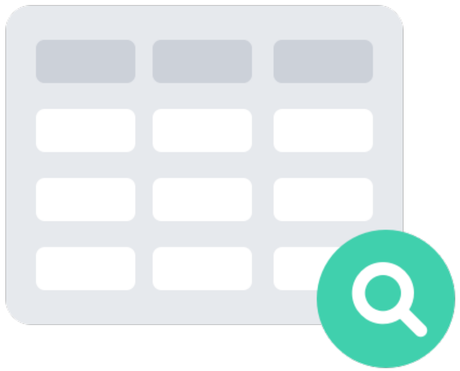
What is structured data?

Data

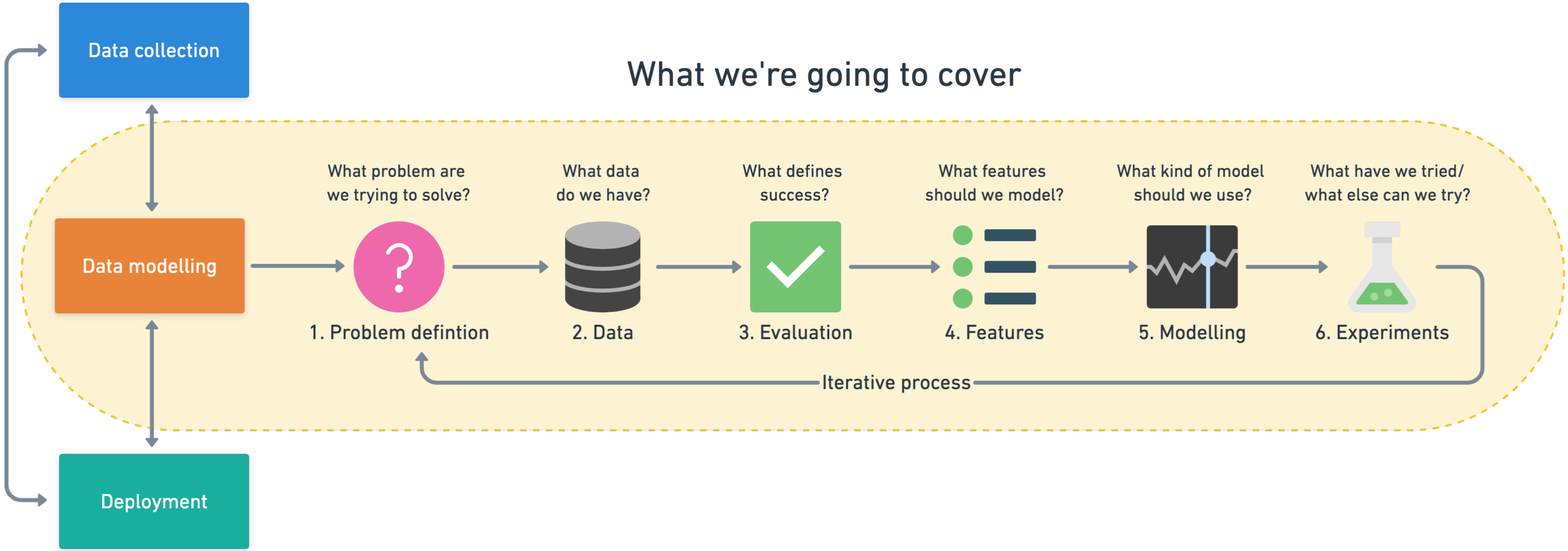


Feature variables | Target

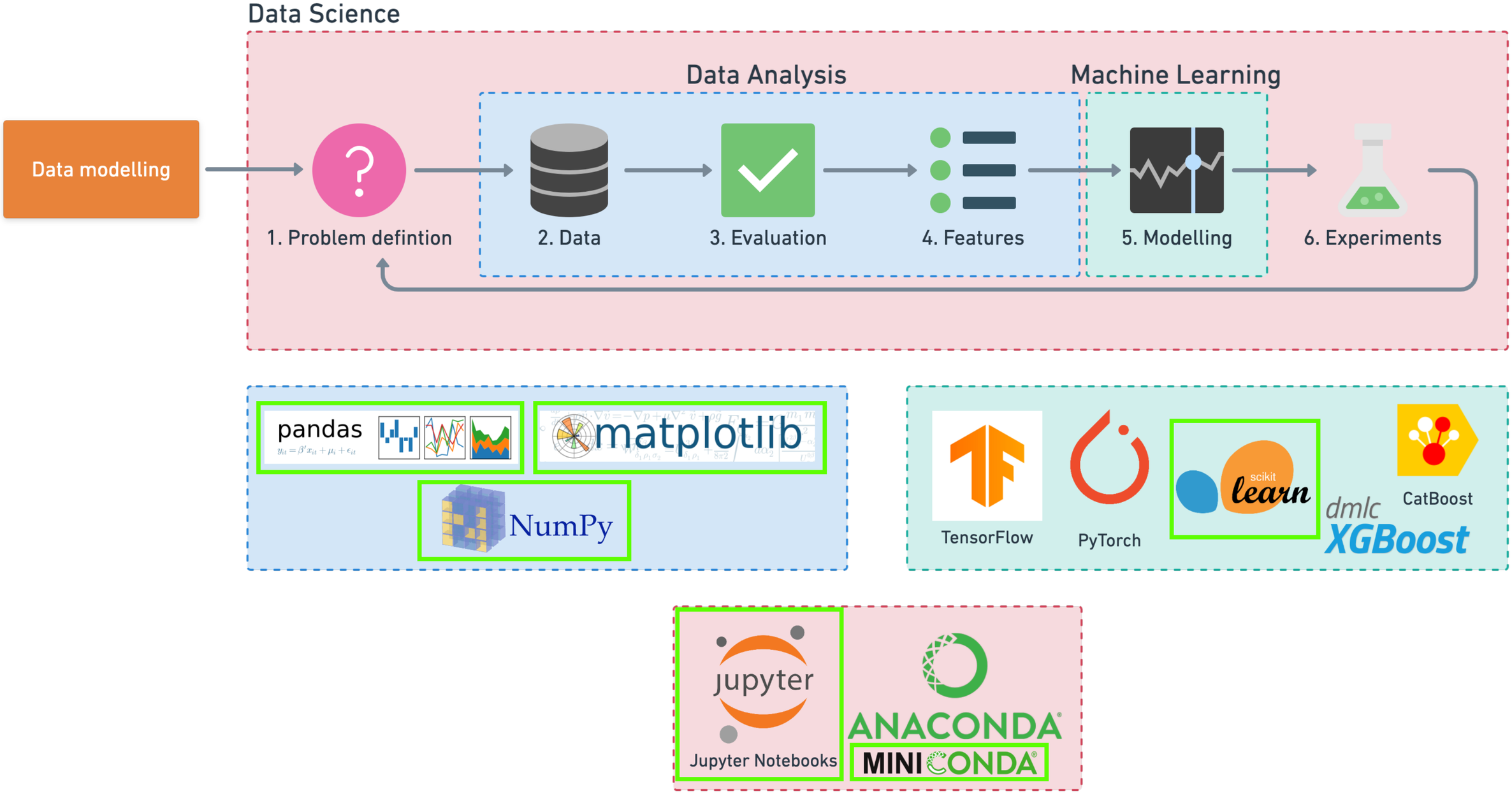
ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110Kg	M	81	4	Yes
5681	64Kg	F	61	1	NO
7911	81Kg	M	57	0	NO

Table 1.0: Patient records

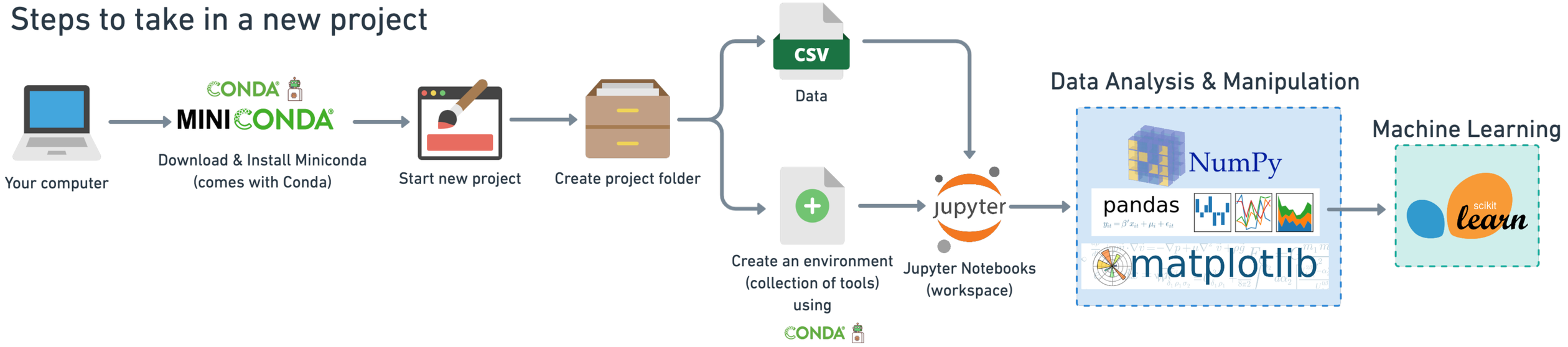
Steps in a full machine learning project



Tools you can use

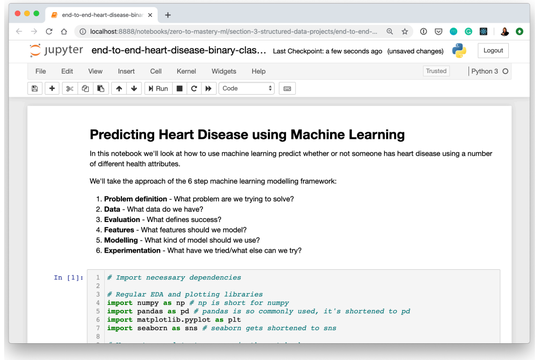


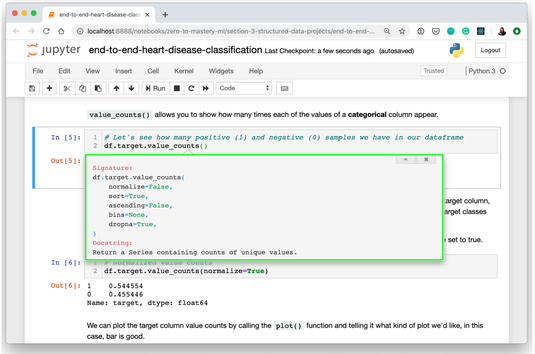
Steps to take in a new project



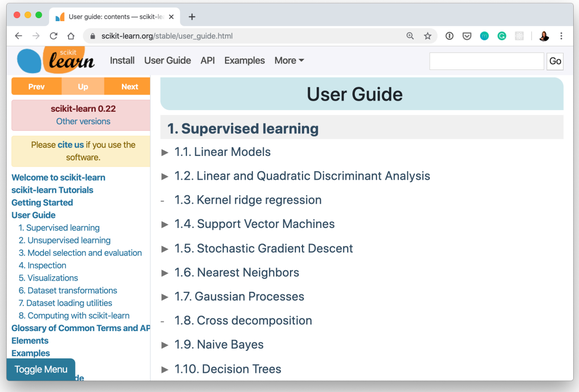
Where can you get help?

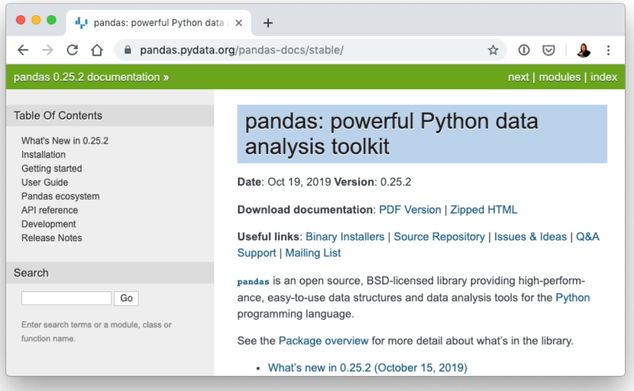
- Follow along with the code
- Try it for yourself
- Press SHIFT + TAB to read the docstring
- Search for it
- Try again
- Ask





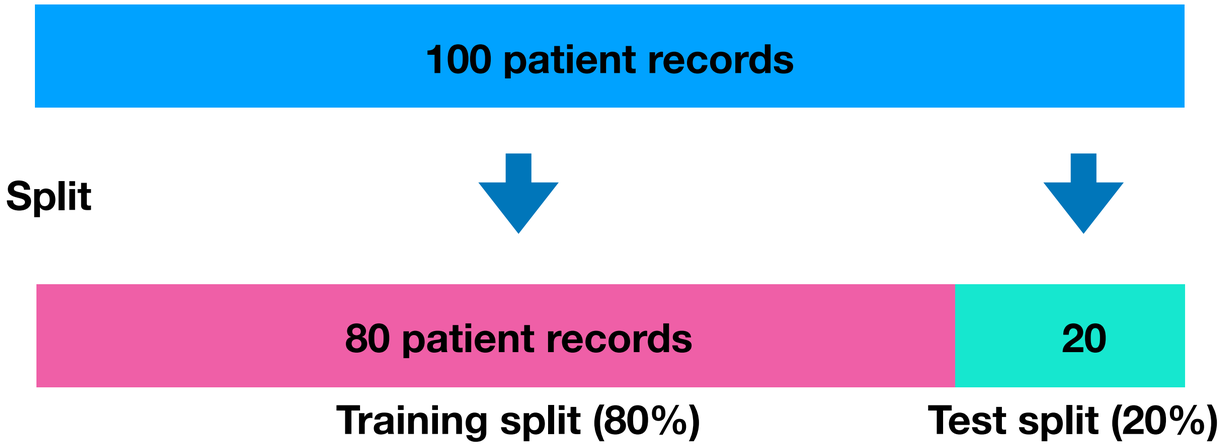
stack overflow





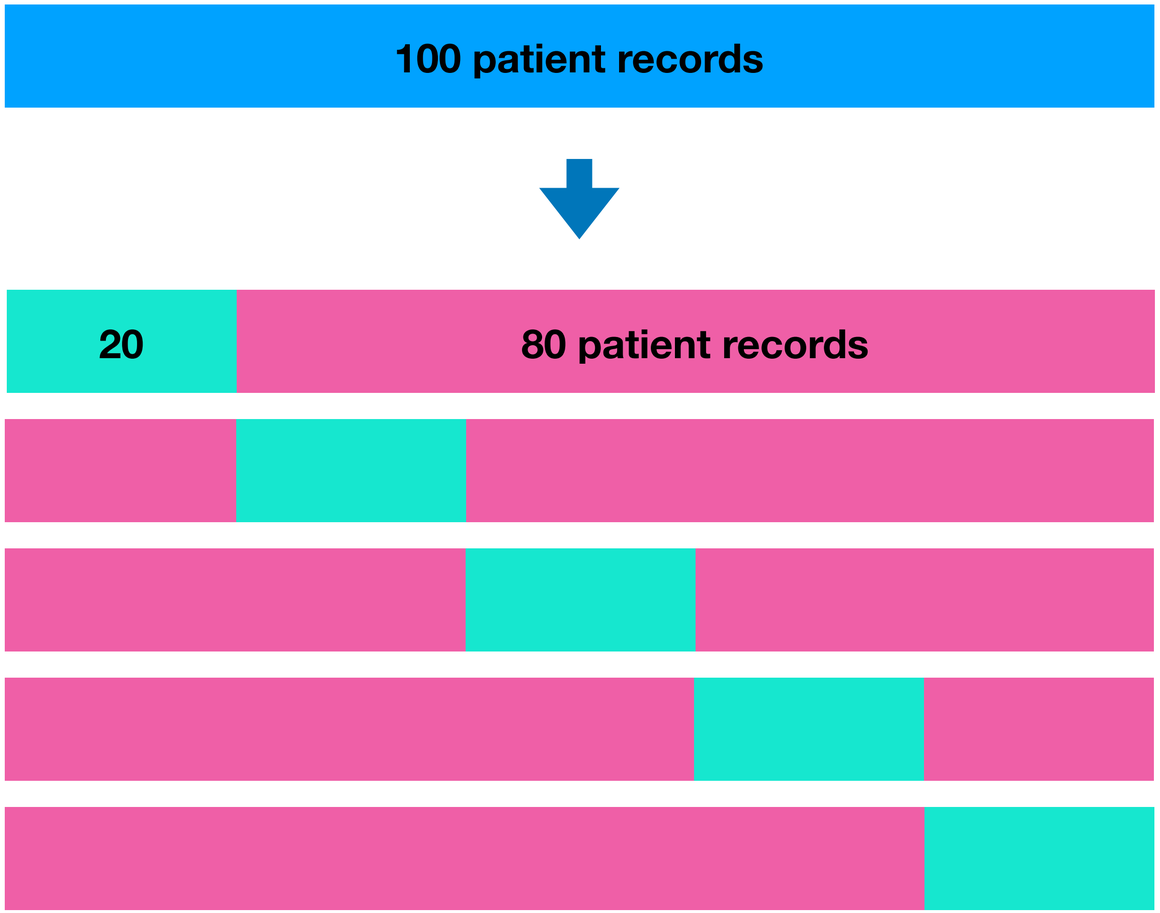
Cross-validation

Normal Train & Test Split



Model is trained on training data, and evaluated on the test data.

5-fold Cross-validation



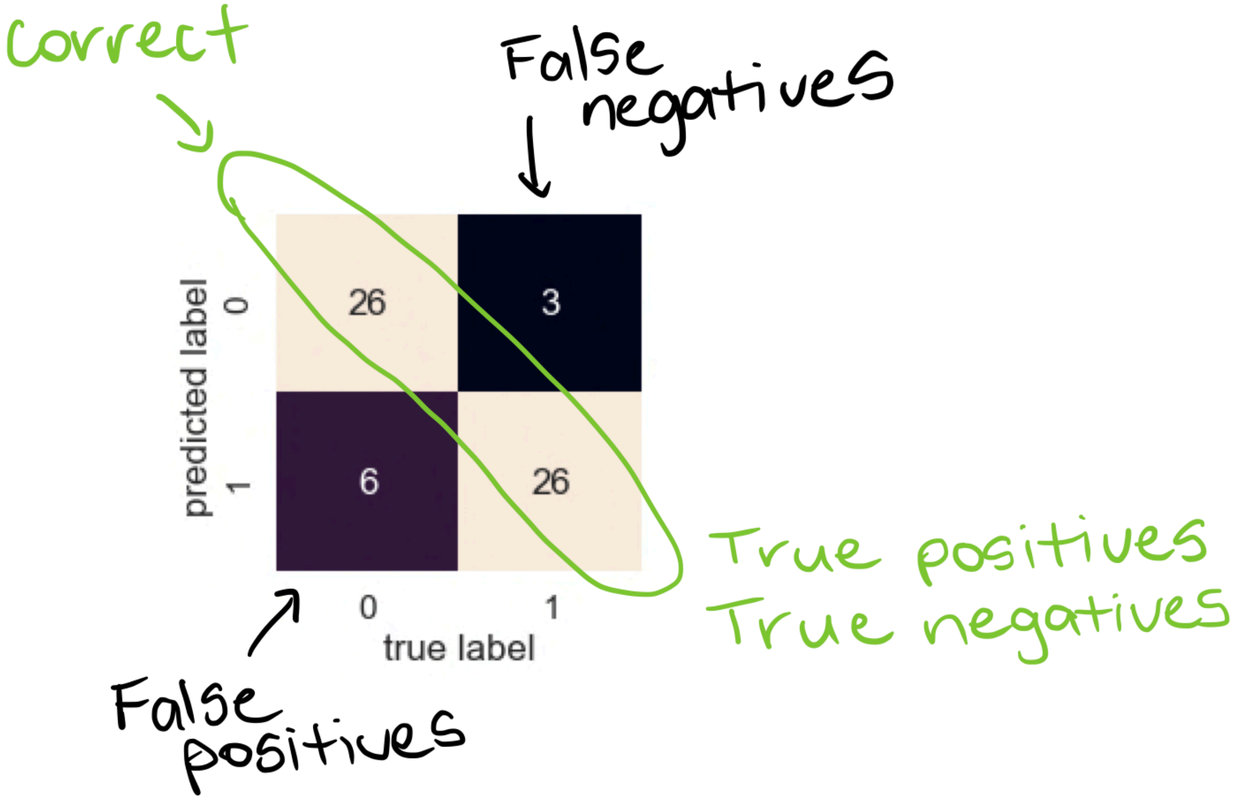
Model is trained on 5 different versions of training data, and evaluated on 5 different versions of the test data.

Classification and Regression metrics

Classification	Regression
Accuracy	**R^2 (r-squared)**
Precision	Mean absolute error (MAE)
Recall	Mean squared error (MSE)
F1	Root mean squared error (RMSE)

Bold = default evaluation in Scikit-Learn

Confusion matrix anatomy



- **True positive = model predicts 1 when truth is 1**
- **False positive = model predicts 1 when truth is 0**
- **True negative = model predicts 0 when truth is 0**
- **False negative = model predicts 0 when truth is 1**

Classification report anatomy

```python
from sklearn.metrics import classification_report

print(classification_report(y_test, y_preds))
```

```
              precision    recall  f1-score   support

           0       0.81      0.90      0.85        29
           1       0.90      0.81      0.85        32

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.86      0.85      0.85        61
```

- **Precision** - Indicates the proportion of positive identifications (model predicted class 1) which were actually correct. A model which produces no false positives has a precision of 1.0.
- **Recall** - Indicates the proportion of actual positives which were correctly classified. A model which produces no false negatives has a recall of 1.0.
- **F1 score** - A combination of precision and recall. A perfect model achieves an F1 score of 1.0.
- **Support** - The number of samples each metric was calculated on.
- **Accuracy** - The accuracy of the model in decimal form. Perfect accuracy is equal to 1.0.
- **Macro avg** - Short for macro average, the average precision, recall and F1 score between classes. Macro avg doesn't class imbalance into effort, so if you do have class imbalances, pay attention to this metric.
- **Weighted avg** - Short for weighted average, the weighted average precision, recall and F1 score between classes. Weighted means each metric is calculated with respect to how many samples there are in each class. This metric will favour the majority class (e.g. will give a high value when one class out performs another due to having more samples).

Which classification metric should you use?

- **Accuracy** is a good measure to start with if all classes are balanced (e.g. same amount of samples which are labelled with 0 or 1).
- **Precision** and **recall** become more important when classes are imbalanced.
- If false positive predictions are worse than false negatives, aim for higher precision.
- If false negative predictions are worse than false positives, aim for higher recall.
- **F1-score** is a combination of precision and recall.