EDA

Exploratory Data Analysis

Committee: Prof. Nguyễn Hữu Tri

Prof. Lư Hồng Phúc

Prof. Lê Thanh Phát

Prof. Đỗ Duy Quý

Advisor: **Prof.** Đỗ Như Tà**i**

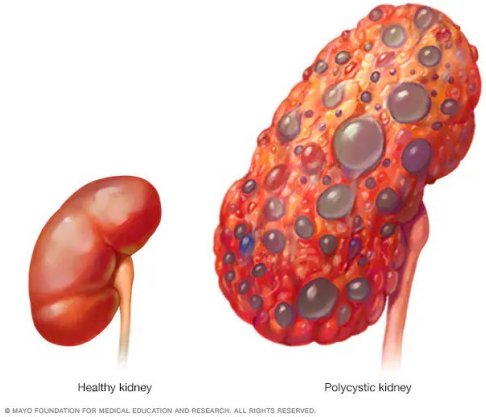
September 29st, 2025

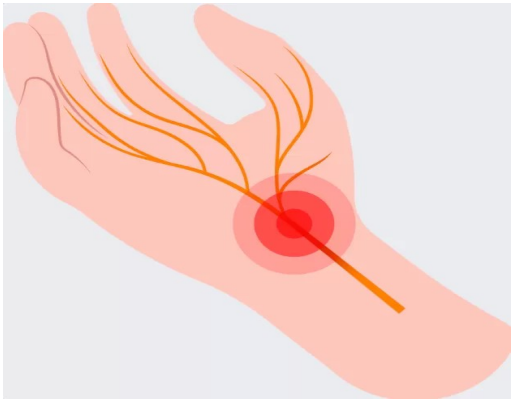
Agenda

1. Introduction
2. Related Works
3. Data and Method
4. EDA
5. Experiments and Discussion
6. Conclusion

1. INTRODUCTION

- **Problem: Diabetes Mellitus**
  - Reason: Lack of insulin
  - Consequence:
    - Cardiovascular
    - Kidney failure
    - Blindness
    - Nerve damage
    - …





1. INTRODUCTION

Early detection and accurate prediction of disease risk

- **Input**: Pima Indians Diabetes dataset
- **Output**:

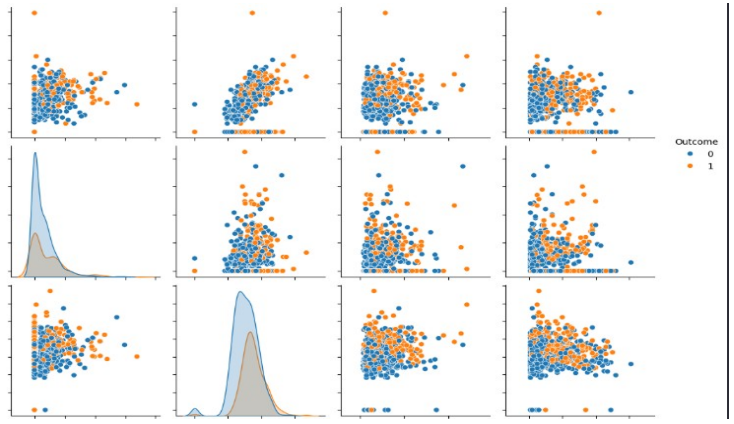
  - Deeply analysis about the dataset

    - Tell a ‘data story’

To lay the foundation for buiding a ML model

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
0	6	148	72	35	0
1	1	85	66	29	0
2	8	183	64	0	0
3	1	89	66	23	94
4	0	137	40	35	168
...	...	...	...	...	...
763	10	101	76	48	180
764	2	122	70	27	0
765	5	121	72	23	112
766	1	126	60	0	0
767	1	93	70	31	0

768 rows × 9 columns



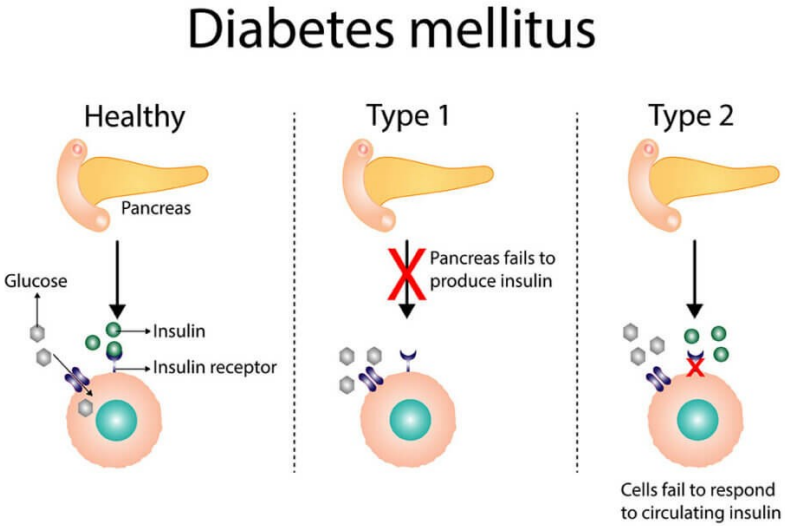
2. RELATED WORKS

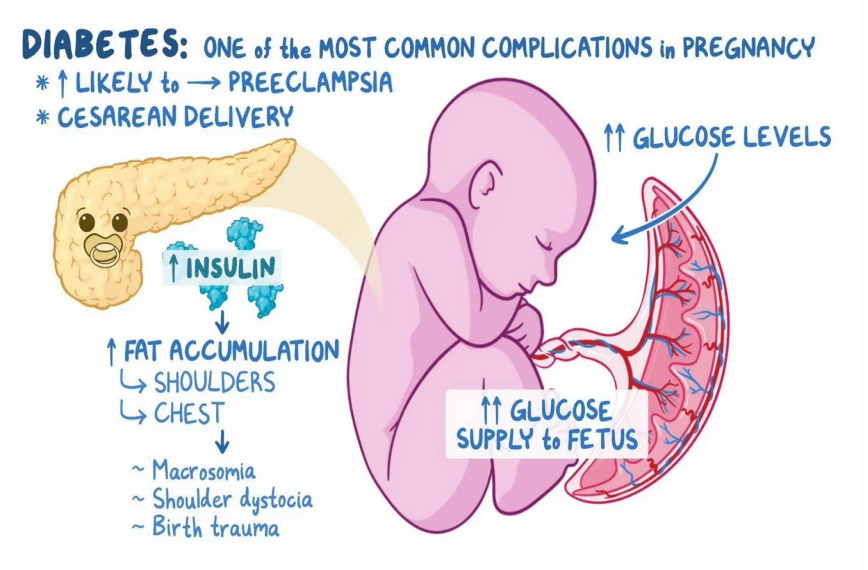
1. Type of diabetes mellitus:

Type 1

Type 2

Gestational Diabetes





2. RELATED WORKS

2. Pima Indian Community:

- a population with an unusually high incidence of type 2 diabetes

3. Pima Indians Diabetes:

• **Sample size**: 768 female patients.

• **Subjects:** Women aged 21 years and older of the Pima ethnic group.

• **Number of attributes:** 8 input features and 1 target variable (Outcome).

• **Target variable:** *Outcome* is a binary variable, with 1 being diabetic and 0 being non-diabetic

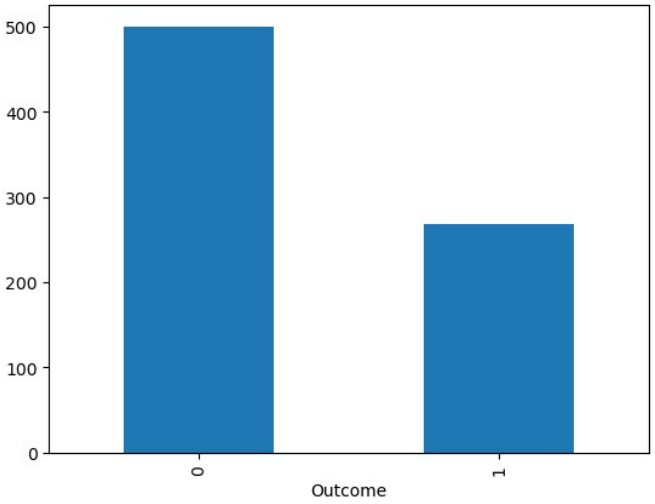
2. RELATED WORKS

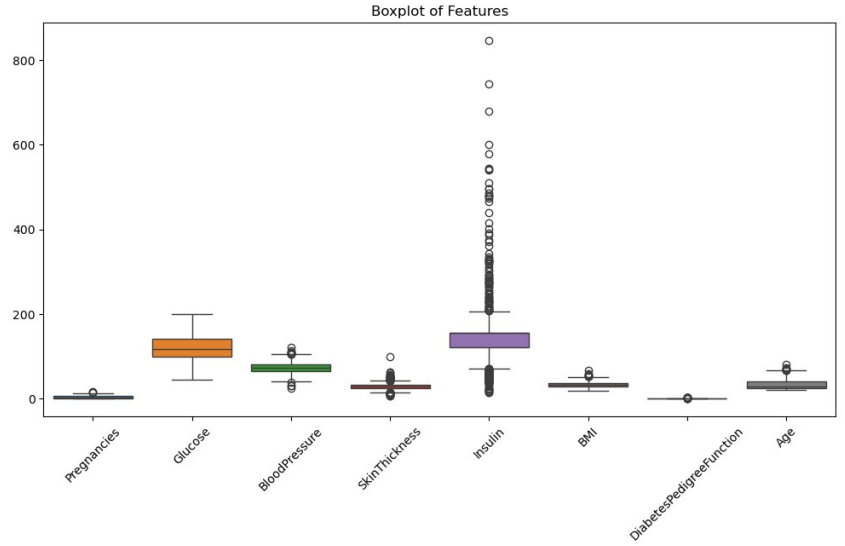
4. Meaning of features in dataset:

- 1. **Pregnancies** (Number of pregnancies)
- 2. **Glucose** (2-hour plasma glucose concentration in the oral glucose tolerance test - OGTT)
- 3. **Blood Pressure** (Diastolic blood pressure - mm Hg)
- 4. **Skin Thickness** (Triceps skin fold thickness - mm)
- 5. **Insulin** (2-hour serum insulin concentration - mu U/ml)
- 6. **BMI** (Body mass index - kg/m^2)
- 7. **DiabetesPedigreeFunction** (Genetic risk index)
- 8. **Age** (Patient's age - years)

3. DATA AND METHODS

- **Data and preprocessing issue:**
  - **Unreasonable Values:** Glucose, BloodPressure, SkinThickness, Insulin, và BMI have values 0.
  - **Class Imbalance:** The number of class 1 is almost double the number of class 2
  - **Outliers:** Insulin features have higher values than others





3. DATA AND METHODS

- **Method for EDA:**
  - **Descriptive statistics**: centrality and dispersion
  - **Univariate analysis**: distribution of each variable using *histograms and density plots*
  - **Bivariate analysis**: differences of 2 features using boxplots and *violin plots*.
  - **Multivariate analysis**: correlation between variables using *heatmaps*.

- **Environment and tools:**

  Python and some library: Pandas, Matplotlib và Seaborn

4. Exploratory Data Analysis

- **Descriptive statistics:**

features

```
df
```

Python

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Shape:(768, 9)

4. Exploratory Data Analysis

- **Descriptive statistics:**

description of each columns

description

	preg	plas	pres	skin	test	mass	pedi	age	class
count	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000	768.0000
mean	3.8451	120.8945	69.1055	20.5365	79.7995	31.9926	0.4719	33.2409	0.3490
std	3.3696	31.9726	19.3558	15.9522	115.2440	7.8842	0.3313	11.7602	0.4770
min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0780	21.0000	0.0000
25%	1.0000	99.0000	62.0000	0.0000	0.0000	27.3000	0.2437	24.0000	0.0000
50%	3.0000	117.0000	72.0000	23.0000	30.5000	32.0000	0.3725	29.0000	0.0000
75%	6.0000	140.2500	80.0000	32.0000	127.2500	36.6000	0.6262	41.0000	1.0000
max	17.0000	199.0000	122.0000	99.0000	846.0000	67.1000	2.4200	81.0000	1.0000

Data quality problem: presence of a minimum (min) value of 0 in some columns

4. Exploratory Data Analysis

Missing values

```
df.isna().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

no missing values

Unreasonable values

```
# Show all row have values 0.0 in the dataset
data_zero = data[(data == 0.0).any(axis=1)]
data_zero
```

0.0s

	preg	plas	pres	skin	test	mass	pedi	age	class
0	6	148	72	35	0	33.6000	0.6270	50	1
1	1	85	66	29	0	26.6000	0.3510	31	0
2	8	183	64	0	0	23.3000	0.6720	32	1
3	1	89	66	23	94	28.1000	0.1670	21	0
4	0	137	40	35	168	43.1000	2.2880	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9000	0.1710	63	0
764	2	122	70	27	0	36.8000	0.3400	27	0
765	5	121	72	23	112	26.2000	0.2450	30	0
766	1	126	60	0	0	30.1000	0.3490	47	1
767	1	93	70	31	0	30.4000	0.3150	23	0

657 rows × 9 columns

a lot of unreasonable values: 0

4. Exploratory Data Analysis

- **Univariate (Non-Grapical):**

correlation between ‘mass’ and plas

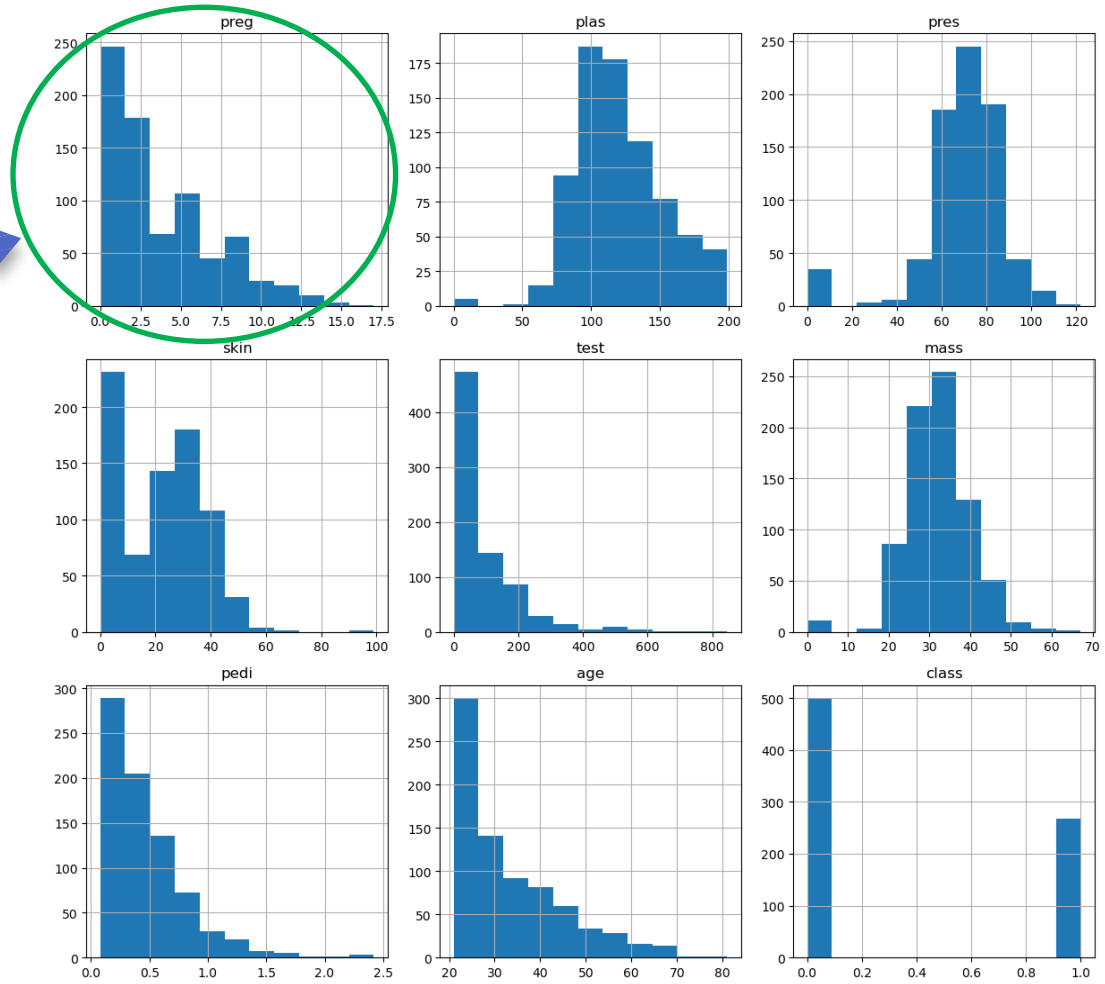
```
data_corr = data.corr(method='pearson')
data_corr
```

✓ 0.0s

	preg	plas	pres	skin	test	mass	pedi	age	class
preg	1.0000	0.1295	0.1413	-0.0817	-0.0735	0.0177	-0.0335	0.5443	0.2219
plas	0.1295	1.0000	0.1526	0.0573	0.3314	0.2211	0.1373	0.2635	0.4666
pres	0.1413	0.1526	1.0000	0.2074	0.0889	0.2818	0.0413	0.2395	0.0651
skin	-0.0817	0.0573	0.2074	1.0000	0.4368	0.3926	0.1839	-0.1140	0.0748
test	-0.0735	0.3314	0.0889	0.4368	1.0000	0.1979	0.1851	-0.0422	0.1305
mass	0.0177	0.2211	0.2818	0.3926	0.1979	1.0000	0.1406	0.0362	0.2927
pedi	-0.0335	0.1373	0.0413	0.1839	0.1851	0.1406	1.0000	0.0336	0.1738
age	0.5443	0.2635	0.2395	-0.1140	-0.0422	0.0362	0.0336	1.0000	0.2384
class	0.2219	0.4666	0.0651	0.0748	0.1305	0.2927	0.1738	0.2384	1.0000

4. Exploratory Data Analysis

- **Univariate(Grapical): Histogram**



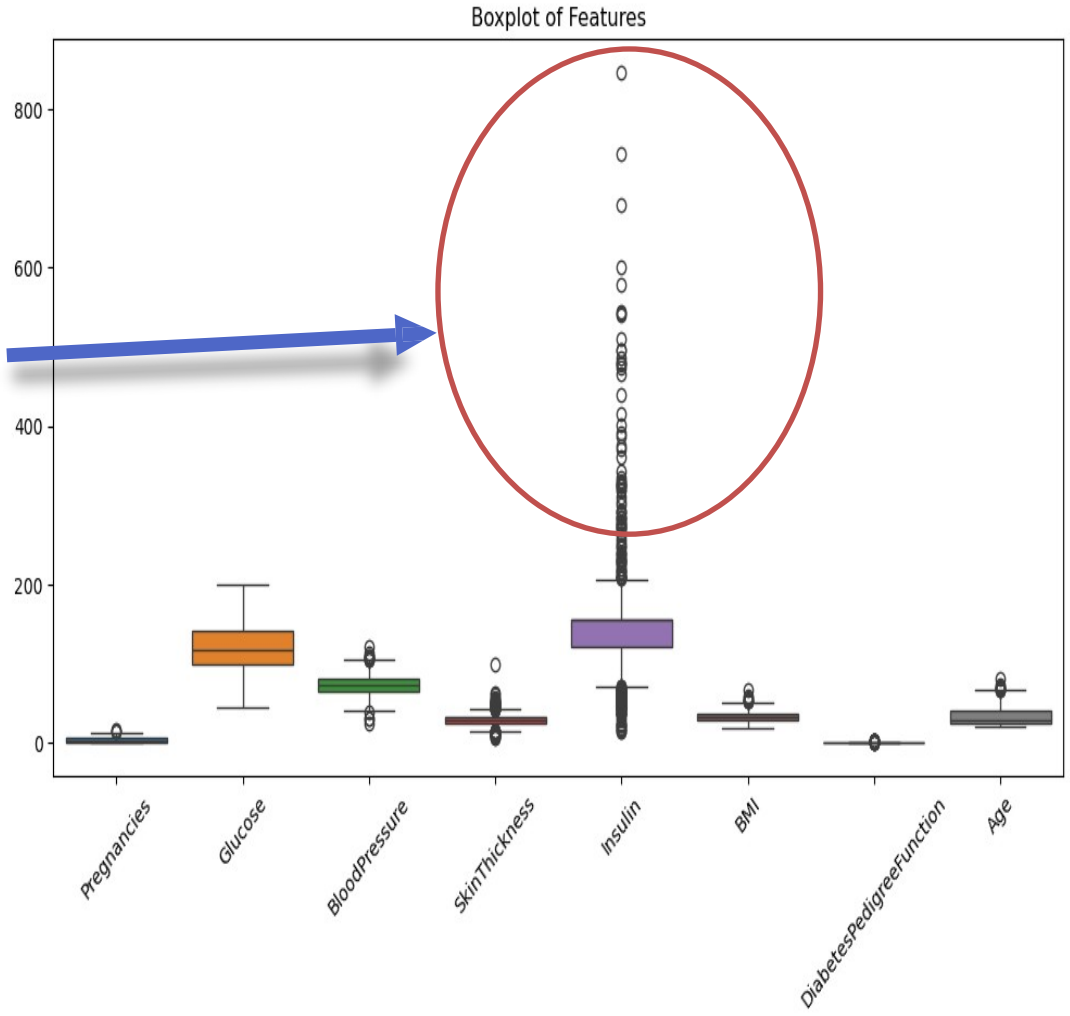
preg has a *long tail to the right*, confirming the presence of very high values.

4. Exploratory Data Analysis

- **Univariate(Grapical): Boxplot**

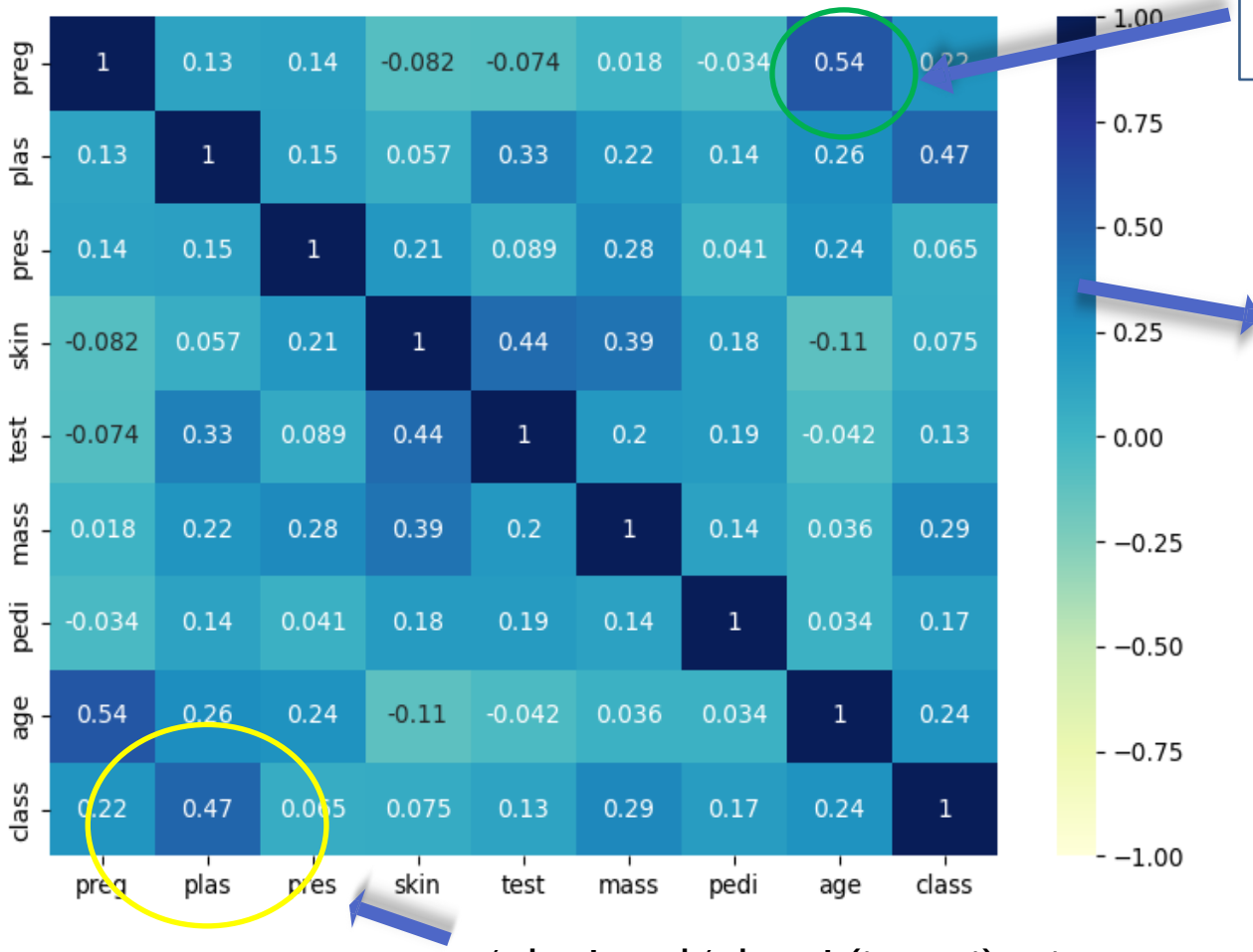
insulin: the most *outlier* property

values far exceeding the upper threshold (above 300 µU/mL).



4. Exploratory Data Analysis

- **Multivariate (Grapical): Heatmap**



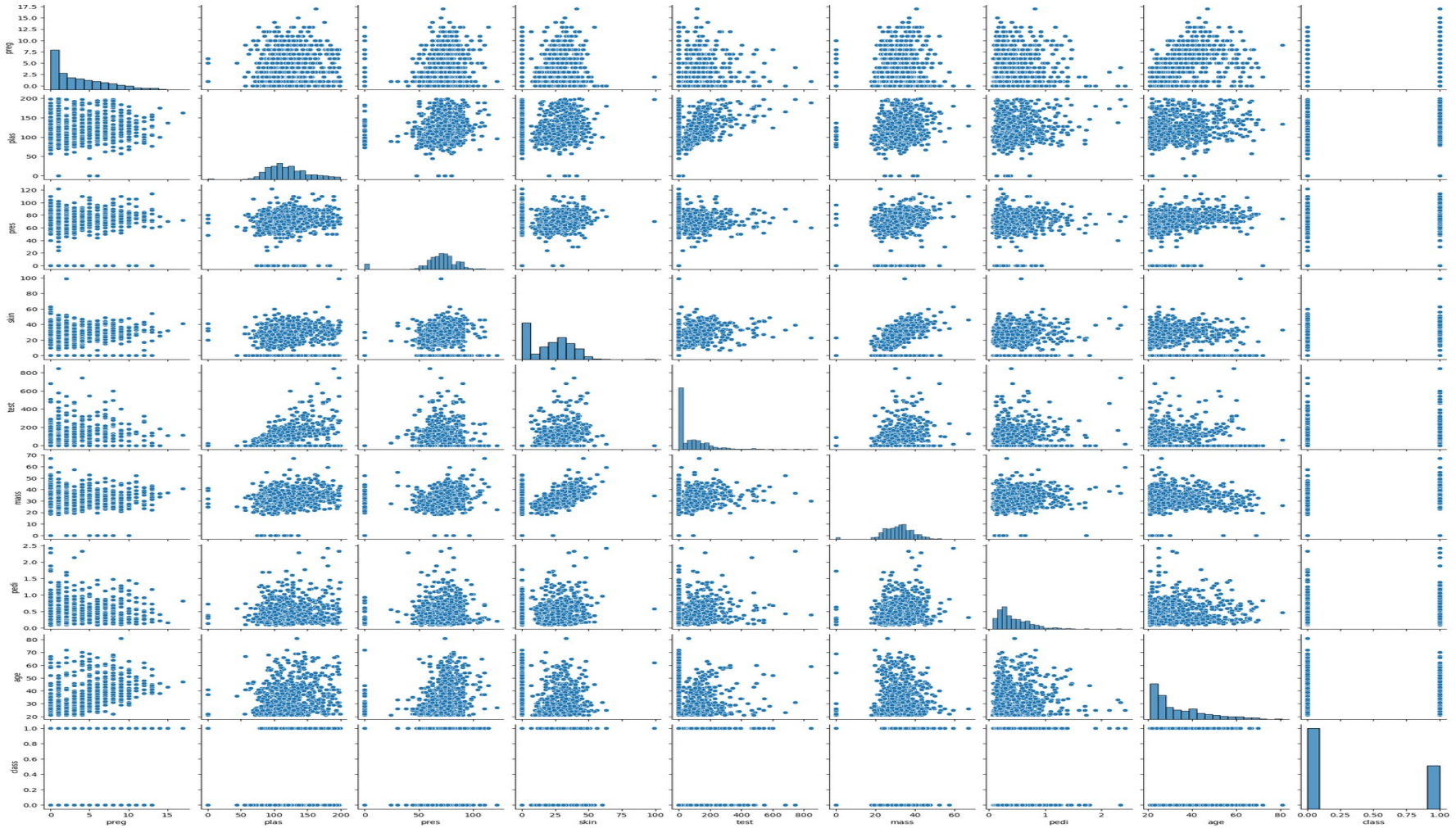
correlation between ‘preg’ and ‘age’

the higher values the higher correlation

‘plas’ and ‘class’ (target): strong correlation

4. Exploratory Data Analysis

- **Multivariate (Grapical): Scatter plot**



4. EXPERIMENTS AND ISCUSSION

- **Medical Implications**

  - **Glucose** is a key predictor

  - Role of **Obesity** and **Age**: BMI and Age are the next two most highly correlated factors.

  - Influence of **Genetics** and Pregnancy: diabetesPedigreeFunction and Pregnancies also showed significant associations. This highlights the importance of genetics and metabolic changes during pregnancy (gestational diabetes)

- **Practical Implications**

  - Foundation for early screening tools

  - Guidance for public health strategies

  - Support for clinical decision making

5. CONCLUSIONS

The main results identified *glucose levels, BMI, and age* as the three most important predictors of type 2 diabetes. In addition, the study also found serious data quality issues, including a high rate of hidden missing values and a clear class imbalance in the target variable.

- **Data Limitations**
  - Only 8 features and 768 samples of dataset, it's to small.
  - Including the unrepresentativeness of the study population (only Pima women)
  - Unreasonable zero values
  - The lack of important lifestyle variables

- **Research and Application Directions**
  - Building binary classification models

- **Final Conclusion**

This report has successfully illustrated the power of Exploratory Data Analytics in unraveling complex relationships in medical data. By connecting statistical findings with medical knowledge, we not only gain a deeper understanding of diabetes risk factors. This is an important step in the effort to apply data science to solve public health challenges.

THANK YOU
FOR LISTENING