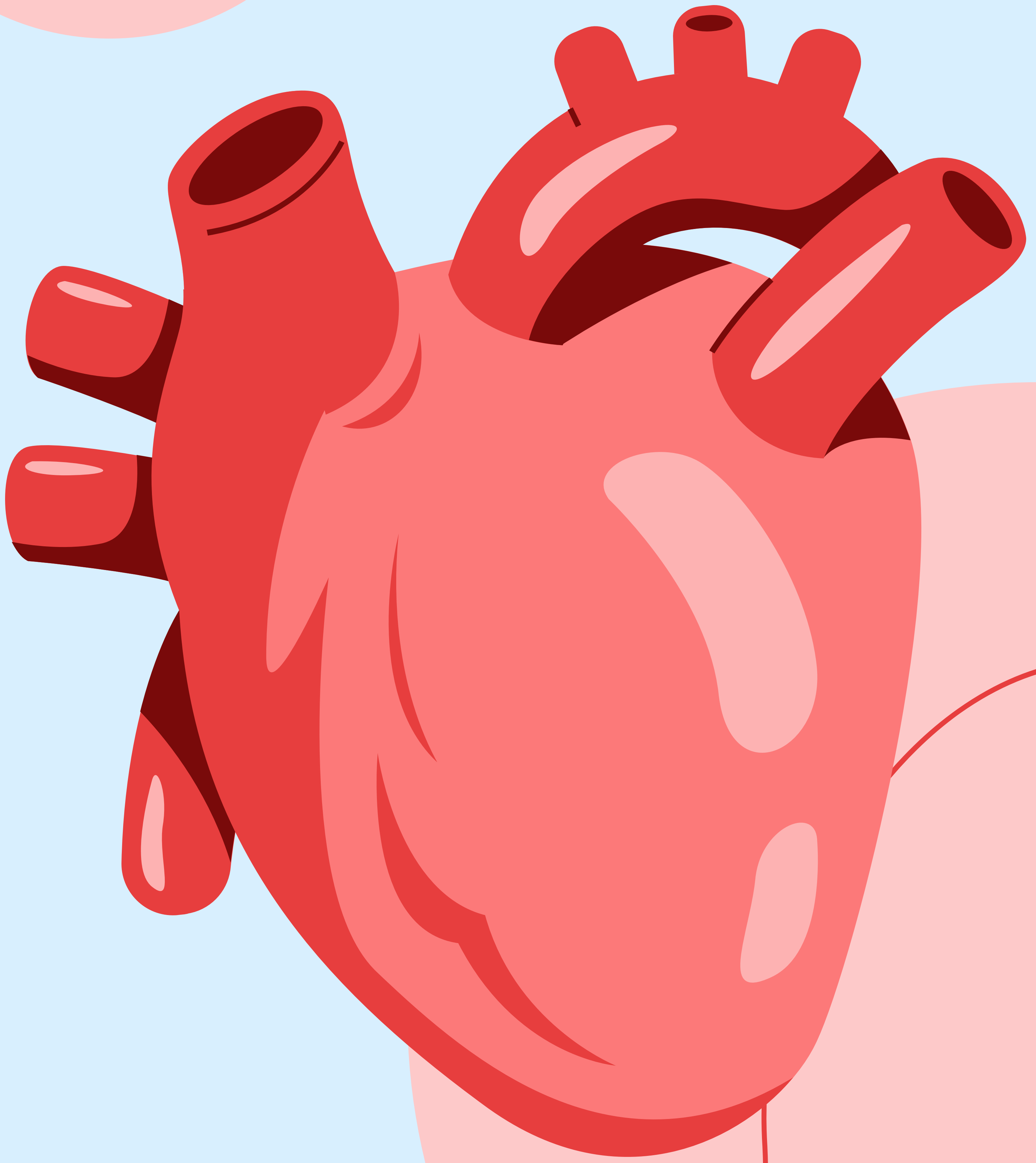
Heart Disease Prediction



group 17
Qixuan Hu, Ledi Wang, Xuehan Wang, Jiaming Xu, Xiaoyi Zhang

Background Introduction

Importance of Heart Disease Prediction

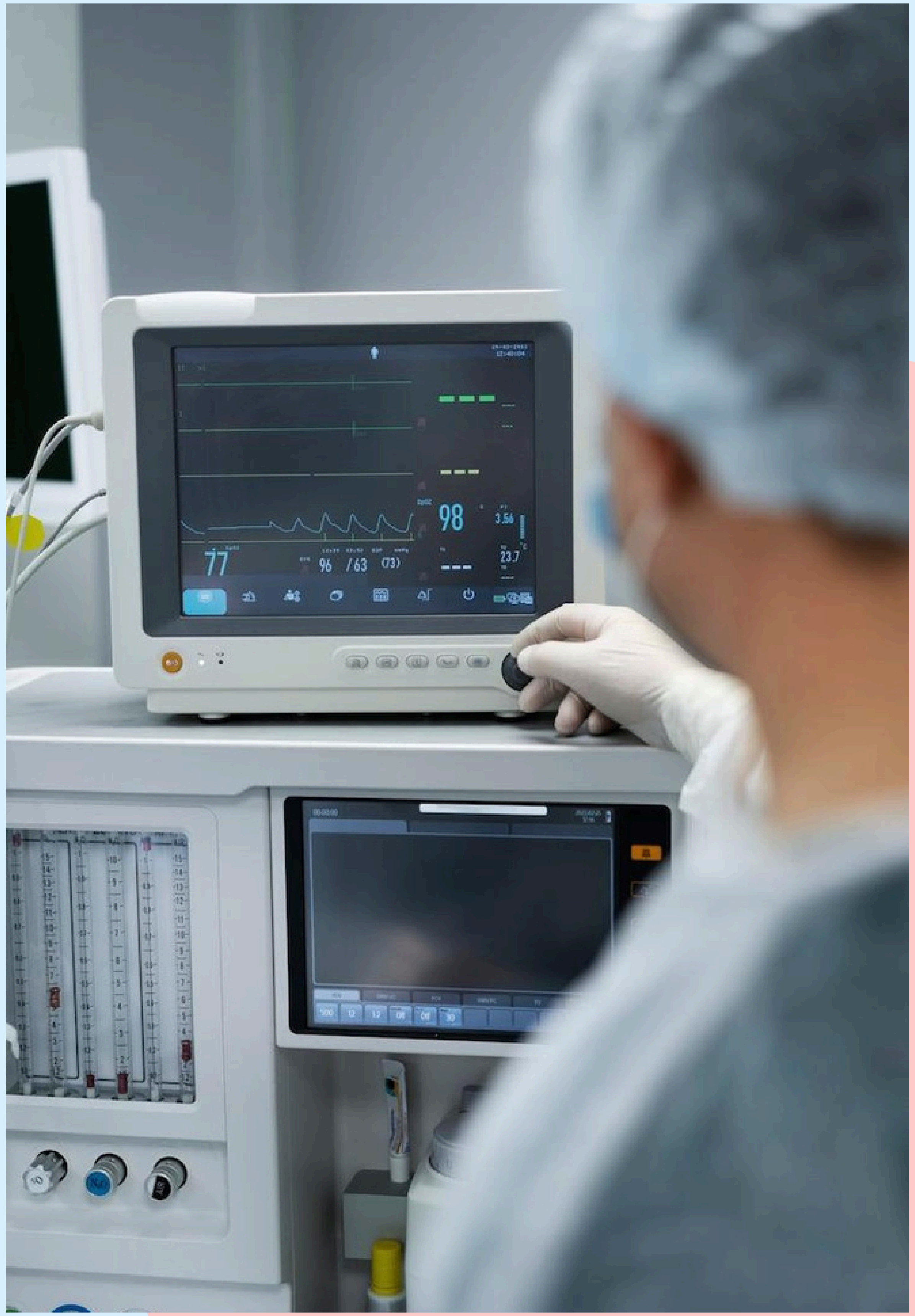
- Leading causes of death workdwide
- Early detection&intervention are critical

Challenges in Traditional Prediction

- Rely on expertise of medical professionals (time-consuming, expensive)
- Quality and quantity of data gathered

Purpose of ML project

- Early detection
- Cost efficiency
- Targeted interventions



Dataset Introduction

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	\
0	No	16.60	Yes	No	No	3	
1	No	20.34	No	No	Yes	0	
2	No	26.58	Yes	No	No	20	
3	No	24.21	No	No	No	0	
4	No	23.71	No	No	No	28	

	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	\
0	30	No	Female	55-59	White	Yes	
1	0	No	Female	80 or older	White	No	
2	30	No	Male	65-69	White	Yes	
3	0	No	Female	75-79	White	No	
4	0	Yes	Female	40-44	White	No	

	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	Yes	Very good	5	Yes	No	Yes
1	Yes	Very good	7	No	No	No
2	Yes	Fair	8	Yes	No	No
3	No	Good	6	No	No	Yes
4	Yes	Very good	8	No	No	No

Target variable

HeartDisease: binary categorical variable for whether a person has heart disease

17 features

Variables related to physical health

BMI:
Reflecting the degree of obesity

PhysicalHealth:
Number of days with physical health in the past 30 days

MentalHealth:
Number of days with mental health in the past 30 days

SleepTime:
Average sleep time per night (hours)

Variables related to lifestyle habits
(all are binary variable)

Smoking
AlcoholDrinking
PhysicalActivity

Functional limitation related variables

DiffWalking
Whether there is difficulty walking (binary variable: Yes/No).

17 features

Variables related to disease history
(binary variable)
Stroke
Diabetic
(Borderline means blood sugar is close to the borderline of diabetes.)
Asthma
KidneyDisease
SkinCancer

Demographic variables
Sex
AgeCategory:
"18-24", "25-29", "30-34", "35-39", "40-44", "45-49", "50-54", "55-59", "60-64", "65-69", "70-74", "75-79", "80 or older"
Race:
White,Hispanic,Black,Other, Asian,AmericanIndian/Alaskan Native
GenHealth:
Evaluation of general health status (5-level classification: Excellent, Very good, Good, Fair, Poor).

Data Processing

1. Binary Coding

- 'Yes', 'No' to 1 and 0
- 'Female', 'Male' to 1 and 0
- 'AgeCategory' to integers (e.g. 50-54 to 50, 55-59 to 55)

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth \
126167	0	23.44	0	0	0	10.0
207506	0	32.49	0	0	0	0.0
274544	0	21.93	0	0	0	0.0
121049	0	26.58	0	0	0	0.0
260961	0	19.02	1	0	0	2.0

	MentalHealth	DiffWalking	Sex	AgeCategory	Diabetic \
126167	20.0	1	1	80	0
207506	4.0	0	0	40	0
274544	0.0	0	0	60	0
121049	2.0	0	0	45	0
260961	2.0	0	1	80	0

	PhysicalActivity	SleepTime	Asthma	KidneyDisease	SkinCancer
126167	0	6.0	0	0	0
207506	1	8.0	0	0	0
274544	1	7.0	0	0	0
121049	1	7.0	0	0	0
260961	1	6.0	0	0	0

```
(array([0, 1]), array([5838,  558]))
```

2. Random Sampling

- Original: 319769 patients
- Sampled 2% of the dataset
- make analysis faster and manageable
- ensures fairness and avoids bias

Preparing Data for Analysis

1. Balancing data (RandomOverSampler)

- Made sure patients with and without heart disease are fairly represented using oversampling

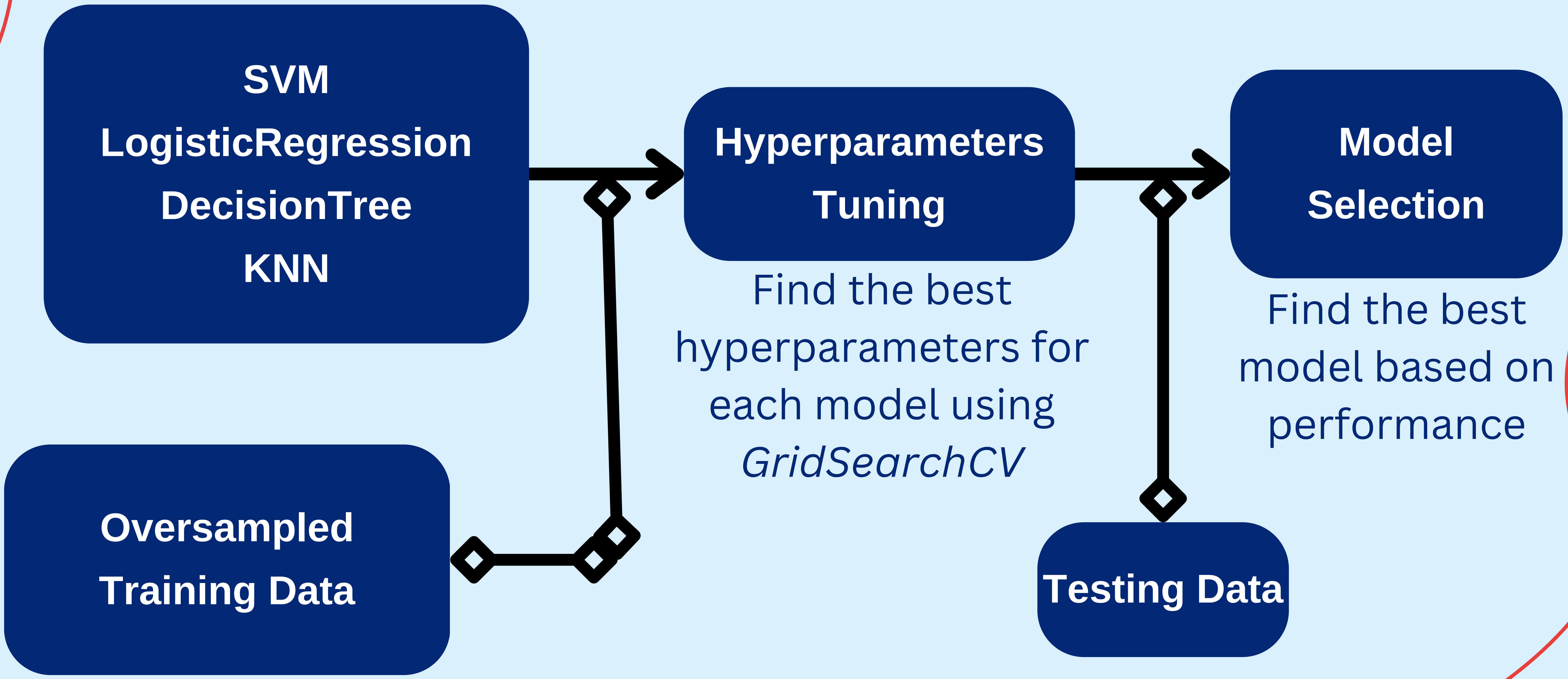
2. Scaling (MinMaxScaler)

- Convert different scales to a fixed range of [0,1]
- Make them contribute equally to the model

3. Splitting Data (train_test_split)

- Divided into training (80%) and testing (20%) sets to ensure accurate predictions

Models Comparison



Evaluating Metrics

F1 Score:

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

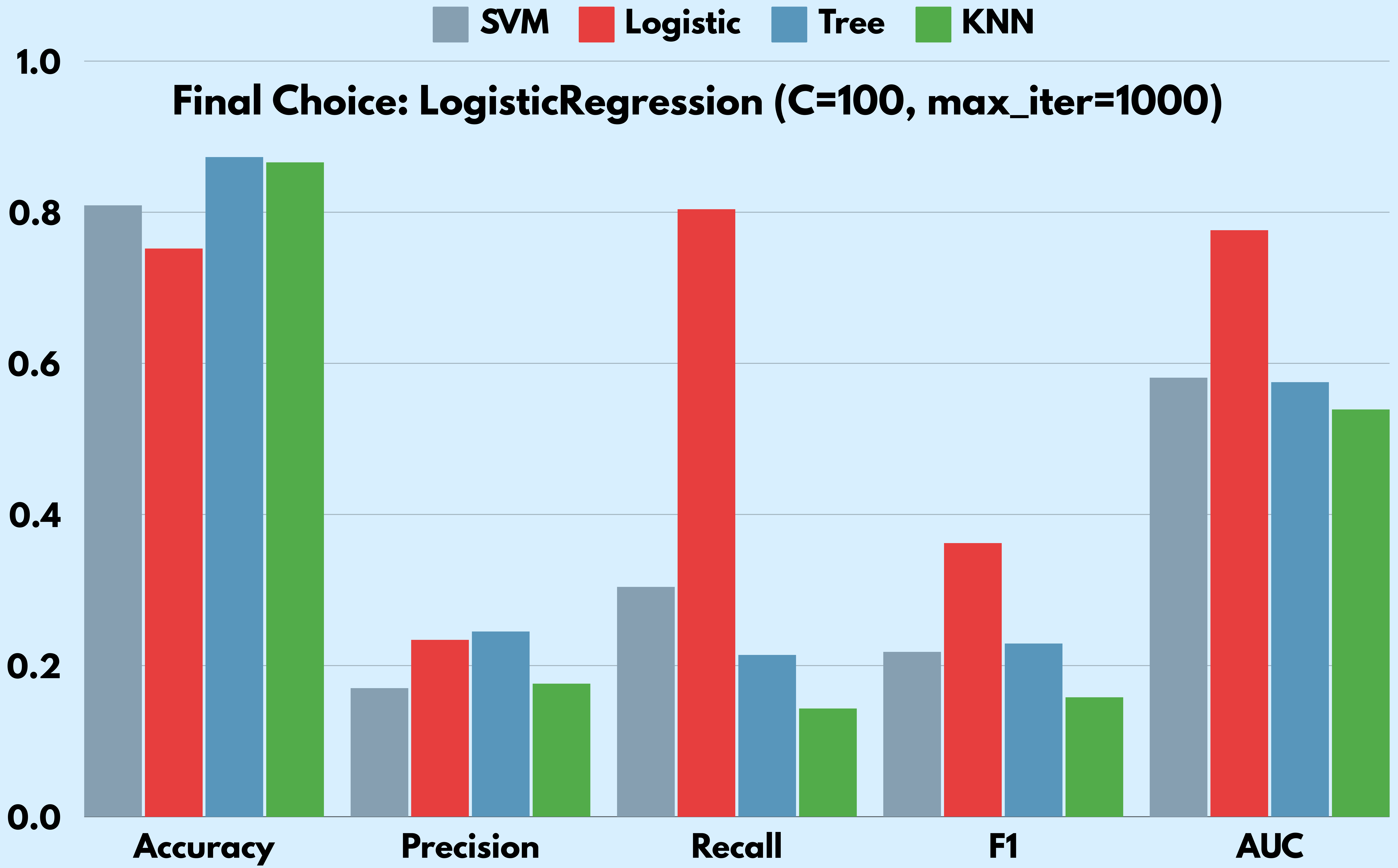
Hyperparameters Tuning → **Model Selection**

F1 only

F1
Precision
Recall
Accuracy
AUC

Heart Disease
versus
No Heart Disease

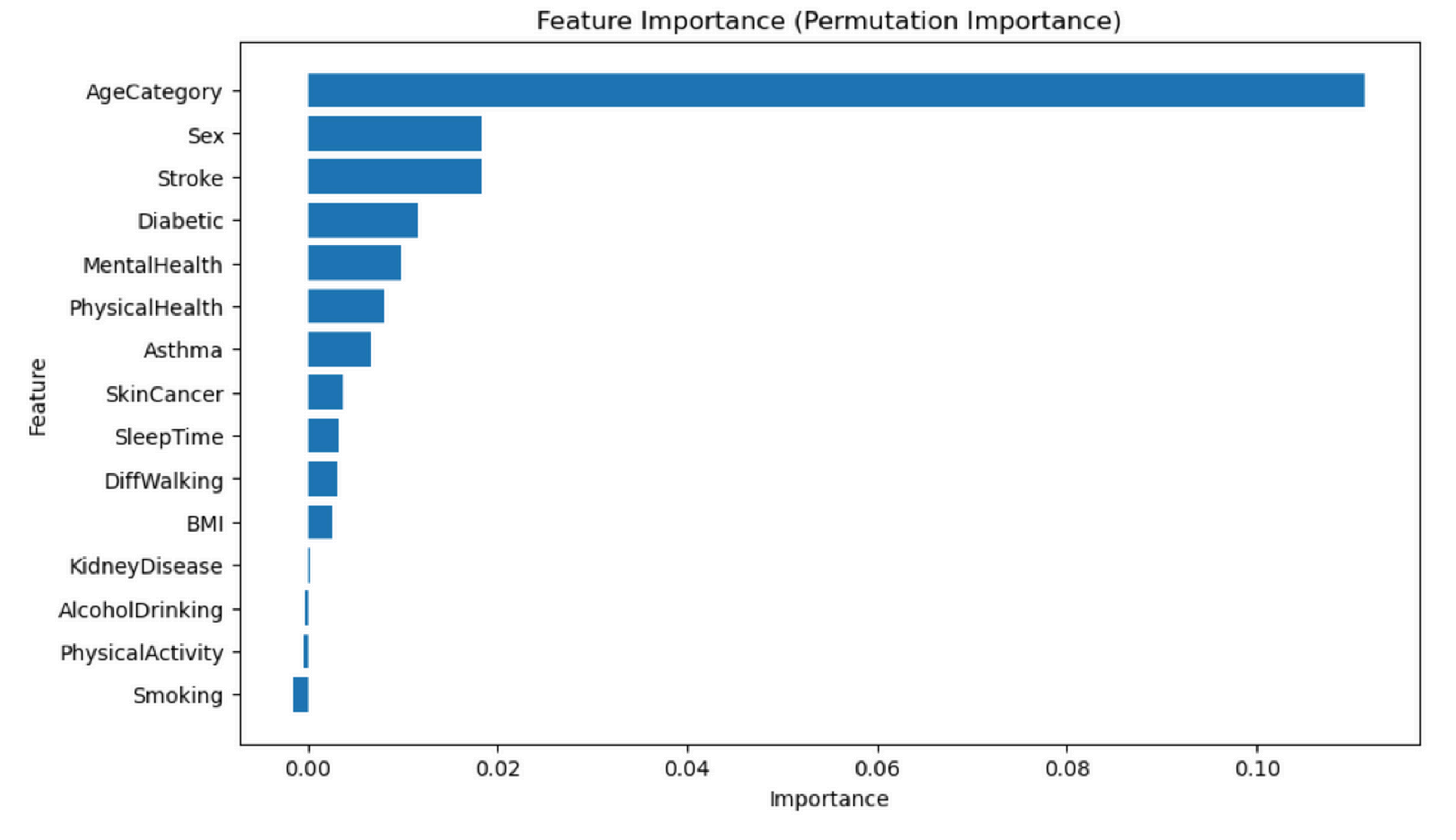
$\approx$ 1 : 14



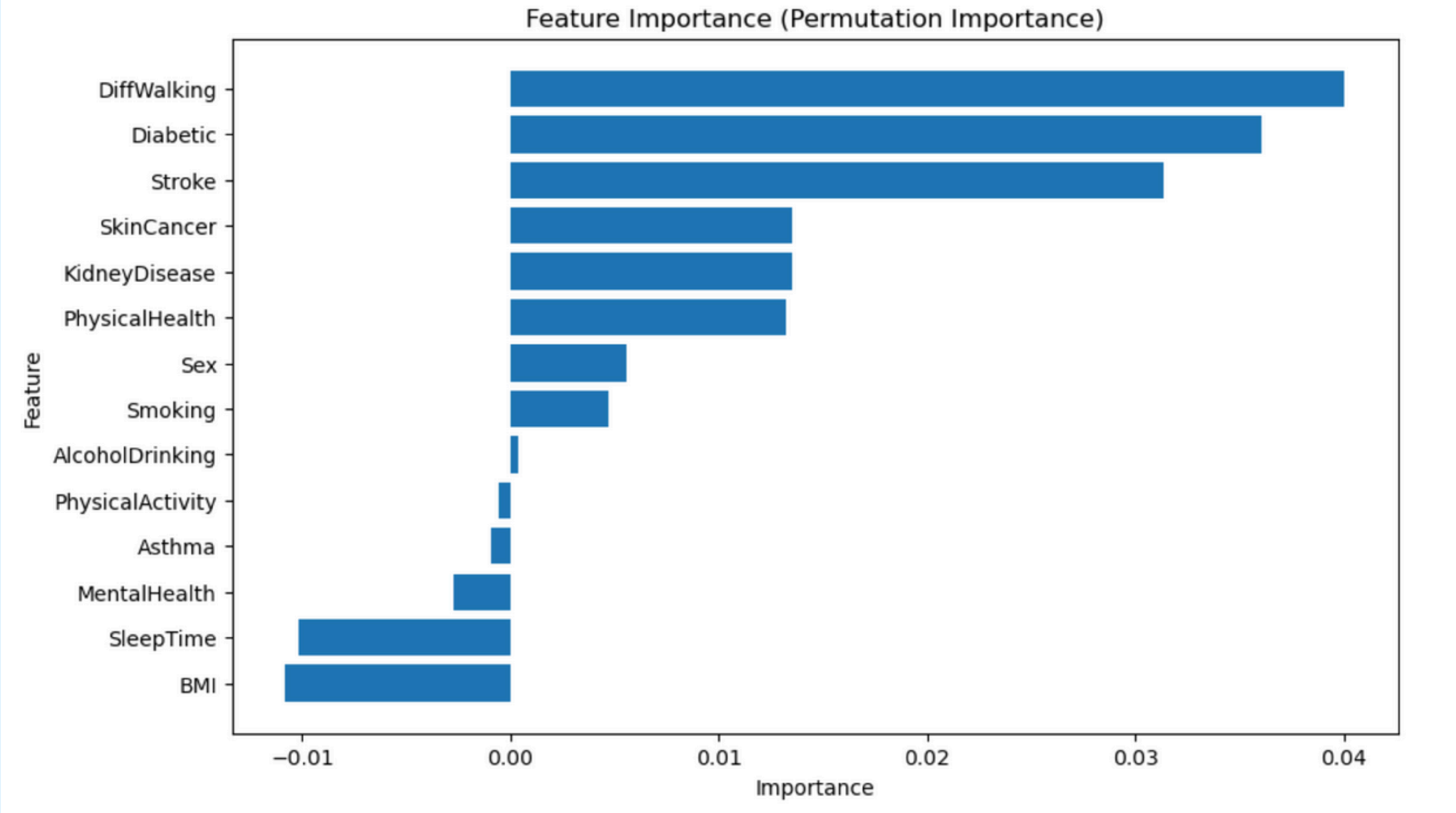
SVM
Logistic
Tree
KNN
1.0
0.8
0.6
0.4
0.2
0.0
Final Choice: LogisticRegression (C=100, max_iter=1000)
Accuracy
Precision
Recall
F1
AUC

Feature Evaluation

Feature importance with age



Feature importance without age



Thanks