

Final project

136010-1 Introduction to DH Tools and Methods

Nesytova Olga

Motivation

- In this Project I am exploring diabetes prediction dataset.
- Predictive medicine is currently in high demand. By leveraging individual health data, such as body mass index and blood pressure measurements, we can predict the risk of various diseases for an individual. This enables early detection of illnesses or prescription of preventive therapies, ultimately extending life and improving the quality of life for individuals.
- MedTech is the field where I worked for two years and where I aim to continue my career.

Example of case

- Suppose I work in an IT company that collaborates with the city administration. They are planning medical events and provide our company with non-personal data obtained from an insurance company or medical organization.
- It is necessary to assist the clients in developing therapeutic and preventive plans

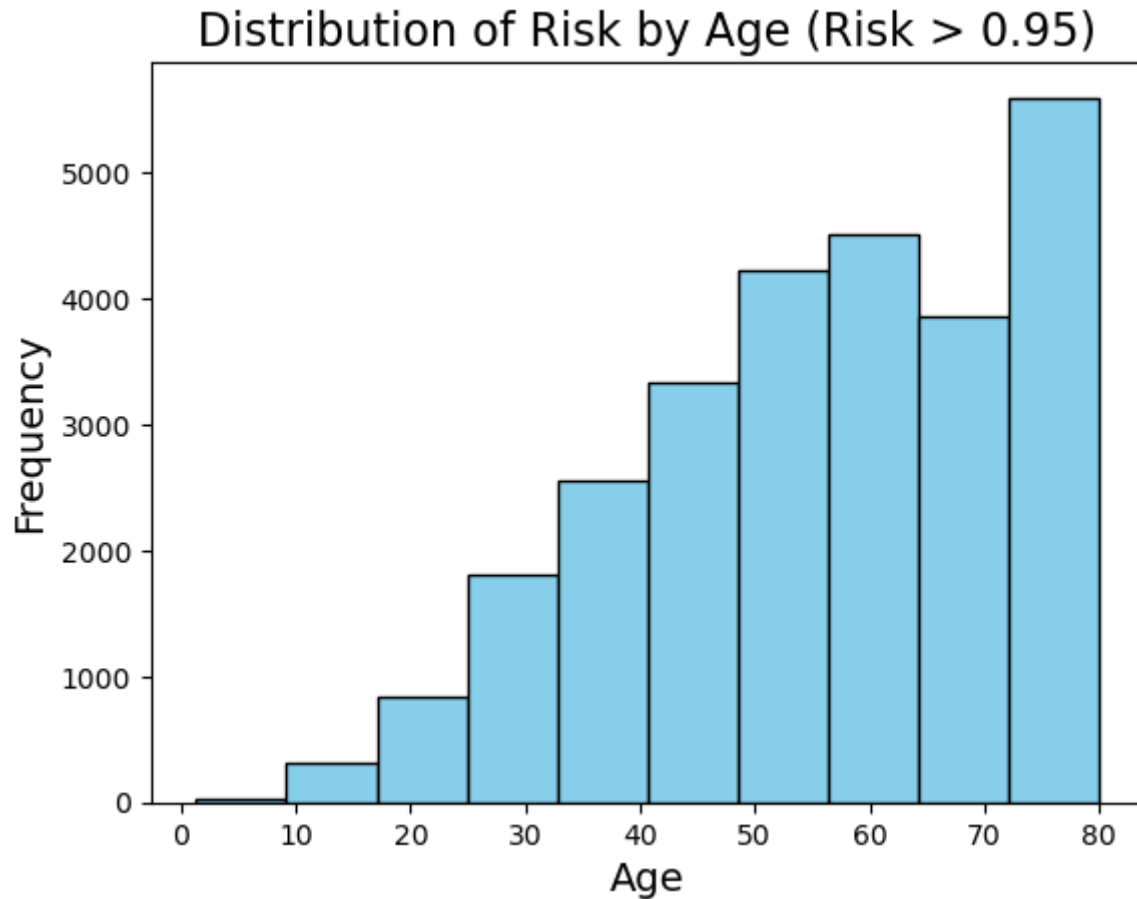
Research questions

- In which age groups is the risk of cardiovascular diseases higher?
- How many people are smokers?
- Who is more affected by obesity: men or women? In which age groups?

Methodology

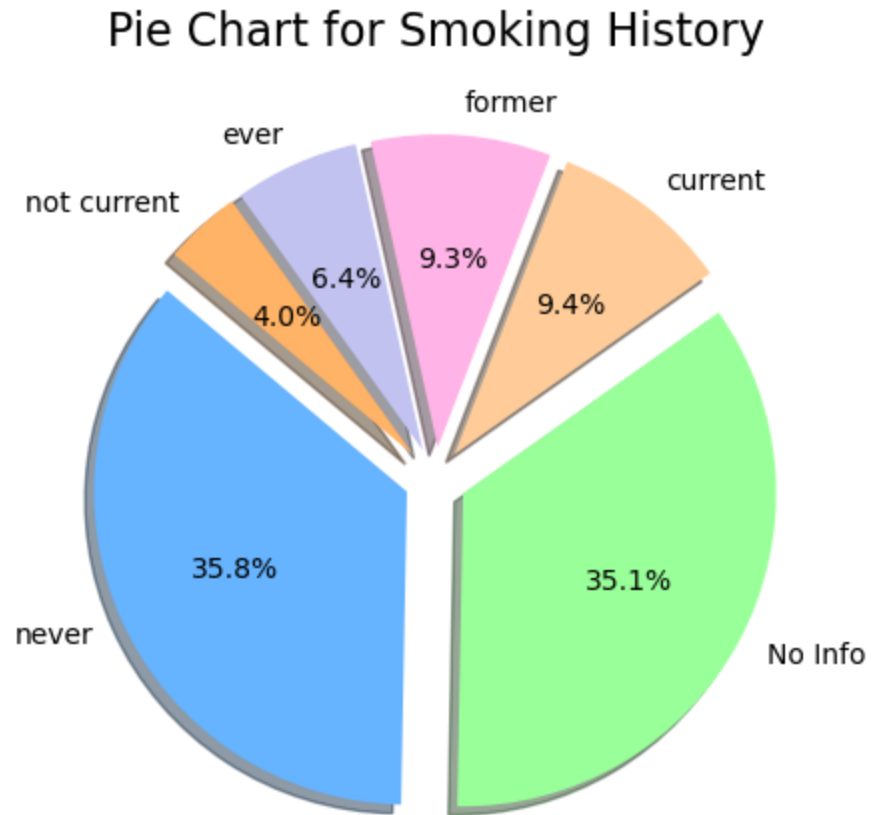
1. Obtain the data and analyze its composition.
2. Calculate the risk of cardiovascular diseases for each row using the formula provided by the client.
3. Generate graphs based on the conditions specified by the client.
4. Analyze the obtained information.
5. Submit the diagrams to the client for developing plans for therapeutic and preventive activities

Results



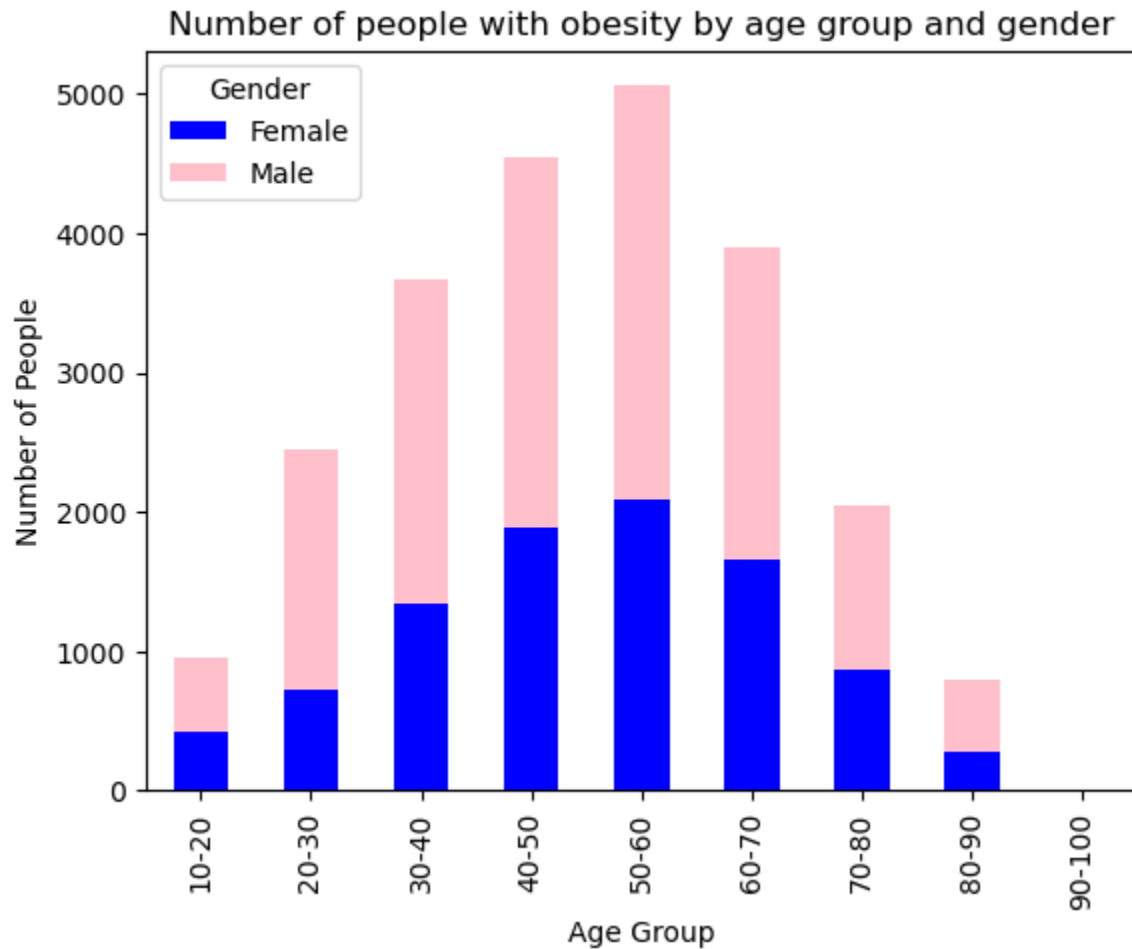
- People aged 49 to 64 and 72 to 80 are most susceptible to the risk of cardiovascular diseases
- It is necessary to implement therapeutic and preventive activities for individuals over 50 years old.

Results



- The percentage of smokers is 9.4%
- A significant portion of rows lacks information. Additional research is required to draw conclusions.

Results



- Men aged 40 to 60 are most susceptible to obesity.
(Obesity is diagnosed when the body mass index is over 30)

Conclusion

- In which age groups is the risk of cardiovascular diseases higher?

People aged 49 to 64 and 72 to 80

- How many people are smokers?

Insufficient data for a reliable conclusion

- Who is more affected by obesity: men or women? In which age groups?

Men aged 40 to 60

Data processing steps

Data processing steps

1. At first import libraries and create data frame from csv-file

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

medical_df = pd.read_csv('C:\pythonhomework\diabetes_prediction_dataset.csv')
```

Data processing steps

2. Display a few rows on the screen to understand the composition of the data. Output information about the data types in the columns


```
medical_df.head()
```

✓ 0.1s

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Data processing steps

3. Output information about the data types in the columns

```
medical_df.info()   
✓ 0.1s  
<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
```

Data processing steps

4. For subsequent calculations, it is necessary to represent the 'gender' column in the format of '0' and '1' and change the data type of certain columns

```
medical_df['gender'] = medical_df['gender'].map({'Male': 0, 'Female': 1})
medical_df['hypertension'] = medical_df['hypertension'].astype(float)
medical_df['heart_disease'] = medical_df['heart_disease'].astype(float)
medical_df['blood_glucose_level'] = medical_df['blood_glucose_level'].astype(float)
medical_df['diabetes'] = medical_df['diabetes'].astype(float)

medical_df.to_csv('C:\pythonhomework\diabetes_prediction_dataset.csv', index=False)
```

Data processing steps

5. Verify the changes. The code executed successfully

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	1.0	80.0	0.0	1.0	never	25.19	6.6	140.0	0.0
1	1.0	54.0	0.0	0.0	No Info	27.32	6.6	80.0	0.0
2	0.0	28.0	0.0	0.0	never	27.32	5.7	158.0	0.0
3	1.0	36.0	0.0	0.0	current	23.45	5.0	155.0	0.0
4	0.0	76.0	1.0	1.0	current	20.14	4.8	155.0	0.0

Data processing steps

6. To answer the question of which age groups have a higher risk of cardiovascular diseases, it is necessary to create a new column in the file and calculate this risk using the formula provided by the client

```
HR = 0.07

medical_df['Probability'] = 1 - (1 - HR)**(medical_df['gender'] + \
0.1 * medical_df['age'] + medical_df['heart_disease'] + \
medical_df['bmi'] + medical_df['HbA1c_level'] + medical_df['diabetes'])

medical_df.to_csv('C:\pythonhomework\diabetes prediction dataset.csv', index=False)
```


Data processing steps

7. Verify the changes.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	Probability
0	1.0	80.0	0.0	1.0	never	25.19	6.6	140.0	0.0	0.951816
1	1.0	54.0	0.0	0.0	No Info	27.32	6.6	80.0	0.0	0.946392
2	0.0	28.0	0.0	0.0	never	27.32	5.7	158.0	0.0	0.925688
3	1.0	36.0	0.0	0.0	current	23.45	5.0	155.0	0.0	0.909142
4	0.0	76.0	1.0	1.0	current	20.14	4.8	155.0	0.0	0.912316

Data processing steps

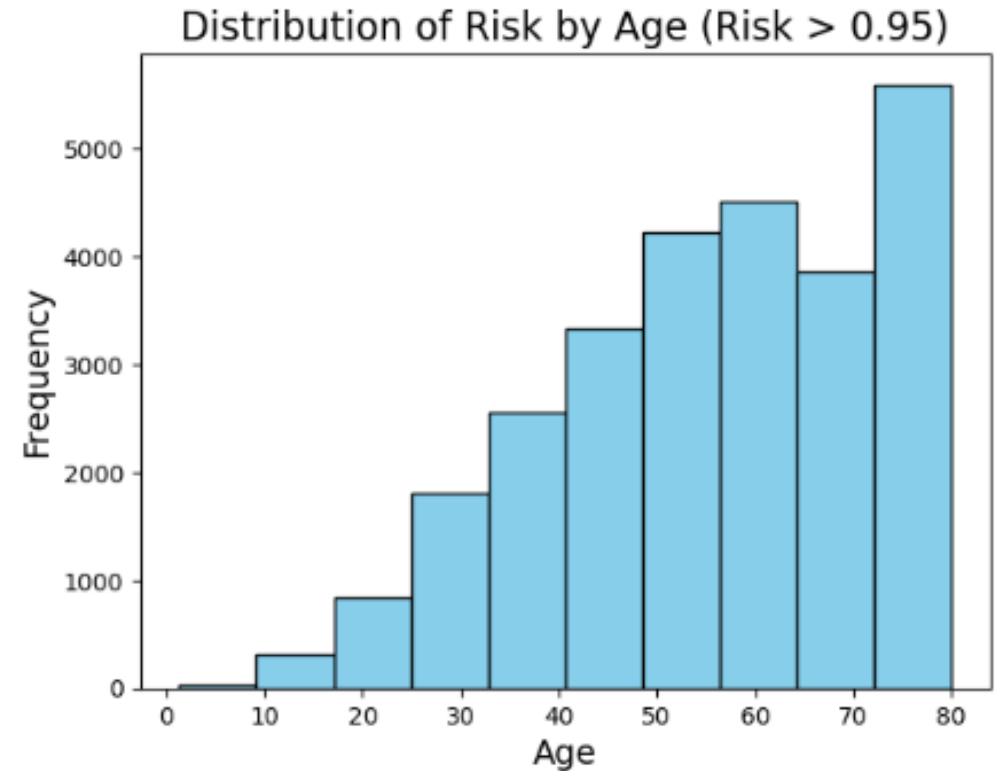
8. Construct a diagram illustrating the distribution of high risk by age (risk above 0.95)

```
high_risk_df = medical_df[medical_df['Probability'] > 0.95]

plt.hist(high_risk_df['age'], bins=10, color='skyblue', edgecolor='black')

plt.title('Distribution of Risk by Age (Risk > 0.95)', fontsize=16)
plt.xlabel('Age', fontsize=14)
plt.ylabel('Frequency', fontsize=14)

plt.show()
```



Data processing steps

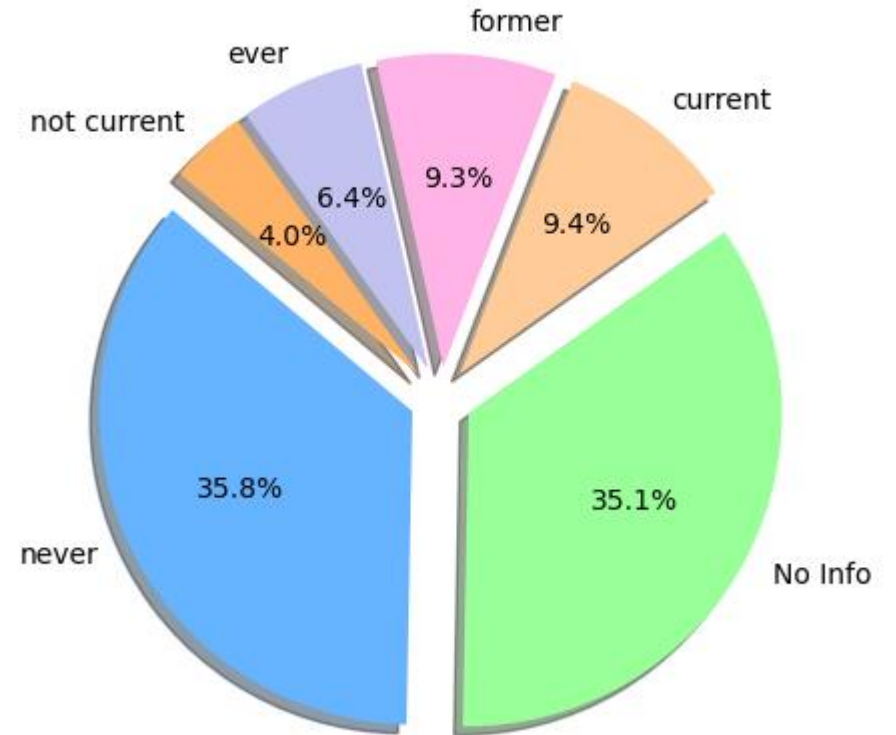
9. Create a pie chart showing the percentage distribution of smokers

```
colors = ['#66b3ff', '#99ff99', '#ffcc99', '#ffb3e6', '#c2c2f0', '#ffb366']
explode = [0.1 for category in medical_df['smoking_history'].unique()]

plt.pie(medical_df['smoking_history'].value_counts(), labels=medical_df['smoking_history'].unique(), /
        autopct='%1.1f%%', colors=colors, explode=explode, shadow=True, startangle=140)
plt.title('Pie Chart for Smoking History', fontsize=16, pad=20)
plt.axis('equal')

plt.show()
```

Pie Chart for Smoking History



Data processing steps

10. Construct a bar chart illustrating the age category with a higher number of individuals with obesity, segmented by gender

```
bins = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
labels = ['10-20', '20-30', '30-40', '40-50', '50-60', '60-70', '70-80', '80-90', '90-100']
medical_df['Age Group'] = pd.cut(medical_df['age'], bins=bins, labels=labels, right=False)

df_obese = medical_df[medical_df['bmi'] > 30]

grouped_data = df_obese.groupby(['Age Group', 'gender']).size().unstack().fillna(0)

grouped_data.plot(kind='bar', stacked=True, color=['blue', 'pink'])

plt.title('Number of people with obesity by age group and gender')
plt.xlabel('Age Group')
plt.ylabel('Number of People')
plt.legend(title='Gender', loc='upper left', labels=['Female', 'Male'])
plt.show()
```

