In [3]:
```
import pandas as  pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
df=pd.read_csv(r"C:\Users\Admin\Downloads\diabetes.csv")
```

In [5]:
```
df.head()
```

Out[5]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeF
0	6	148	72	35	0	33.6	
1	1	85	66	29	0	26.6	
2	8	183	64	0	0	23.3	
3	1	89	66	23	94	28.1	
4	0	137	40	35	168	43.1	

Pregnancies: This refers to the number of times a person has been pregnant. It could be a factor in diabetes risk, especially for women who have had gestational diabetes.

Glucose: The plasma glucose concentration after a 2-hour oral glucose tolerance test. This test measures the body's ability to process glucose and can indicate the presence of diabetes or prediabetes.

BloodPressure: The diastolic blood pressure (in mm Hg), which measures the pressure in the arteries when the heart is resting between beats. High blood pressure is a risk factor for diabetes and other health conditions.

SkinThickness: The triceps skinfold thickness (in mm) is a measure of body fat. High body fat is a known risk factor for developing type 2 diabetes.

Insulin: This refers to the 2-hour serum insulin level (in micro-units per milliliter), which indicates how much insulin the body is producing after a glucose load. High insulin levels may suggest insulin resistance, a precursor to diabetes.

BMI (Body Mass Index): The body mass index, which is a measure of body fat based on height and weight. A higher BMI is associated with a greater risk of developing type 2 diabetes.

DiabetesPedigreeFunction: A function that takes into account family history and genetic factors to assess the likelihood of diabetes based on ancestry. A higher value indicates a greater genetic risk.

Age: The age of the individual in years. Age is a significant factor, as the risk of diabetes increases as people get older.

Outcome: The class variable indicating whether the individual has diabetes (1) or not (0). This is the target variable for prediction models.

In [6]:
```
df.shape
```

Out[6]:
```
(768, 9)
```

In [7]:
```
df.columns
```

Out[7]:
```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

In [8]:
```
df.isnull().sum()
```

Out[8]:
```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

In [9]:
```
df.nunique()
```

Out[9]:
```
Pregnancies                  17
Glucose                     136
BloodPressure                47
SkinThickness                51
Insulin                     186
BMI                         248
DiabetesPedigreeFunction    517
Age                          52
Outcome                       2
dtype: int64
```

In [10]:
```
df.describe()
```

Out[10]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000

In [11]:
```
df['Outcome'].value_counts()
```

Out[11]:

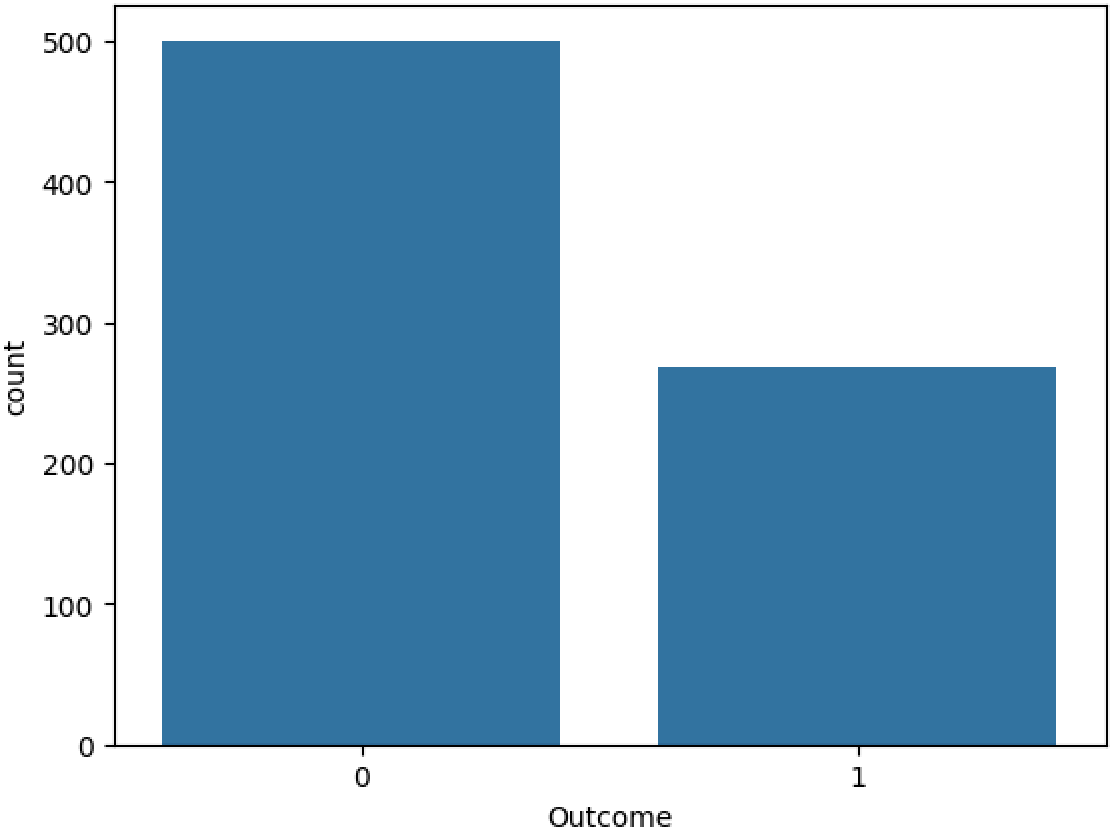
```
Outcome
0    500
1    268
Name: count, dtype: int64
```

In [12]:

```
sns.countplot(data=df,x='Outcome')
```

Out[12]:

```
<Axes: xlabel='Outcome', ylabel='count'>
```

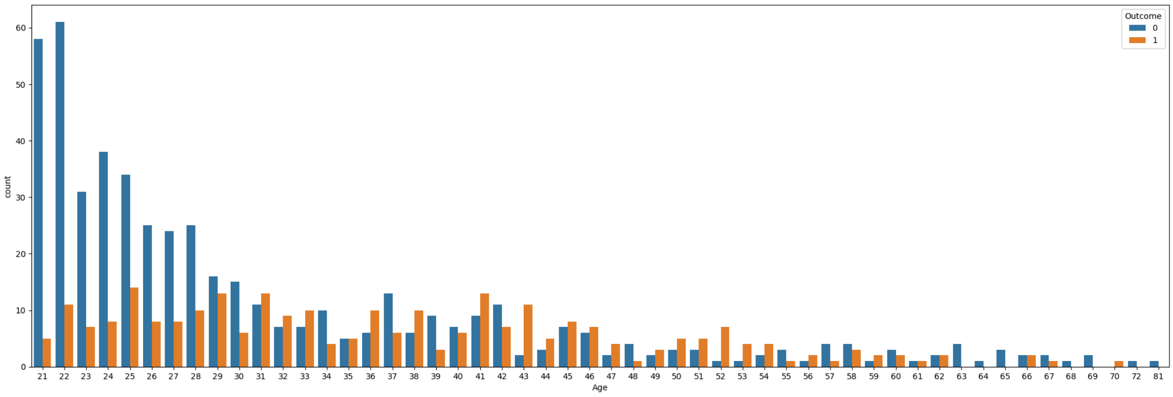


In [13]:

```
plt.subplots(figsize=(25,8))
sns.countplot(data=df,x='Age',hue='Outcome')
```

Out[13]:

```
<Axes: xlabel='Age', ylabel='count'>
```



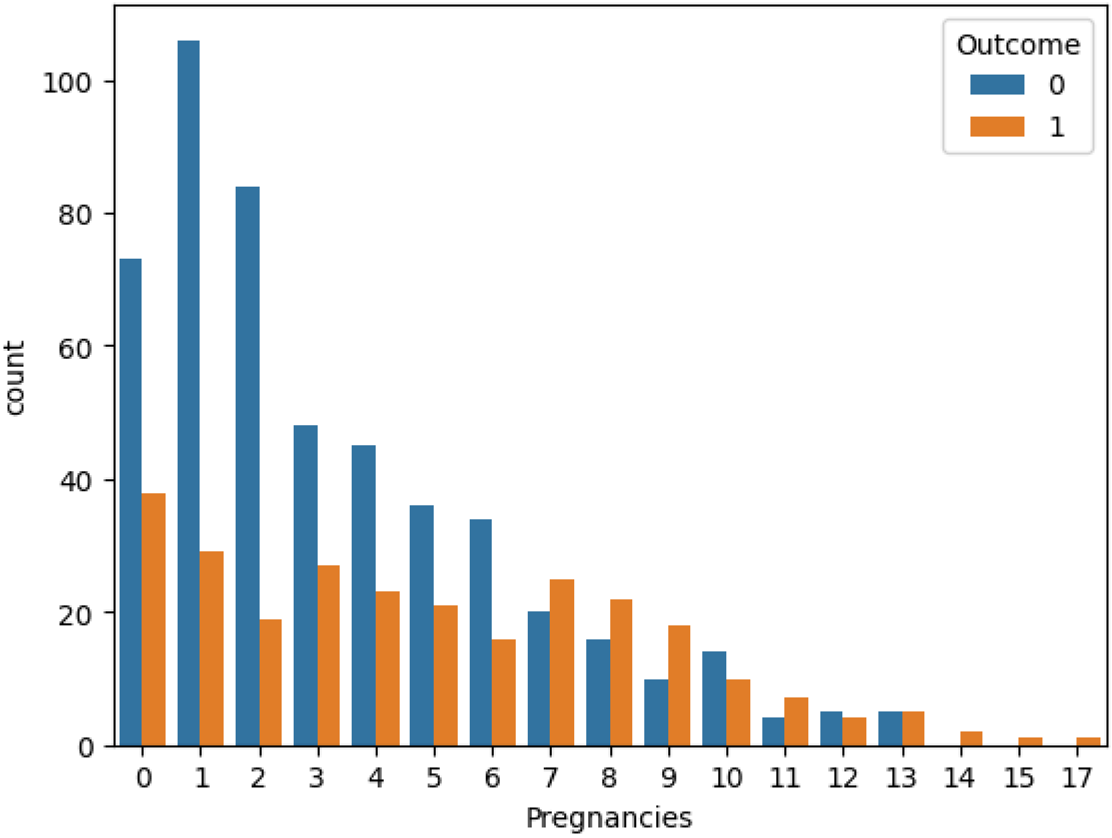
INSIGHTS: The chart suggests that the prevalence of diabetes might increase with age. The distribution for "Outcome 1" is shifted towards older ages, indicating a higher likelihood of diabetes in older individuals.

In [14]:

```
sns.countplot(x='Pregnancies',hue='Outcome',data=df)
```

Out[14]:

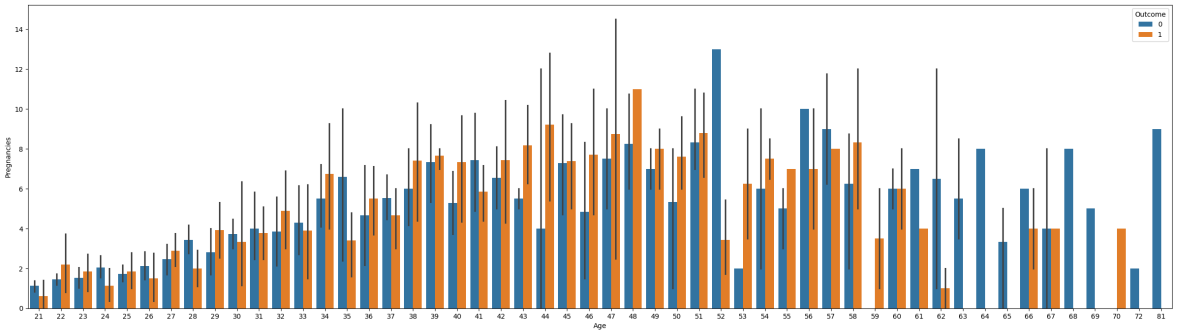
```
<Axes: xlabel='Pregnancies', ylabel='count'>
```



INSIGHTS: The chart suggests that the prevalence of diabetes might increase with the number of pregnancies. The distribution for "Outcome 1" is shifted towards individuals with more pregnancies, indicating a higher likelihood of diabetes in women with more pregnancies.

```
In [15]: plt.subplots(figsize=(30,8))
         sns.barplot(x='Age',y='Pregnancies',hue='Outcome',data=df,capsize=0)
```

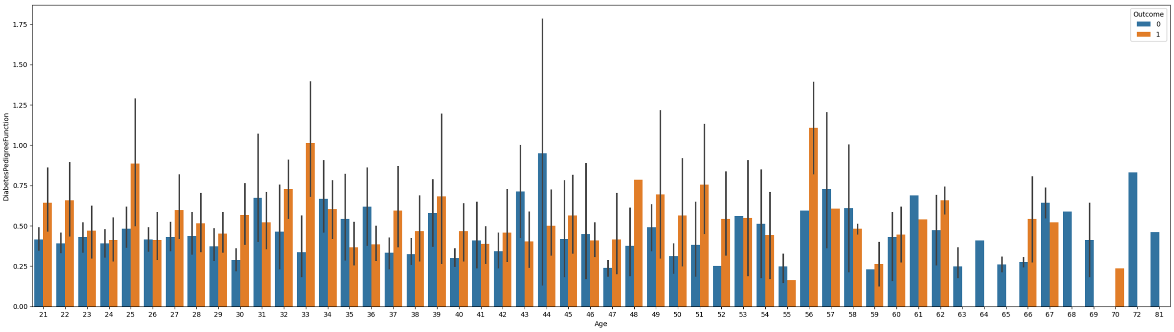
```
Out[15]: <Axes: xlabel='Age', ylabel='Pregnancies'>
```



INSIGHTS: The chart suggests a complex relationship between age, pregnancies, and diabetes risk. While the average number of pregnancies increases with age for both outcomes, the rate of increase appears to be higher for individuals with diabetes.

```
In [17]: plt.subplots(figsize=(30,8))
         sns.barplot(x='Age',y='DiabetesPedigreeFunction',hue='Outcome',data=df)
```

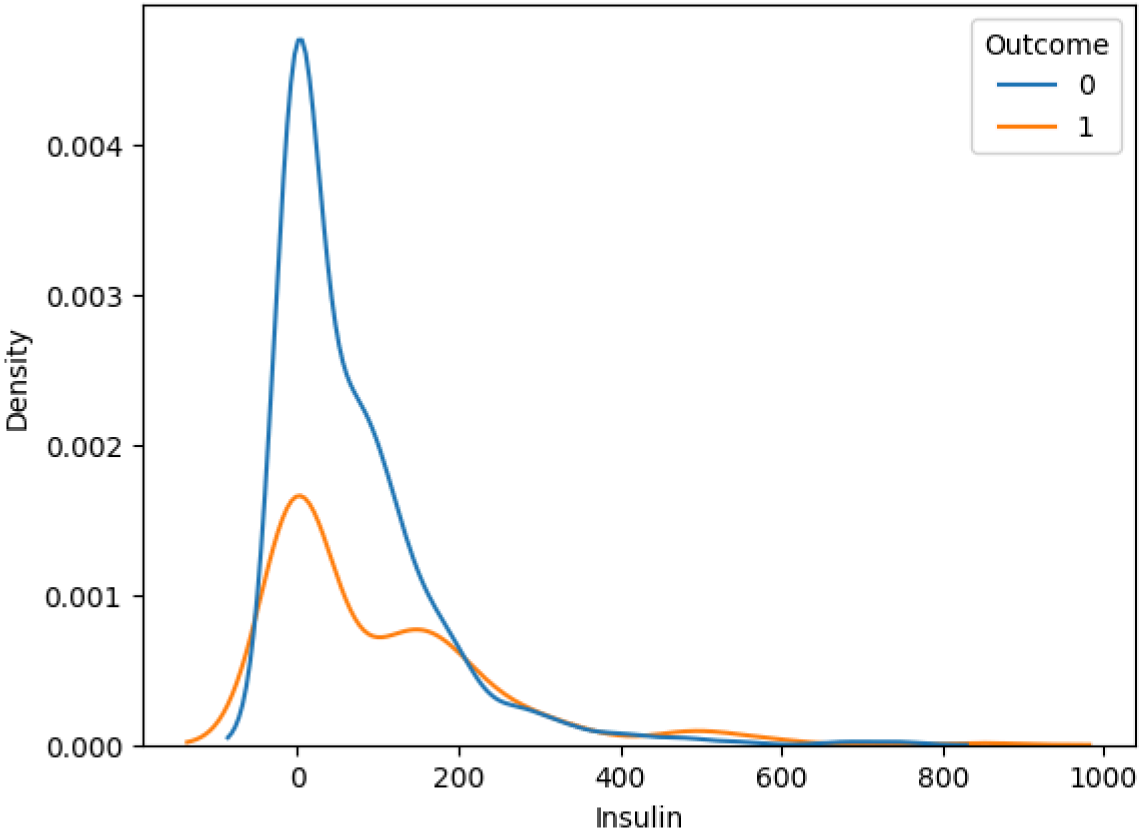
```
Out[17]: <Axes: xlabel='Age', ylabel='DiabetesPedigreeFunction'>
```



INSIGHTS: DPF generally increases with age: In both outcomes, there's a tendency for the DPF to be higher in older individuals. This suggests a potential correlation between age and DPF. Variability in DPF: The vertical lines extending from the top of each bar represent the variability in DPF within each age group. This variability seems to increase with age, suggesting that DPF becomes more diverse as people get older. Outcome 0 (likely no diabetes) has lower DPF overall: The blue bars, representing outcome 0, are generally lower than the orange bars (outcome 1), indicating that individuals without diabetes tend to have lower DPF values compared to those with diabetes. There are some individuals with high DPF values even at younger ages, and some older individuals with relatively low DPF values. These outliers could be due to genetic predisposition, lifestyle factors, or other underlying health conditions.

```
In [18]: sns.kdeplot(x='Insulin',hue='Outcome',data=df)
```

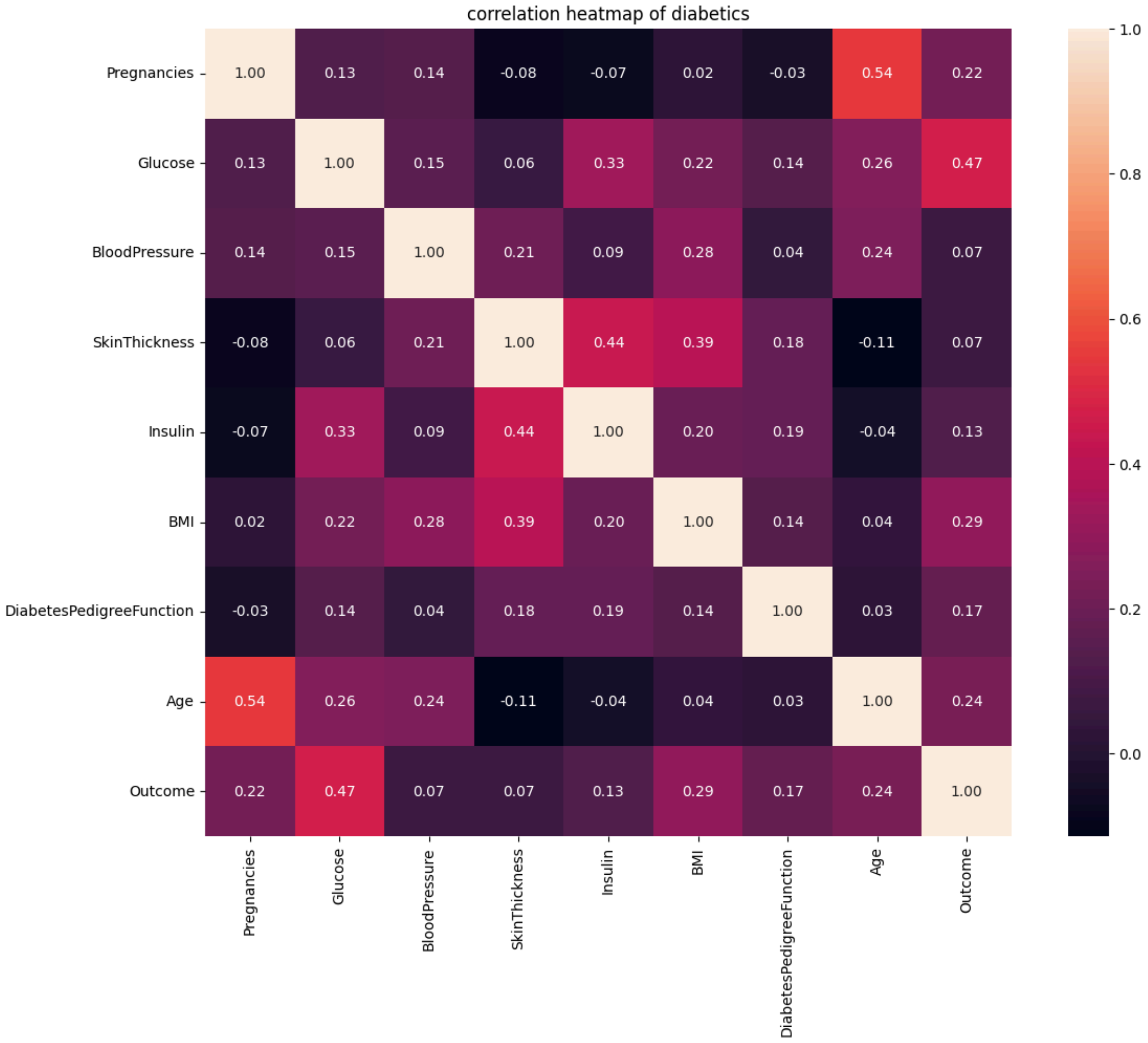
```
Out[18]: <Axes: xlabel='Insulin', ylabel='Density'>
```



INSIGHTS: Outcome 0 (likely no diabetes) has a lower insulin level distribution: The blue curve, representing outcome 0, is generally shifted to the left compared to the orange curve (outcome 1). This suggests that individuals without diabetes tend to have lower insulin levels compared to those with diabetes. Outcome 1 (likely diabetes) has a higher insulin level distribution: The orange curve, representing outcome 1, is shifted to the right, indicating that individuals with diabetes generally have higher insulin levels. There is some overlap between the two curves, meaning that some individuals with low insulin levels might still have diabetes, and some individuals with high insulin levels might not have diabetes. This indicates that insulin level alone might not be a definitive predictor of diabetes. The curves are not perfectly symmetrical. The distribution for outcome 0 seems to be more skewed to the right, while the distribution for outcome 1 appears to be more skewed to the left. This suggests that the variability in insulin levels is different for the two outcomes.

```
In [19]: correlation=df.corr()
In [20]: plt.subplots(figsize=(14,10))
         plt.title('correlation heatmap of diabetics')
         sns.heatmap(correlation,square=True,annot=True,fmt='.2f',linecolor='white')
```

```
Out[20]: <Axes: title={'center': 'correlation heatmap of diabetics'}>
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
SkinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00

OVERALL INTERPRETATION OF THE HEATMAP :

The heatmap visualizes the correlation coefficients between different features (Pregnancies, Glucose, Blood Pressure, etc.) and the outcome (whether a person has diabetes or not). Correlation coefficients range from -1 to 1, where:

1: Perfect positive correlation (as one variable increases, the other increases proportionally) 0: No correlation -1: Perfect negative correlation (as one variable increases, the other decreases proportionally) Specific Observations:

Glucose: Shows a strong positive correlation with the outcome (0.47). This suggests that higher glucose levels are significantly associated with an increased risk of diabetes. BMI: Has a moderate positive correlation with the outcome (0.29). This indicates that individuals with higher BMIs are more likely to develop diabetes. Age: Shows a moderate positive correlation with the outcome (0.24). This suggests that the risk of diabetes increases with age. Pregnancies: Has a weak positive correlation with the outcome (0.22). This implies that having more pregnancies might slightly increase the risk of diabetes.

Diabetes Pedigree Function: Has a weak positive correlation with the outcome (0.17). This suggests that a family history of diabetes might slightly increase the risk. Insulin: Shows a weak positive correlation with the outcome (0.13). This indicates that higher insulin levels might be associated with a slightly increased risk of diabetes. Blood Pressure: Shows a very weak positive correlation with the outcome (0.07). This suggests that blood pressure might have a minimal impact on diabetes risk. Skin Thickness: Shows a very weak positive correlation with the outcome (0.07). This indicates that skin thickness might have a minimal impact on diabetes risk. Inferences:

Glucose and BMI: These factors seem to be the most significant predictors of diabetes risk based on their strong positive correlations. Age, Pregnancies, and Diabetes Pedigree Function: These factors also contribute to the risk but to a lesser extent. Insulin, Blood Pressure, and Skin Thickness: These factors appear to have a minimal impact on diabetes risk based on this analysis.

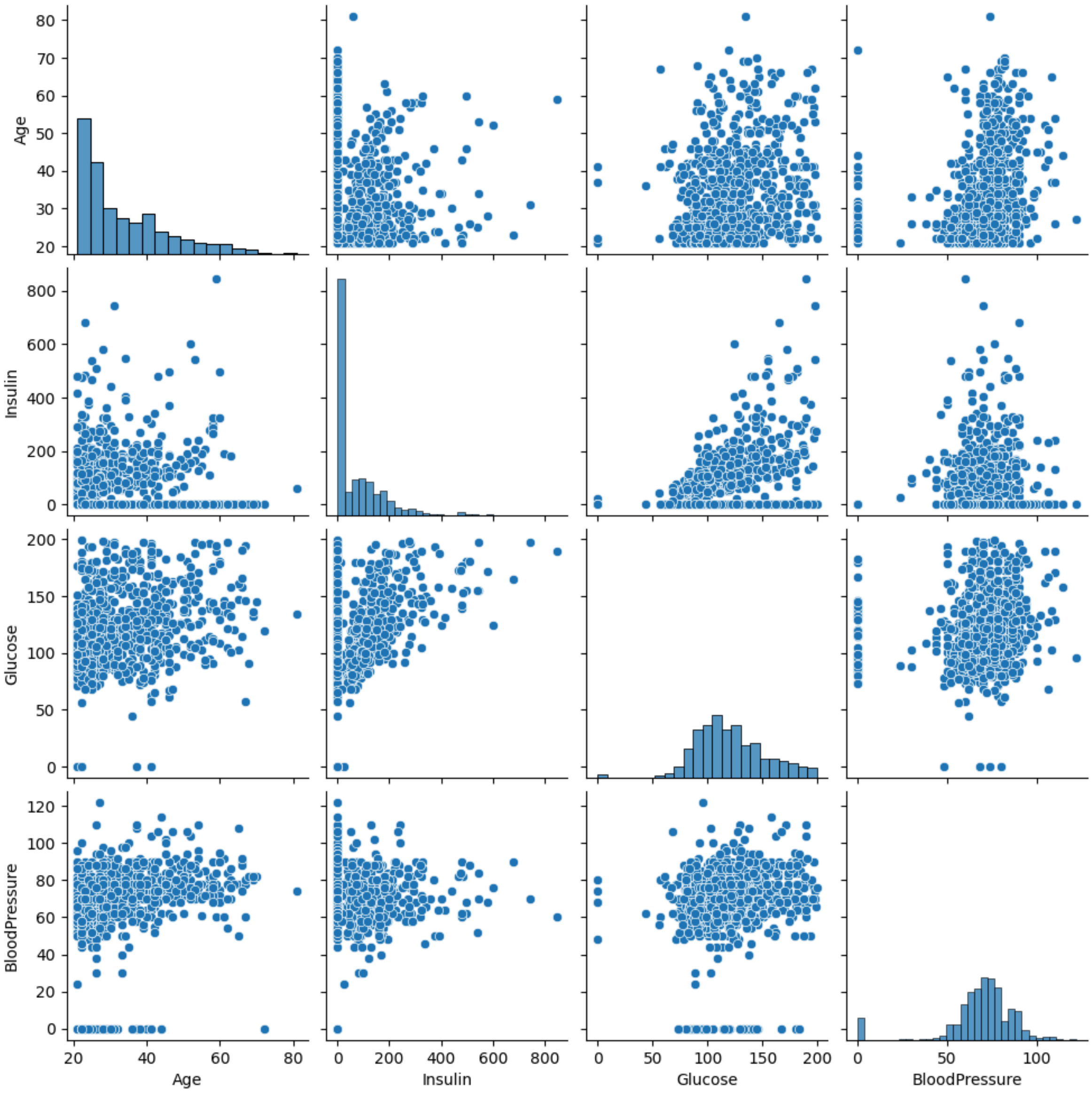
This heatmap only shows correlations, not causation. It doesn't necessarily mean that higher glucose levels directly cause diabetes. Other factors not included in this analysis might also influence diabetes risk.

In [21]:

```
mul_var=['Age','Insulin','Glucose', 'BloodPressure']
sns.pairplot(df[mul_var],kind='scatter',diag_kind='hist')
```

Out[21]:

```
<seaborn.axisgrid.PairGrid at 0x2857498bb30>
```



INSIGHTS OF THE PAIRPLOTS:

Age vs. Insulin: There seems to be a slight positive correlation between age and insulin levels. As age increases, insulin levels tend to increase slightly. Age vs. Glucose: There appears to be a weak positive correlation between age and glucose levels. As age increases, glucose levels tend to increase slightly. Age vs. Blood Pressure: There appears to be a weak positive correlation between age and blood pressure. As age increases, blood pressure tends to increase slightly. Insulin vs. Glucose: There appears to be a moderate positive correlation between insulin and glucose levels. As insulin levels increase, glucose levels tend to increase. Insulin vs. Blood Pressure: There appears to be a weak positive correlation between insulin and blood pressure. As insulin levels increase, blood pressure tends to increase slightly. Glucose vs. Blood Pressure: There appears to be a weak positive correlation between glucose and blood pressure. As glucose levels increase, blood pressure tends to increase slightly.