Aprendizado de Máquina - 2023.02

Escolher um dataset do kaggle e avaliar os atributos usando medidas de posição e dispersão gerando histogramas e boxplot

Aluno: Diego Vasconcelos Schardosim de Matos

DRE: 120098723

Dataset escolhido

Selecionei o dataset Heart Disease Classification Dataset pois se encaixa bem aos requesitos do trabalho e também me chamou atenção pois um de seus usos seria aplicações voltadas a saúde pública

In [28]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
```

In [29]:
```
df = pd.read_csv("Heart Attack.csv")
```

In [30]:
```
df.head()
```

Out[30]:

	age	gender	impluse	pressurehight	pressurelow	glucose	kcm	troponin	class
0	64	1	66	160	83	160.0	1.80	0.012	negative
1	21	1	94	98	46	296.0	6.75	1.060	positive
2	55	1	64	160	77	270.0	1.99	0.003	negative
3	64	1	70	120	55	270.0	13.87	0.122	positive
4	55	1	64	112	65	300.0	1.08	0.003	negative

In [31]:
```
df.shape
```

Out[31]:
```
(1319, 9)
```

In [32]:
```
df.duplicated().sum()
```

Out[32]:
```
0
```

In [33]:
```
df.isnull().any()
```

Out[33]:
```
age             False
gender          False
impluse         False
pressurehight   False
pressurelow     False
glucose         False
kcm             False
troponin        False
class           False
dtype: bool
```

Este é um bom dataset pois além de possuir colunas bem explicativas categorizadas em 2 classes distintas, não possui instâncias duplicadas ou nulas.

In [34]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1319 entries, 0 to 1318
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   age            1319 non-null   int64  
 1   gender         1319 non-null   int64  
 2   impluse        1319 non-null   int64  
 3   pressurehight  1319 non-null   int64  
 4   pressurelow    1319 non-null   int64  
 5   glucose        1319 non-null   float64
 6   kcm            1319 non-null   float64
 7   troponin       1319 non-null   float64
 8   class          1319 non-null   object 
dtypes: float64(3), int64(5), object(1)
memory usage: 92.9+ KB
```

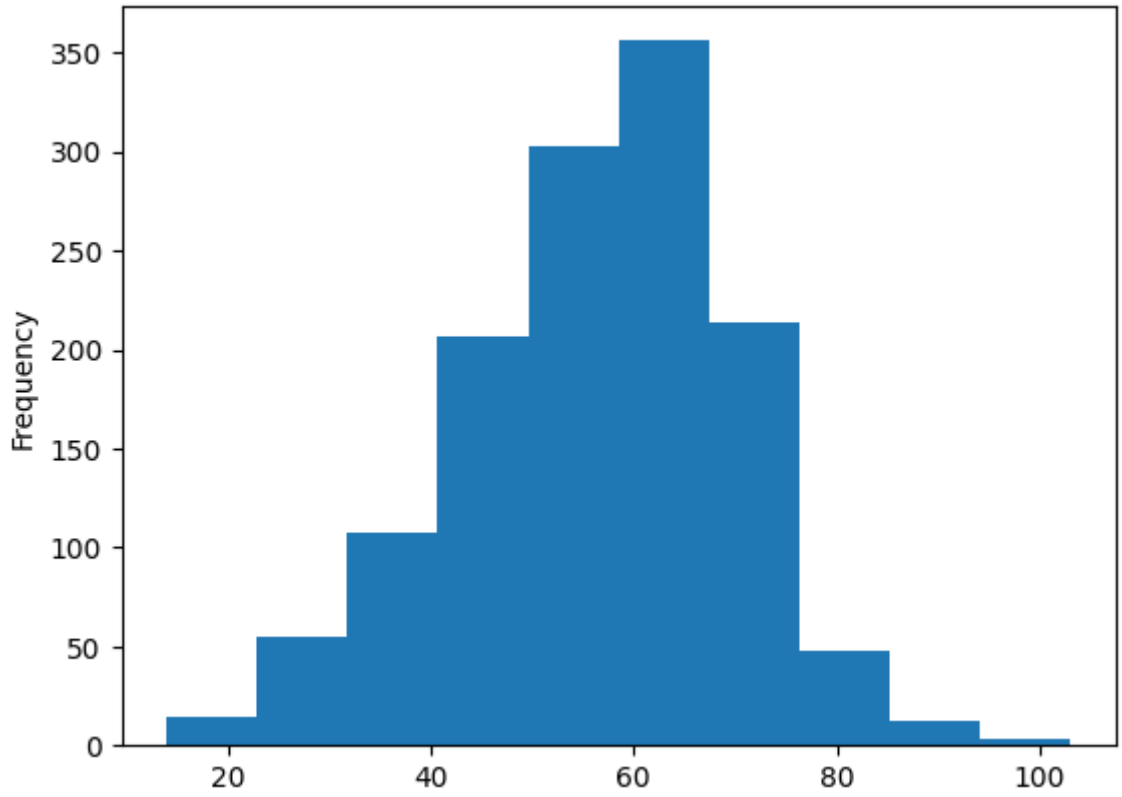
In [35]:
```
df.describe()
```

Out[35]:

	age	gender	impluse	pressurehight	pressurelow	glucose	
count	1319.000000	1319.000000	1319.000000	1319.000000	1319.000000	1319.000000	1319.00
mean	56.191812	0.659591	78.336619	127.170584	72.269143	146.634344	15.27
std	13.647315	0.474027	51.630270	26.122720	14.033924	74.923045	46.32
min	14.000000	0.000000	20.000000	42.000000	38.000000	35.000000	0.32
25%	47.000000	0.000000	64.000000	110.000000	62.000000	98.000000	1.65
50%	58.000000	1.000000	74.000000	124.000000	72.000000	116.000000	2.85
75%	65.000000	1.000000	85.000000	143.000000	81.000000	169.500000	5.80
max	103.000000	1.000000	1111.000000	223.000000	154.000000	541.000000	300.00

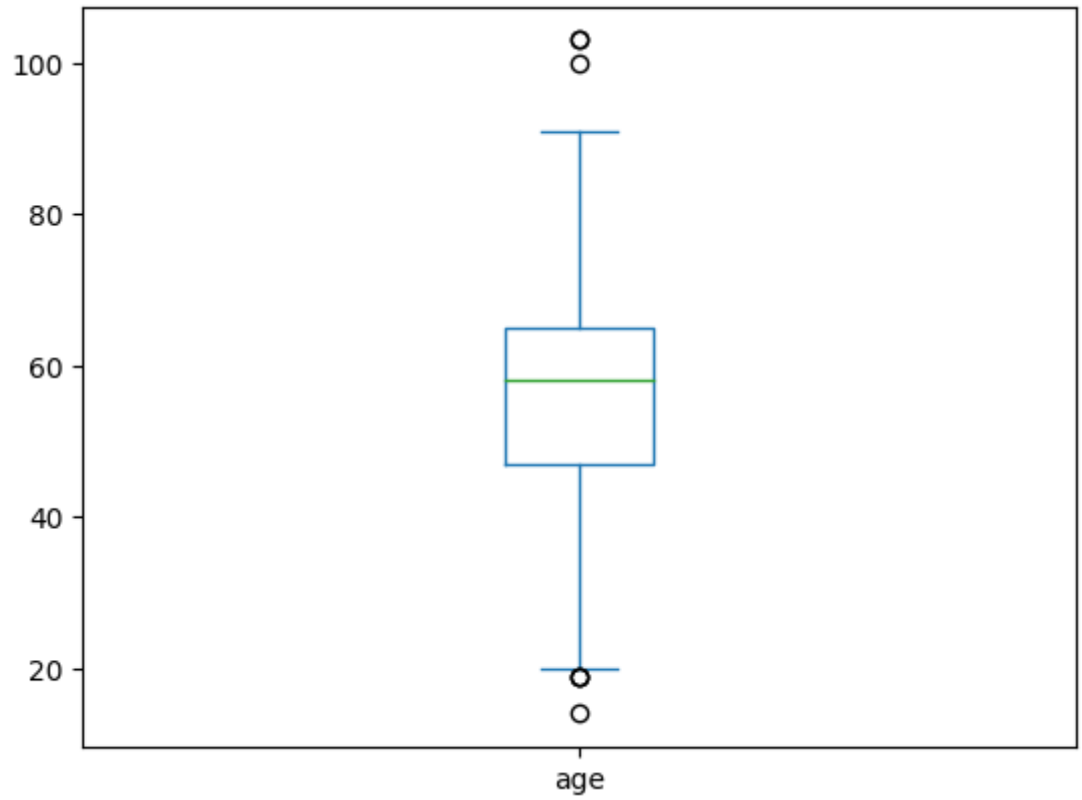
In [36]:
```
df['age'].plot.hist()
```

Out[36]:
```
<Axes: ylabel='Frequency'>
```



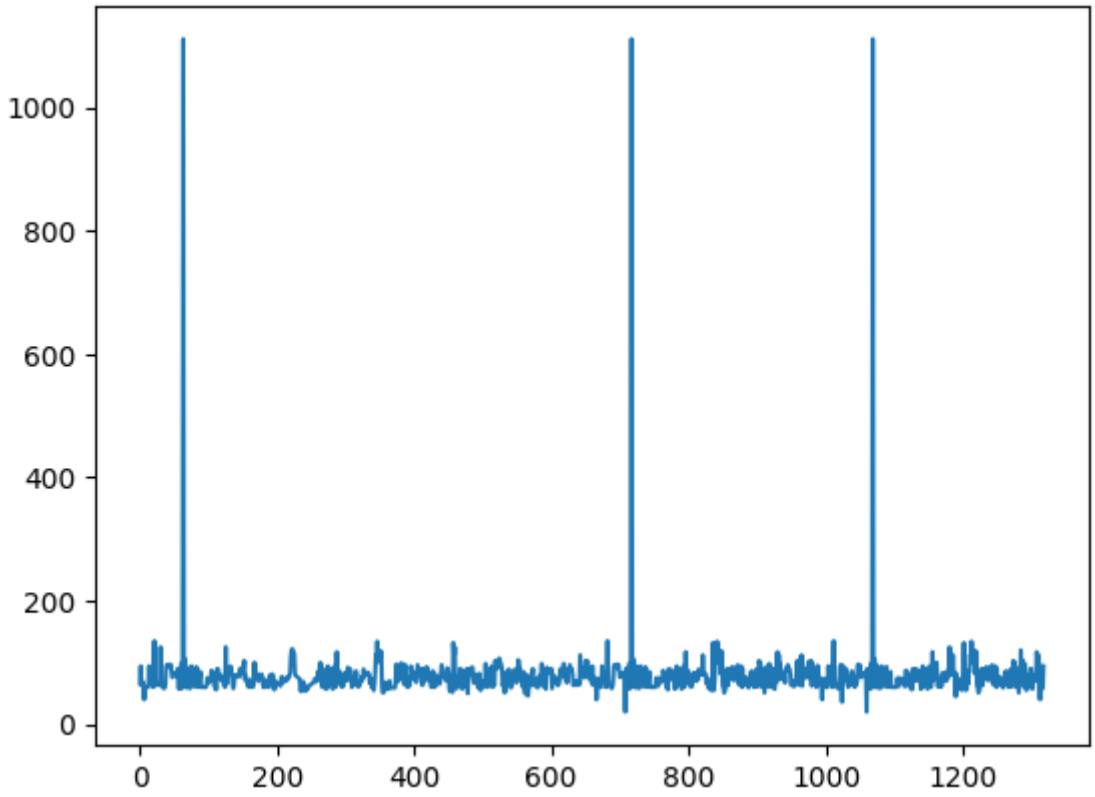
In [37]:
```
df['age'].plot.box()
```

Out[37]: <Axes: >



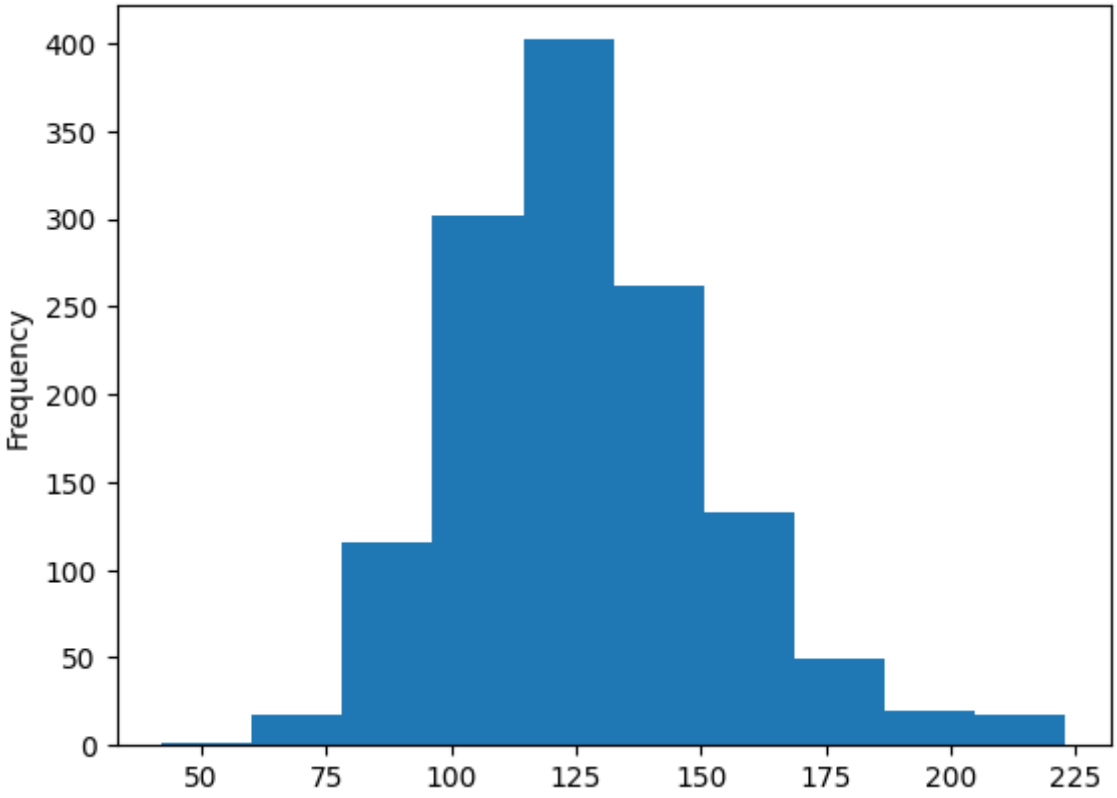
In [38]:
```
df['impluse'].plot()
```

Out[38]: <Axes: >



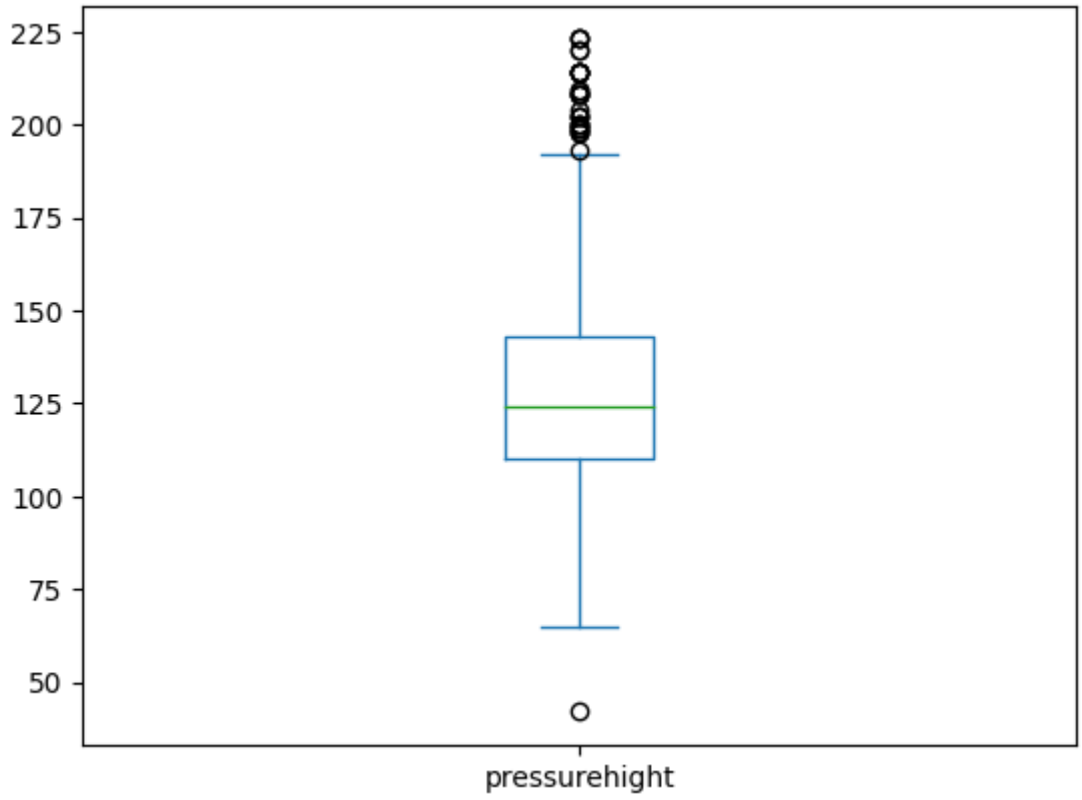
In [39]:
```
df['pressurehight'].plot.hist()
```

Out[39]: <Axes: ylabel='Frequency'>



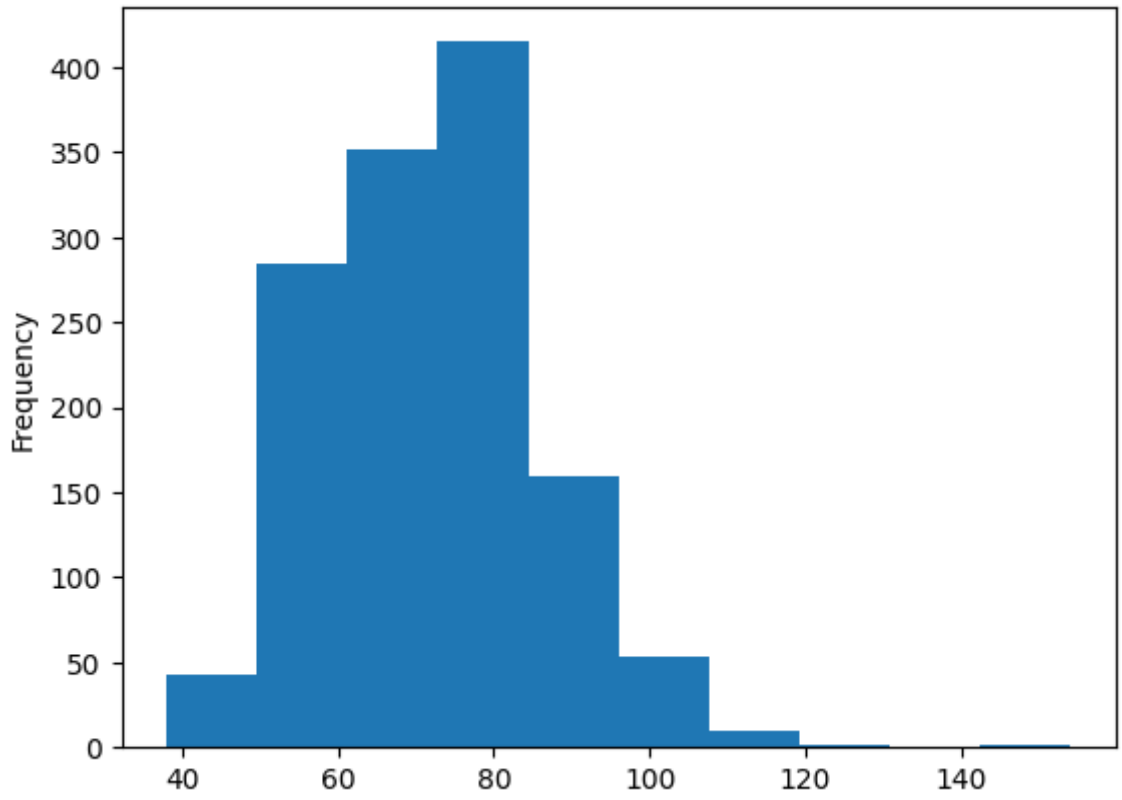
In [40]:
```
df['pressurehight'].plot.box()
```

Out[40]: <Axes: >



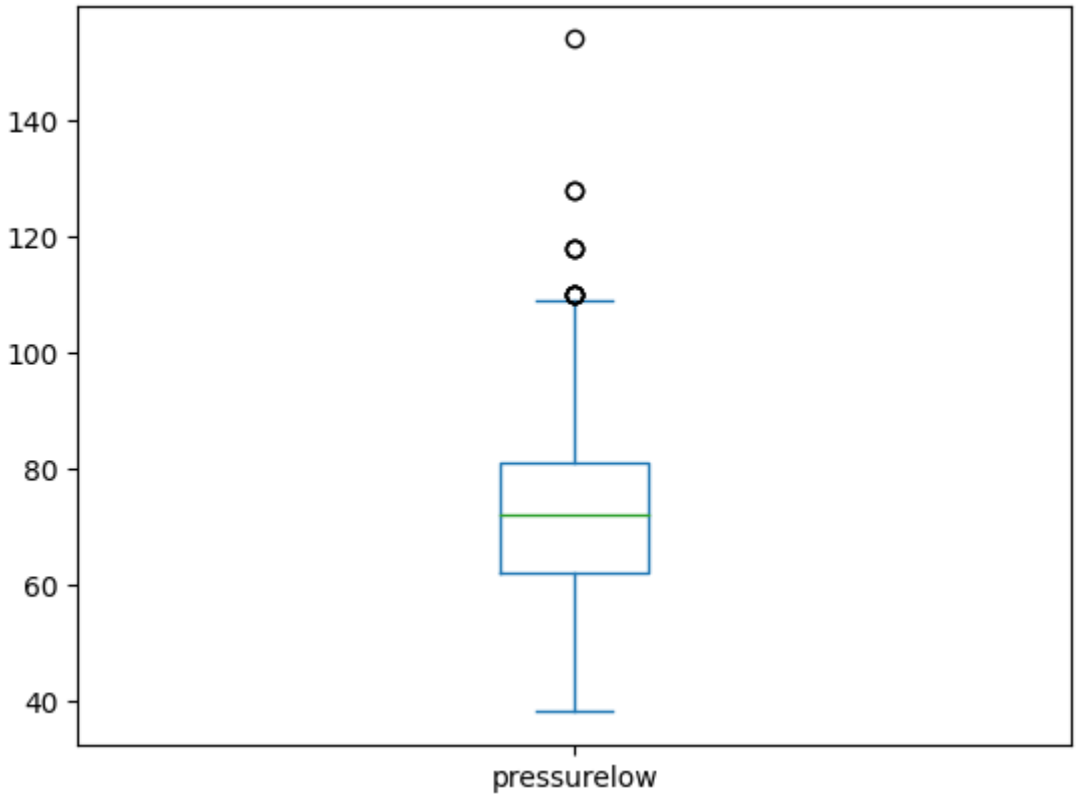
In [41]:
```
df['pressurelow'].plot.hist()
```

Out[41]: <Axes: ylabel='Frequency'>



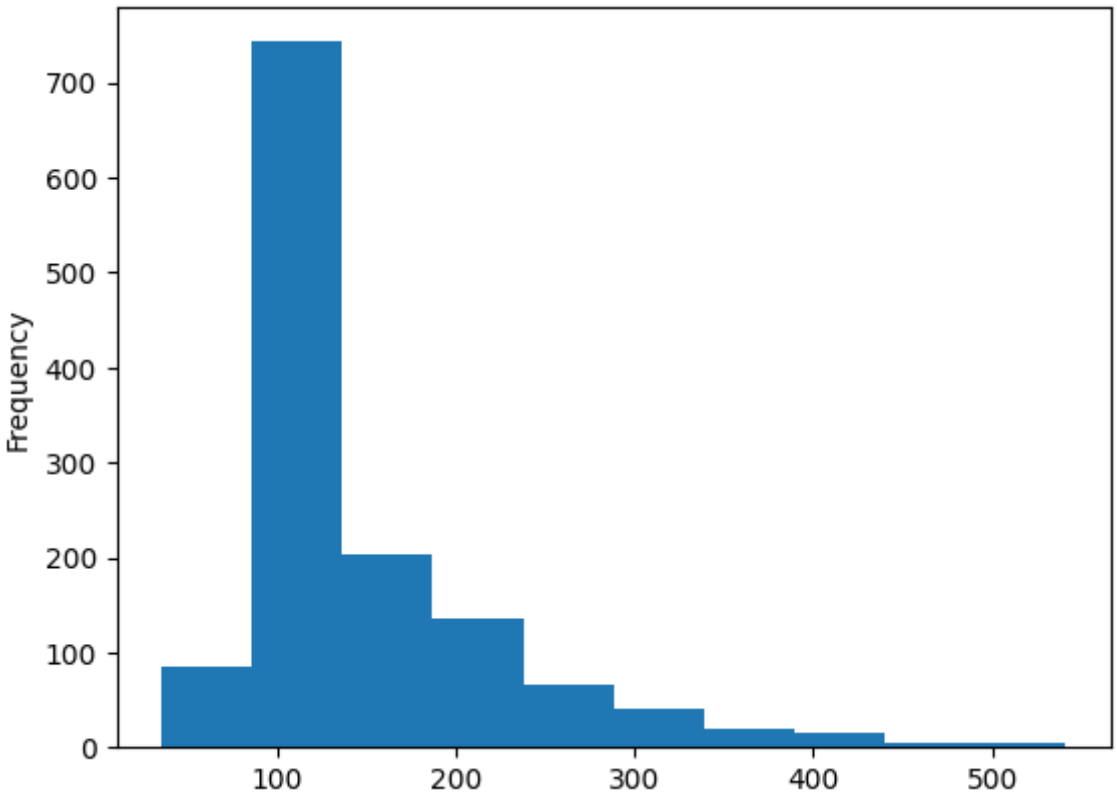
In [42]:
```
df['pressurelow'].plot.box()
```

Out[42]: <Axes: >



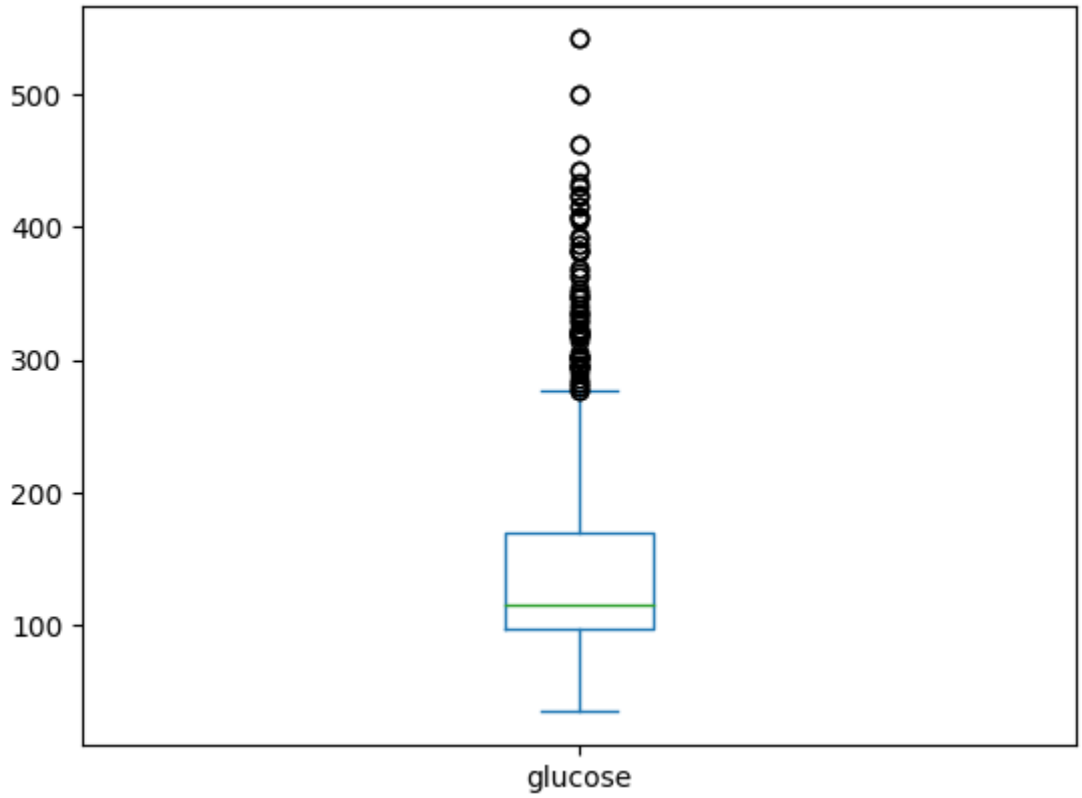
In [43]:
```
df['glucose'].plot.hist()
```

Out[43]: <Axes: ylabel='Frequency'>



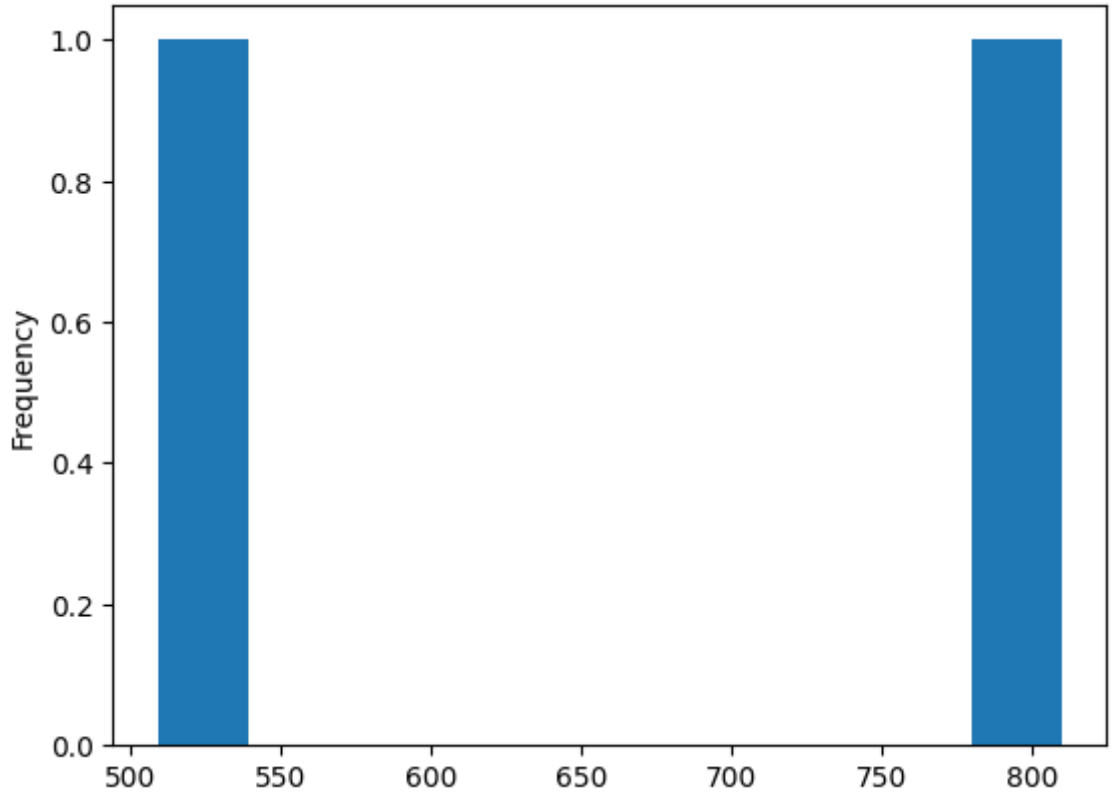
In [44]:
```
df['glucose'].plot.box()
```

Out[44]: <Axes: >



In [62]:
```
df['class'].value_counts().plot.hist()
```

Out[62]:
```
<Axes: ylabel='Frequency'>
```



Conclusão

Existem bastante outliers nos campos de glucose e pressão sanguínea sendo necessário testar com diferentes modelos para saber o impacto em remover essas instancias. Mas

no geral deve ser possível gerar um bom modelo classificador.

OBS: Não conseguir "girar" o último gráfico, mas a ideia seria termos uma noção da distribuição das classes.