Diabetes dataset analysis

In [6]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [70]:
```
df=pd.read_csv(r"C:\Users\Jaswanth Reddy\Downloads\diabetes_data_upload.csv")
df.head()
```

Out[70]:

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	A
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	

In [66]:
```
df.isnull().sum()
```

Out[66]:
```
Age                 0
Gender              0
Polyuria            0
Polydipsia          0
sudden weight loss  0
weakness            0
Polyphagia          0
Genital thrush      0
visual blurring     0
Itching             0
Irritability        0
delayed healing     0
partial paresis     0
muscle stiffness    0
Alopecia            0
Obesity             0
class               0
dtype: int64
```

From the above cell it is clear that there are no missing values in the dataset

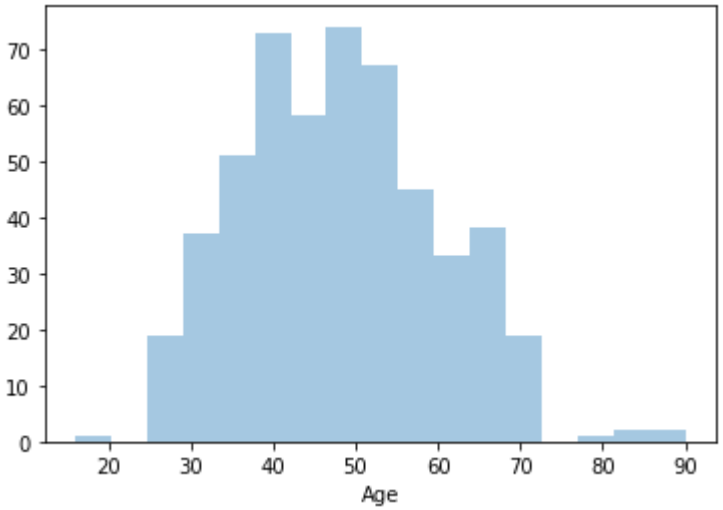
In [71]:
```
df.describe()
```

Out[71]:

	Age
count	520.000000
mean	48.028846
std	12.151466
min	16.000000
25%	39.000000
50%	47.500000
75%	57.000000
max	90.000000

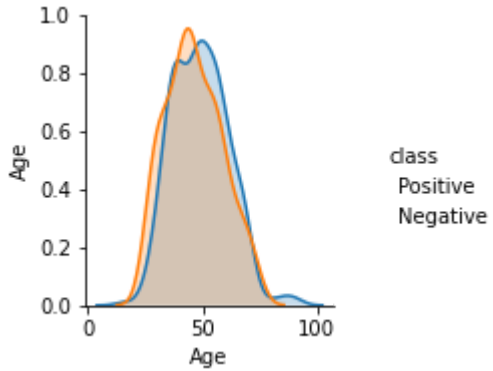
In [67]:
```
a=['Age']
for i in a:
    print(df[i].describe())
    sns.distplot(df[i],kde=False)
    plt.show()
```

```
count    520.000000
mean      48.028846
std       12.151466
min       16.000000
25%       39.000000
50%       47.500000
75%       57.000000
max       90.000000
Name: Age, dtype: float64
```



In [12]:
```
sns.pairplot(df,hue='class')
```

Out[12]:
```
<seaborn.axisgrid.PairGrid at 0x2aafe404e50>
```



In [31]:
```
a=df.groupby('class')['Age'].value_counts()
a.head(50)
```

Out[31]:

class	Age	
Negative	30	18
	43	15
	47	11
	53	11
	40	10
	45	8
	38	7
	46	7
	54	7
	58	7
	72	7
	27	6
	42	6
	57	6
	34	5
	36	5
	44	5
	50	5
	60	5
	67	5
	28	4
	37	4
	56	4
	64	4
	35	3
	51	3
	59	3
	65	3
	32	2
	33	2
	55	2
	70	2
	26	1
	29	1
	39	1
	49	1
	52	1
	61	1

```
          66      1
          68      1
Positive  48     28
          35     27
          55     20
          39     15
          40     14
          38     13
          50     13
          58     11
          43     10
          45     10
Name: Age, dtype: int64
```

Conclusion:- Age group around 35 to 60 are a high chances of getting diabetes

In [32]:
```
a=df.groupby('class')['Polyuria'].value_counts()
a.head()
```

Out[32]:
```
class     Polyuria
Negative  No          185
          Yes          15
Positive  Yes         243
          No           77
Name: Polyuria, dtype: int64
```

conclusion:- From the above cell we can conclude that the majority of patients who suffer from polyuria may also suffer from diabetes and some patients even without polyuria also suffer from diabetes are more in number

In [33]:
```
a=df.groupby('class')['Polydipsia'].value_counts()
a.head()
```

Out[33]:
```
class     Polydipsia
Negative  No            192
          Yes             8
Positive  Yes           225
          No             95
Name: Polydipsia, dtype: int64
```

conclusion:- From the above cell we can conclude that the majority of patients who suffer from Polydipsia may also suffer from diabetes and some patients even without Polydipsia also suffer from diabetes are more in number

In [34]:
```
a=df.groupby('class')['sudden weight loss'].value_counts()
a.head()
```

Out[34]:
```
class     sudden weight loss
Negative  No                   171
          Yes                   29
Positive  Yes                  188
          No                   132
Name: sudden weight loss, dtype: int64
```

conclusion:- The patients weight loss depends on diabets condition

In [35]:
```
a=df.groupby('class')['weakness'].value_counts()
a.head()
```

Out[35]:
```
class     weakness
Negative  No          113
          Yes          87
Positive  Yes         218
          No          102
Name: weakness, dtype: int64
```

conclusion:- The patient who are diabetes are few chances of feeling body weakness

In [36]:
```
a=df.groupby('class')['Polyphagia'].value_counts()
a.head()
```

Out[36]:
```
class     Polyphagia
Negative  No           152
          Yes           48
Positive  Yes          189
          No           131
Name: Polyphagia, dtype: int64
```

conclusion:- Polyphagia effects much

In [77]:
```
a=df.groupby('class')['Genital thrush'].value_counts()
a.head()
```

Out[77]:
```
class     Genital thrush
Negative  No               167
          Yes               33
Positive  No               237
          Yes               83
Name: Genital thrush, dtype: int64
```

conclusion:- Genital thrush also doesn't effect much

In [68]:
```
df['class'=pd.get_dummies(df['class'])
df.head()
```

Out[68]:

	Negative	Positive
0	0	1
1	0	1
2	0	1
3	0	1
4	0	1

In [79]:
```
a=df.groupby('class')['visual blurring'].value_counts()
a.head()
```

Out[79]:
```
class     visual blurring
Negative  No                142
          Yes                58
Positive  Yes               175
          No                145
Name: visual blurring, dtype: int64
```

conclusion:- There is a 80 percent of chance of getting visual blurring

In [88]:
```
a=df.groupby('class')['Itching'].value_counts()
a.head()
```

Out[88]:
```
class     Itching
Negative  No         101
          Yes         99
Positive  No         166
          Yes        154
Name: Itching, dtype: int64
```

conclusion:- 60percent chance of getting itches during diabetes

In [90]:
```
a=df.groupby('class')['Irritability'].value_counts()
a.head()
```

Out[90]:
```
class     Irritability
Negative  No              184
          Yes              16
Positive  No              210
          Yes             110
Name: Irritability, dtype: int64
```

conclusion:- diabetes patients suffer itching a lot

In [92]:
```
a=df.groupby('class')['delayed healing'].value_counts()
a.head()
```

Out[92]:
```
class     delayed healing
Negative  No                114
          Yes                86
Positive  No                167
          Yes               153
Name: delayed healing, dtype: int64
```

conclusion:- Delayed in healing during diabetes

In [94]:
```
a=df.groupby('class')['partial paresis'].value_counts()
a.head()
```

Out[94]:
```
class     partial paresis
Negative  No                168
          Yes                32
Positive  Yes               192
          No                128
Name: partial paresis, dtype: int64
```

conclusion:- There is a high chance of partial paresis

In [96]:
```
a=df.groupby('class')['muscle stiffness'].value_counts()
a.head()
```

Out[96]:
```
class     muscle stiffness
Negative  No                 140
          Yes                 60
Positive  No                 185
          Yes                135
Name: muscle stiffness, dtype: int64
```

conclusion:- Decrease in muscle stifness happens during diabetes

In [98]:
```
a=df.groupby('class')['Obesity'].value_counts()
a.head()
```

Out[98]:
```
class     Obesity
Negative  No         173
          Yes         27
Positive  No         259
          Yes         61
Name: Obesity, dtype: int64
```

conclusion:- Obseity doesn't matters much

From all above conclusion i can conclude that age group around 35 to 60 are a high chance of diabetes.

some diseases like polyuria,polydipsia also effects in diabetes positive and they suffer from weight loss,visual blurring,itches, delayed in healing,partial paresis,decrease in muscle strength(common symptoms what they feel during diabetes)

In [ ]:
```

```