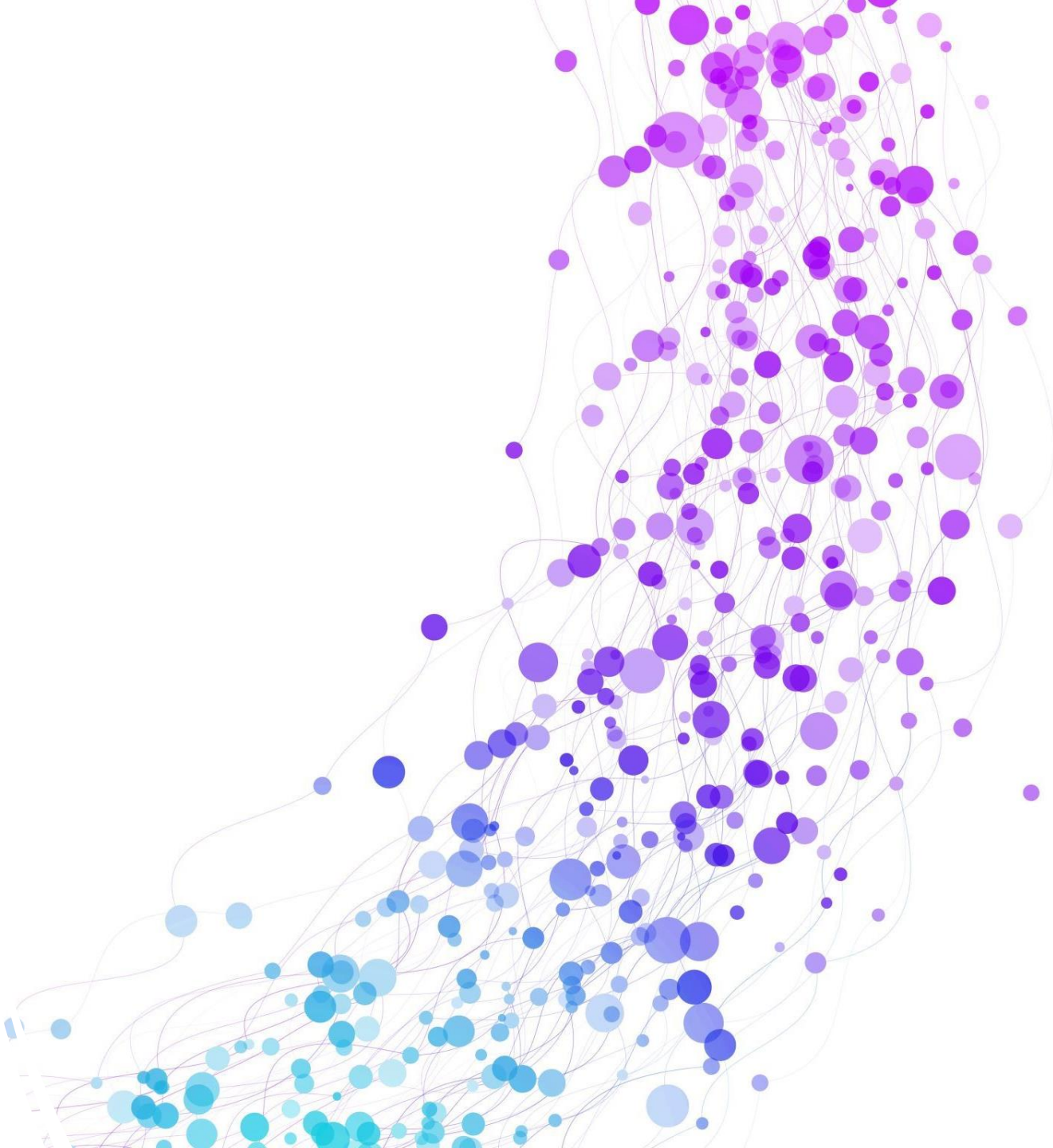
Predicting Coronary Artery Disease in the U.S. Population

Hiram G. Menendez

Springboard Fellow



Coronary Artery Disease (CAD)

- It is the #1 cause of death in the United States.
- Many don't know they have it until they have a heart attack.
- Being able to predict can lead to early intervention.

Image by Kenny Eliason from Unsplash

Purpose

Build machine learning models that can predict whether an individual has heart disease.

The Data

- Data used was the Behavioral Risk Factor Surveillance System for the year 2015.
- Contains responses to hundreds of health-related questions about:
  - General health
  - Hypertension
  - Arthritis
  - Depression
  - Demographics
  - Seatbelt use

Images from https://www.cdc.gov/

How this is done

- Machine learning algorithms "learn" from data.
- After learning, the models should be able to make predictions about data it has "never seen".

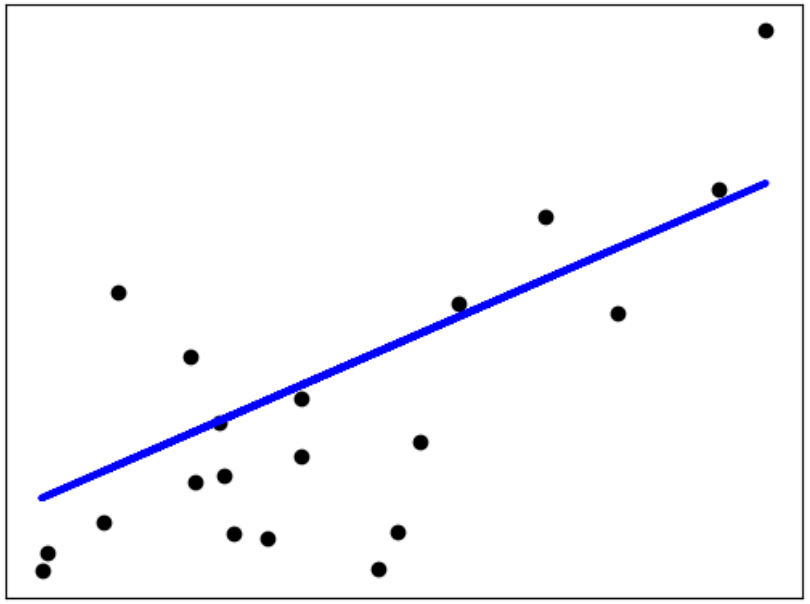


Image from scikit-learn.org

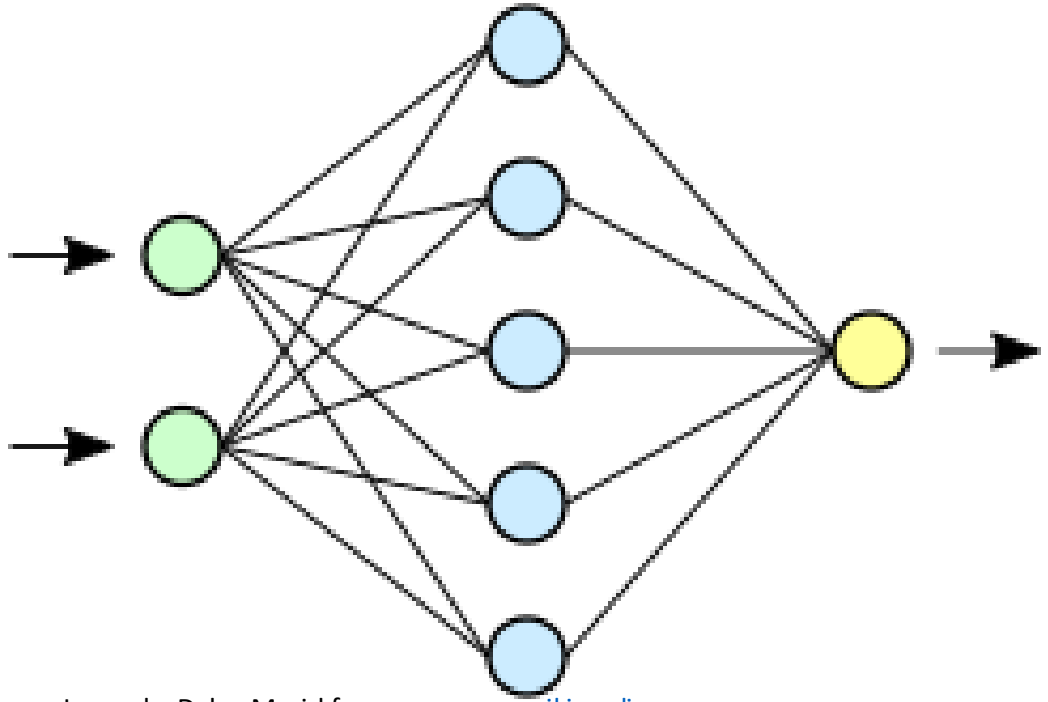
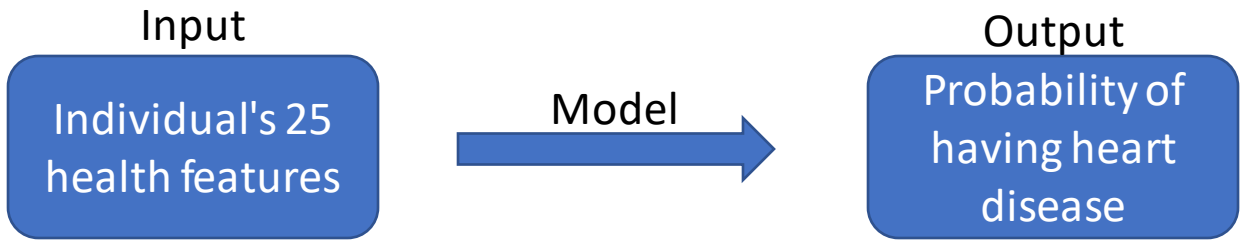


Image by Dake, Mysid from commons.wikimedia.org

What Information is used from the Data?

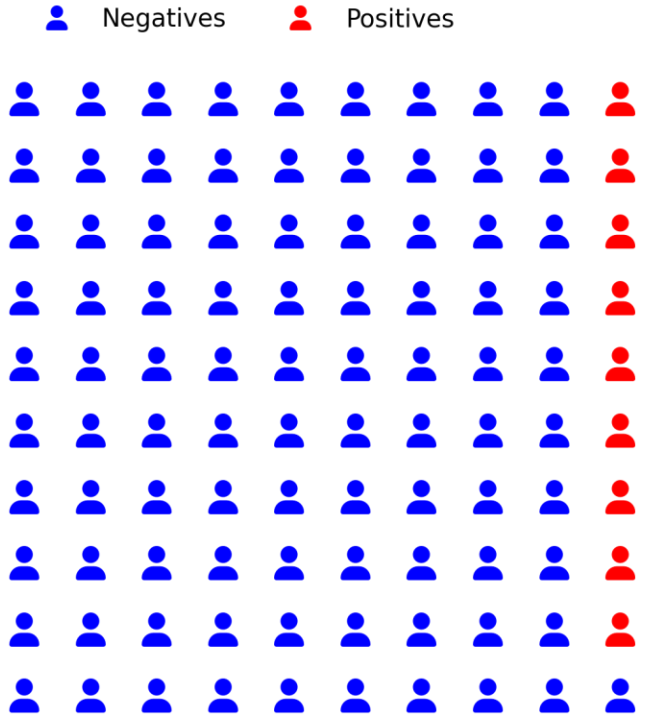
- 25 "features" from each sample in the data were used.
- All of them are categorical: they say to what category an individual belongs to.
- Some of the features used in the models:
  - Good General Health
  - Hypertension
  - High Cholesterol
  - BMI
  - Smoker
  - Heavy Drinker

	Good_Health	Health_Insurance	Hypertension	High_Cholesterol	Asthma_Status	Arthritis	Race	Age_Cat	BMI_Cat	Education_Level	...
0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	5.0	4.0	2.0	...
1	1.0	2.0	2.0	2.0	3.0	2.0	1.0	4.0	3.0	4.0	...
2	2.0	9.0	2.0	1.0	3.0	1.0	1.0	6.0	2.0	2.0	...
3	2.0	1.0	1.0	1.0	3.0	1.0	1.0	5.0	3.0	2.0	...
4	2.0	1.0	2.0	2.0	3.0	1.0	1.0	5.0	2.0	3.0	...
...	...	...	...	...	...	...	...	...	...	...	...
441451	2.0	9.0	1.0	1.0	3.0	1.0	5.0	6.0	1.0	1.0	...
441452	1.0	1.0	2.0	2.0	3.0	2.0	5.0	2.0	3.0	3.0	...
441453	2.0	9.0	1.0	1.0	3.0	2.0	5.0	6.0	4.0	2.0	...
441454	1.0	1.0	1.0	2.0	3.0	2.0	5.0	4.0	2.0	3.0	...
441455	1.0	1.0	1.0	1.0	3.0	2.0	5.0	5.0	2.0	4.0	...

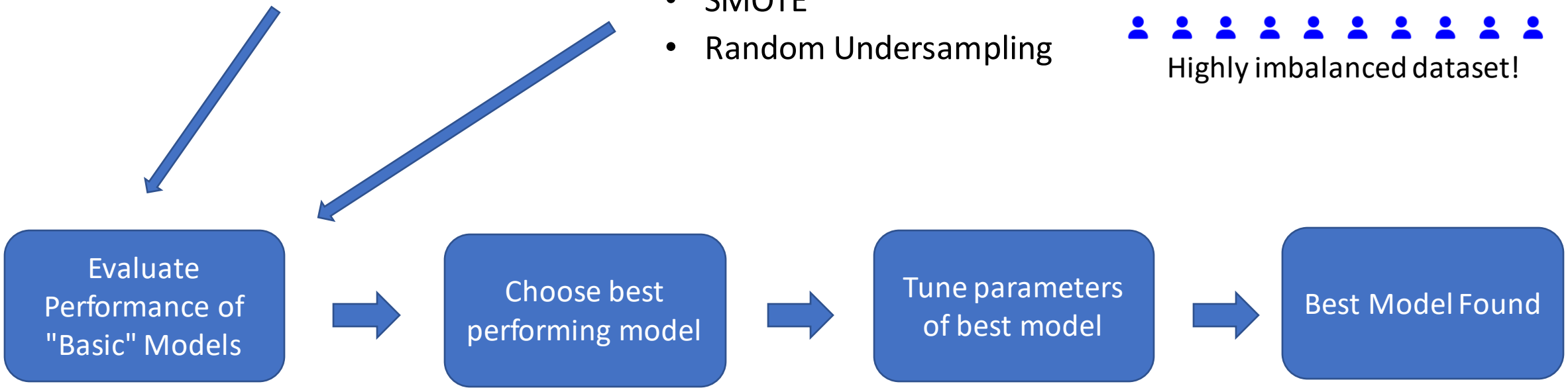


Modeling

- Options to build model:
  - 3 algorithms:
    - Logistic Regression
    - Random Forest
    - XGBoost
  - 5 resampling techniques:
    - No resampling
    - Class Weights
    - Random Oversampling
    - SMOTE
    - Random Undersampling

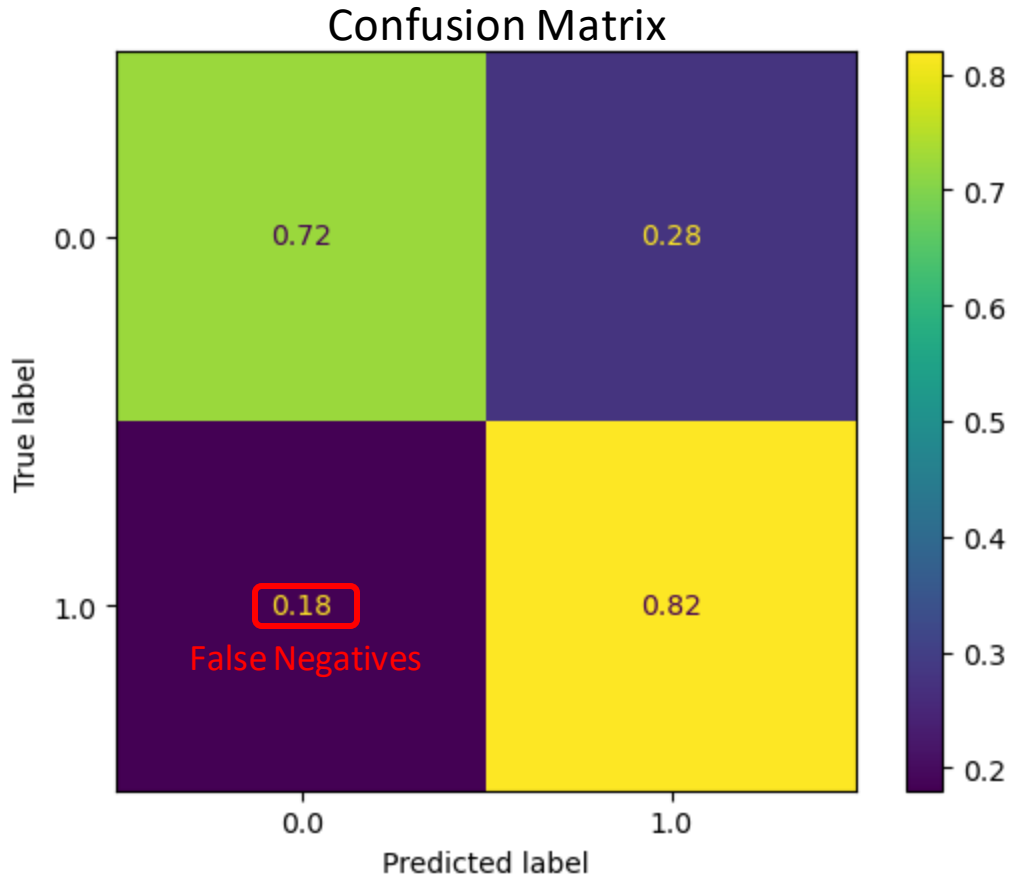


Highly imbalanced dataset!



Evaluating Models

- The accuracy score is not an ideal metric for imbalanced datasets.
- For heart disease, minimizing false negatives should be a priority.
- The recall score is therefore the most suitable metric.
- **Warning: be aware of model's low precision score.**

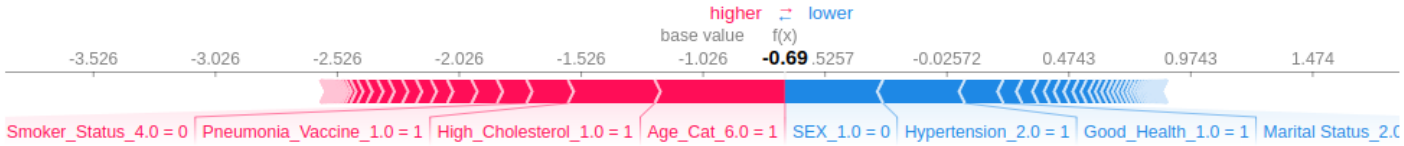


"Given that the individual has heart disease, the best model correctly predicts this 82% of the time"

Classification Report:

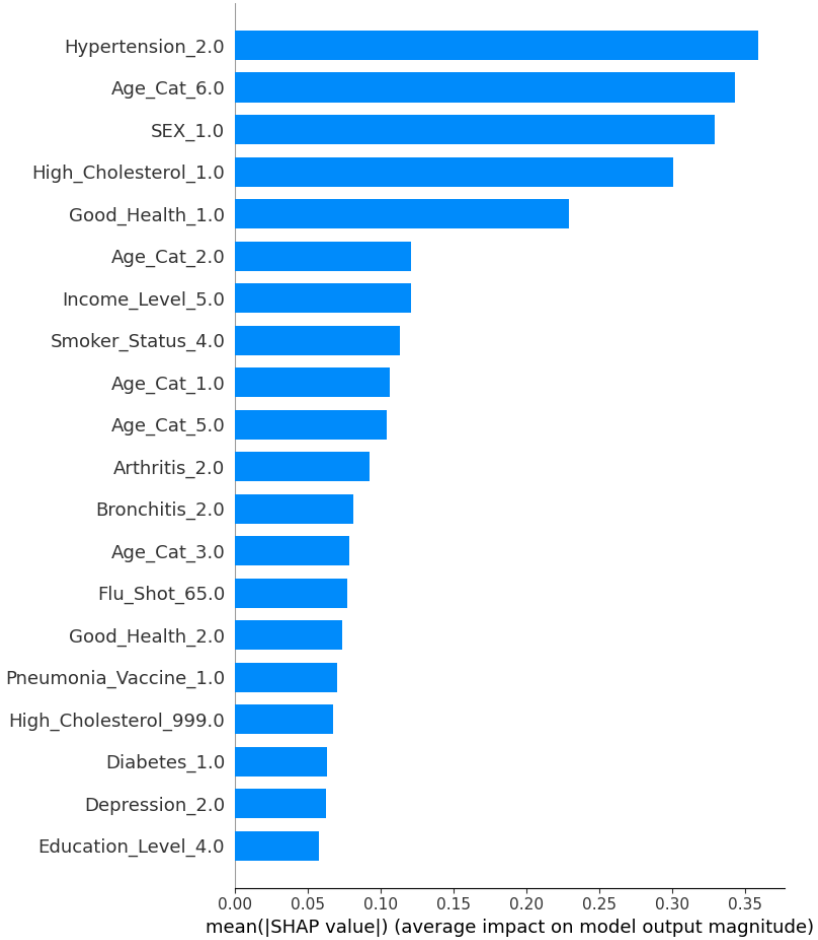
	precision	recall	f1-score	support
0.0	0.98	0.72	0.83	119665
1.0	0.22	0.82	0.35	11590
accuracy			0.73	131255
macro avg	0.60	0.77	0.59	131255
weighted avg	0.91	0.73	0.79	131255

Model Interpretability



For each prediction, SHAP values can tell you which features contributed the most to the prediction.

Average SHAP values tell you, on average, which features are the "most important"



Guidelines when Using the Model

1. Gather patient info on 25 features and input into model.
2. SHAP value force graph can be generated to explain to patient what factors contributed to decision.
3. Those predicted to have heart disease are the best candidates to refer to more accurate testing.

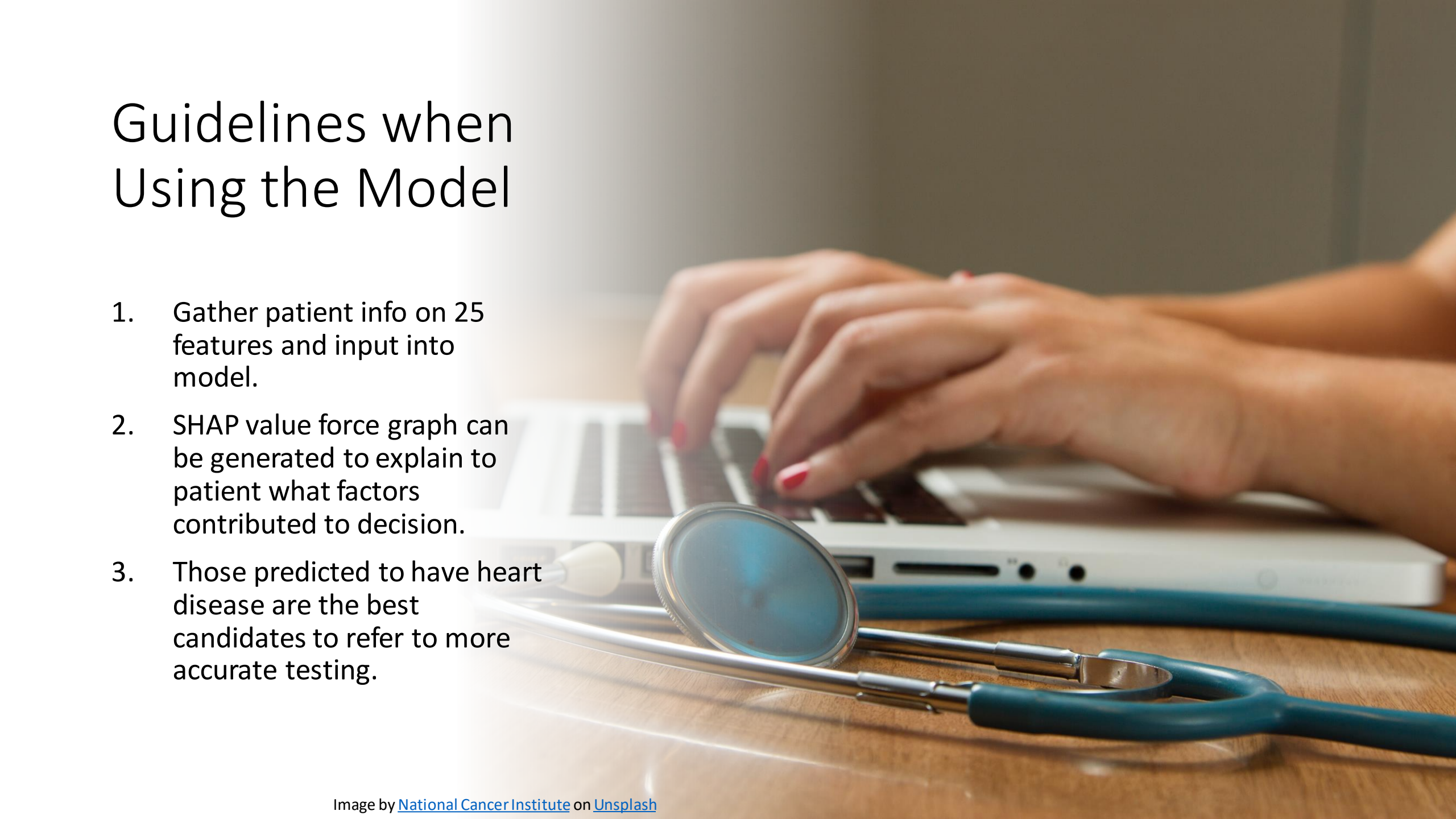


Image by National Cancer Institute on Unsplash


Thank you!