

# Project - Diabetes Prediction

**Nir Chauser:**

**Rina Beloborodov:**

```
In [1]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import matplotlib.image as mpimg
```

Once you have chosen your dataset:

- 1) State which dataset you chose.
- 2) Provide a brief (2-4 sentences) description of the dataset. What is this dataset about?
- 3) List the features in the dataset and their types.
- 4) List the number of records in the dataset.

## Q1 + Q2

מסד הנתונים אותו בחרנו הוא על סוכרת. סוכרת שאינה מכוקרת מכילה לעודף סוכר בدم, ועם הזמן היא עלולה לגרום לפגיעה ניכרת במערכות הגוף. שיעורי התחלואה בעולם גדים בהתקופה, והמחלה מהווה נטול כלכלי ממשמעות על המשק. לפיכך חיזוי מראש וחקר המחלת יכול להוביל לתוצאות חיוביות הן על המשק והן על החולים העתידיים.

המידע לקוח מהאתר:

<https://www.tasmc.org.il/Be-Well/InterestAreas/diabet/Pages/whatisdiabetes.aspx>

```
In [2]: df_0= pd.read_csv("diabetes_prediction_dataset.csv")
df_0
display(df_0)
display(df_0.describe())
display(df_0.describe(include=[ '0']))
```

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glu
<b>0</b>	Female	80.0	0	1	never	25.19	6.6	
<b>1</b>	Female	54.0	0	0	No Info	27.32	6.6	
<b>2</b>	Male	28.0	0	0	never	27.32	5.7	
<b>3</b>	Female	36.0	0	0	current	23.45	5.0	
<b>4</b>	Male	76.0	1	1	current	20.14	4.8	
...	...	...	...	...	...	...	...	...
<b>99995</b>	Female	80.0	0	0	No Info	27.32	6.2	
<b>99996</b>	Female	2.0	0	0	No Info	17.37	6.5	
<b>99997</b>	Male	66.0	0	0	former	27.83	5.7	
<b>99998</b>	Female	24.0	0	0	never	35.42	4.0	
<b>99999</b>	Female	57.0	0	0	current	22.43	6.6	

100000 rows × 9 columns

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_l
<b>count</b>	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
<b>mean</b>	41.885856	0.07485	0.039420	27.320767	5.527507	138.058
<b>std</b>	22.516840	0.26315	0.194593	6.636783	1.070672	40.708
<b>min</b>	0.080000	0.00000	0.000000	10.010000	3.500000	80.000
<b>25%</b>	24.000000	0.00000	0.000000	23.630000	4.800000	100.000
<b>50%</b>	43.000000	0.00000	0.000000	27.320000	5.800000	140.000
<b>75%</b>	60.000000	0.00000	0.000000	29.580000	6.200000	159.000
<b>max</b>	80.000000	1.00000	1.000000	95.690000	9.000000	300.000

	gender	smoking_history
<b>count</b>	100000	100000
<b>unique</b>	3	6
<b>top</b>	Female	No Info
<b>freq</b>	58552	35816



## Q3 + Q4

In [3]: df\_0.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   gender            100000 non-null   object  
 1   age                100000 non-null   float64 
 2   hypertension       100000 non-null   int64   
 3   heart_disease     100000 non-null   int64   
 4   smoking_history    100000 non-null   object  
 5   bmi                100000 non-null   float64 
 6   HbA1c_level        100000 non-null   float64 
 7   blood_glucose_level 100000 non-null   int64   
 8   diabetes           100000 non-null   int64  
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

נשים לב שישנם משתנים שהטיפ שלהם הוא "אינט" או "פלואט" אבל הם קטגוריאליים.

משתנים נומריים: .age, bmi, HbA1c\_level, blood\_glucose\_level

. משתנים קטגוריאליים: gender, hypertension, heart\_disease, smoking\_history, diabetes

יש 100,000 תצפויות בדטה.

## Feature Explanations:

- gender : מגדר
- age : גיל
- hypertension : יתר לחץ-דם
- heart\_disease : קיום מחלת לב
- smoking\_history : היסטוריה עישון
- bmi : bmi
- HbA1c\_level : סוג של המוגלבין
- blood\_glucose\_level : רמות סוכר בدم
- diabetes : קיום סוכרת

In [4]: df\_0.isnull().sum()

```
Out[4]: gender      0
         age        0
         hypertension 0
         heart_disease 0
         smoking_history 0
         bmi        0
         HbA1c_level 0
         blood_glucose_level 0
         diabetes     0
         dtype: int64
```

In [5]: print("מבחן כפילוית")
duplicate\_rows = df\_0[df\_0.duplicated()]
duplicate\_rows
df = df\_0.drop\_duplicates()
df

מבחן כפילוית

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glu
0	Female	80.0	0	1	never	25.19	6.6	
1	Female	54.0	0	0	No Info	27.32	6.6	
2	Male	28.0	0	0	never	27.32	5.7	
3	Female	36.0	0	0	current	23.45	5.0	
4	Male	76.0	1	1	current	20.14	4.8	
...	...	...	...	...	...	...	...	...
99994	Female	36.0	0	0	No Info	24.60	4.8	
99996	Female	2.0	0	0	No Info	17.37	6.5	
99997	Male	66.0	0	0	former	27.83	5.7	
99998	Female	24.0	0	0	never	35.42	4.0	
99999	Female	57.0	0	0	current	22.43	6.6	

96146 rows × 9 columns

לאחר מחיקת הכפליות, יש 96,146 תצפיות.

## Part 2: Exploratory data analysis

In this part, you will do an initial exploration of the dataset you chose. This part should serve the next parts. That is, you should look at variables that can influence your analyses for parts (3) and (4). Of course, you can (and probably should) also explore further, and/or use this as a way to motivate questions for parts (3) and (4). You should explain why you are exploring the particular variables you chose.

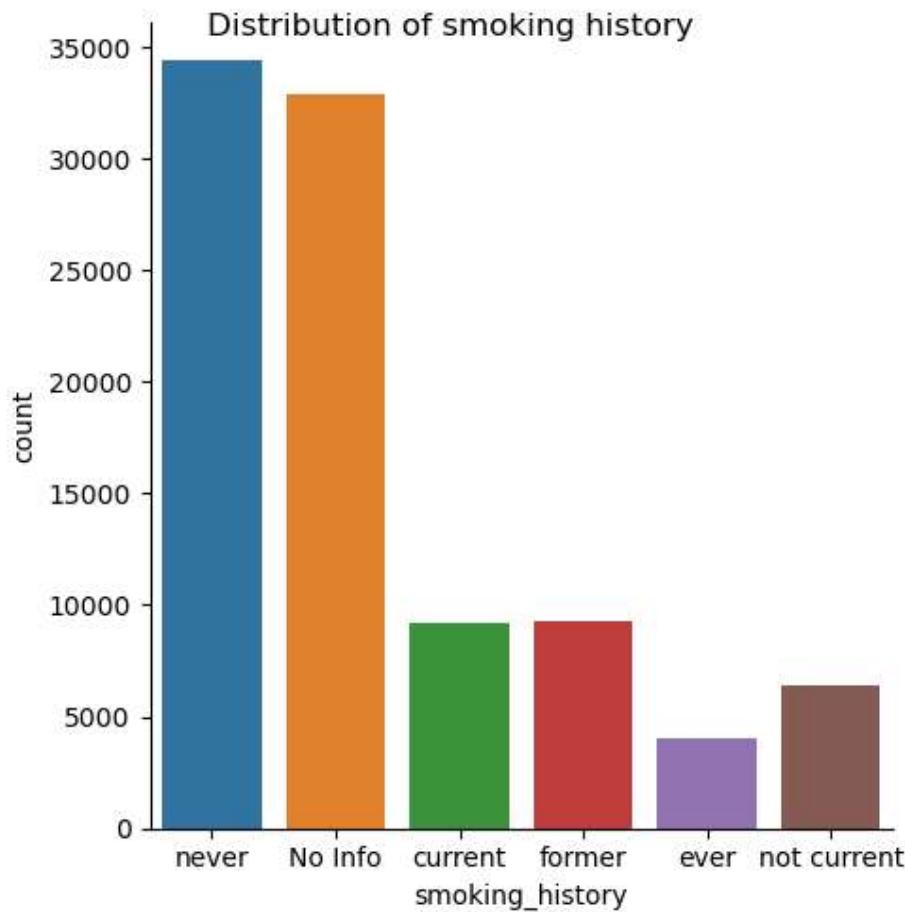
- 1) Show plots illustrating the distribution of at least 5 variables in your dataset. Comment on anything interesting you observe.
- 2) Show plots illustrating bivariate relationships for at least 2 pairs of variables. Explain what you observe (e.g., positive/negative correlation, no correlation, etc.).

## Q1

### תרשיי עמודות עבור התפלגות משתנים קטגוריאליים

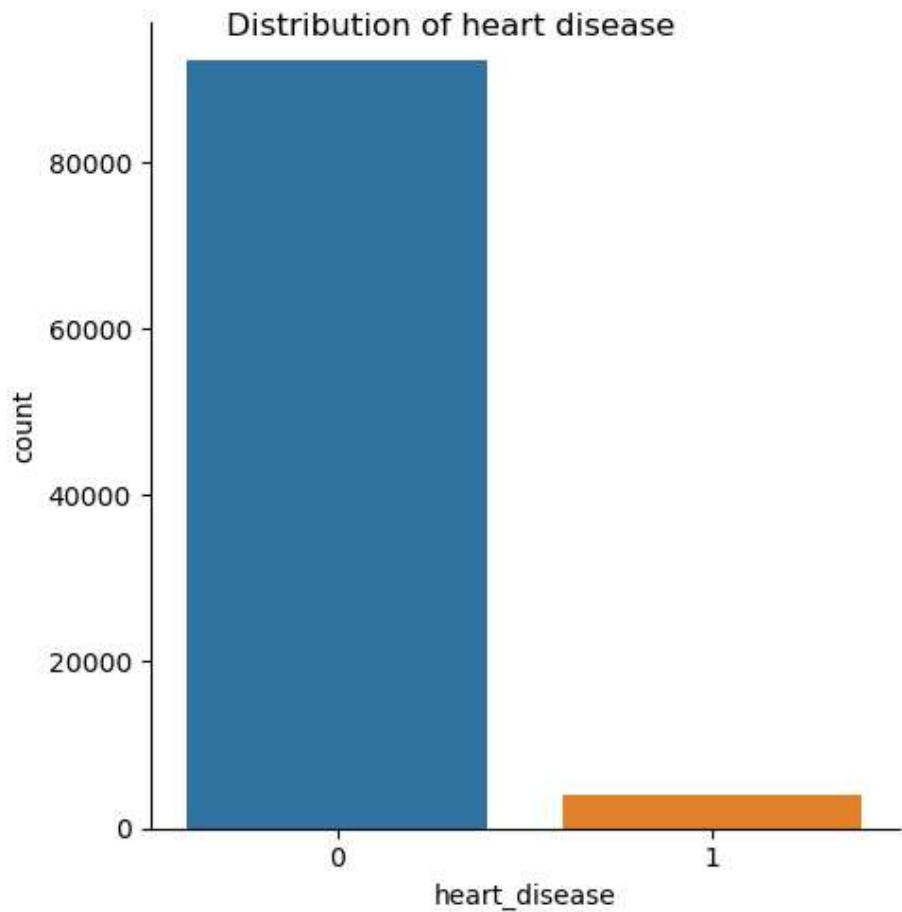
```
In [6]: facetgrid_obj=sns.catplot(x='smoking_history',kind='count',data=df)
facetgrid_obj.fig.suptitle('Distribution of smoking history ')
```

```
Out[6]: Text(0.5, 0.98, 'Distribution of smoking history ')
```



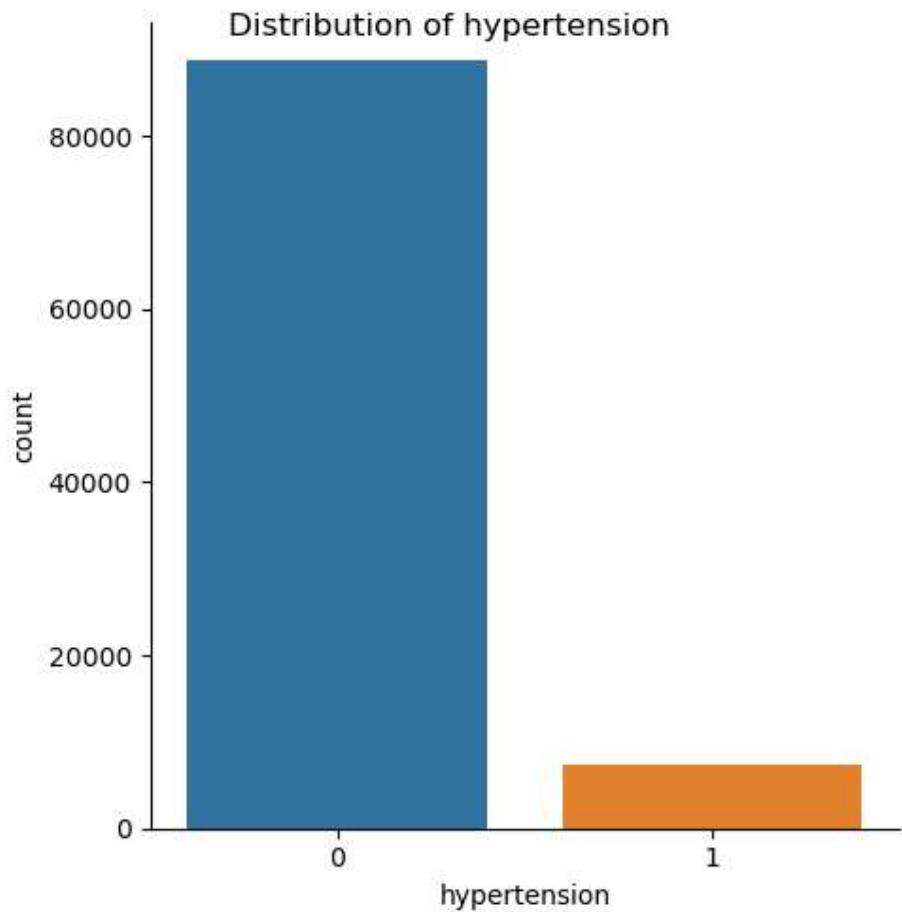
```
In [7]: facetgrid_obj=sns.catplot(x='heart_disease',kind='count',data=df)
facetgrid_obj.fig.suptitle('Distribution of heart disease ')
```

```
Out[7]: Text(0.5, 0.98, 'Distribution of heart disease ')
```



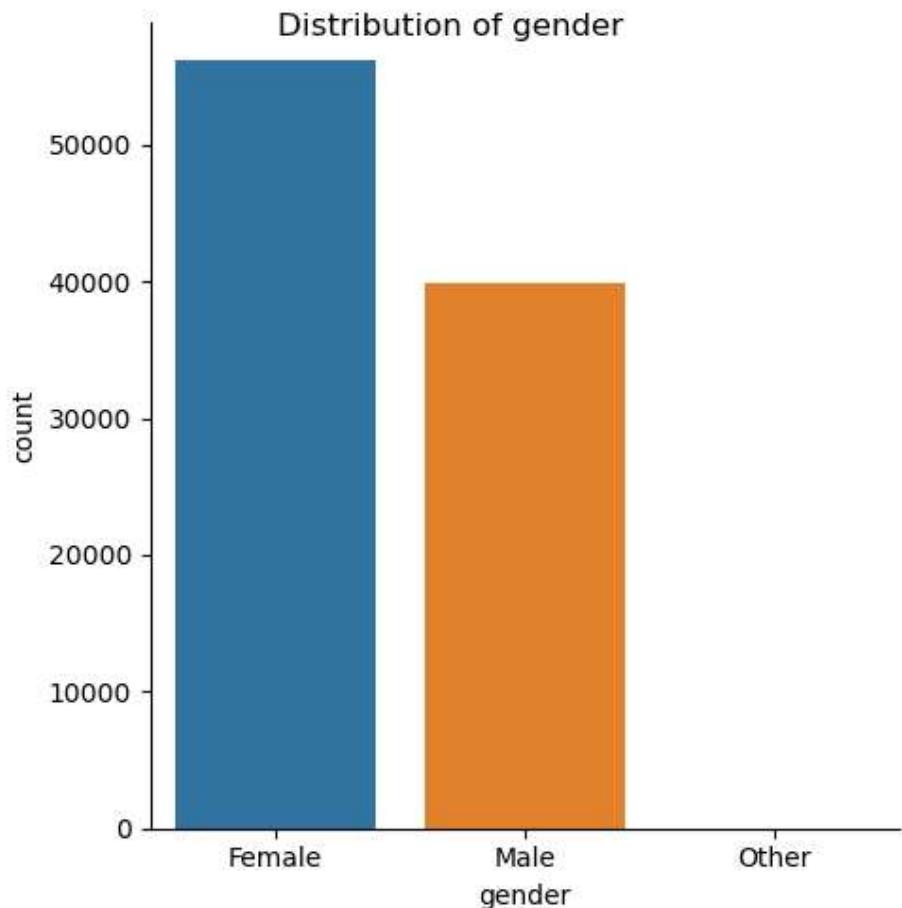
```
In [8]: facetgrid_obj=sns.catplot(x='hypertension',kind='count',data=df)
facetgrid_obj.fig.suptitle('Distribution of hypertension ')
```

```
Out[8]: Text(0.5, 0.98, 'Distribution of hypertension ')
```



```
In [9]: facetgrid_obj=sns.catplot(x='gender',kind='count',data=df)
facetgrid_obj.fig.suptitle('Distribution of gender ')
```

```
Out[9]: Text(0.5, 0.98, 'Distribution of gender ')
```



ראינו שהתקבלה עמודה נוספת ורוצים לוודא את כמות ה"אחרים".

```
In [10]: df['gender'].value_counts()
```

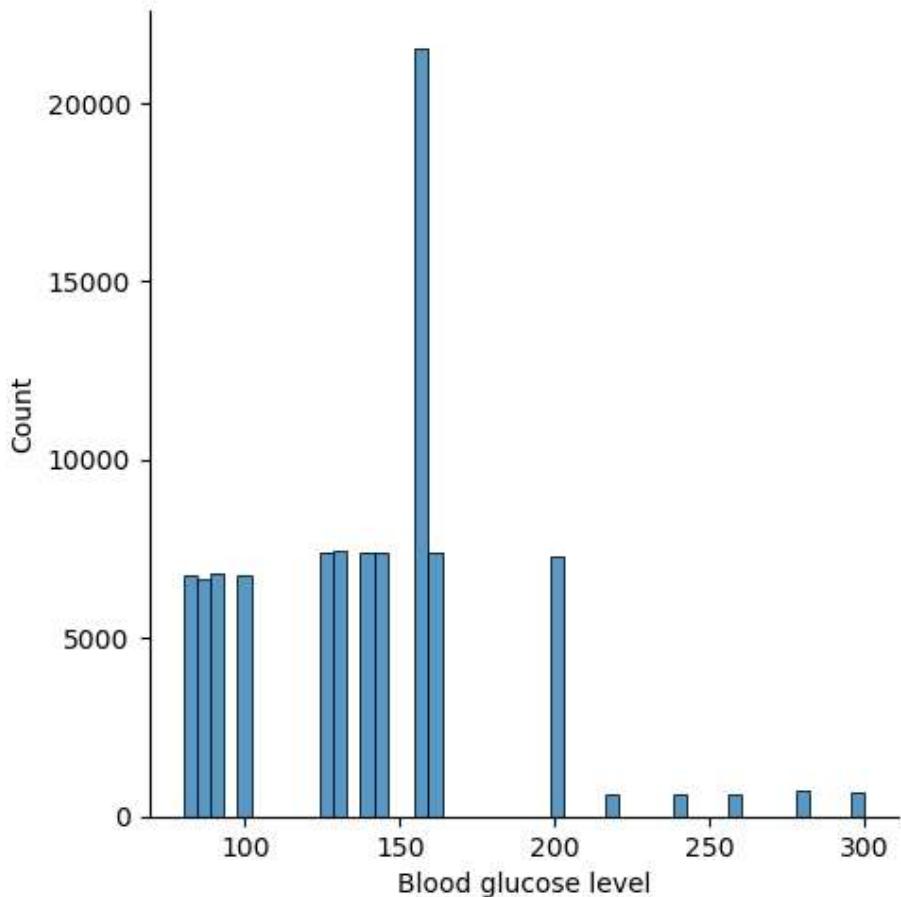
```
Out[10]: Female    56161
          Male     39967
          Other      18
          Name: gender, dtype: int64
```

כמות האחרים זניחה.

### תרשיימי עמודות עבור התפלגות משתנים נומריים

```
In [11]: ax = sns.displot(data = df, x = 'blood_glucose_level', bins = 50)
          ax.set(xlabel="Blood glucose level", title='Distribution Of Blood Glucose Level')
          plt.show()
```

### Distribution Of Blood Glucose Level



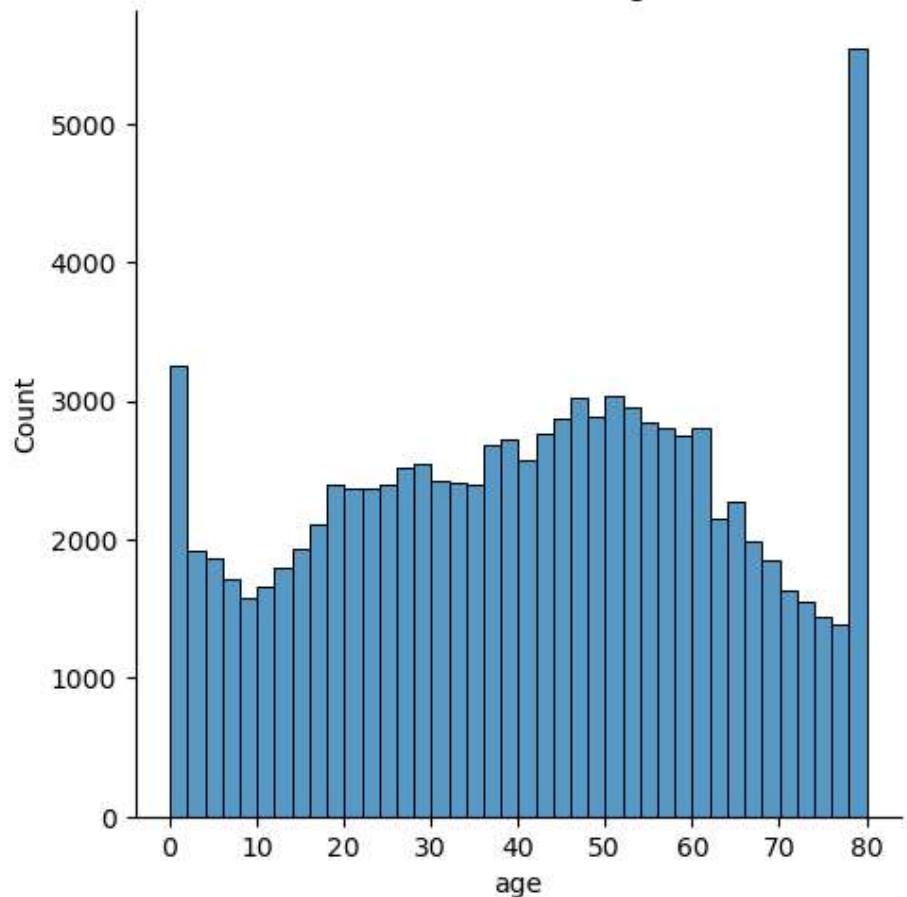
טווח ערכים של רמת גליקוז תקינה נع בין 72 ל-100 בזמינים בהם לא אוכלים. כאשר נמצאים עד שעתיים אחרי ארוחה הרמה יכולה להגיע ל-140. לפי המידע הנ"ל ניתן להניח שבדיקה הדם נעשו כשבועיים לאחר ארוחה שכן, מרבית הנבדקים נמצאים סביבה 140 שהוא תרומה לאחר שעתיים מהאכילה.

המידע לקוח מאתר:

<https://www.oleniklaw.co.il/%D7%A8%D7%9E%D7%AA-%D7%A1%D7%95%D7%9B%D7%A8-%D7%AA%D7%A7%D7%99%D7%A0%D7%94/>

```
In [12]: ax = sns.displot(data = df, x = 'age', bins = 40)
ax.set(xlabel="age", title='Distribution Of age')
plt.show()
```

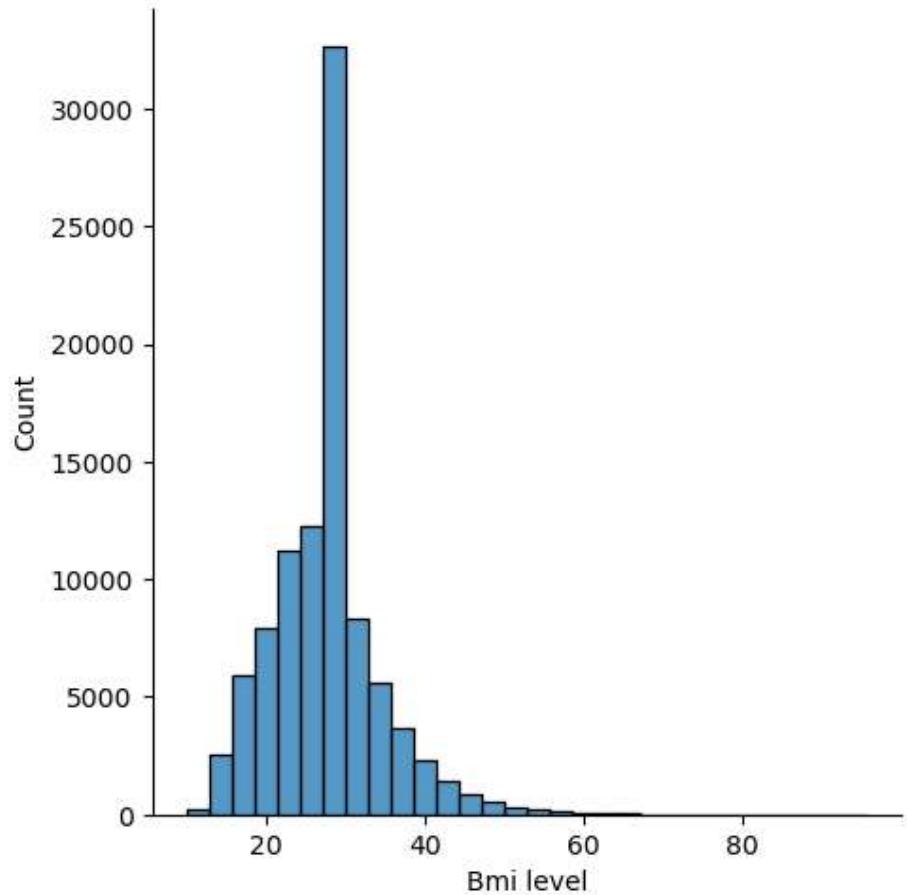
### Distribution Of age



ניתן לראות שקכחת הגיל 78-80 היא קבוצת הגיל הגדולה ביותר.

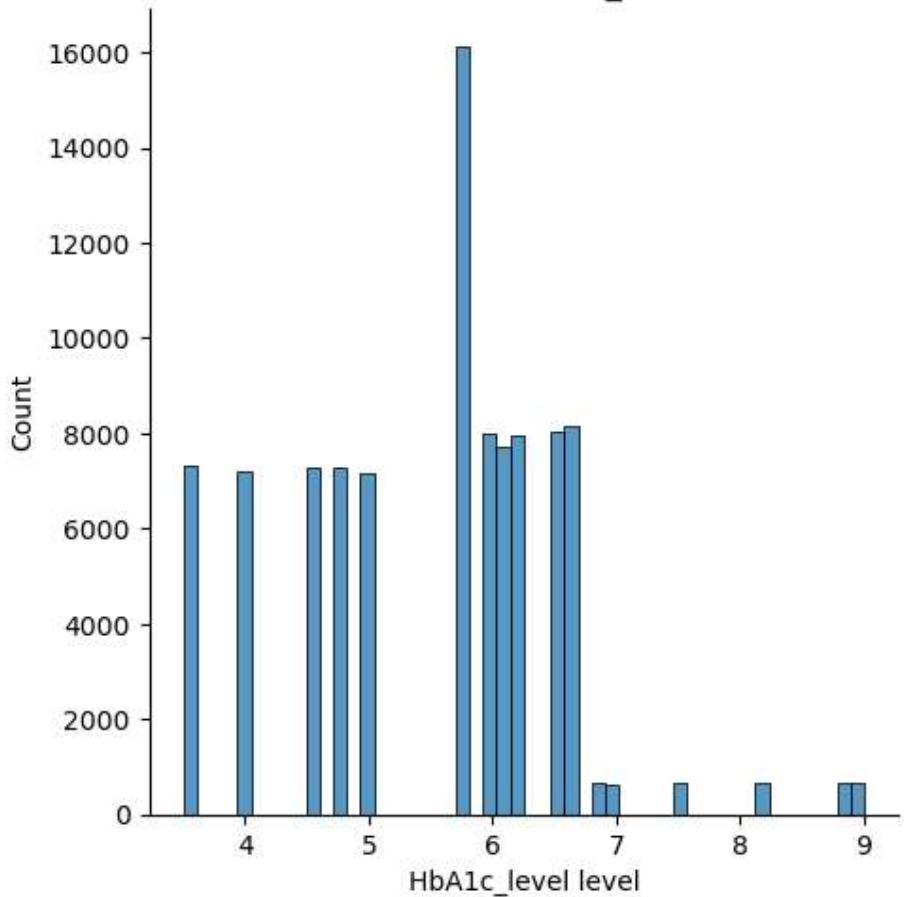
```
In [13]: ax = sns.displot(data = df, x = 'bmi', bins = 30)
ax.set(xlabel="Bmi level", title='Distribution Of bmi')
plt.show()
```

### Distribution Of bmi



```
In [14]: ax = sns.displot(data = df, x = 'HbA1c_level', bins = 50)
ax.set(xlabel="HbA1c_level level", title='Distribution Of HbA1c_level level')
plt.show()
```

### Distribution Of HbA1c\_level level

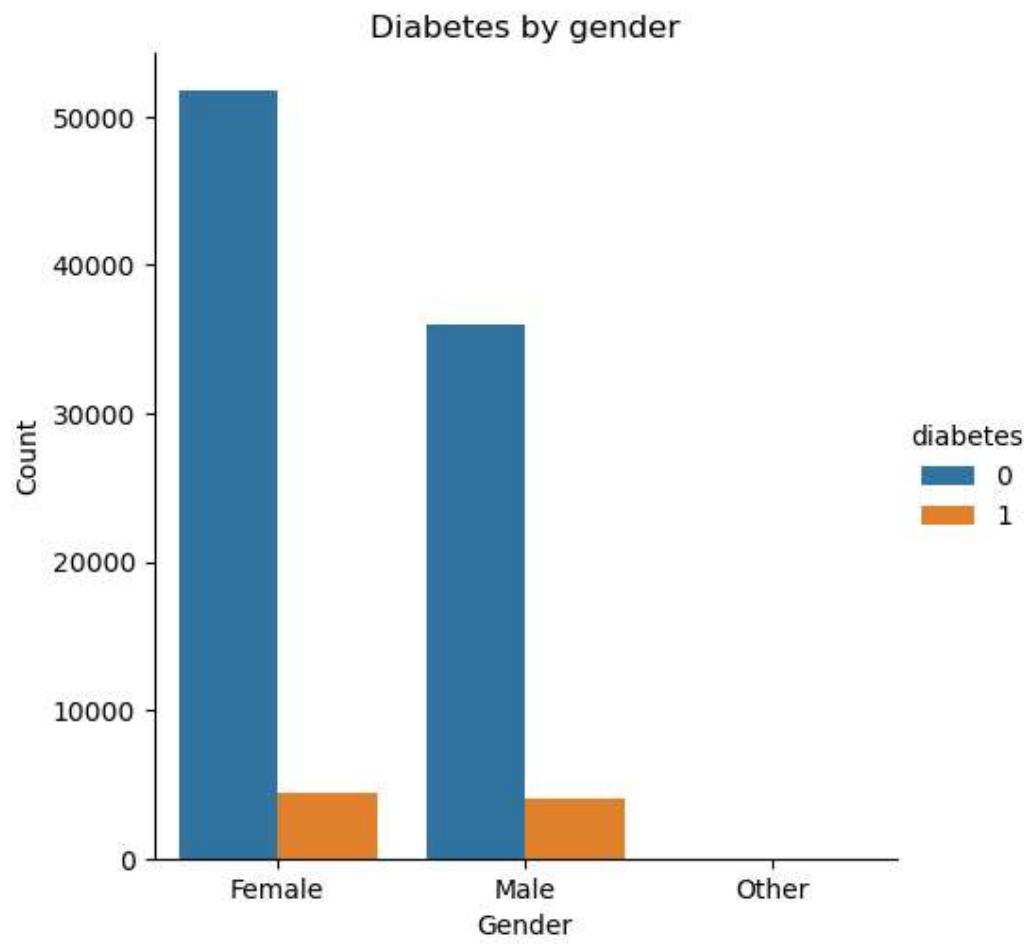


תרשים זה מייצג את התפלגות רמות הימולובין A1C. ערך תקין של חלבון זה הינו פחות מ-5.7% ורחוק מ-6.7% האדם מוגדר כחולה סוכרת. בגרף ניתן לראות שmericה האנשים במצב של טרום סוכרת.

## Q2

### תרשיי קשרים בין משתנים

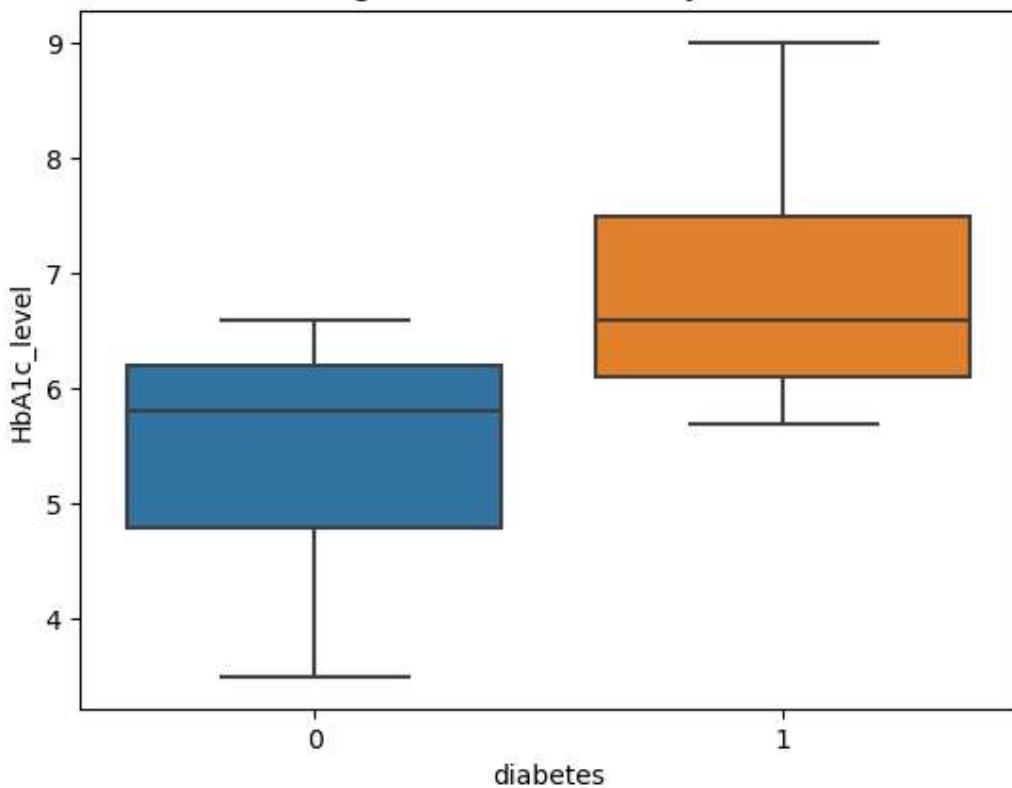
```
In [15]: ax = sns.catplot(kind='count', x='gender', hue='diabetes', data=df)
ax.set(xlabel='Gender', ylabel='Count', title='Diabetes by gender');
```



ניתן לראות שאין קשר בין מגן הנבדק לבין הימצאותו כחולה בסוכרת.

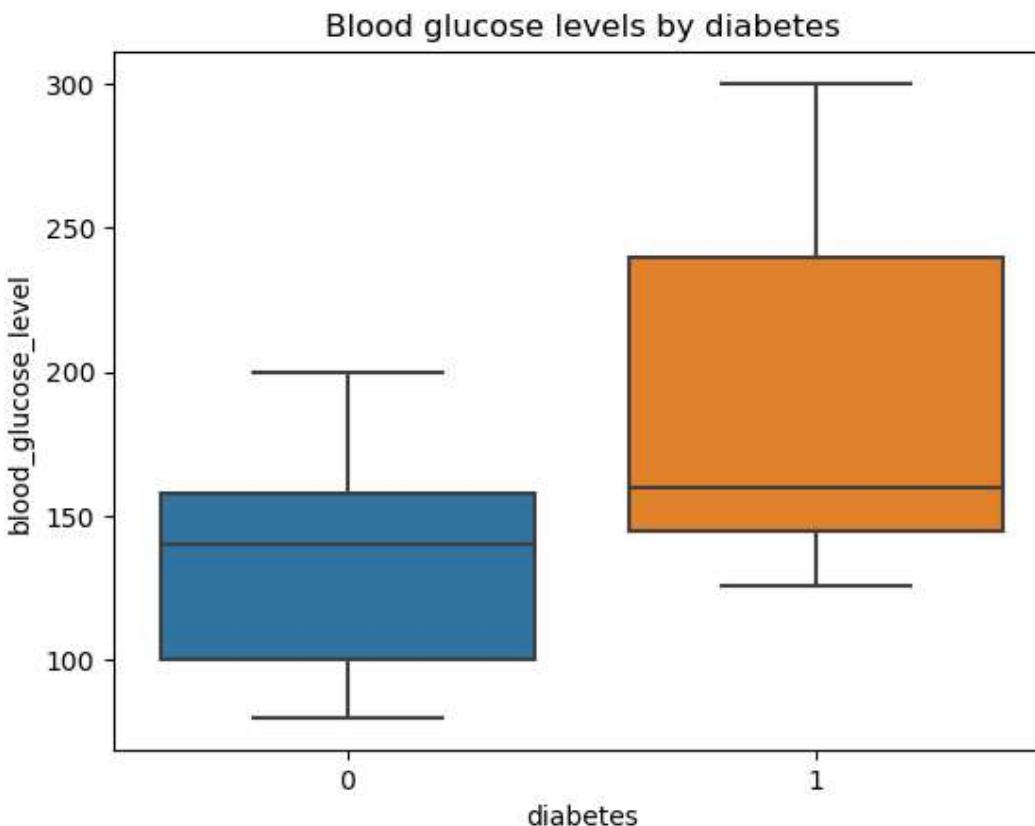
```
In [16]: sns.boxplot(x='diabetes', y='HbA1c_level', data=df)
plt.title(" Hemoglobin - A1C Levels by diabetes")
plt.show()
```

### Hemoglobin - A1C Levels by diabetes



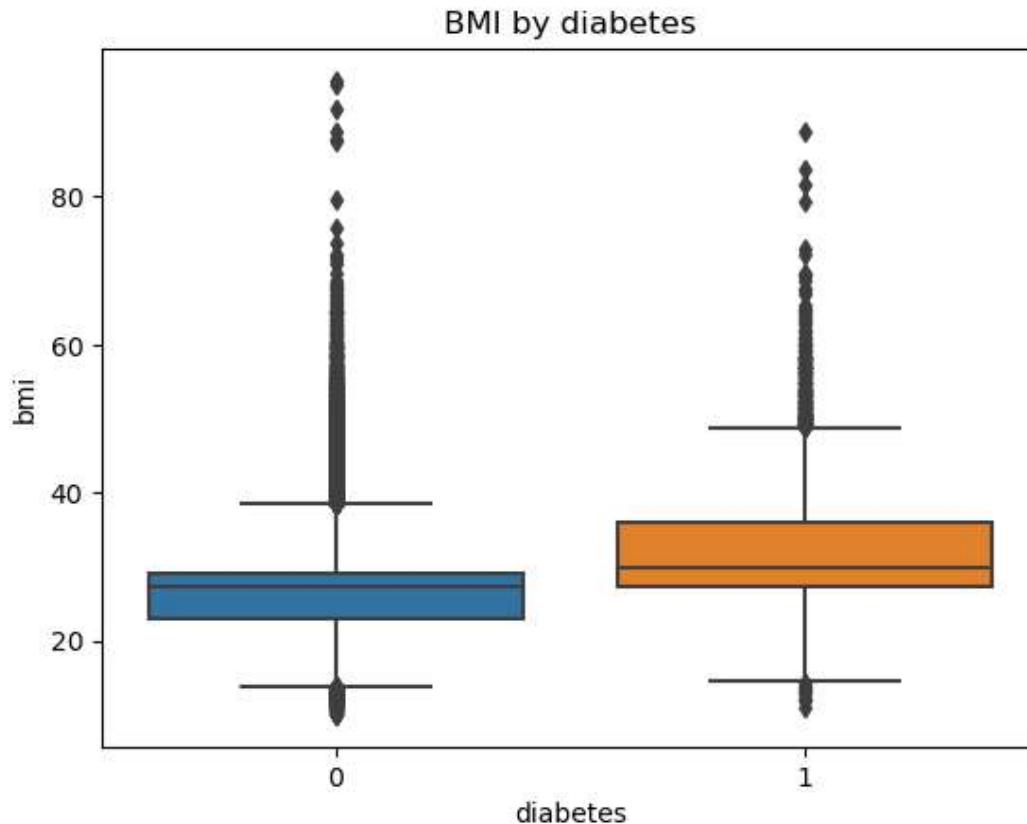
ניתן לראות שיש קשר בין קיומ סוכרת לרמות גבוהות של המוגלובין A1C בدم.

```
In [17]: sns.boxplot(x='diabetes', y='blood_glucose_level', data=df)
plt.title("Blood glucose levels by diabetes")
plt.show()
```



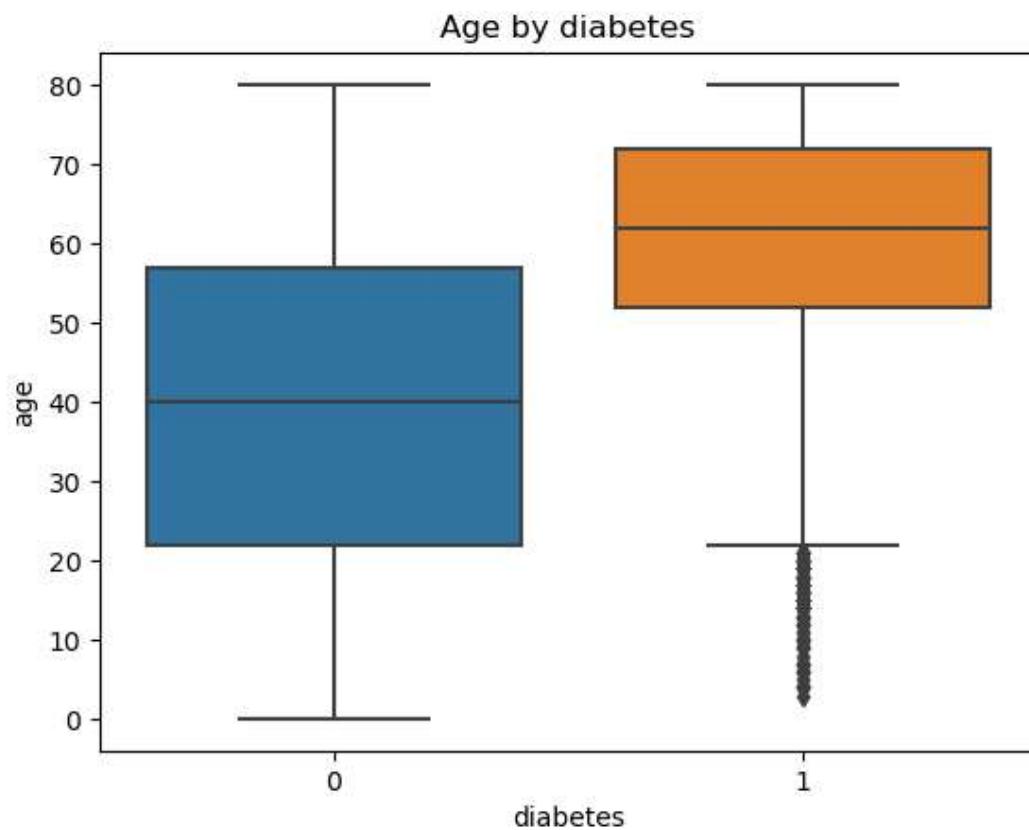
ניתן לראות שיש קשר בין קיומ סוכרת לרמות גבוחות של האלקוז בدم.

```
In [18]: sns.boxplot(x='diabetes', y='bmi', data=df)
plt.title("BMI by diabetes")
plt.show()
```



ניתן לראות שיש קשר בין קיומ סוכרת לרמות גבוחות של מזד ה bmi.

```
In [19]: sns.boxplot(x='diabetes', y='age', data=df)
plt.title("Age by diabetes")
plt.show()
```



ניתן לראות שיש קשר בין קיומ סוכרת לגיל מבוגר יותר.

### **Part 3: Estimation and hypothesis testing**

In this part, you will formally test a hypothesis using your data.

- 1) What is the question you want to explore? Why is it interesting to you?
- 2) Clearly state your null hypothesis and alternative hypothesis.
- 3) Run a test and report the results in a comprehensive way.

## **Q1**

האם יש חידש בתחלואה הסוכרת בין לגברים לנשים?

לפי הגרף של סוכרת לפחות אנו רואים שאין חידש מובהק בין כמות הגברים והנשים החולמים בסוכרת. רצינו לבדוק האם דבר זה נכון גם לכל האוכלוסייה ולא רק במדגם הנתון.

## **Q2**

OH: אין חידש בין שיעור התחלואה של גברים בסוכרת לעומת שיעור התחלואה של נשים באוכלוסייה.

H1 : יש הבדל בין שיעור תחלואה של גברים בסוכרת לעומת שיעור תחלואה של נשים באוכלוסייה.

סטטיסטי: הפרש הממוצעים בין שיעור תחלואה הגברים לשיעור תחלואה הנשים.

## Q3

In [20]: `df['gender'].value_counts()`

```
Out[20]: Female    56161
          Male     39967
          Other      18
          Name: gender, dtype: int64
```

בנתונים שלנו ישנו ישנה קטגוריה של "אחר". בניתוח ההשערות לא נתיחס אליה משום שהוא לא מושם לנו בזיהויים את ההבדל בין גברים לנשים וקטgoriyah ה"אחר" אינה מוגדרת כמספר.

In [21]: `def diff_of_avgs(df, column_name, groupby_var):
 grpby_var = df.groupby(groupby_var)
 avgs = grpby_var[column_name].mean()
 return avgs.loc['Male'] - avgs.loc['Female']

def bootstrap_mean_difference(original_sample, column_name, grouping_var, num_repli:
 '''This function returns an array of bootstrapped differences between two samp
 original_sample: df containing the original sample
 column_name: name of column containing the variable to average
 grouping_var: name of variable according to which to group
 num_replications: number of bootstrap samples'''
 original_sample_size = original_sample.shape[0] # we need to replicate with the
 original_sample_cols_of_interest = original_sample[[column_name, grouping_var]
 bstrap_mean_diffs = np.empty(num_replications)
 for i in range(num_replications):
 bootstrap_sample = original_sample_cols_of_interest.sample(original_sample_s
 resampled_mean_diff = diff_of_avgs(bootstrap_sample, column_name, grouping_
 bstrap_mean_diffs[i] = resampled_mean_diff

 return bstrap_mean_diffs`

In [22]: `# bootstrap procedure
bstrap_diffs = bootstrap_mean_difference(df, 'diabetes', 'gender', 5000) # 5000 simulations

# the 95% confidence interval
left_end = np.percentile(bstrap_diffs, 2.5, interpolation='higher')
right_end = np.percentile(bstrap_diffs, 97.5, interpolation='higher')
print('The 95% bootstrap confidence interval for difference between population mea

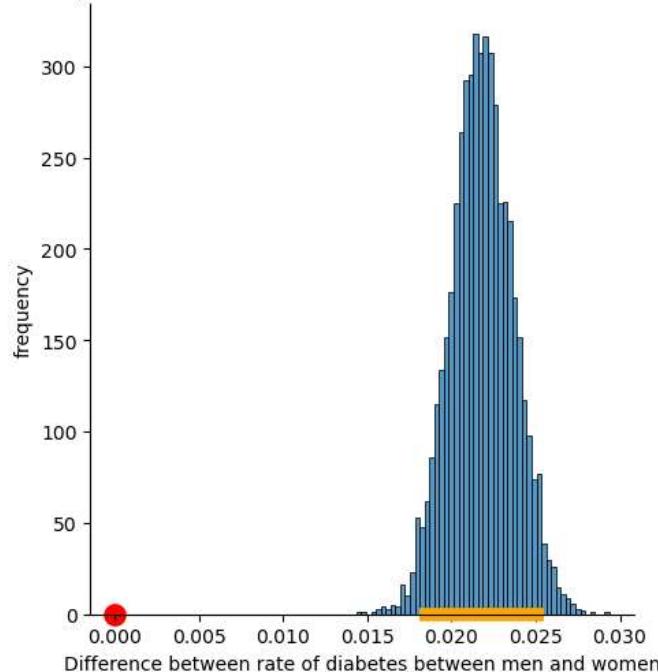
# visualize results
ax = sns.displot(bstrap_diffs)
plt.hlines(y=0, xmin=left_end, xmax=right_end, colors='orange', linestyles='solid'
ax.set(xlabel='Difference between rate of diabetes between men and women', ylabel=
 title='Distribution of bootstrap estimates for the difference between diabetes rates')

plt.scatter(0, 0, color='red', s=120, clip_on=False)`

The 95% bootstrap confidence interval for difference between population means [0.018128272481010885, 0.02548488150061151]

```
C:\Users\ Nir \AppData\Local\Temp\ipykernel_13444\2647064481.py:5: DeprecationWarning: the `interpolation=' argument to percentile was renamed to `method='`, which has additional options.
  Users of the modes 'nearest', 'lower', 'higher', or 'midpoint' are encouraged to review the method they used. (Deprecated NumPy 1.22)
    left_end = np.percentile(bstrap_diffs, 2.5, interpolation='higher')
C:\Users\ Nir \AppData\Local\Temp\ipykernel_13444\2647064481.py:6: DeprecationWarning: the `interpolation=' argument to percentile was renamed to `method='`, which has additional options.
  Users of the modes 'nearest', 'lower', 'higher', or 'midpoint' are encouraged to review the method they used. (Deprecated NumPy 1.22)
    right_end = np.percentile(bstrap_diffs, 97.5, interpolation='higher')
<matplotlib.collections.PathCollection at 0x20263758550>
Out[22]:
```

Distribution of bootstrap estimates for the difference between diabetes rate of men and women



יצרנו התחפוגות בעזרת 5,000 סימולציות על הפרשי שיעורי תחלואה הסכרת בין גברים לנשים. יצרנו רוח סמרק ברמת בטחון של 95% והאפס לא נכלל בתוך רוח הסמרק. כמובן, ברמת בטחון של 95% (ובכל רמת בטחון אחרת, כי אפס לא נמצא בכלל בהתפלוגות) ניתן לדוחות את השערת האפס. ולפי מזגם זה שאלינו נתיחס כמגדם מיצג נקבע שלגברים שיעור התחלואה גבוהה יותר מאשר לנשים.

#### Part 4: Prediction/clustering

In this part, you will see how well you can address a classification problem or a clustering problem using your data. Choose one of the following two options.

##### Option 1: classification

- 1) What do you want to try to classify? Why? What is a potential application of an algorithm that classifies your target variable?
- 2) Clearly state what is the target variable (class) you are trying to predict, which variables (features) you are using to predict the class, and why you chose these variables..
- 3) Use kNN for the classification task and report the results.

## Q1

אנחנו רוצים ליצור מסזאג אשר יעזר לנו לחזות האם אדם יהיה בסוכרת על סמך תכונות כמו: מחלות לב, יתר לחץ דם, גיל, מין וכו'. ונבדוק אחר כך אילו תכונות הן בעלות קורלציה גבוהה עם המטרה שהיא מחלת הסוכרת. האלגוריתם ימליץ לאדם להיבדק לסוכרת. כדי שנדע על סמך אילו נתוניים להתביסס נבדוק את הקורלציה ביניהם לתחלוואה בסוכרת.

## Q2

המטרה היא האם אדם חוליה או לא. כאשר 1 זה אדם חוליה ו 0 אדם שהוא לא חוליה. את הפיצ'רים נבחר באמצעות התבוננות בהיטמאפ.

## Q3

```
In [23]: encoded_df = pd.get_dummies(df, columns=['gender'], drop_first=True)
#הופכים את המין למשתנה נומרי#
encoded_df.sample(10)
```

	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level
69915	8.0	0	0	never	17.56	6.6	159
17864	29.0	0	0	never	30.50	6.0	84
90122	48.0	1	0	never	49.41	6.0	126
535	80.0	0	0	never	20.93	6.6	126
78682	54.0	0	0	never	24.68	6.5	160
74225	29.0	0	0	not current	28.67	5.0	100
71433	71.0	1	1	former	27.32	6.0	90
15542	21.0	0	0	never	21.25	6.2	100
81945	25.0	0	0	ever	27.32	6.6	159
83792	77.0	0	0	No Info	19.18	5.8	200

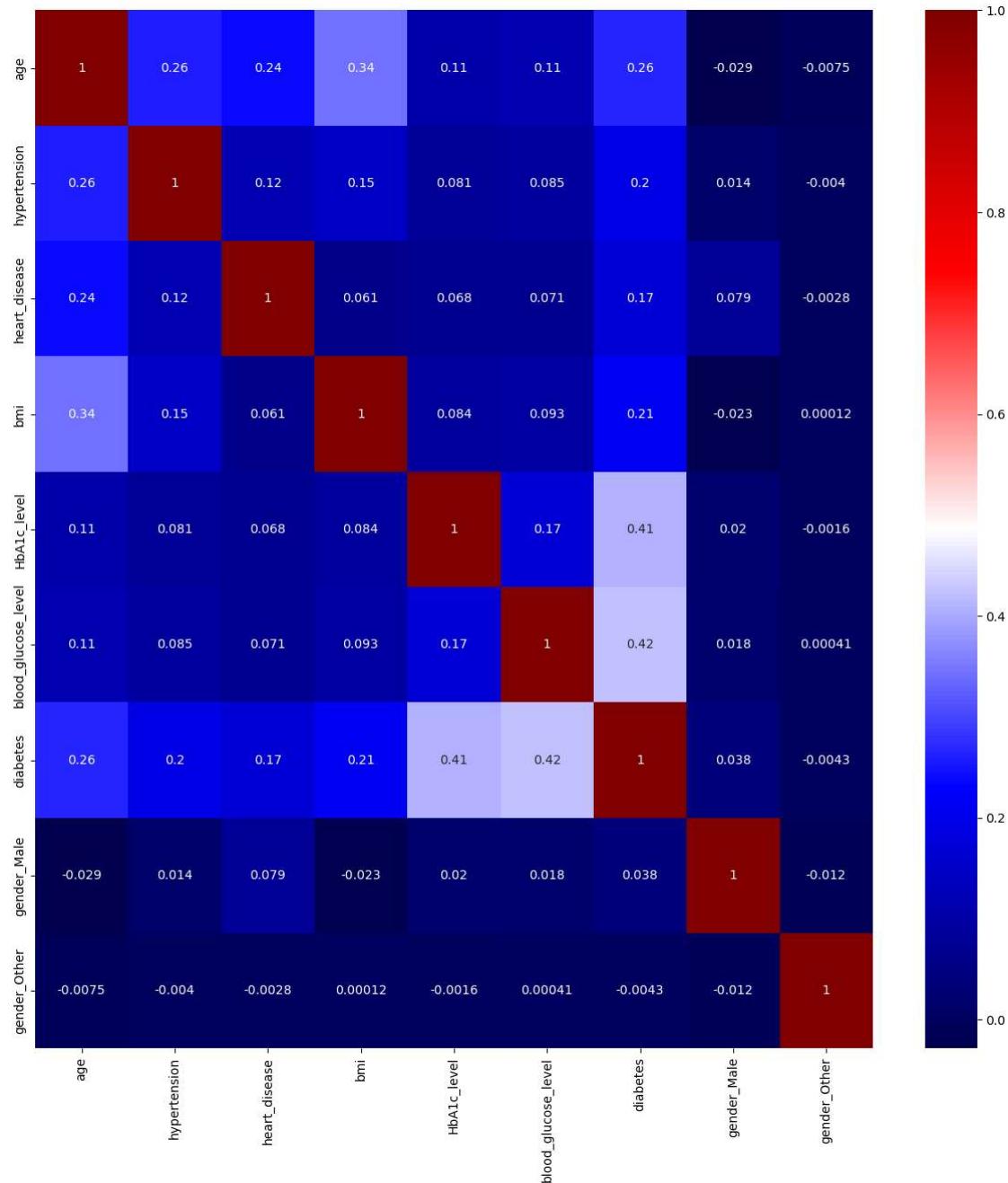
**הערה:** לא התייחסנו לקטגוריות העישון (ולכן לא הפכנו אותה למשתנה נומרי) כי אנחנו לא יודעים איך מי שלקח את הדגימות סיוג את מזדי העישון לקטגוריות. יכול להיות שבאותה קטgorיה יש אנשים שנחחשפו לעישון בפרק זמן ממושך יותר ובכמות רבה יותר. למשל: אדם שעישן בעבר היה יכול לעשן סיגריה אחת ביום, חבילה ביום או מצית ביום. בנוסף, גם אדם שלא מעשן סיגריות יכול להיחשף לנזקי עישון אם הוא נמצא בקרבת מעשנים. ככלומר, מעשן פסיבי.



In [24]:

```
# compute correlation between each pair of variables in data frame
correlations = encoded_df.corr(numeric_only=True)

#plot heat map
plt.figure(figsize=(15,15))
g=sns.heatmap(correlations,annot=True, cmap="seismic")
```



In [25]: `correlations.reindex(correlations[ "diabetes" ].abs().sort_values().index)[ "diabetes" ]`

Out[25]:

gender_Other	-0.004256
gender_Male	0.037883
heart_disease	0.170711
hypertension	0.195710
bmi	0.214932
age	0.264927
HbA1c_level	0.406408
blood_glucose_level	0.424336
diabetes	1.000000

Name: diabetes, dtype: float64

הפייצ'רים: `blood_glucose_level`, `HbA1c_level`, נמצאים בקורסיה הגדולה ביותר עם תחלאות הסוכרת.

In [26]: `#we need to split our data to train and test sets`  
`from sklearn.model_selection import train_test_split, cross_val_score`  
`from sklearn.neighbors import KNeighborsClassifier`

```
from sklearn.preprocessing import StandardScaler, MinMaxScaler

#Let's keep only the variables we're interested in
knn_df = encoded_df[['HbA1c_level', 'blood_glucose_level', 'diabetes']]
knn_df = knn_df.sample(frac=1)
knn_df
```

Out[26]:

	HbA1c_level	blood_glucose_level	diabetes
<b>37411</b>	6.0	145	0
<b>84537</b>	7.5	140	1
<b>98384</b>	6.5	145	0
<b>61010</b>	6.1	140	0
<b>21156</b>	6.0	158	0
...	...	...	...
<b>24582</b>	4.5	200	0
<b>81610</b>	6.2	85	0
<b>8227</b>	4.8	140	0
<b>73069</b>	6.0	158	0
<b>86852</b>	6.1	90	0

96146 rows × 3 columns

In [27]:

```
# Split to X and Y
X = knn_df.loc[:, knn_df.columns != 'diabetes'] # features
Y = knn_df.loc[:, 'diabetes'].values # labels

# Split to train and test
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20)
X_train
X_test
display(X_test, X_train, Y_test, Y_train)
```

	HbA1c_level	blood_glucose_level
<b>86784</b>	6.1	90
<b>17514</b>	5.7	90
<b>39224</b>	4.5	130
<b>31633</b>	6.6	126
<b>83488</b>	5.0	159
...	...	...
<b>74300</b>	5.8	90
<b>39086</b>	3.5	100
<b>51879</b>	6.1	260
<b>88917</b>	5.0	145
<b>29444</b>	4.0	85

19230 rows × 2 columns

	HbA1c_level	blood_glucose_level
<b>9562</b>	6.5	200
<b>15058</b>	3.5	80
<b>15833</b>	5.0	130
<b>79797</b>	6.2	158
<b>59742</b>	6.0	145
...	...	...
<b>16160</b>	6.8	126
<b>78682</b>	6.5	160
<b>62526</b>	5.0	80
<b>72174</b>	5.8	90
<b>86487</b>	6.2	160

76916 rows × 2 columns

```
array([0, 0, 0, ..., 1, 0, 0], dtype=int64)
array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

```
In [28]: # We need to scale the variables to be on the same scale
# We choose here to standardize using z-scores.
df_columns = X_train.columns
scaler = StandardScaler()
scaled_X_train = scaler.fit_transform(X_train)
scaled_X_test = scaler.transform(X_test)

scaled_df = pd.DataFrame(scaled_X_train, columns=df_columns)
scaled_df.describe()
```

Out[28]:

	HbA1c_level	blood_glucose_level
<b>count</b>	7.691600e+04	7.691600e+04
<b>mean</b>	3.264676e-16	-5.344128e-17
<b>std</b>	1.000007e+00	1.000007e+00
<b>min</b>	-1.889763e+00	-1.425462e+00
<b>25%</b>	-6.795346e-01	-9.354093e-01
<b>50%</b>	2.514106e-01	4.469672e-02
<b>75%</b>	6.237887e-01	5.102471e-01
<b>max</b>	3.230435e+00	3.965121e+00

המדד החשוב עבור המסוווג שלנו הוא Accuracy חשב לנו למצוא אנשים שהם חולמים בסוכרת ולסוווג אותם נכון. כמו כן סיוג של אדם שאינו חולם בסוכרת כאחד שכן חולמה עלול להביא לנטילת תרופות המגבירות אינסולין ולכטוף גם לפגוע בו. לפיכך הדיקח שוכ בנסיבות זה.



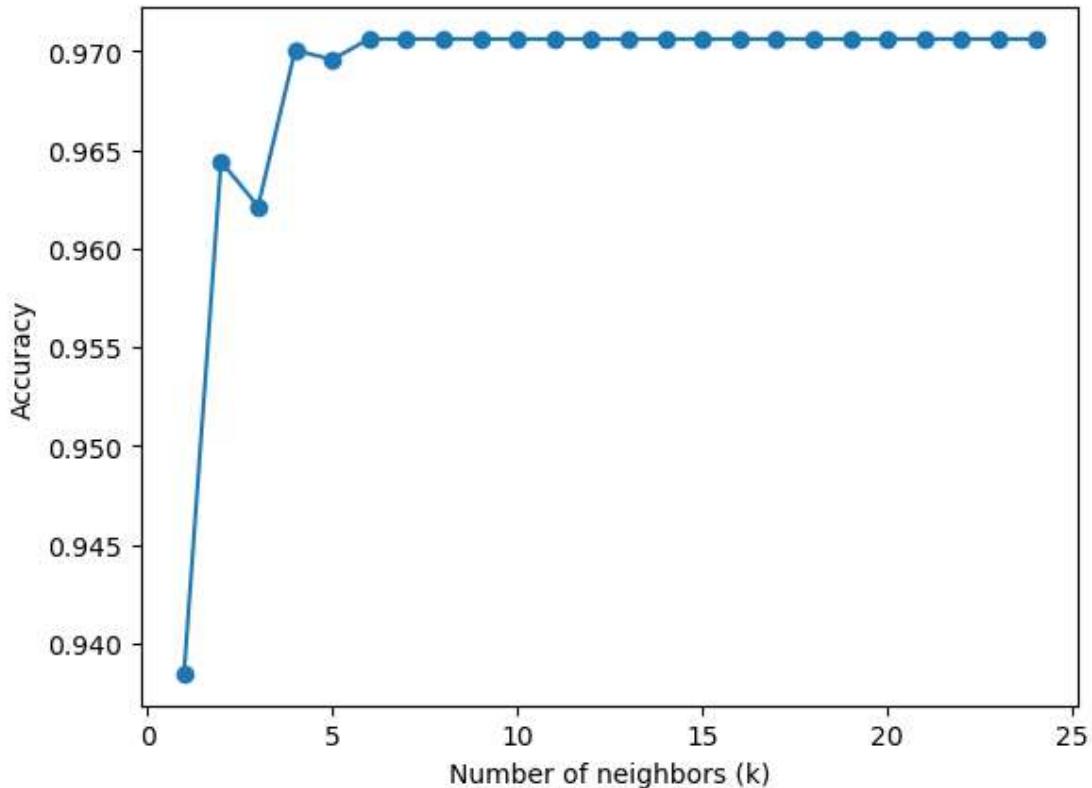
In [39]:

```
# Run CV to find optimal k - using only the training data
from sklearn.model_selection import cross_val_score

mean_cv_scores = []
k_list = range(1, 25)
for nn in k_list:
    knn_cv = KNeighborsClassifier(n_neighbors=nn)
    cv_scores = cross_val_score(knn_cv, scaled_X_train, Y_train, cv=10)
    mean_cv_scores.append(cv_scores.mean())

# output results
best_k = mean_cv_scores.index(max(mean_cv_scores))+1 # gets index of best performing
print('Highest accuracy is obtained for k =', best_k, 'and equals', max(mean_cv_scores))
plt.plot(k_list, mean_cv_scores, '-o')
plt.xlabel('Number of neighbors (k)')
plt.ylabel('Accuracy');
```

Highest accuracy is obtained for k = 6 and equals 0.9706043046038575



```
In [40]: # Retrain our chosen kNN on the whole training data and test its accuracy on the test data
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score

knn_classifier = KNeighborsClassifier(n_neighbors=best_k)
knn_classifier.fit(scaled_X_train, Y_train)
print('accuracy of the classifier is', knn_classifier.score(scaled_X_test, Y_test))

# Compute a confusion matrix
predictions = knn_classifier.predict(X=scaled_X_test) # get the classifier's predictions

accuracy of the classifier is 0.9713988559542381

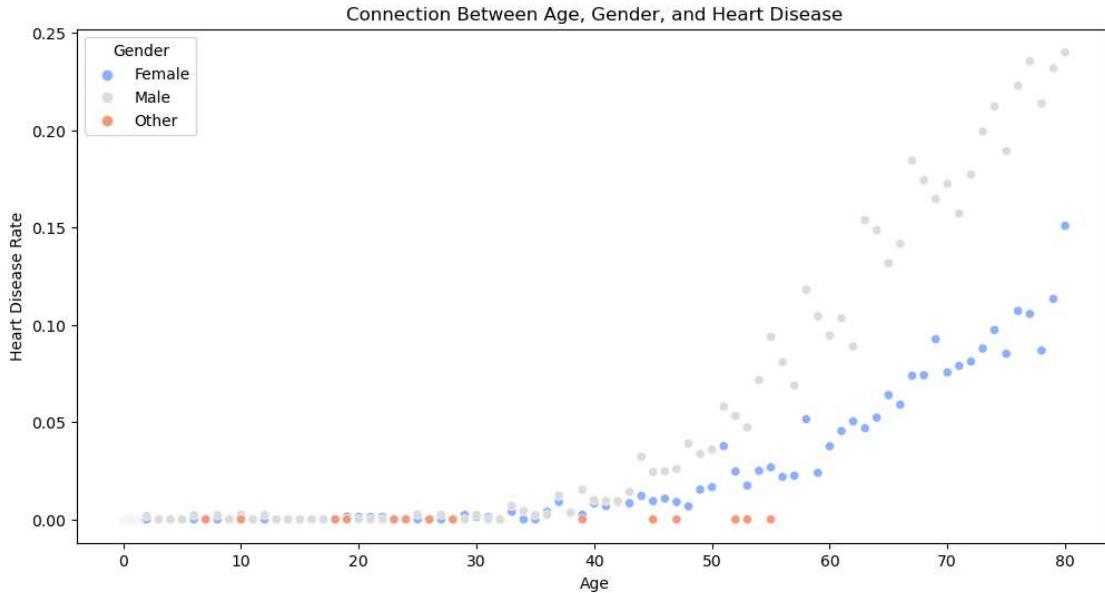
המסויוג שיצרנו הצליח מושם ממשודד הדיק של גבוי.
```

### חקר נוספת:

נרצה לבדוק קשר בין גיל ומגדר לסיכון לתחלואה במחלת לב.

```
In [31]: # Analyze the connection between age, gender, and heart disease
age_gender_heart = df.groupby(['age', 'gender'])['heart_disease'].mean().reset_index()

# Visualize the connection
plt.figure(figsize=(12, 6))
sns.scatterplot(data=age_gender_heart, x='age', y='heart_disease', hue='gender', palette='viridis')
plt.title('Connection Between Age, Gender, and Heart Disease')
plt.xlabel('Age')
plt.ylabel('Heart Disease Rate')
plt.legend(title='Gender')
plt.show()
```



ניתן לראות שהסיכוי לחלהות לב בקרב גברים מתחילה מוקדם יותר מאשר אצל נשים  
והסיכון גבוה יותר.

נרצה ליצור מסויים שיגיד אם אדם הוא בקבוצת סיכון לחלהות לב.

```
In [32]: df_old = df.loc[df['age'] >= 40]
```

```
In [33]: # Functions to apply the condition and create new columns:
def apply_condition(row):
    if row['gender'] == 'Male' and row['age'] > 40 and row['heart_disease'] == 1:
        return 1
    else:
        return 0

# Apply the function to create new columns
df_old['A man in a risk group'] = df_old.apply(apply_condition, axis=1)

def apply_condition(row):
    if row['gender'] == 'Female' and row['age'] > 50 and row['heart_disease'] == 1:
        return 1
    else:
        return 0

# Apply the custom function to create new columns
df_old['A woman in a risk group'] = df_old.apply(apply_condition, axis=1)
df_old
```

```
C:\Users\ Nir צ' אופר\AppData\Local\Temp\ipykernel_13444\1149132936.py:9: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
    df_old['A man in a risk group'] = df_old.apply(apply_condition, axis=1)
C:\Users\ Nir צ' אופר\AppData\Local\Temp\ipykernel_13444\1149132936.py:18: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
    df_old['A woman in a risk group'] = df_old.apply(apply_condition, axis=1)
```

Out[33]:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glu
0	Female	80.0	0	1	never	25.19	6.6	
1	Female	54.0	0	0	No Info	27.32	6.6	
4	Male	76.0	1	1	current	20.14	4.8	
6	Female	44.0	0	0	never	19.31	6.5	
7	Female	79.0	0	0	No Info	23.86	5.7	
...	...	...	...	...	...	...	...	
99984	Male	80.0	1	0	No Info	20.96	6.6	
99986	Female	63.0	0	0	never	29.01	4.8	
99993	Female	40.0	0	0	never	40.69	3.5	
99997	Male	66.0	0	0	former	27.83	5.7	
99999	Female	57.0	0	0	current	22.43	6.6	

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glu
0	Female	80.0	0	1	never	25.19	6.6	
1	Female	54.0	0	0	No Info	27.32	6.6	
4	Male	76.0	1	1	current	20.14	4.8	
6	Female	44.0	0	0	never	19.31	6.5	
7	Female	79.0	0	0	No Info	23.86	5.7	
...	...	...	...	...	...	...	...	
99984	Male	80.0	1	0	No Info	20.96	6.6	
99986	Female	63.0	0	0	never	29.01	4.8	
99993	Female	40.0	0	0	never	40.69	3.5	
99997	Male	66.0	0	0	former	27.83	5.7	
99999	Female	57.0	0	0	current	22.43	6.6	

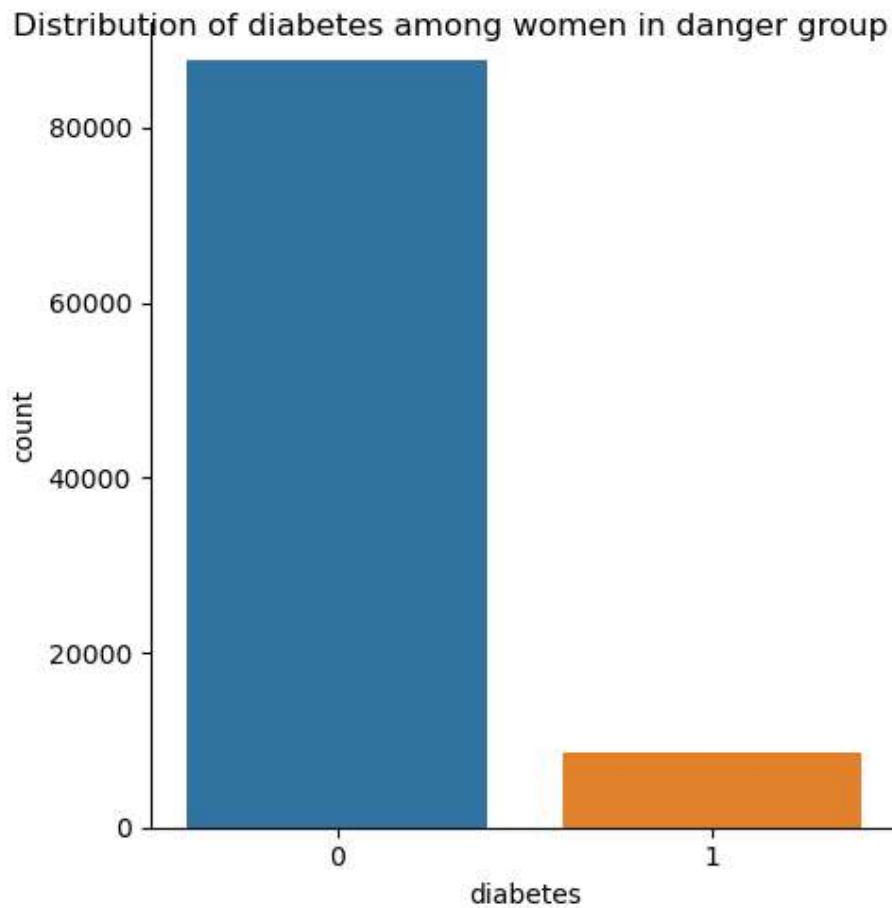
52451 rows × 11 columns

In [34]: #אנתה פרים נפלע עבור נשים  
`women_df = df_old[df_old['gender'] == 'Female']  
 כמה נמצאות בקבוצת סיכון()=  
 fulfill_condition = women_df['A woman in a risk group'].sum()  
 do_not_fulfill_condition = len(women_df) - fulfill_condition  
 print("Number of women in danger group:", fulfill_condition)  
 print("Number of women not in danger group:", do_not_fulfill_condition)  
 facetgrid_obj=sns.catplot(x='diabetes',kind='count',data=df)  
 facetgrid_obj.fig.suptitle('Distribution of diabetes among women in danger group ')`

Number of women in danger group: 1435

Number of women not in danger group: 29283

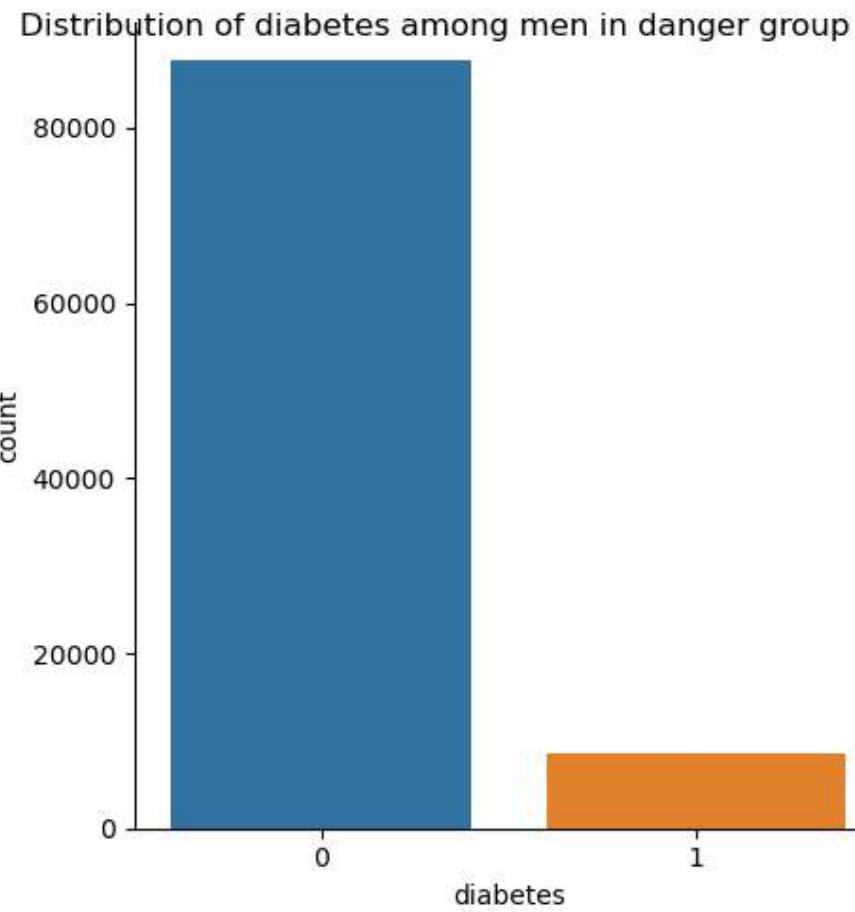
Out[34]: Text(0.5, 0.98, 'Distribution of diabetes among women in danger group ')



```
In [35]: #אטחה פרויים נפרד עבור גברים
men_df = df_old[df_old['gender'] == 'Male']
נהה נמצאים בקבוצת סיכון#sum()
fulfill_condition = men_df['A man in a risk group'].sum()
do_not_fulfill_condition = len(men_df) - fulfill_condition
print("Number of men in danger group:", fulfill_condition)
print("Number of men not in danger group:", do_not_fulfill_condition)
facetgrid_obj=sns.catplot(x='diabetes',kind='count',data=df)
facetgrid_obj.fig.suptitle('Distribution of diabetes among men in danger group ')
```

Number of men in danger group: 2331  
 Number of men not in danger group: 19396

Out[35]: Text(0.5, 0.98, 'Distribution of diabetes among men in danger group ')



לא נוכל ליצור מסוג בעזרת הדאטה שיש לנו בגלל חוסר פרופורציה בין כמות האנשים בכל מין  
שחם בקבוצת סיכון לבין אלו שלא.

נרצה לחקור סכנה לסכנת הריאן בגין מבוגר.

```
In [36]: # ייצרים דאנט פרים עבור נשים בגילאי עד 45
women_between_35_45 = df[(df['gender'] == 'Female') & (df['age'] >= 35) & (df['age'] <= 45)
print("Number of women between ages 35 and 45:", len(women_between_35_45))

Number of women between ages 35 and 45: 8881
```

```
In [37]: women_between_35_45
```

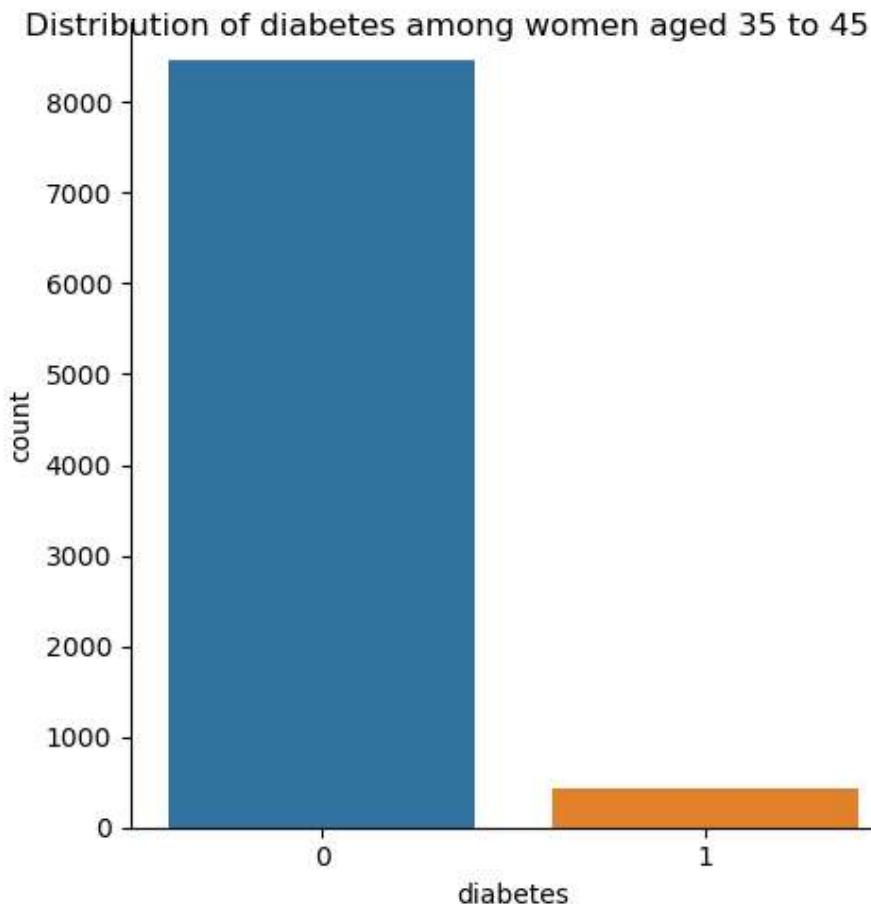
Out[37]:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glu
3	Female	36.0	0	0	current	23.45	5.0	
6	Female	44.0	0	0	never	19.31	6.5	
17	Female	42.0	0	0	never	24.48	5.7	
18	Female	42.0	0	0	No Info	27.32	5.7	
32	Female	41.0	0	0	current	22.01	6.2	
...	...	...	...	...	...	...	...	...
99958	Female	38.0	0	0	not current	23.58	4.8	
99966	Female	38.0	0	0	No Info	27.32	5.8	
99967	Female	45.0	0	0	former	32.81	4.5	
99993	Female	40.0	0	0	never	40.69	3.5	
99994	Female	36.0	0	0	No Info	24.60	4.8	

8881 rows × 9 columns

In [38]: `facetgrid_obj=sns.catplot(x='diabetes',kind='count',data=women_between_35_45)  
facetgrid_obj.fig.suptitle('Distribution of diabetes among women aged 35 to 45')  
women_between_35_45['diabetes'].value_counts()`

Out[38]: 0 8458  
1 423  
Name: diabetes, dtype: int64



לא נוכל ליצור מסוווג בעזרה הדעתה שיש לנו בגלל חוסר פרופורציה בין כמות הנשים בגילאי - 35  
45 שחולות בסוכרת וכאליה שלא חולות בסוכרת.

## עבודה מסכמת בקורס מבוֹא לניתוח נתונים-מוד נתונים על סוכרת

מגישיים:

רינה בלובורודוב

ניר צ'אוסר

סיכום:

مبין כל הנושאים שניתנו למחקר לעבודה בחרכנו במסד הנתונים של הסוכרת. כיום שיעורי ההימצאות של הסוכרת בישראל יציבים, אך על פי נתוני הדרציה הבינלאומית לסוכרת, שיעור הסוכרת לגיל בני 20-79 בישראל, הוא 9.7% - לעומת גבולה הממוצע (6.3%) במדינות אירופה. נתונים אלה מדויקים ולכך כחוקרי נתונים אנו מرجשים צורך לעזרו ואולי לקדם את התופעה לידי מודעות או לפחות עזרה מסוימת בניתור הבעיה טרם החמרה.

[/https://publichealth.doctorsonly.co.il/2021/11/243235](https://publichealth.doctorsonly.co.il/2021/11/243235)

מדוע דוחא בחרנו לחקר סוכרת?

סוכרת היא נושא בריאותי נפוץ וצומח ברחבי העולם. הבנת הסיבות, גורמי הסיכון ואסטרטגיות הנהול יכולה לשמש בטיפול בנטול הגובר של סוכרת והשפעתה על ייחידים. כמו כן קיימות השלכות בריאותיות: סוכרת עלולה להוביל לשיבוכים שונים, כגון מחלות לב וכלי דם, מחלת כלות, נזק עצבי וביעות ראייה. חקר הסוכרת מאפשר לנו לקבל תובנות לגבי המנגנונים הבסיסיים, גילי מוקדם ומונעה של סיבוכים אלו, ובוטפו של דבר לשפר את איכות החיים של הנפגעים. לימוד סוכרת מסייע לחוקרם לפתח אסטרטגיות טיפול חדשות ויעילות יותר, התקדמות בטכנולוגיה לניטור גלוקוז ומtan אינסולין, וגישה מותאמת אישית לטיפול רפואי ומחקר בתחום זה יכול להביא לשינויים משמעותיים בחיהם של אלו החולים עם המצב.

השאלות ששאלנו:

האם קיים הבדל בין גברים ונשים באחוזי התחלואה במחלת הסוכרת?

האם קיימת קטגוריה מסוימת שבה אחוזי התחלואה גבוהים במיוחד?

האם הגיל הוא גורם משפיע לתחלואה במחלת הסוכרת?

מהם גורמי הסיכון המשותפים לחולים והאם ניתן לאתר אותם טרם התפרצות המחלה?

תיאור מערכת הנתונים:

מערכת הנתונים שלנו כולל אוסף של נתונים רפואיים של החולים, יחד עם מצב הסוכרת שלהם (חיובי או שלילי). הנתונים כוללים מאפיינים כמו גיל, מגן, מדד מסת הגוף (BMI), יתר לחץ דם, מחלות לב, היסטוריה עשון, רמת HbA1c ורמת הגלוקוז בدم.

מערכת נתונים זה יכול לשמש לבניית מודלים של מידת מכונה כדי לחזות סוכרת בחולים על סמך ההיסטוריה הרפואית שלהם. הוא יכול להיות שימושי עבור אנשי מקצוע בתחום הבריאות בזיהוי החולים שעולים להיות בסיכון לפתח סוכרת ובפיתוח תוכניות טיפול מותאמות אישית. בנוסף, מערכת הנתונים יכול לשמש חוקרים כדי לחזור את הקשרים בין גורמים רפואיים שונים ואת הסבירות לפתח סוכרת.

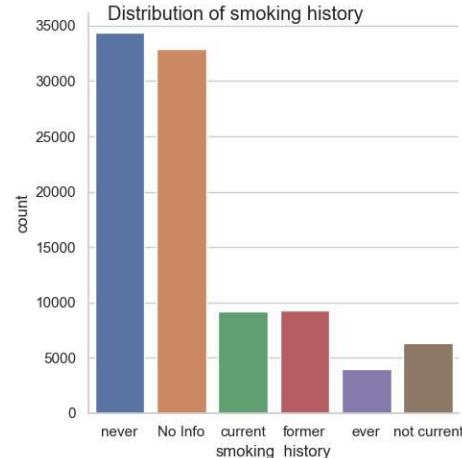
בחלק מהmarker נאלצנו לבצע התאמות לסוג השאלה שלנות ולכיווני המחקר שרצינו לחזור, כמו שילוב בין מידעות, מחיקת רשומות כפולות והפיקת משתנים קטגוריאליים לנוראים לצורך נוחות.

כפי שנכתב בתיאור, מערכת רשומות מספק מידע מספק במקרה לחזור קשרים שעיניינו אותן כמו מגדר והימצאות מחלת, וזה השאלה שבה נתונים נתמקד במחקר.

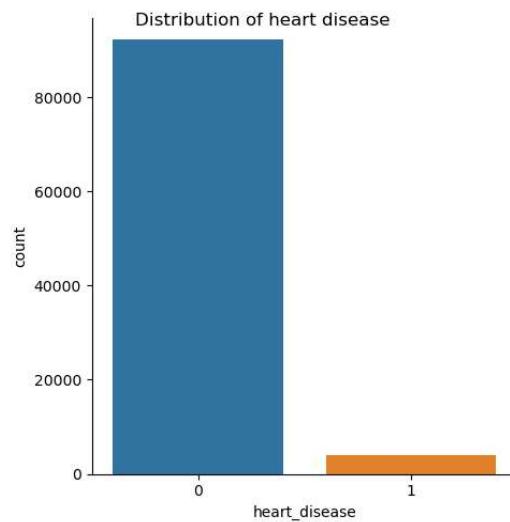
כמו כן במהלך המחקר נתקלנו בשאלות נוספות שעינינו אותן ומערכת הנתונים סייפק לנו מידע בשbill לענות עליהן( בעיית שנטקלנו בהן נפרט מדויק לא הצלחנו להגיע לתשובה סופית לחילק המשאלות).

### ייצוג התפלגיות במדגם שלנו:

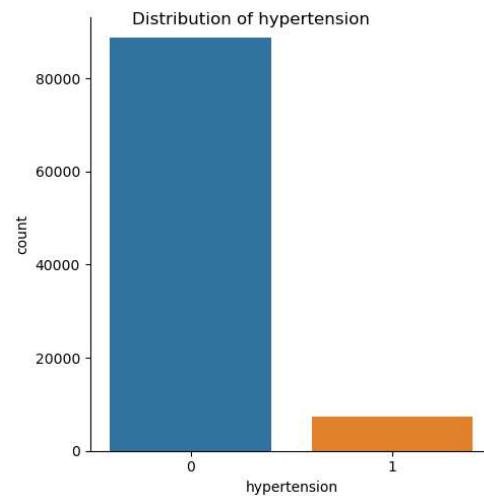
בתרשים זהה אנו יכולים לראות את ההתפלגות של האנשים לקטגוריות עישון שונות. על עבר שלוש מהנדגים אין מידע אודות הרגלי העישון שלהם. בנוסף ההבדלים בין הקטגוריות לא ברורים, למשל: לנוכח העובדה ש `not_current` מוגדרת כלא עישון, אנחנו לא יודעים איך מי שליך את הדגימות סיווג את מדדי העישון לקטגוריות. יכול להיות שבאותה קטgorיה יש אנשים שנחשפו לעישון בפרק זמן ממושך יותר ובכמות רבה יותר. למשל: אדם שעישן בעבר היה יכול לעשן סיגריה אחת ביום, חביבה ביום או מזית ביום. בנוסף, גם אדם שלא מעשן סיגריות יכול להיחשף לפחות עישון אם הוא נמצא בקרבם מעשנים. לעומת זאת, מעשן פסיבי.



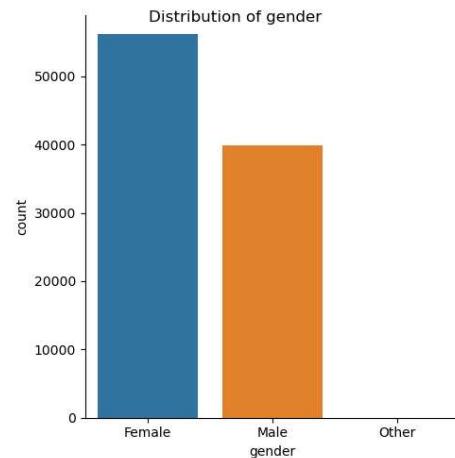
תרשים זה מייצג התפלגות של חוליות מחלת לב במאגר הנתונים שלנו. ניתן לראות שבמאגר כמעט אין חוליות במחלות לב.



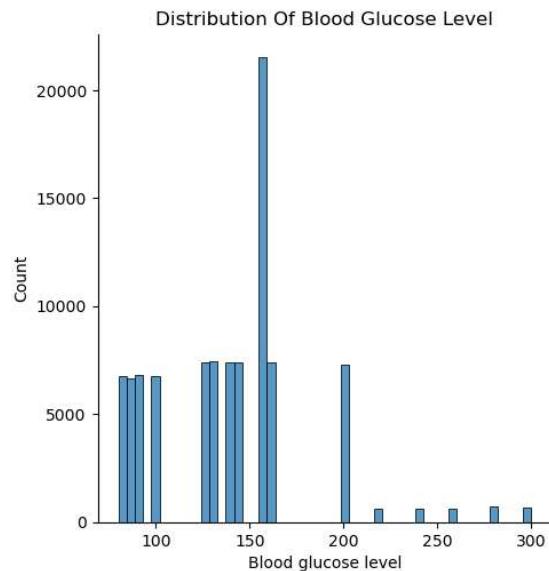
תרשים זה מייצג התפלגות של אנשים בעלי יתר לחץ-דם במאגר הנתונים שלנו. ניתן לראות שבמאגר כמעט אין אנשים עם לחץ דם גבוה.



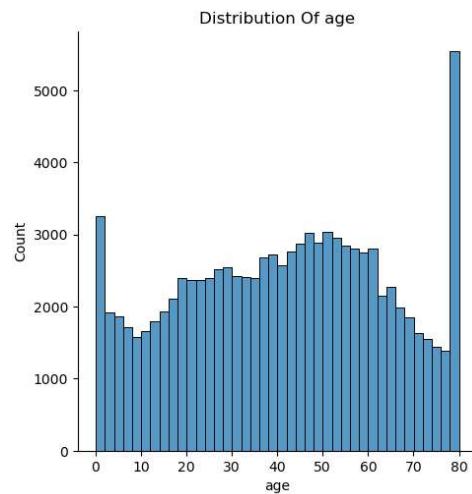
בתרשים זה אנחנו רואים שכמות הנשים גבוהה כרבע מכמות הגברים.



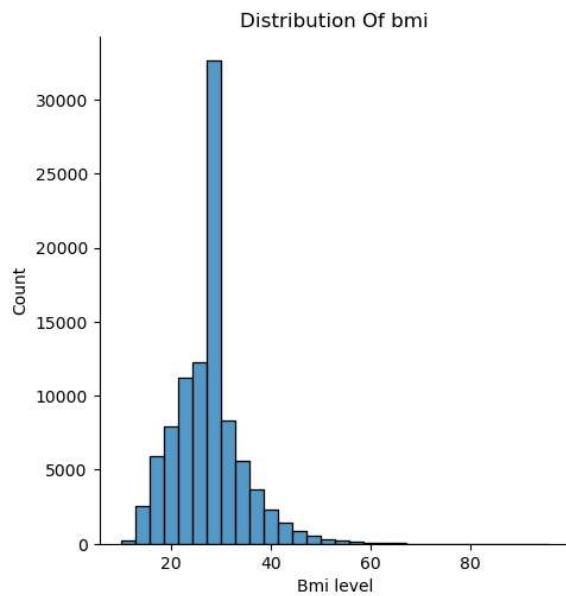
בתרשים ניתן לראות שאצל רוב הנדגמים רמות הגלוקוז היו מעל הנורמה (הנורמה היא 100-126 י"ח). נראה ישם הרבה אנשים שלא צמו לפני הבדיקה / או חולים בסוכרת.



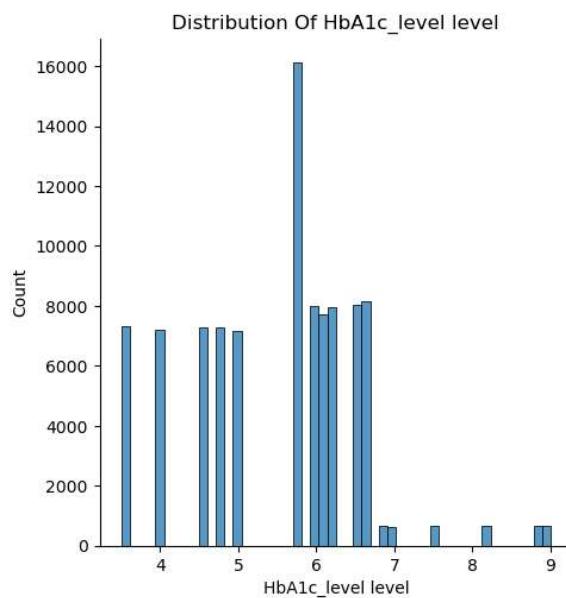
תרשים זה מייצג את התפלגות הגילאים. ניתן לראות שקבוצת הגיל הגדולה ביותר במדגם היא גילאי 78-80.



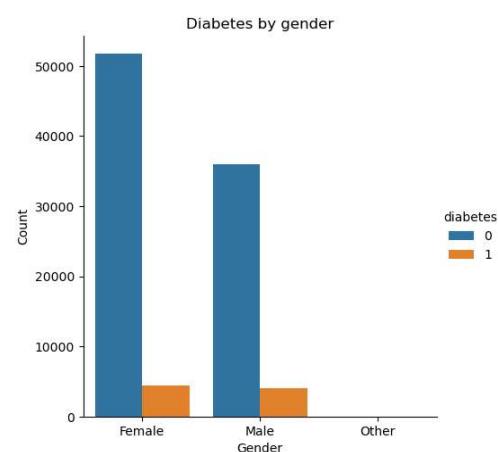
תרשים זה מייצג את התפלגות מדד `bmi` בקרוב אוכלוסיית המדגם.



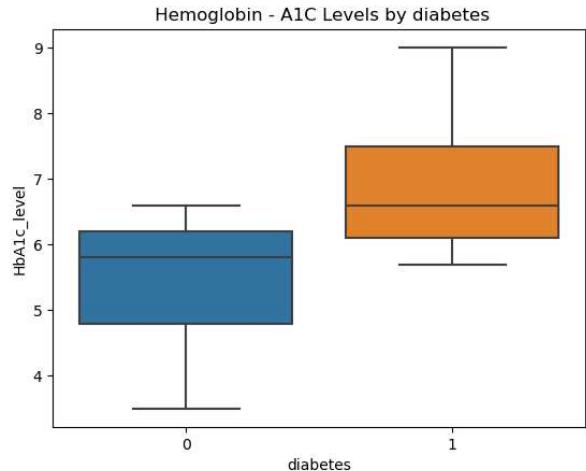
תרשים זה מייצג את התפלגות רמות הולבון המוגולביון A1C. ערך תקין של חלבון זה הינו פחות מ 5.7%. והחול מ 6.7% האדם מוגדר כחולה סוכרת. בגרף ניתן לראות ש מרבית האנשים במצב של טרום סוכרת.



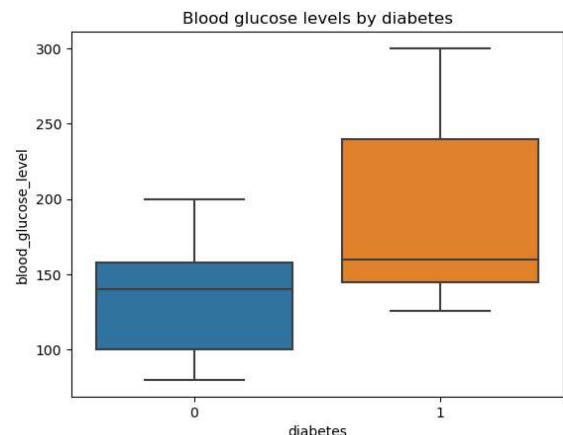
תרשים זה מראה את התפלגות חוליות הסוכרת בכלמין. ניתן לראות שבממוצע הנשים אינן הבדל משמעותית בין כמות הגברים שחוליות בסוכרת לבין כמות הנשים שחוליות בסוכרת. בנוסף, בממוצע הנשים יש יותר נשים שאינן חוליות מאשר גברים שאינם חוליות.



בתרשים זה ניתן לראות שיש קשר בין קיומ סוכרת לרמות גבוחות של המוגולוביין A1C בדם. כלומר, בקרב חולי הסוכרת רמות המוגולוביין זהה גבוחות יותר.

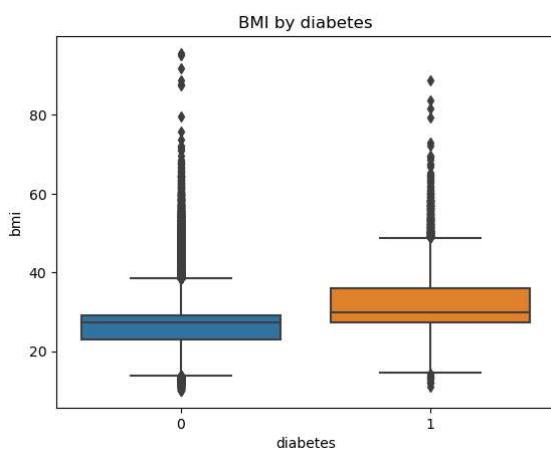


בתרשים זה ניתן לראות שקיים קשר בין תחלואה בסוכרת לבין רמות גבוחות של גליקוז. כלומר, בקרב חולי הסוכרת רמות הגלוקוז בדם גבוהות יותר מאשר שלא חולים.

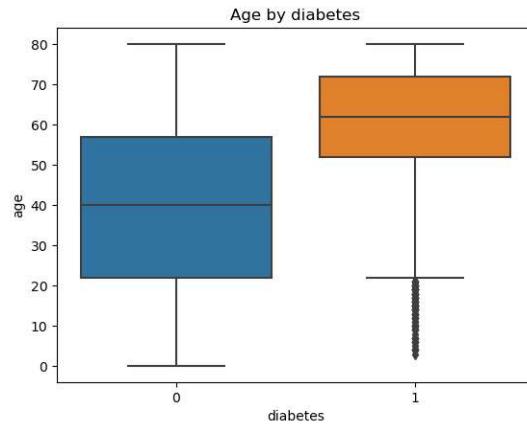


[https://www.maccabi4u.co.il/5737-he/Maccabi.aspx?TabId=5739\\_5740\\_5743](https://www.maccabi4u.co.il/5737-he/Maccabi.aspx?TabId=5739_5740_5743)

לפי תרשימים זה ניתן לראות שיש קשר בין קיומ סוכרת לרמות גבוחות של מדד ה *bmi*. כלומר, בקרב חולי הסוכרת מדד *bmi* גבוה יותר.



בתרשים זה ניתן לראות שיש קשר בין קיומם סוכרת לגיל מבוגר יותר. כלומר, מחלת הסוכרת נפוצה יותר בקרב אנשים מבוגרים.



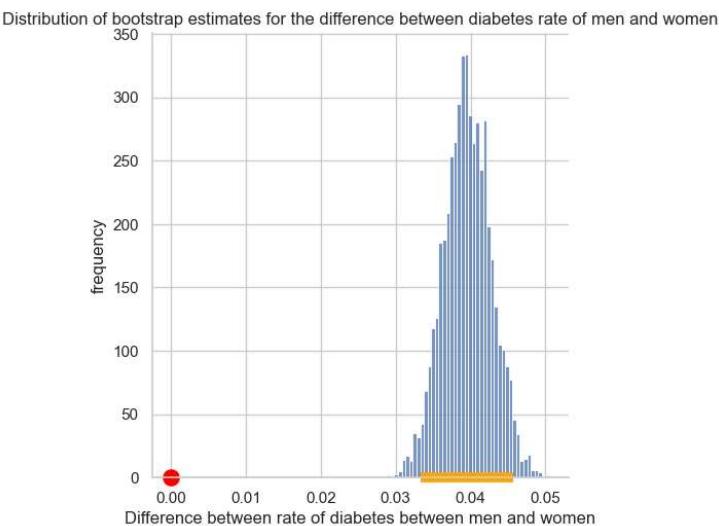
#### חיקירת השערות:

לפי הגרף של סוכרת לפי מיין אנו רואים שאין הבדל מובהק בין כמות הגברים והנשים החולים בסוכרת. רצינו לבדוק האם דבר זה נכון גם לכל האוכלוסייה ולא רק במדגם הנוכחי. אנו משתמשים בשיטת הקראטראפbootstrap לחישוב מרוחה ביטחון של 95% ונבדוק את ההפרש בין שיעור הנשים החולים לבין שיעור הגברים החולים וזאת על מנת לבדוק את ההשערה שלנו. שיטה זאת משמשת אותנו כאשר אין גישה לאוכלוסייה הכללית. ניתן להשתמש בה כאשר מניחים שהמדגם מייצג את האוכלוסייה הכללית. אנחנו לא יודעים כיצד נדגמו הננתונים ולכן אין לנו ברירה אלא לקוות שהמדגם אכן מייצג.

0H: אין הבדל בין שיעור התחלואה של גברים בסוכרת לעומת שיעור התחלואה של נשים באוכלוסייה.  
H1: יש הבדל בין שיעור התחלואה של גברים בסוכרת לעומת שיעור התחלואה של נשים באוכלוסייה.

סטטיסטי: הפרש הממוצעים בין שיעור תחלואה הגברים לשיעור תחלואה הנשים.

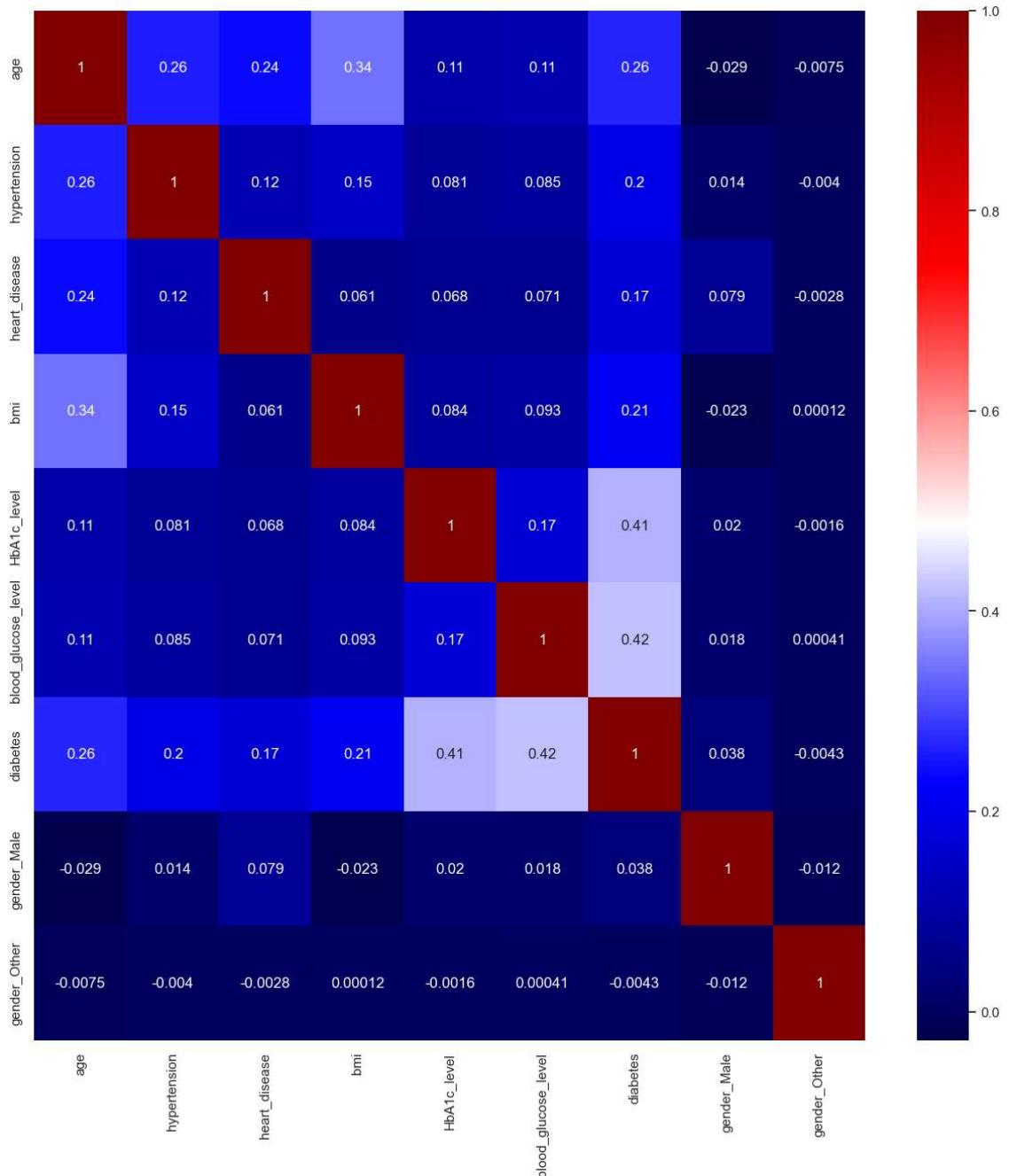
רווח הסマー שהתקבל הוא ערך [0.0333, 0.0457]  
עבור מרוחה ביטחון של 95%. ניתן לראות כי 0 אינו נכלל בטוחה זה (0) מייצג שאין הבדל בין שיעור התחלואה של גברים לזו של נשים) ולכן ניתן לדחות את השערת האפס. כלומר, יש הבדל בין שיעור הנשים החולים לבין שיעור הגברים החולים בסוכרת.



#### חיזוי:

רצינו למצוא מהן התכונות אשר עשויות להשפיע על מחלת הסוכרת וליצור בהן שימוש כדי לחזות האם אדם חוליה או לא.

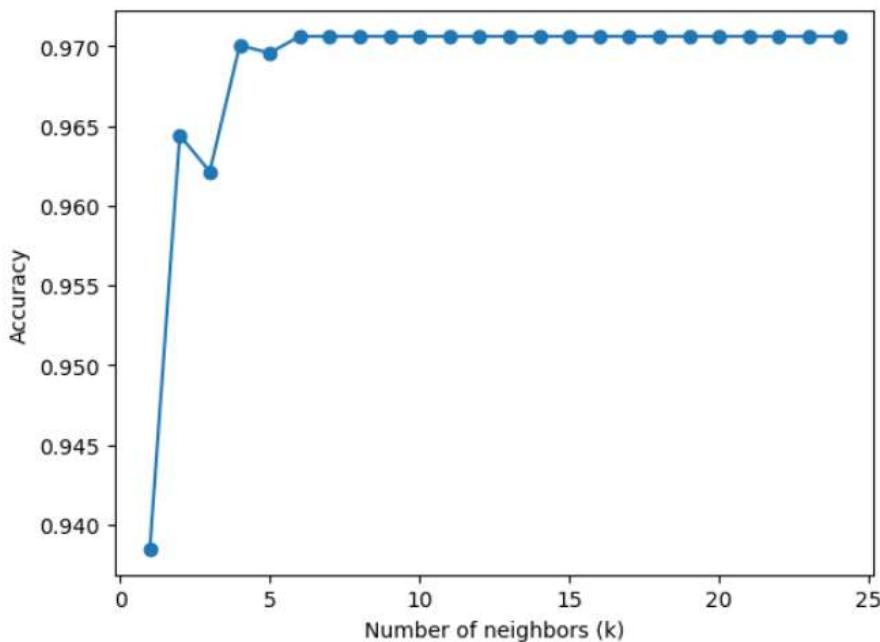
נשתמש בheatmap כדי למצוא את גורמי הסיכון אשר נמצאים בקורלציית הגבהות ביותר עם תחלואה בסוכרת.



אנו רואים שהפיזרים: `blood_glucose_level`, `HbA1c_level`, נמצאים בקורלציה גבוהה ביותר עם תחלואה הסוכרת. בתרשימים הקודמים רأינו שיש קשר בין מחלת הסוכרת ולן נרצה לבדוק אם הם יכולים לסייע בחיזוי.

על מנת לאמן את המסוווג שלנו נשתמש בשיטת ה - NNk. בהתחלה שמרנו את המשתנים שבחורנו ('`HbA1c_level`', '`blood_glucose_level`', '`diabetes`') וחילקנו את הנתונים כך ש 20% הם סט המבחן ו- 80% הם סט האימון.

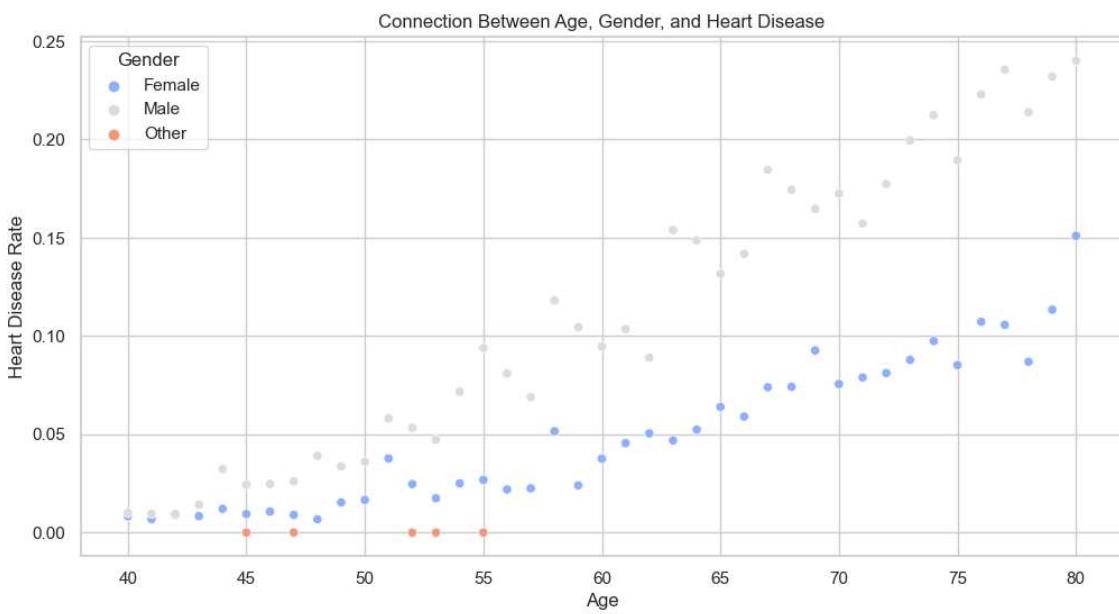
Highest accuracy is obtained for k = 6 and equals 0.9706043046038575



קיבלנו שהדיק יהה הטוב ביותר כאשר מספר השכנים יהיה 6.

רצינו לבדוק אם יש קשרים נוספים שהם מעניינים ואפשר לחקור עליהם.

1. קשר בין גיל וgendר לבין תחלואה במחלת לב (ולבנות מסווג במידת האפשר).



ניתן לראות שהטיצוי לחילות לב בקרב גברים מתחילה מוקדם יותר מאשר אצל נשים והטיצוי גבוה יותר. וכן ניתן לראות אם ניתן ליצור מסווג שיגיד אם אדם הוא בקרב ציכון לחילות לב.

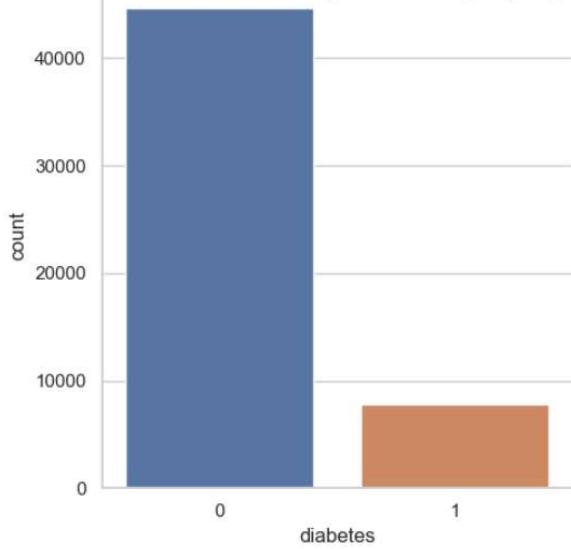
הוסףנו למאגר נתונים אודוט השתייכות לקבוצות סיכון. הגדרנו שגבר יהיה בקבוצת סיכון אם גילו מעל 40 שנה והוא חולה במחלת לב ואישה תהיה בקבוצת סיכון אם גילה מעל 50 שנה והוא חולה במחלת לב.

קיבלו את התפלגיות הבאות:

Number of men in danger group: 2331  
Number of men not in danger group: 19396

Text(0.5, 0.98, 'Distribution of diabetes among men in danger group ')

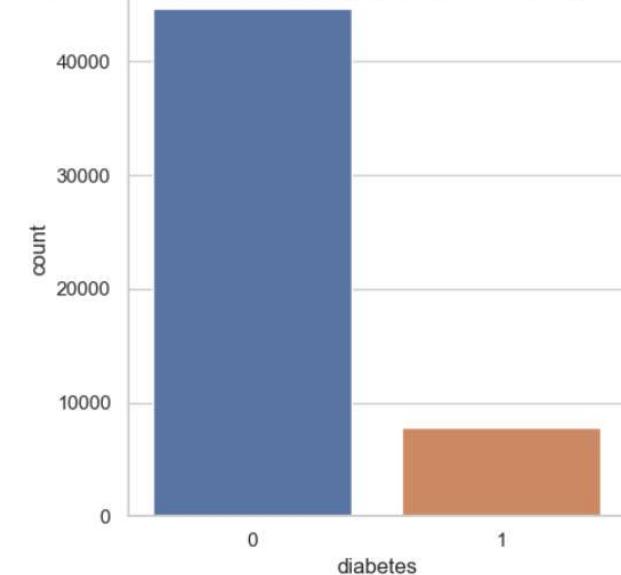
Distribution of diabetes among men in danger group



Number of women in danger group: 1435  
Number of women not in danger group: 29283

Text(0.5, 0.98, 'Distribution of diabetes among women in danger group ')

Distribution of diabetes among women in danger group

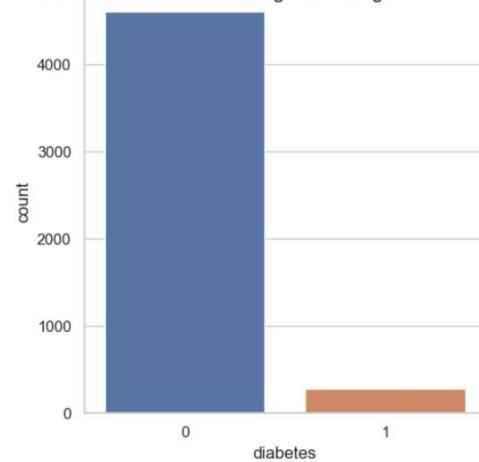


לא נוכל ליצור מסוג בעזרת הדטה שיש לנו בגלל חוסר פרופורציה בין כמות האנשים בכל מין מהם בקבוצת סיכון לבין אלו שלא.

2. קשר נוסף שרצינו לבדוק (ולבנות מסווג במידת האפשר) הוא סכנה לסתורת הרין בגיל מבוגר (35-45).

0 4609  
1 277  
Name: diabetes, dtype: int64

Distribution of diabetes among women aged 35 to 45



גם כאן לא נוכל ליצור מסוג בעזרת הדטה שיש לנו בגלל חוסר פרופורציה בין כמות הנשים בקבוצת גיל זאת לבין אלו שלא.

### מגבלות והטיות

גורמים דמוגרפיים יכולים להיות פקטור משמעותי בסוכרת. ההבדלים הקיימים בתהליכי הפטופיזיולוגיים בקרב מטופלים קשורים למוצא, לתרבות, לשנת עליה, لأنמנה משפחתיות, למחלות רקע ולהתנהגות בריאות. [https://cdn.doctorsonly.co.il/2018/06/06\\_margarita-preymoviz.pdf](https://cdn.doctorsonly.co.il/2018/06/06_margarita-preymoviz.pdf)

הטיות שלולות להיווצר:

#### הטיית בחירה:

יכול להיווצר הטיה בחירה ברישום של משתתפים למחקר, מכאן נקבע מוגם לא מייצג. שיטת הבוטסטראפ מתיחסת למוגם מייצג ואילו אם אין זה המשקנות שלנו עלולות להיות שגויות.

#### גורמים מבלבלים:

מצב סוציאו-אקונומי ומחלות נלוות, דברים שאנו לא חושפים אליהם מראש אלא רק למידע הרובוטי המוצע באקסל. כפי שצוין במאמר בראש הדף, אי התחשבות בגורמים המבלבלים הללו בניתוח הנתונים עלול להוביל לתוצאות מוטעות ולמסקנות לא מדויקות.

#### שאיות מדידה:

שאיות במדידה של משתנים הקשורים לסוכרת, כמו רמות הגלוקוז בدم, ממד מסת הגוף או שימוש בתרופות. מדידות לא מדויקות עלולות להכניס רעש ולהשפע על מהימנות הנתונים.

#### הטייה באיסוף נתונים:

שיטת איסוף כמו דיווח עצמי ומילוי שאלונים או סקירות של טבלאות רפואיות, עלולות ליצור הטיה. ישם מטופלים שעולים לדוח דיווח כזה על תסמן או התנהגות מסוימת בעקבות לחיצים או הטיה סביבתיות. כמו כן, ניתן מצב בו הרשות הרפואית לא כוללת את כל רמידע הנדרש, מה שיכל להוביל לנתחים לא מלאים או לא מדויקים.

### כיוונים עתידיים:

במהלך המחקר עלינו מספר שאלות אשר יכולות להיחקר בעתיד.

#### סוכרת באוכלוסיות מיוחדות:

נרצה להתמקד בסוכרת באוכלוסיות ספציפיות, כמו ילדים, מתבגרים ומבוגרים. נתשמש במידע שקיים במאסד הנתונים כדי להסביר מסקנות לגבי דפוסים ספציפיים לגיל מסוים, ומשם לפיק (באמצעות רפואיים) גישות טיפול וטיפול ארוכות טווח באוכלוסיות אלו.

#### גורמי סיכון סביבתיים:

אם יש השפעה של גורמים סביבתיים, כמו אורח חיים מסוים, סביבה זיהומית וכו', על התפתחות סוכרת והתרדרותה? נרצה להוסיף למסד הנתונים שלו מידע הנוגע לצורת התזונה (תזונה מבוקרת, תזונה שבוססת על טבעונות, דיאטות למיניהן ועוד), פעילות גופנית, מתח והשפעות סביבתיות אחרות וכן לבדוק האם יש לכל אלו השפעות חיוביות/שליליות על סוכרת.

באותה צורה בה אספנו את הנתונים הקודמים ניתן להוסיף רובייקות נוספת במילוי השאלהים לצורך פרמטרים אלה.

### בריאות וכלכלה:

ונכל לשאול כיצד שיעור התחלואה יכול להשפיע על הנטל הכלכלי של סוכרת על אנשים, מערכות הבריאות והחברה. נבדוק מה שיעור התחלואה בקרב האוכלוסייה ורק מערכות הבריאות הציבוריות יכולו להיערך בהתאם לעזרת מודלי הניבוי שניצור. כדי לספק תשובה לשאללה זו נצטרך לחקור מהן ההשפעות הכלכליות של התחלואה על הכלכלת. לאחר מכן ליצור טווחי הוצאות וכמוון ליצור מדרג מתאים עבורן.