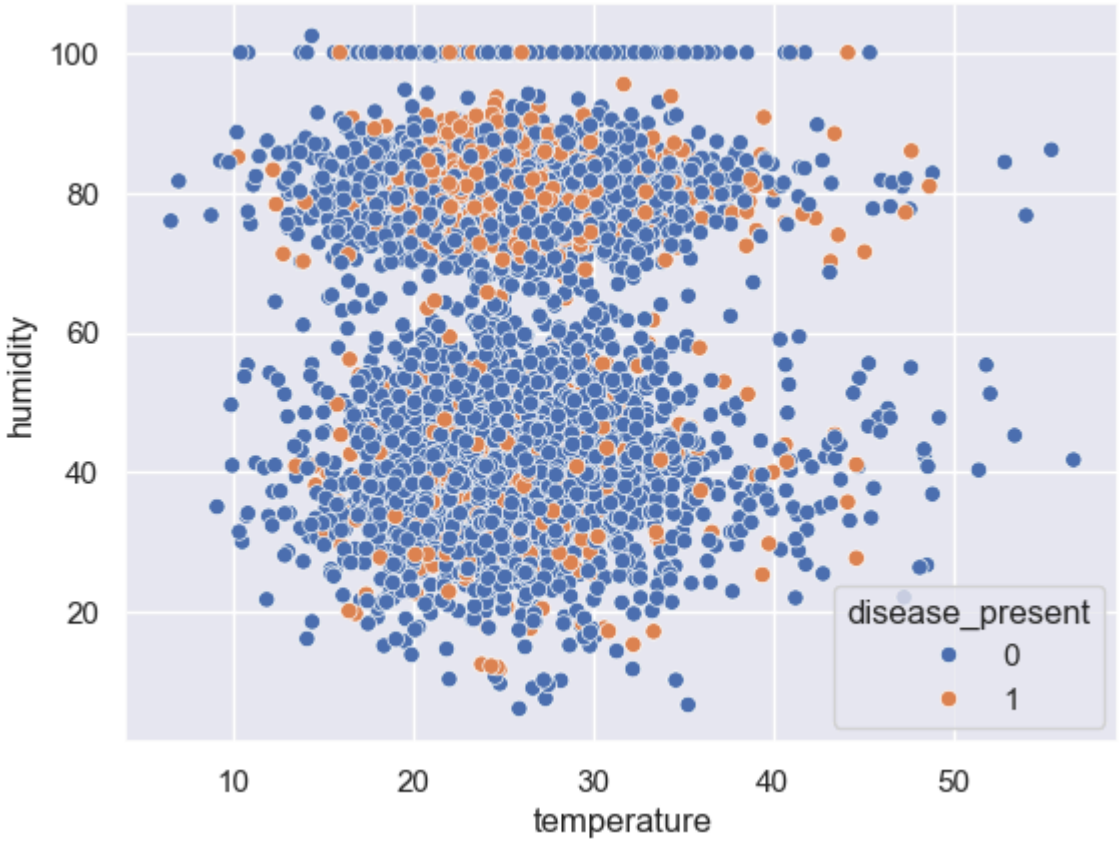
this data is pretty simple, and even humans can read it.
so we can use simple models like decision tree classifiers to analise and predict them

```
<Axes: xlabel='temperature', ylabel='humidity'>
```



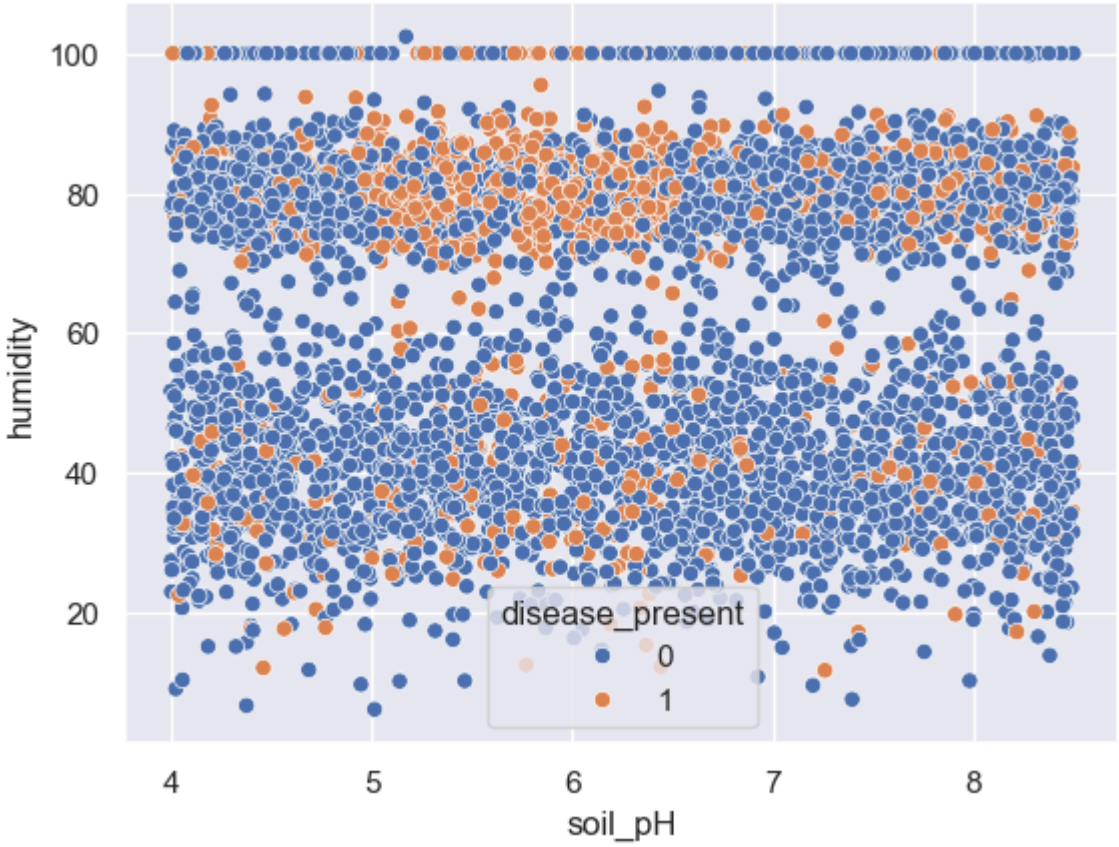
Humidity and Disease Presence

Humidity plays a significant role in the presence of plant disease.
The scatter plot reveals **two distinct clusters** based on humidity:

- **low humidity (~40%)**
- and **high humidity (~80%)**

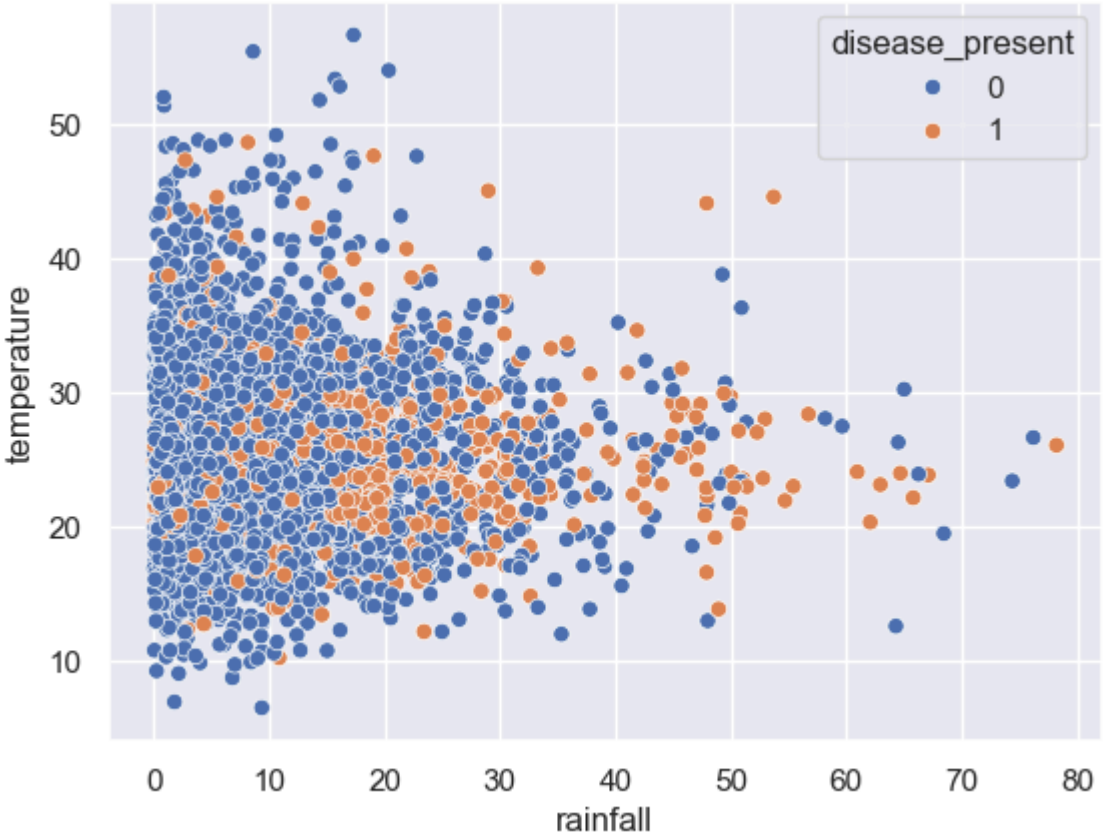
The **high humidity cluster** shows a much higher rate of disease presence.
This might suggests that **elevated humidity** creates favorable conditions for disease

```
<Axes: xlabel='soil_pH', ylabel='humidity'>
```



- the area with ph level between **5 and 6.5** in the high humidity cluster
- has a pretty high chance of disease presence

```
<Axes: xlabel='rainfall', ylabel='temperature'>
```



rainfall kinda behaves similarly to humidity,
sample size is small, but anything over **20mm** has high **disease** risk
it's a good idea to get more data on that range.

well i think it's time to write some models, we'll train a decision tree classifier on the data

```
      temperature   humidity   rainfall   soil_pH  disease_present
6252    38.189409  81.236450   1.821797  5.189183                0
4684    19.814670  32.527782   8.650504  5.983636                1
1731    23.253416  64.457696   8.218361  4.022305                0
4742    25.008095  44.829299   2.177695  8.153636                0
4521    30.371665  41.536491  14.861256  7.288489                0
Accuracy: 0.77
```

well i was wrong, this data seems to be a little tougher than i expected we're goonna need bigger guns, a **random forest** model might do the trick

```
Accuracy with Random Forest: 0.86
              precision    recall  f1-score   support

           0       0.87      0.96      0.91      1147
           1       0.79      0.54      0.64       353

    accuracy                           0.86      1500
   macro avg       0.83      0.75      0.78      1500
weighted avg       0.85      0.86      0.85      1500
```

way better results, but the model has high false negative rates,