                      NaN               Immediately

  Had fever in last two weeks  \
0                          No
1                          No
2                          No
3                          No
4                          No

   Hemoglobin level adjusted for altitude (g/dl - 1 decimal) Anemia level.1  \
0                                                  NaN                   NaN
1                                                  NaN                   NaN
2                                                  NaN                   NaN
3                                                114.0           Not anemic
4                                                  NaN                   NaN

  Taking iron pills, sprinkles or syrup
0                                   Yes
1                                    No
2                                    No
3                                    No
4                                    No
```

[183]:
```
df.isna().sum()
```

[183]:
```
Age in 5-year groups                                                    0
Type of place of residence                                              0
Highest educational level                                               0
Wealth index combined                                                   0
Births in last five years                                               0
Age of respondent at 1st birth                                          0
Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)   20788
Anemia level                                                            20788
Have mosquito bed net for sleeping (from household questionnaire)       0
Smokes cigarettes                                                       0
Current marital status                                                  0
Currently residing with husband/partner                                 1698
When child put to breast                                                12756
Had fever in last two weeks                                             3211
Hemoglobin level adjusted for altitude (g/dl - 1 decimal)               23742
Anemia level.1                                                          23742
Taking iron pills, sprinkles or syrup                                   3211
dtype: int64
```

[184]:
```
df.shape
```

[184]:
```
(33924, 17)
```

[185]:
```
df.duplicated().sum()
```

[185]:
```
4678
```

[186]:
```
df=df.drop_duplicates()
```

[187]:
```
df.columns
```

[187]:
```
Index(['Age in 5-year groups', 'Type of place of residence',
       'Highest educational level', 'Wealth index combined',
       'Births in last five years', 'Age of respondent at 1st birth',
       'Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)',
       'Anemia level',
       'Have mosquito bed net for sleeping (from household questionnaire)',
       'Smokes cigarettes', 'Current marital status',
       'Currently residing with husband/partner', 'When child put to breast',
       'Had fever in last two weeks',
       'Hemoglobin level adjusted for altitude (g/dl - 1 decimal)',
       'Anemia level.1', 'Taking iron pills, sprinkles or syrup'],
      dtype='object')
```

[188]:
```
df['Anemia level.1']
```

[188]:
```
0              NaN
1              NaN
2              NaN
3        Not anemic
4              NaN
            …
33919    Not anemic
33920    Not anemic
33921    Not anemic
33922      Moderate
33923           NaN
Name: Anemia level.1, Length: 29246, dtype: object
```

[189]:
```
df.dropna(subset=['Anemia level.1'], inplace=True)
```

[190]:
```
df.reset_index(drop=True, inplace=True)
```

[191]:
```
df.head()
```

[191]:

	Age in 5-year groups	Type of place of residence	Highest educational level
0	25-29	Urban	Secondary
1	30-34	Urban	Higher
2	35-39	Urban	Secondary
3	20-24	Urban	Secondary
4	25-29	Urban	Higher

	Wealth index combined	Births in last five years
0	Richest	1
1	Richest	1
2	Richest	2
3	Richest	1
4	Richest	1

	Age of respondent at 1st birth
0	25
1	30
2	32
3	19
4	24

	Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)
0	95.0
1	113.0
2	121.0
3	108.0
4	116.0

	Anemia level
0	Moderate
1	Mild
2	Not anemic
3	Moderate
4	Mild

	Have mosquito bed net for sleeping (from household questionnaire)
0	Yes
1	Yes
2	Yes
3	Yes
4	Yes

	Smokes cigarettes	Current marital status
0	No	Married
1	No	Married
2	No	Married
3	No	Married

```
4                No                Married

  Currently residing with husband/partner When child put to breast  \
0                         Living with her                    105.0
1                         Living with her                      NaN
2                         Living with her              Immediately
3                         Living with her              Immediately
4                         Living with her                 Days: 1

  Had fever in last two weeks  \
0                          No
1                          No
2                          No
3                          No
4                          No

   Hemoglobin level adjusted for altitude (g/dl - 1 decimal) Anemia level.1  \
0                                              114.0             Not anemic
1                                              119.0             Not anemic
2                                              102.0                   Mild
3                                              113.0             Not anemic
4                                              109.0                   Mild

  Taking iron pills, sprinkles or syrup
0                                    No
1                                    No
2                                   Yes
3                                   Yes
4                                    No
```

[192]:
```
df.shape
```

[192]:
```
(10171, 17)
```

[193]:
```
df.isna().sum()
```

[193]:
```
Age in 5-year groups                                                    0
Type of place of residence                                              0
Highest educational level                                               0
Wealth index combined                                                   0
Births in last five years                                               0
Age of respondent at 1st birth                                          0
Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)     120
Anemia level                                                          120
Have mosquito bed net for sleeping (from household questionnaire)       0
Smokes cigarettes                                                       0
Current marital status                                                  0
```

```
Currently residing with husband/partner                                  518
When child put to breast                                                3806
Had fever in last two weeks                                                0
Hemoglobin level adjusted for altitude (g/dl - 1 decimal)                  0
Anemia level.1                                                             0
Taking iron pills, sprinkles or syrup                                      0
dtype: int64
```

[194]:
```
hemo_level_adjusted = df['Hemoglobin level adjusted for altitude and smoking (g/
↪dl - 1 decimal)']
hemo_level_adjusted
```

[194]:
```
0         95.0
1        113.0
2        121.0
3        108.0
4        116.0
         ...
10166    120.0
10167    120.0
10168    120.0
10169    149.0
10170    123.0
Name: Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal),
Length: 10171, dtype: float64
```

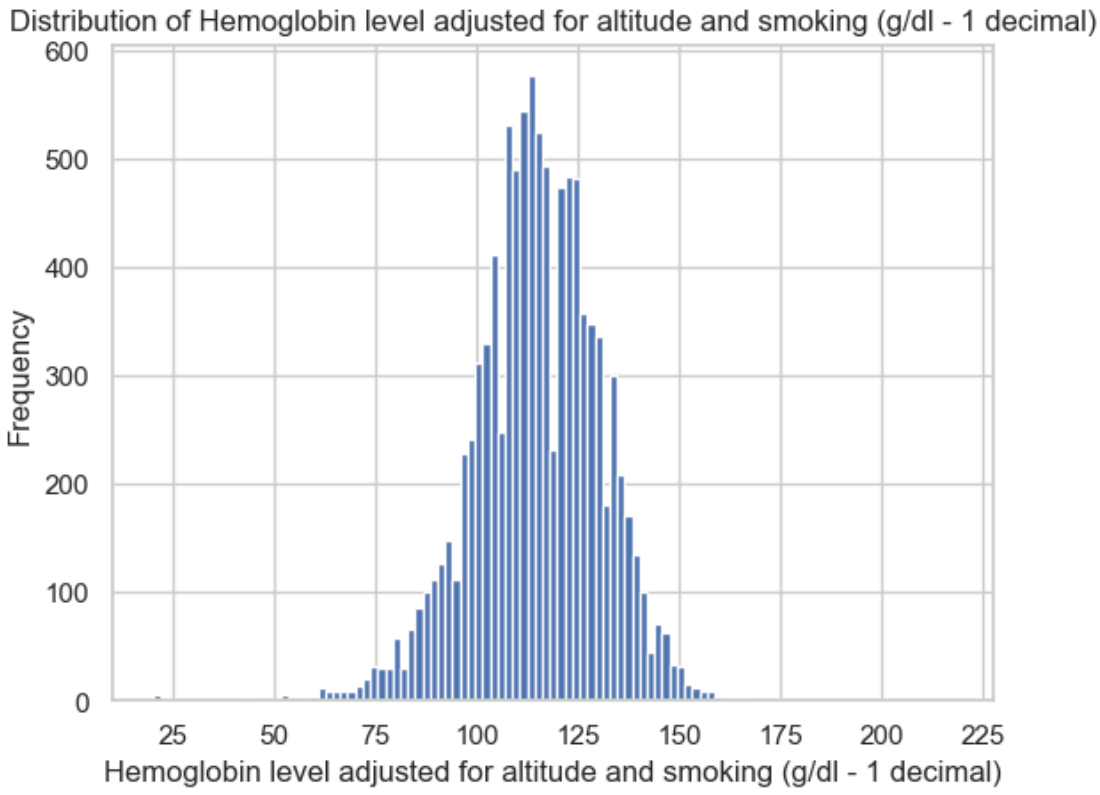
[195]:
```
hemo_level_adjusted.isna().sum()
```

[195]:
```
120
```

[196]:
```
hemo_level_adjusted.var()
```

[196]:
```
247.79435994834301
```

[197]:
```
plt.hist(df['Hemoglobin level adjusted for altitude and smoking (g/dl - 1␣
↪decimal)'], bins='auto')
plt.title('Distribution of Hemoglobin level adjusted for altitude and smoking␣
↪(g/dl - 1 decimal)')
plt.xlabel('Hemoglobin level adjusted for altitude and smoking (g/dl - 1␣
↪decimal)')
plt.ylabel('Frequency')
plt.show()
```



[198]:
```
hemo_level_adjusted.fillna(hemo_level_adjusted.mean(), inplace=True)
```

[199]:
```
df.drop(columns=['Currently residing with husband/partner','Anemia␣
↪level'],inplace=True)
```

[200]:
```
breast_feed = df['When child put to breast']
breast_feed
```

[200]:
```
0              105.0
1                NaN
2        Immediately
3        Immediately
4           Days: 1
              …
10166    Immediately
10167            NaN
10168       Hours: 1
10169       Hours: 1
10170    Immediately
Name: When child put to breast, Length: 10171, dtype: object
```

[201]:
```
breast_feed.unique()
```

[201]:
```
array(['105.0', nan, 'Immediately', 'Days: 1', 'Hours: 1', '103.0',
       '203.0', '102.0', '111.0', '106.0', '104.0', '202.0', '107.0',
       '108.0', '120.0', '123.0', '110.0', '112.0', '207.0', '109.0',
       '113.0', '205.0', '115.0', '117.0', '212.0', '114.0', '204.0',
       '119.0', '211.0', '121.0', '214.0', '206.0', '118.0', '210.0',
       '208.0', '116.0', '223.0', '220.0'], dtype=object)
```

[202]:
```
breast_feed.shape
```

[202]:
```
(10171,)
```

[203]:
```
breast_feed.isna().sum()
```

[203]:
```
3806
```

[204]:
```
df.drop(columns='When child put to breast',inplace=True)
```

[205]:
```
df.isna().sum()
```

[205]:
```
Age in 5-year groups                                                0
Type of place of residence                                          0
Highest educational level                                           0
Wealth index combined                                               0
Births in last five years                                           0
Age of respondent at 1st birth                                      0
Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)    0
Have mosquito bed net for sleeping (from household questionnaire)       0
Smokes cigarettes                                                   0
Current marital status                                              0
Had fever in last two weeks                                         0
Hemoglobin level adjusted for altitude (g/dl - 1 decimal)           0
Anemia level.1                                                      0
Taking iron pills, sprinkles or syrup                               0
dtype: int64
```

[206]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10171 entries, 0 to 10170
Data columns (total 14 columns):
 #   Column                                                      Non-
Null Count  Dtype
---  ------
--------------  -----
 0   Age in 5-year groups
```

```
10171 non-null  object
 1   Type of place of residence
10171 non-null  object
 2   Highest educational level
10171 non-null  object
 3   Wealth index combined
10171 non-null  object
 4   Births in last five years
10171 non-null  int64
 5   Age of respondent at 1st birth
10171 non-null  int64
 6   Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)
10171 non-null  float64
 7   Have mosquito bed net for sleeping (from household questionnaire)
10171 non-null  object
 8   Smokes cigarettes
10171 non-null  object
 9   Current marital status
10171 non-null  object
 10  Had fever in last two weeks
10171 non-null  object
 11  Hemoglobin level adjusted for altitude (g/dl - 1 decimal)
10171 non-null  float64
 12  Anemia level.1
10171 non-null  object
 13  Taking iron pills, sprinkles or syrup
10171 non-null  object
dtypes: float64(2), int64(2), object(10)
memory usage: 1.1+ MB
```

Separating Dependent And Independent Variables

[207]:
```
y = df['Anemia level.1']
x = df.drop(columns='Anemia level.1',axis=1)
```

[208]:
```
x.shape
```

[208]:
```
(10171, 13)
```

[209]:
```
y
```

[209]:
```
0        Not anemic
1        Not anemic
2              Mild
3        Not anemic
4              Mild
            ...
10166          Mild
```

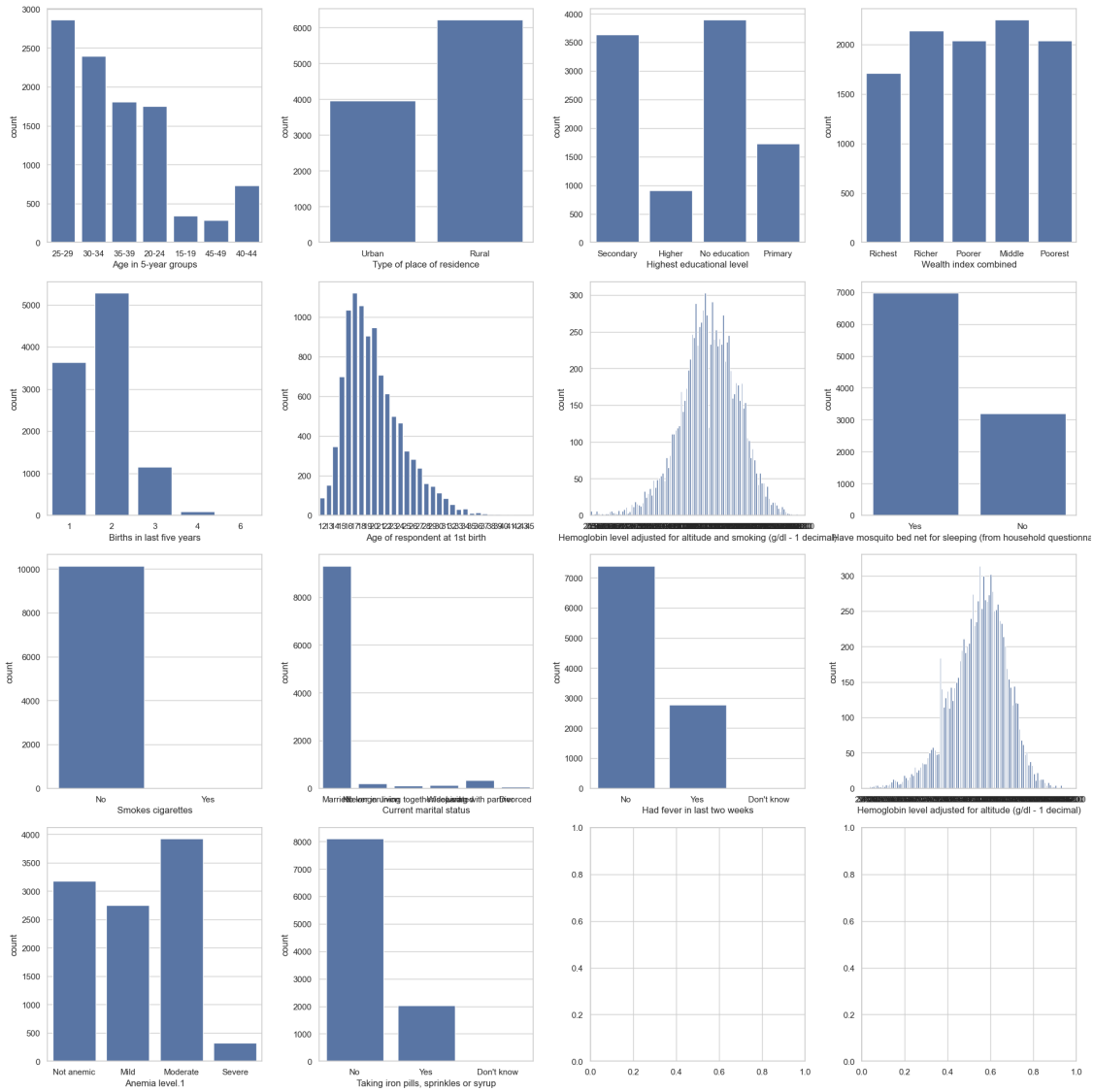
```
10167    Not anemic
10168    Not anemic
10169    Not anemic
10170      Moderate
Name: Anemia level.1, Length: 10171, dtype: object
```

Visualizing the Data

[210]:
```
y.unique()
```

[210]:
```
array(['Not anemic', 'Mild', 'Moderate', 'Severe'], dtype=object)
```

[211]:
```
count=0
fig, ax=plt.subplots(4,4,figsize=(20,20))
ax=ax.flatten()
for i in df.columns:
    sns.countplot(df,x=i,ax=ax[count])
    count+=1
plt.tight_layout()
```



[212]:
```
x['Age in 5-year groups'].unique()
```

[212]:
```
array(['25-29', '30-34', '35-39', '20-24', '15-19', '45-49', '40-44'],
      dtype=object)
```

[213]:
```
x['Age in 5-year groups'] = x['Age in 5-year groups'].apply(lambda x:␣
 ↪sum(map(int, x.split('-'))) / 2 if isinstance(x, str) else x)
```

[214]:
```
x
```

[214]:

	Age in 5-year groups	Type of place of residence	\
0	27.0	Urban	
1	32.0	Urban	

```
2                    37.0                    Urban
3                    22.0                    Urban
4                    27.0                    Urban
…                    …                       …
10166                37.0                    Rural
10167                37.0                    Rural
10168                27.0                    Rural
10169                27.0                    Rural
10170                22.0                    Rural

      Highest educational level Wealth index combined  \
0                     Secondary               Richest
1                        Higher               Richest
2                     Secondary               Richest
3                     Secondary               Richest
4                        Higher               Richest
…                         …                   …
10166                 Secondary                Richer
10167                 Secondary                Richer
10168              No education                Richer
10169                    Higher                Richer
10170                 Secondary                Richer

       Births in last five years  Age of respondent at 1st birth  \
0                              1                              25
1                              1                              30
2                              2                              32
3                              1                              19
4                              1                              24
…                            …                              …
10166                          2                              19
10167                          2                              19
10168                          1                              27
10169                          1                              22
10170                          1                              21

       Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)  \
0                                                   95.0
1                                                  113.0
2                                                  121.0
3                                                  108.0
4                                                  116.0
…                                                  …
10166                                              120.0
10167                                              120.0
10168                                              120.0
10169                                              149.0
```

```
10170                                                 123.0

      Have mosquito bed net for sleeping (from household questionnaire)  \
0                                                       Yes
1                                                       Yes
2                                                       Yes
3                                                       Yes
4                                                       Yes
…                                                       …
10166                                                   Yes
10167                                                   Yes
10168                                                   Yes
10169                                                   Yes
10170                                                   Yes

      Smokes cigarettes Current marital status Had fever in last two weeks  \
0                    No                Married                          No
1                    No                Married                          No
2                    No                Married                          No
3                    No                Married                          No
4                    No                Married                          No
…                   …                  …                               …
10166                No                Married                          No
10167                No                Married                          No
10168                No         Never in union                          No
10169                No                Married                          No
10170                No                Married                          No

       Hemoglobin level adjusted for altitude (g/dl - 1 decimal)  \
0                                                     114.0
1                                                     119.0
2                                                     102.0
3                                                     113.0
4                                                     109.0
…                                                     …
10166                                                 108.0
10167                                                 120.0
10168                                                 120.0
10169                                                 119.0
10170                                                  75.0

      Taking iron pills, sprinkles or syrup
0                                        No
1                                        No
2                                       Yes
3                                       Yes
4                                        No
```

```
…                                       …
10166                                     Yes
10167                                     Yes
10168                                      No
10169                                      No
10170                                     Yes

[10171 rows x 13 columns]
```

[215]:
```
x['Age in 5-year groups'].dtype
```

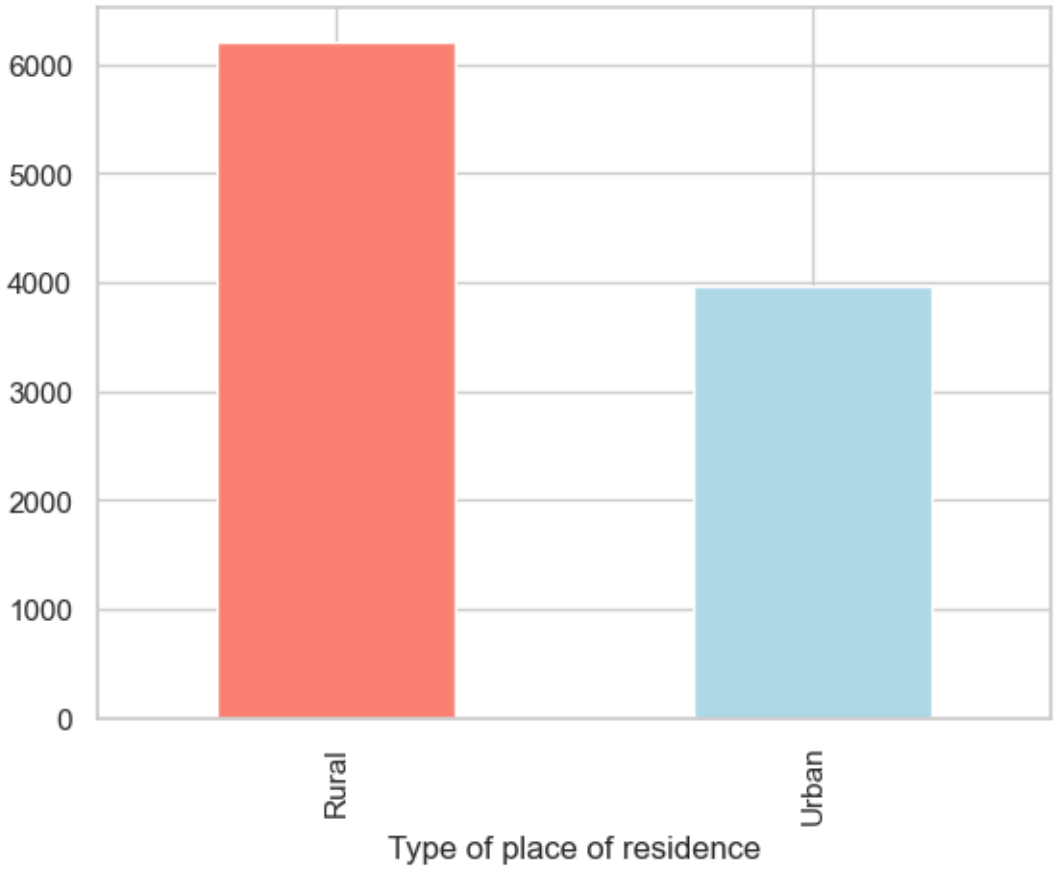
[215]:
```
dtype('float64')
```

[216]:
```
x['Type of place of residence'].unique()
```

[216]:
```
array(['Urban', 'Rural'], dtype=object)
```

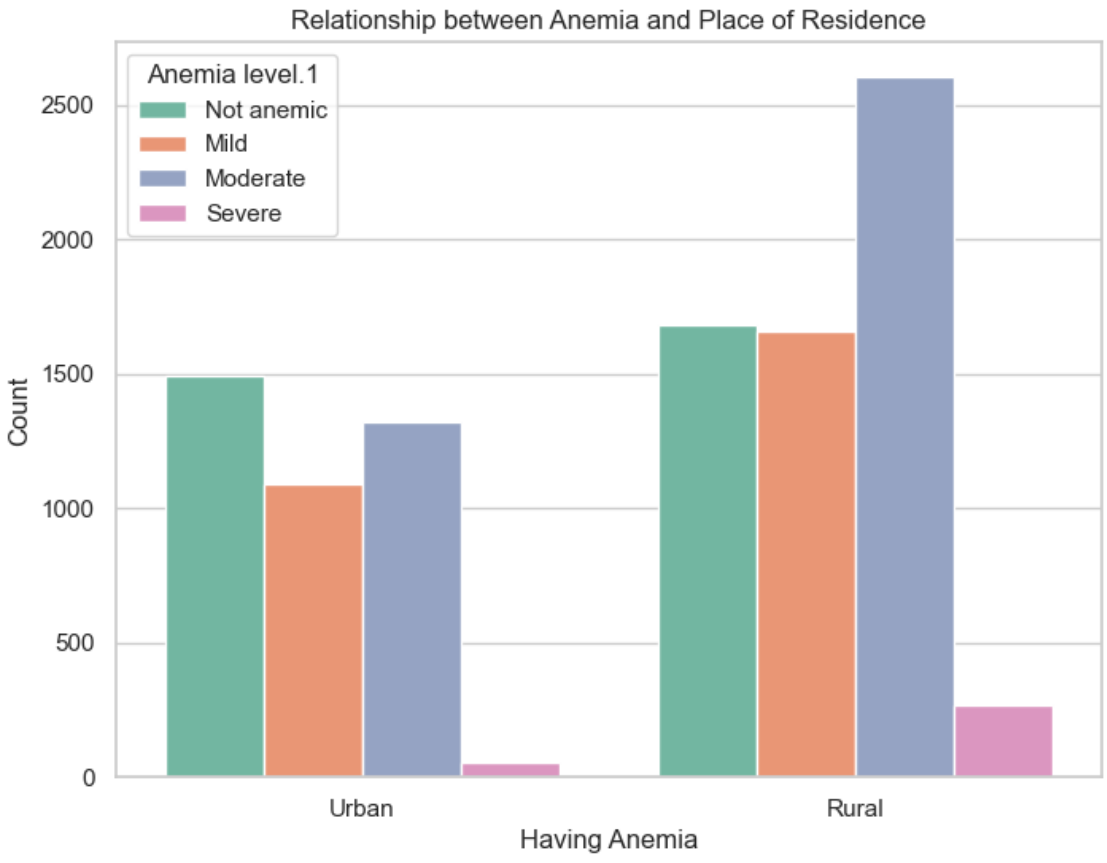
[217]:
```
x['Type of place of residence'].value_counts().plot(kind="bar",␣
↪color=["salmon", "lightblue"])
```

[217]:
```
<Axes: xlabel='Type of place of residence'>
```

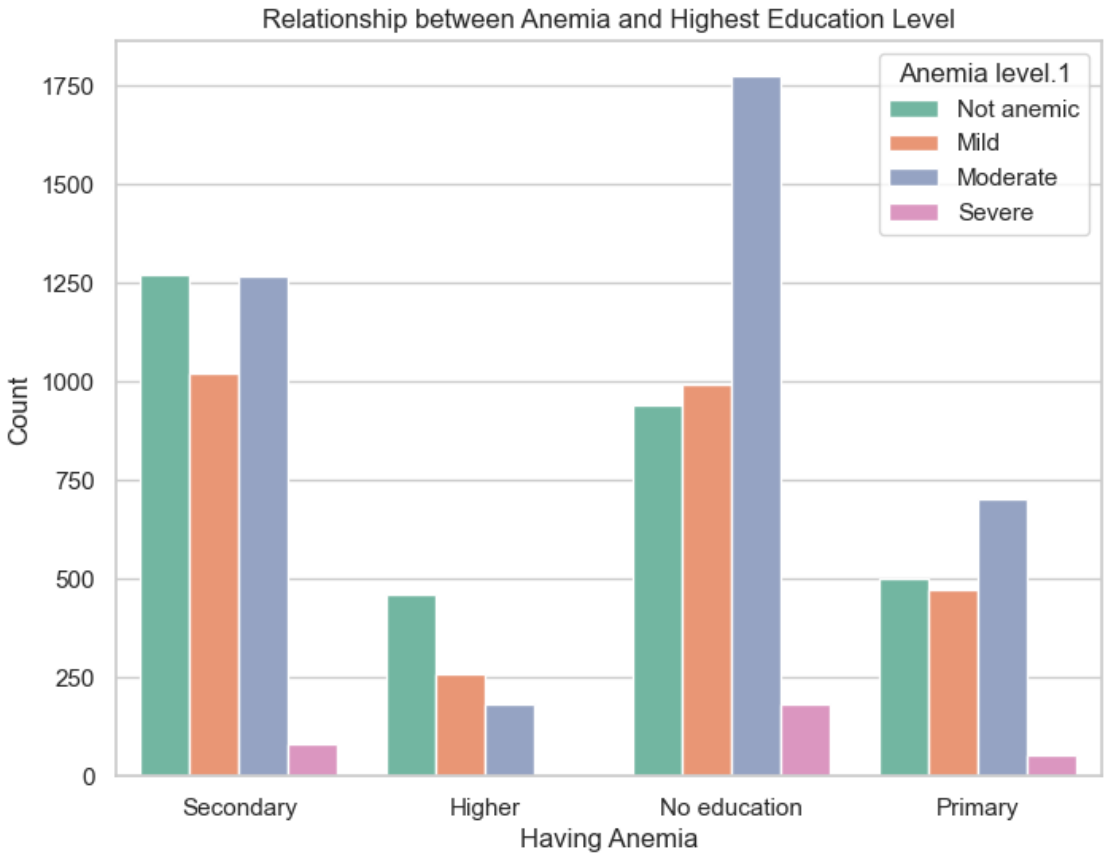


[218]:
```
def plotter(x,y,title):
    sns.set(style="whitegrid")
    plt.figure(figsize=(8, 6))
    sns.countplot(x=x, hue=y,palette="Set2")

    plt.title(title)
    plt.xlabel('Having Anemia')
    plt.ylabel('Count')

    plt.show()
    return
```

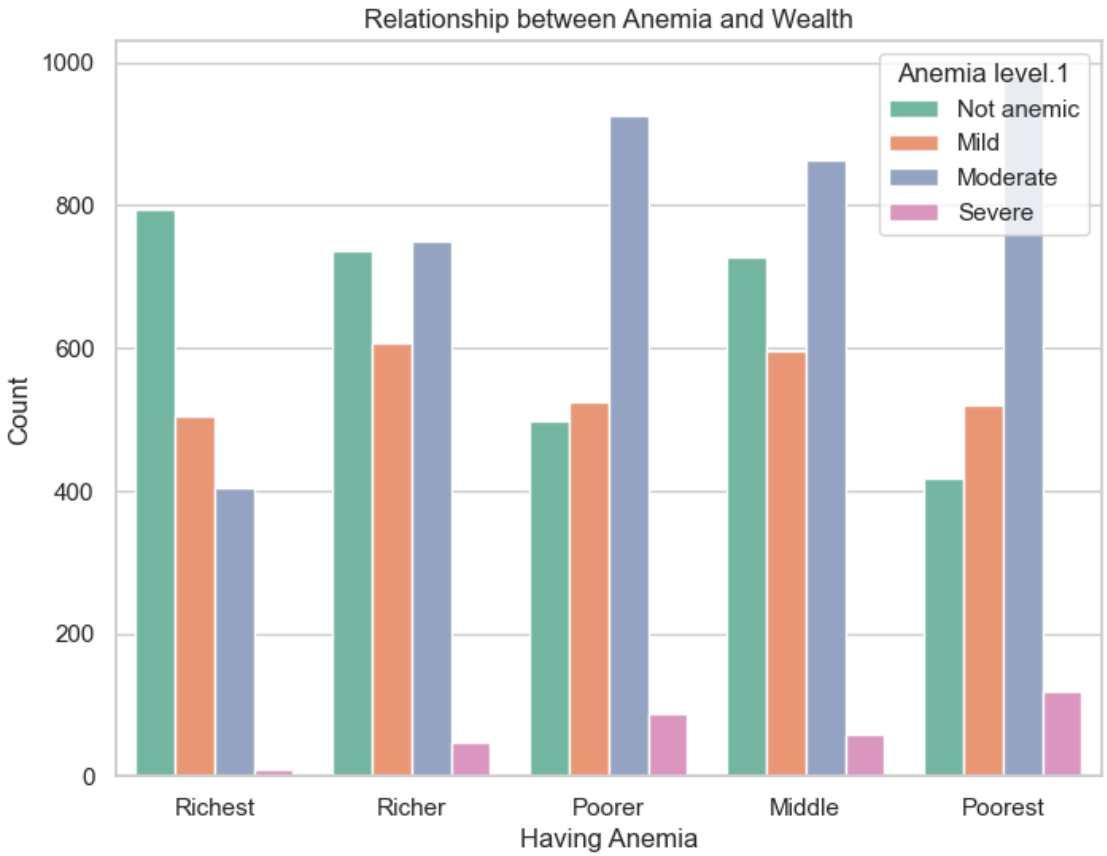
[219]:
```
plotter(x['Type of place of residence'],y,'Relationship between Anemia and␣
↪Place of Residence')
```



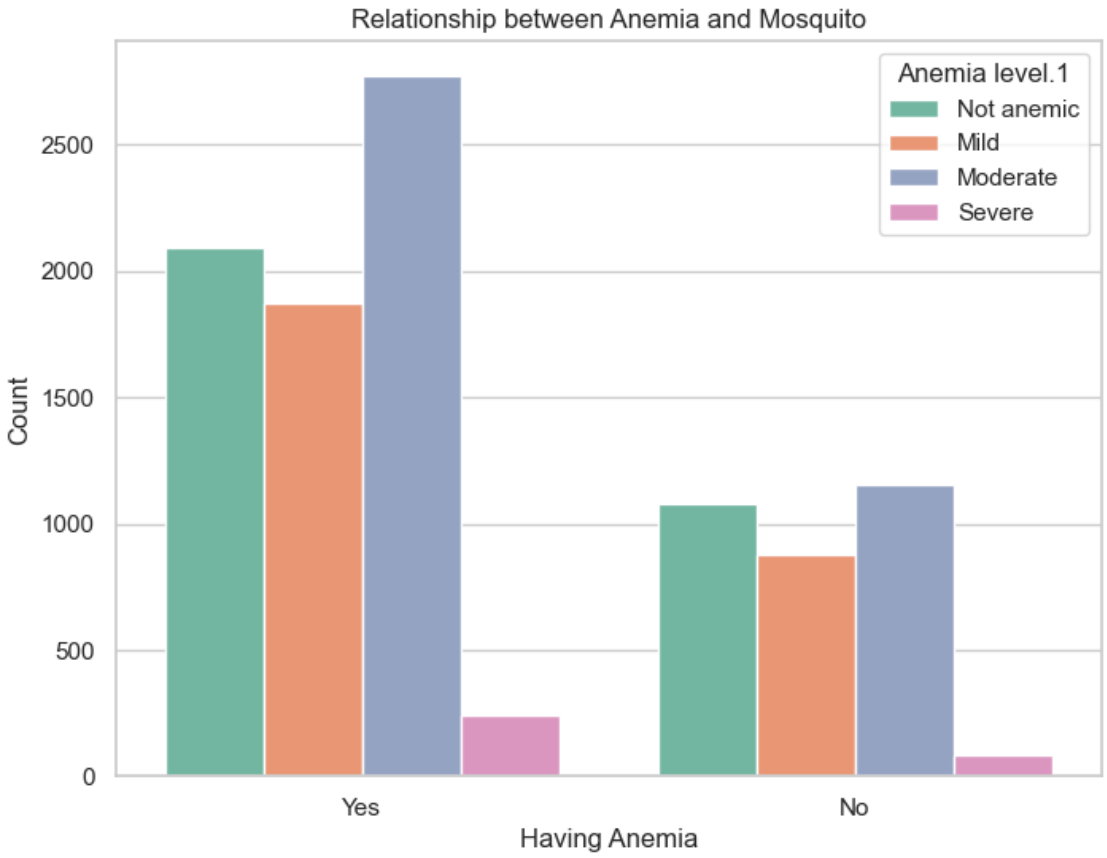
[220]:
```
plotter(x['Highest educational level'],y,'Relationship between Anemia and␣
↪Highest Education Level')
```



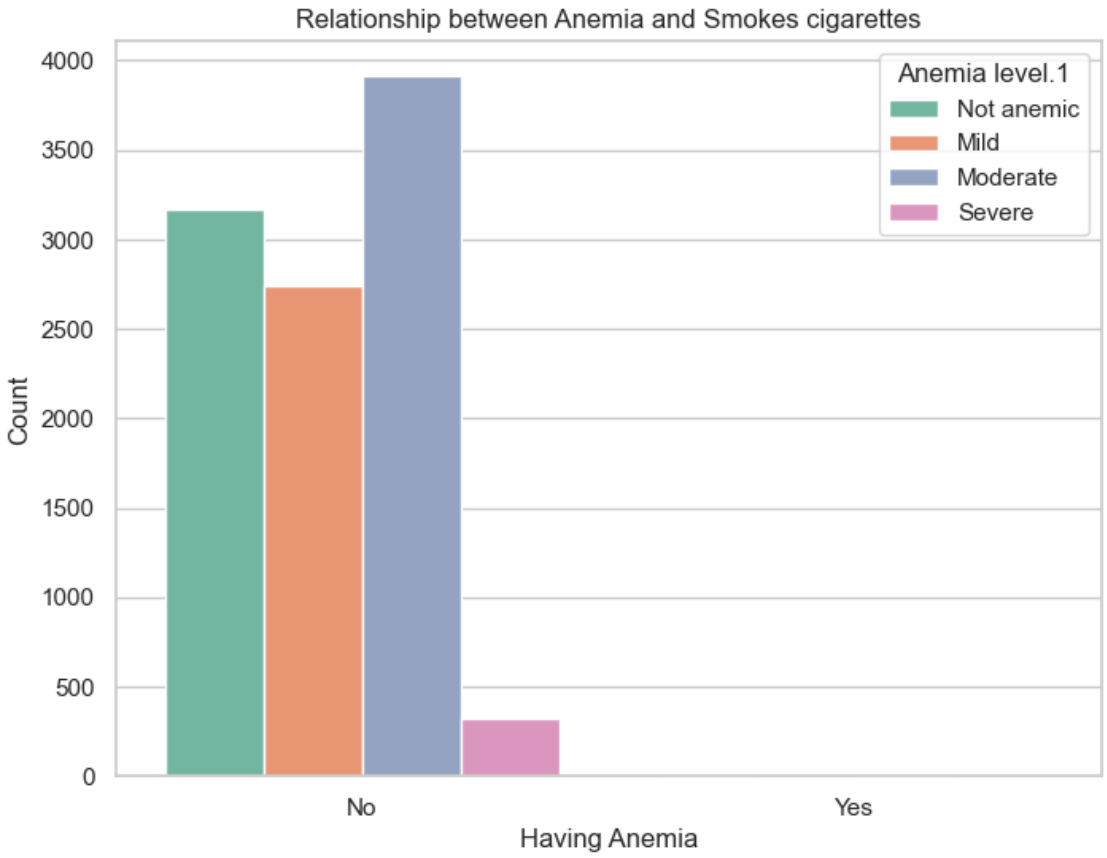
[221]:
```
plotter(x['Wealth index combined'],y,'Relationship between Anemia and Wealth')
```



[222]:
```
plotter(x['Have mosquito bed net for sleeping (from household␣
↪questionnaire)'],y,'Relationship between Anemia and Mosquito')
```



[223]:
```
plotter(x['Smokes cigarettes'],y,'Relationship between Anemia and Smokes␣
↪cigarettes')
```



[224]:
```
x['Smokes cigarettes'].unique()
```

[224]:
```
array(['No', 'Yes'], dtype=object)
```

[225]:
```
x['Smokes cigarettes'].value_counts()
```

[225]:
```
Smokes cigarettes
No     10147
Yes       24
Name: count, dtype: int64
```

I think smoking cigarettes doesn't affect having Anemia or not , so let's drop that column

[226]:
```
x.drop(columns='Smokes cigarettes',axis=1,inplace=True)
```

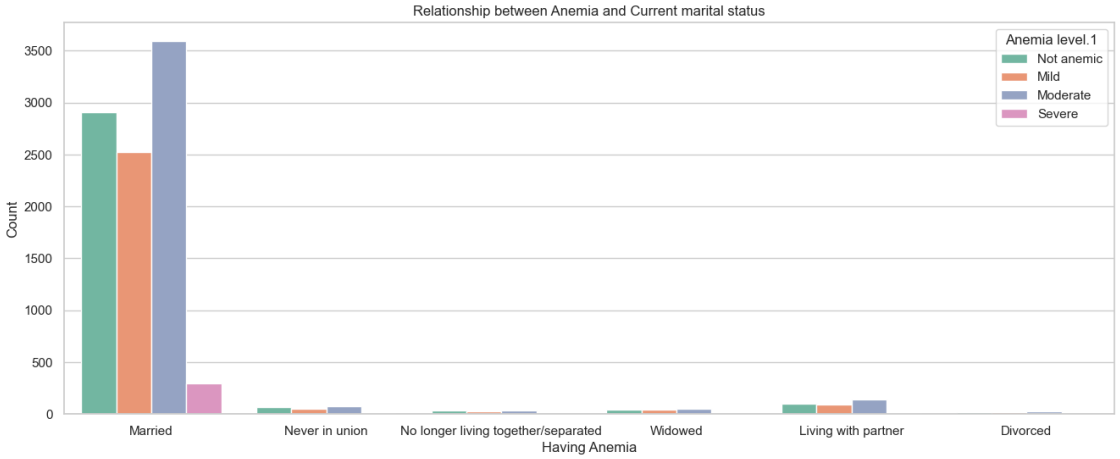
[227]:
```
x.head()
```

[227]:
```
   Age in 5-year groups Type of place of residence Highest educational level  \
0                  27.0                      Urban                 Secondary
1                  32.0                      Urban                    Higher
```

```
2                  37.0                       Urban                  Secondary
3                  22.0                       Urban                  Secondary
4                  27.0                       Urban                     Higher

  Wealth index combined  Births in last five years  \
0               Richest                          1
1               Richest                          1
2               Richest                          2
3               Richest                          1
4               Richest                          1

   Age of respondent at 1st birth  \
0                              25
1                              30
2                              32
3                              19
4                              24

   Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)  \
0                                                95.0
1                                               113.0
2                                               121.0
3                                               108.0
4                                               116.0

  Have mosquito bed net for sleeping (from household questionnaire)  \
0                                                Yes
1                                                Yes
2                                                Yes
3                                                Yes
4                                                Yes

  Current marital status Had fever in last two weeks  \
0                Married                          No
1                Married                          No
2                Married                          No
3                Married                          No
4                Married                          No

   Hemoglobin level adjusted for altitude (g/dl - 1 decimal)  \
0                                               114.0
1                                               119.0
2                                               102.0
3                                               113.0
4                                               109.0

  Taking iron pills, sprinkles or syrup
```

```
0                                    No
1                                    No
2                                   Yes
3                                   Yes
4                                    No
```

[228]:
```
sns.set(style="whitegrid")
plt.figure(figsize=(16, 6))
sns.countplot(x=x['Current marital status'], hue=y,palette="Set2")

plt.title("Relationship between Anemia and Current marital status")
plt.xlabel('Having Anemia')
plt.ylabel('Count')

plt.show()
```



I think Current marital status doesn't affect having Anemia or not , so let's drop that column

[229]:
```
x.drop(columns='Current marital status',axis=1,inplace=True)
```

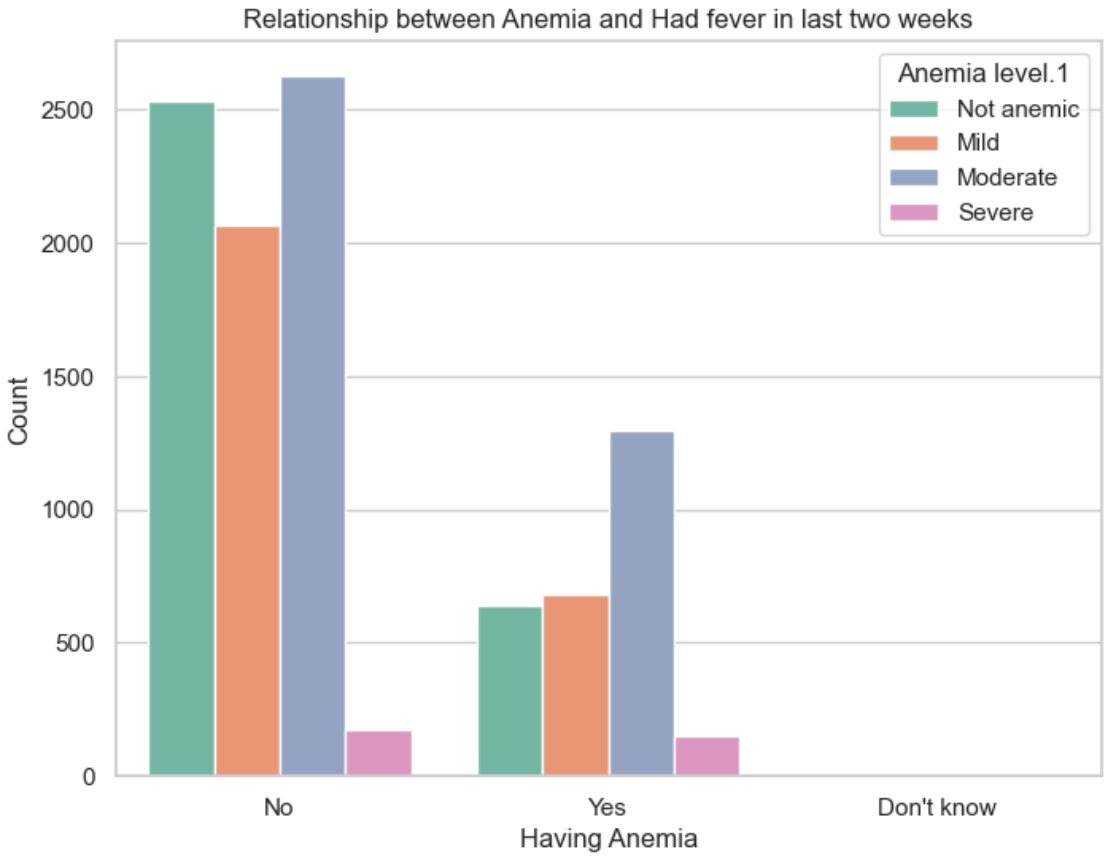
[230]:
```
x.columns
```

[230]:
```
Index(['Age in 5-year groups', 'Type of place of residence',
       'Highest educational level', 'Wealth index combined',
       'Births in last five years', 'Age of respondent at 1st birth',
       'Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)',
       'Have mosquito bed net for sleeping (from household questionnaire)',
       'Had fever in last two weeks',
       'Hemoglobin level adjusted for altitude (g/dl - 1 decimal)',
       'Taking iron pills, sprinkles or syrup'],
      dtype='object')
```

[231]:
```
x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10171 entries, 0 to 10170
Data columns (total 11 columns):
 #   Column                                                                Non-
Null Count  Dtype
---  ------
--------------  -----
 0   Age in 5-year groups
10171 non-null  float64
 1   Type of place of residence
10171 non-null  object
 2   Highest educational level
10171 non-null  object
 3   Wealth index combined
10171 non-null  object
 4   Births in last five years
10171 non-null  int64
 5   Age of respondent at 1st birth
10171 non-null  int64
 6   Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)
10171 non-null  float64
 7   Have mosquito bed net for sleeping (from household questionnaire)
10171 non-null  object
 8   Had fever in last two weeks
10171 non-null  object
 9   Hemoglobin level adjusted for altitude (g/dl - 1 decimal)
10171 non-null  float64
 10  Taking iron pills, sprinkles or syrup
10171 non-null  object
dtypes: float64(3), int64(2), object(6)
memory usage: 874.2+ KB
```

[232]:
```
plotter(x['Had fever in last two weeks'],y,'Relationship between Anemia and Had␣
↪fever in last two weeks')
```



convert our catergorical data to numerical using one hot encoding

[233]:
```
x= x.iloc[:,:].values
x
```

[233]:
```
array([[27.0, 'Urban', 'Secondary', …, 'No', 114.0, 'No'],
       [32.0, 'Urban', 'Higher', …, 'No', 119.0, 'No'],
       [37.0, 'Urban', 'Secondary', …, 'No', 102.0, 'Yes'],
       …,
       [27.0, 'Rural', 'No education', …, 'No', 120.0, 'No'],
       [27.0, 'Rural', 'Higher', …, 'No', 119.0, 'No'],
       [22.0, 'Rural', 'Secondary', …, 'No', 75.0, 'Yes']], dtype=object)
```

[234]:
```
ct =␣
↪ColumnTransformer(transformers=[('encoder',OneHotEncoder(),[1,2,3,7,8,10])],remainder='pass
x = np.array(ct.fit_transform(x))
```

[235]:
```
x
```

[235]:
```
array([[0.0, 1.0, 0.0, …, 25, 95.0, 114.0],
       [0.0, 1.0, 1.0, …, 30, 113.0, 119.0],
       [0.0, 1.0, 0.0, …, 32, 121.0, 102.0],
       …,
       [1.0, 0.0, 0.0, …, 27, 120.0, 120.0],
       [1.0, 0.0, 1.0, …, 22, 149.0, 119.0],
       [1.0, 0.0, 0.0, …, 21, 123.0, 75.0]], dtype=object)
```

let's convert our dependant categroical data into numerical data

[236]:
```
le = LabelEncoder()
y = np.array(le.fit_transform(y))
y
```

[236]:
```
array([2, 2, 0, …, 2, 2, 1])
```

[237]:
```
np.unique(y)
```

[237]:
```
array([0, 1, 2, 3])
```

[238]:
```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.
↪2,random_state=42)

x_train.shape,x_test.shape
```

[238]:
```
((8136, 24), (2035, 24))
```

[239]:
```
y_train.shape,y_test.shape
```

[239]:
```
((8136,), (2035,))
```

[240]:
```
sc = StandardScaler()

x_train[:,[-1,-2,-3]] = sc.fit_transform(x_train[:,[-1,-2,-3]])
x_test[:,[-1,-2,-3]] = sc.fit_transform(x_test[:,[-1,-2,-3]])
x_train
```

[240]:
```
array([[1.0, 0.0, 0.0, …, -1.3459967100539345, 1.3040544531398048,
        -0.08824165576283671],
       [1.0, 0.0, 0.0, …, -1.7976134794623027, 0.019326375584944898,
        1.971827530137993],
       [1.0, 0.0, 0.0, …, -0.21695478653301434, 0.661690414362375,
        2.0362046921973938],
       …,
       [1.0, 0.0, 0.0, …, -1.1201883253497504, 1.3040544531398048,
        -0.08824165576283671],
       [0.0, 1.0, 0.0, …, 0.23466198287535373, 1.689472876406263,
```

```
        1.7786960439597903],
       [1.0, 0.0, 0.0, …, 0.23466198287535373, -1.265401701969915,
        -0.7963904384162469]], dtype=object)
```

[241]:
```
le = LabelEncoder()
y = np.array(le.fit_transform(y))
y
```

[241]:
```
array([2, 2, 0, …, 2, 2, 1], dtype=int64)
```

[242]:
```
x_train,X_test,y_train,y_test = train_test_split(x,y,test_size=0.
↪2,random_state=42)

x_train.shape,X_test.shape,y_train.shape,y_test.shape
```

[242]:
```
((8136, 24), (2035, 24), (8136,), (2035,))
```

[243]:
```
# Put models in a dictionary
models = {"Logistic Regression": LogisticRegression(),
          "KNN": KNeighborsClassifier(n_neighbors=5),
          "Random Forest":␣
↪RandomForestClassifier(n_estimators=100,criterion='entropy'),
          "Kernel SVM" : SVC(kernel='rbf',random_state=0),
          "Support Vector" : SVC(kernel='linear',random_state=0),
          "Naive Based" :GaussianNB(),
          "Decision Tree" : ␣
↪DecisionTreeClassifier(criterion='entropy',random_state=0)
        }
```

[244]:
```
def fit_and_score(models, x_train, x_test, y_train, y_test):
    np.random.seed(42)
    accuracy_scores = {}

    for name, model in models.items():
        model.fit(x_train, y_train)
        y_pred = model.predict(X_test)
        accuracy_scores[name] = accuracy_score(y_test,y_pred)
    return accuracy_scores
```

[245]:
```
model_scores = fit_and_score(models=models, x_train=x_train, x_test=x_test,␣
↪y_train=y_train, y_test=y_test)

model_scores
```

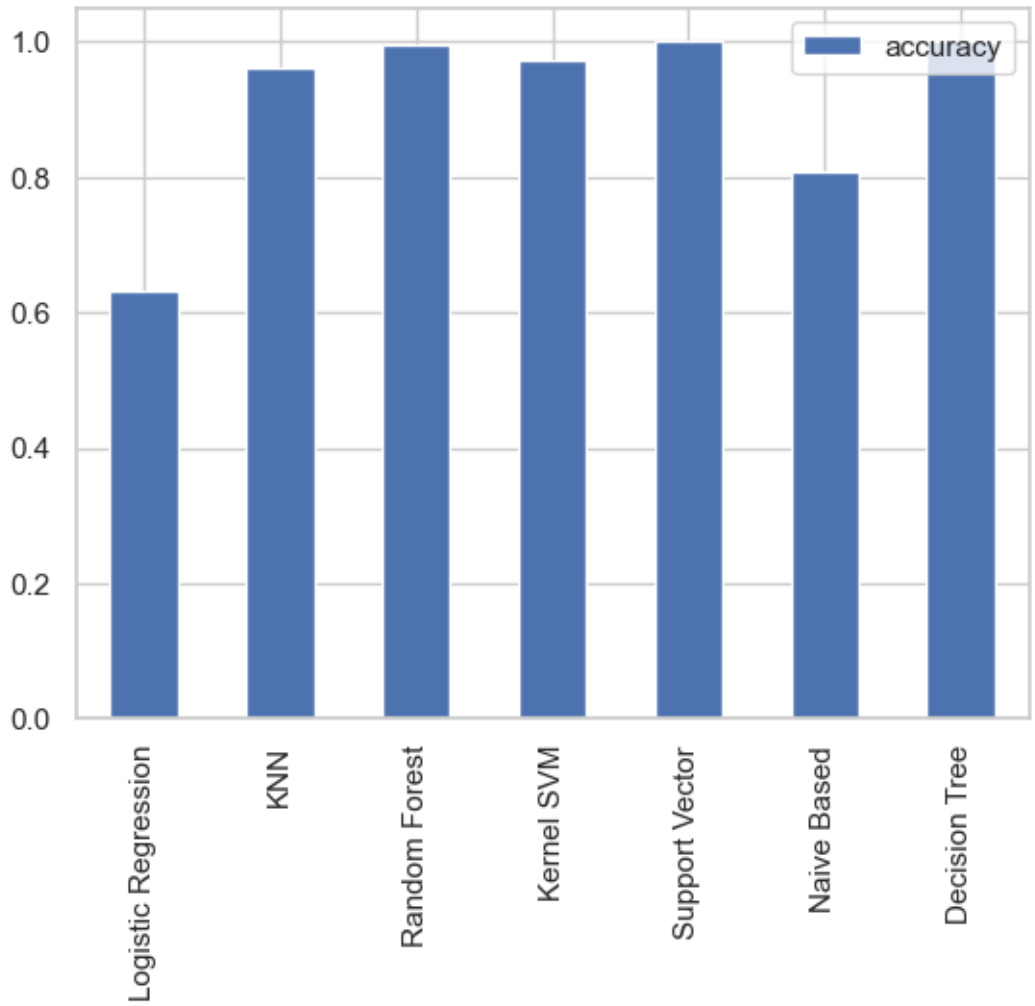
```
c:\Users\risha\AppData\Local\Programs\Python\Python312\Lib\site-
packages\sklearn\linear_model\_logistic.py:460: ConvergenceWarning: lbfgs failed
to converge (status=1):
```

```
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

[245]:
```
{'Logistic Regression': 0.6314496314496314,
 'KNN': 0.9611793611793612,
 'Random Forest': 0.9960687960687961,
 'Kernel SVM': 0.972972972972973,
 'Support Vector': 1.0,
 'Naive Based': 0.8088452088452088,
 'Decision Tree': 1.0}
```

[246]:
```
model_comp = pd.DataFrame(model_scores, index=["accuracy"])
model_comp.T.plot.bar()
```

[246]:
```
<Axes: >
```



Finding Model with highest accuracy score

[247]:
```
best_model = max(model_scores, key=model_scores.get)
best_score = model_scores[best_model]
best_model,best_score
```

[247]:
```
('Support Vector', 1.0)
```

[248]:
```
classifier =  DecisionTreeClassifier(criterion='entropy',random_state=0)
classifier.fit(x_train,y_train)
```

[248]:
```
DecisionTreeClassifier(criterion='entropy', random_state=0)
```

[249]:
```
y_preds = classifier.predict(x_test)
```

[250]:
```
accuracy = accuracy_score(y_test,y_preds)
accuracy
```

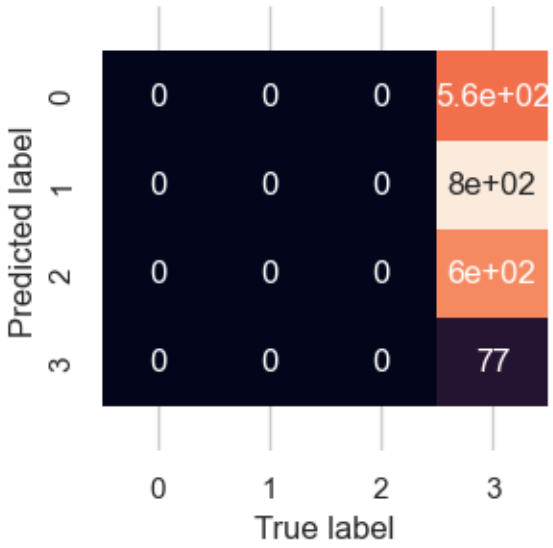
[250]:
```
0.03783783783783784
```

Confusion metrix

[251]:
```
from sklearn.metrics import confusion_matrix,classification_report

print(confusion_matrix(y_test, y_preds))
```

```
[[  0   0   0 556]
 [  0   0   0 798]
 [  0   0   0 604]
 [  0   0   0  77]]
```

[252]:
```
def plot_conf_mat(y_test, y_preds):

    fig, ax = plt.subplots(figsize=(3, 3))
    ax = sns.heatmap(confusion_matrix(y_test, y_preds),
                     annot=True,
                     cbar=False)
    plt.xlabel("True label")
    plt.ylabel("Predicted label")

    bottom, top = ax.get_ylim()
    ax.set_ylim(bottom + 0.5, top - 0.5)

plot_conf_mat(y_test, y_preds)
```



[253]:
```
print(classification_report(y_test, y_preds))
```

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       556
           1       0.00      0.00      0.00       798
           2       0.00      0.00      0.00       604
           3       0.04      1.00      0.07        77

    accuracy                           0.04      2035
   macro avg       0.01      0.25      0.02      2035
weighted avg       0.00      0.04      0.00      2035
```

```
c:\Users\risha\AppData\Local\Programs\Python\Python312\Lib\site-
packages\sklearn\metrics\_classification.py:1471: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
c:\Users\risha\AppData\Local\Programs\Python\Python312\Lib\site-
packages\sklearn\metrics\_classification.py:1471: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
c:\Users\risha\AppData\Local\Programs\Python\Python312\Lib\site-
packages\sklearn\metrics\_classification.py:1471: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

[254]:
```
from collections import Counter
Counter(y_train)
Counter(y_test)
```

[254]:
```
Counter({1: 798, 2: 604, 0: 556, 3: 77})
```

[ ]:
```

```