Lecture 7: Nearest Neighbors and Decision Trees

https://tinyurl.com/cis5190-9-26-2022

Osbert Bastani and Zachary G. Ives

CIS 4190/5190 – Fall 2022

A Different Kind of Learning

To this point: *parametric* learning

Given a predetermined family of functions that maps from input features to prediction, learn a set of *parameters* for this function

.. one way: by optimizing against the *loss function*

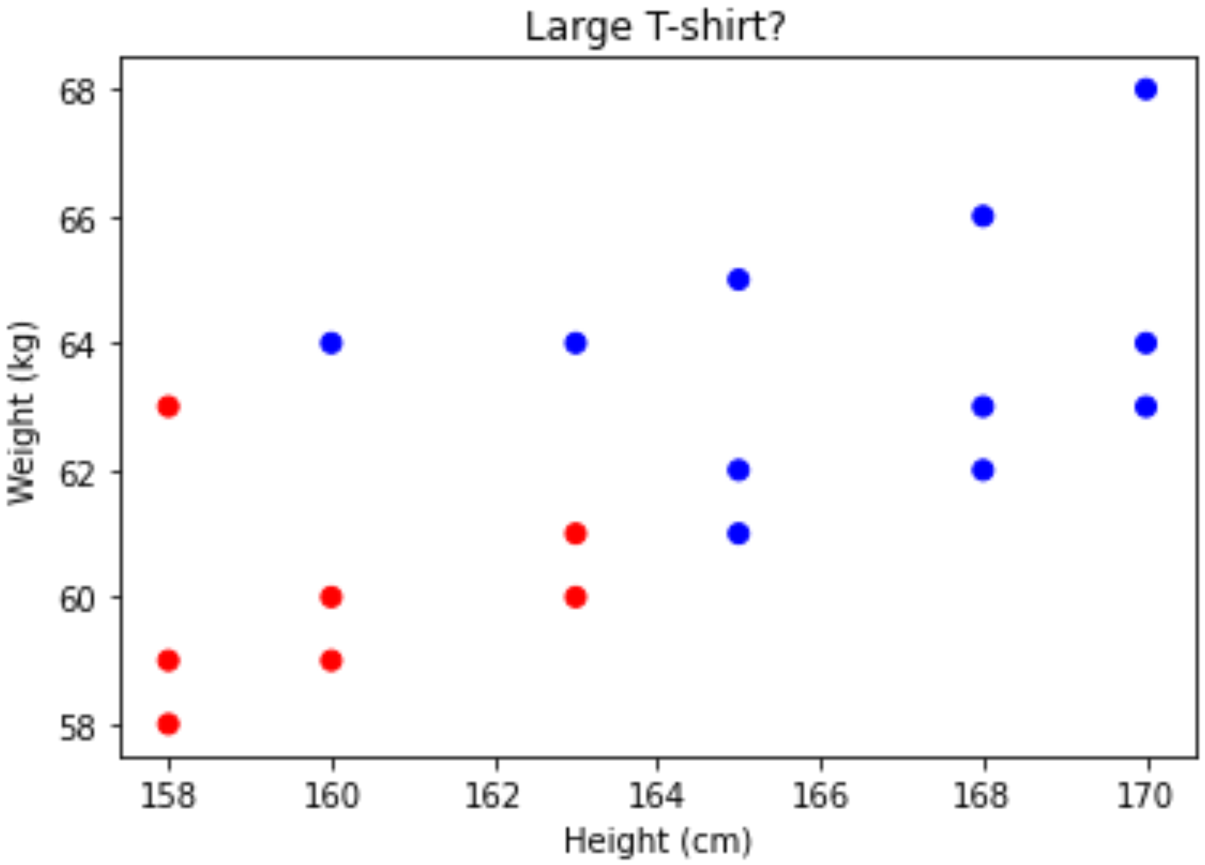
linear regression – continuous-valued output

logistic regression – Boolean-valued output

But this is not the only kind of ML algorithm – now, we'll see two variations on this theme

- k-Nearest Neighbors
- Decision trees

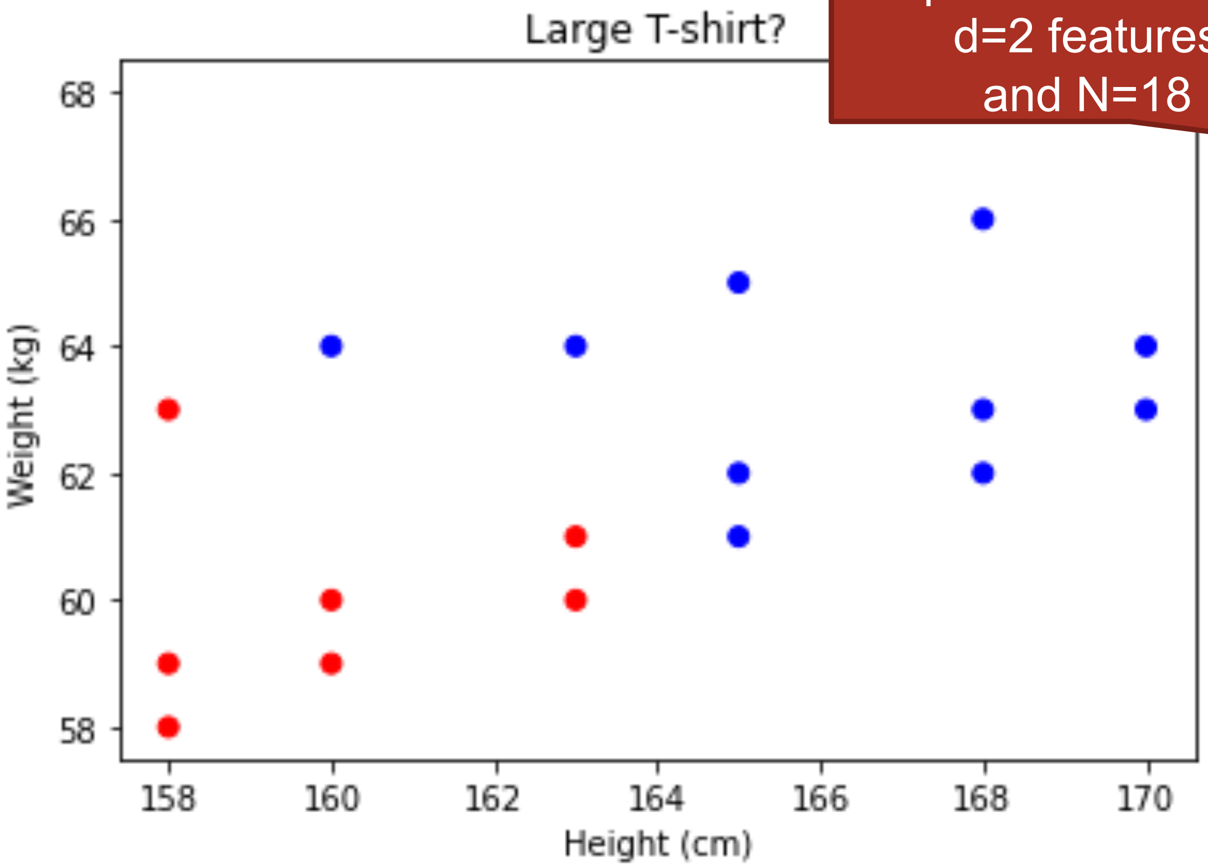
Our Default Setup: Training for Binary Classification



Height (cm)	Weight (kg)	Large (vs Medium) t-shirt?
158	58	F
158	59	F
158	63	F
160	59	F
160	60	F
163	60	F
163	61	F
160	64	T
163	64	T
165	61	T
165	62	T
165	65	T
168	62	T
168	63	T
168	66	T
170	63	T
170	64	T
170	68	T

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

Our Default Setup: Training for Binary Classification



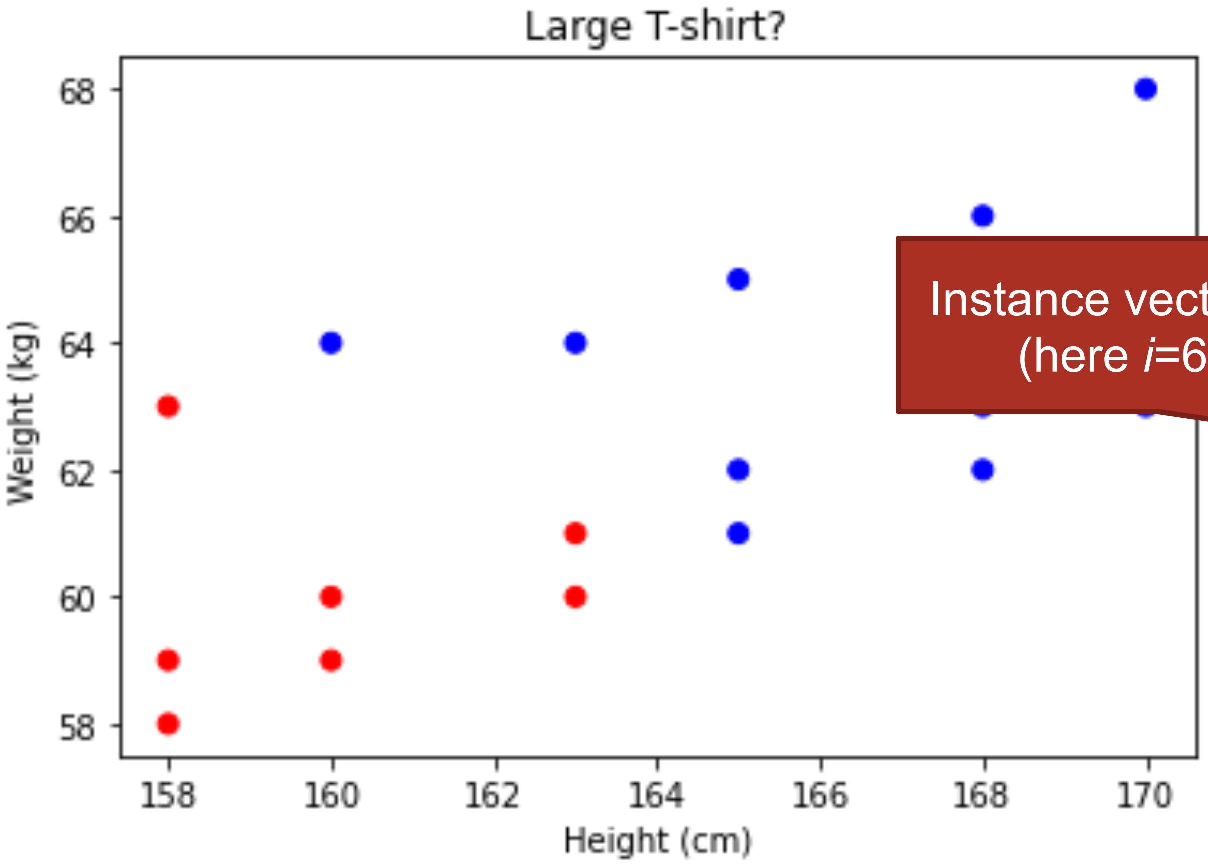
Input matrix $\boldsymbol{X}$ with d=2 features and N=18

Class vector y

Height (cm)	Weight (kg)	Large (vs Medium) t-shirt?
158	58	F
158	59	F
158	63	F
160	59	F
160	60	F
163	60	F
163	61	F
160	64	T
163	64	T
165	61	T
165	62	T
165	65	T
168	62	T
168	63	T
168	66	T
170	63	T
170	64	T
170	68	T

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

Our Default Setup: Training for Binary Classification



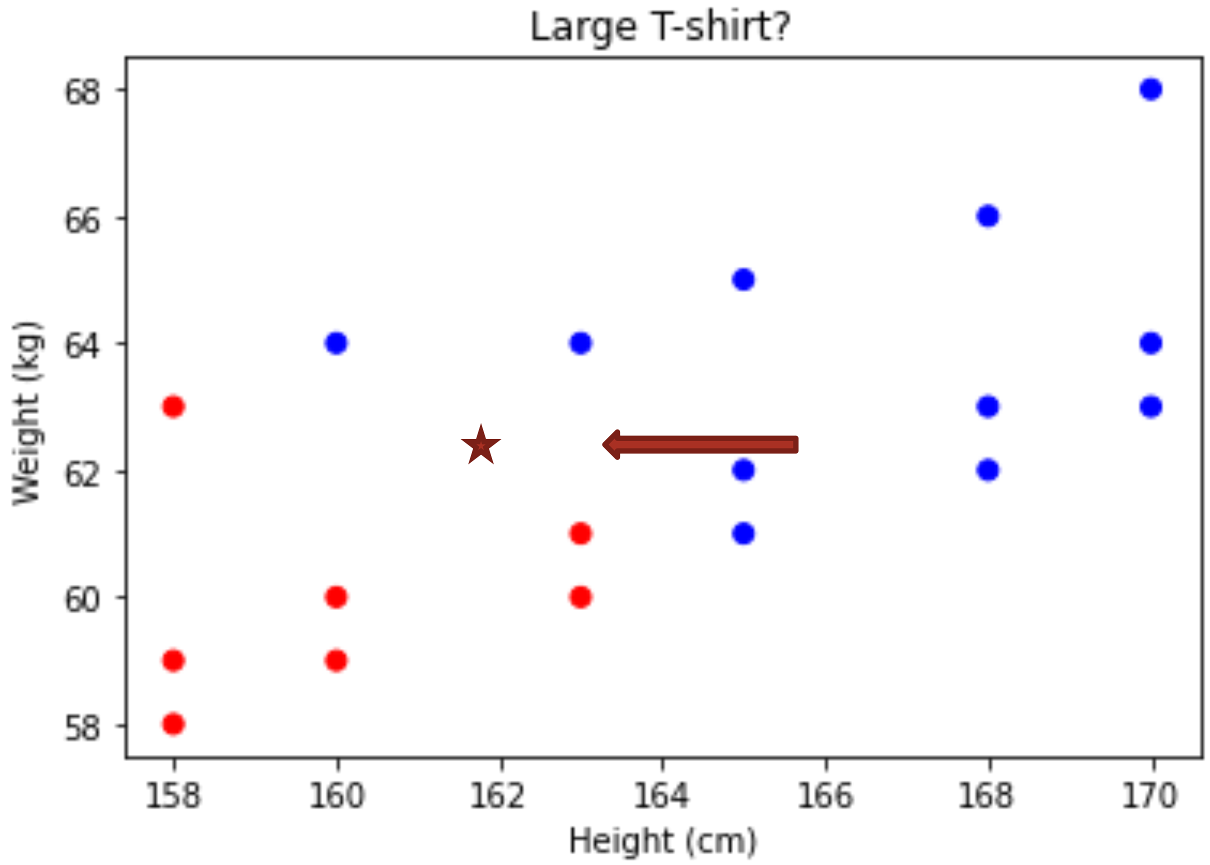
Instance vector x_i, (here i=6)

Class label y_i

Height (cm)	Weight (kg)	Large (vs Medium) t-shirt?
158	58	F
158	59	F
158	63	F
160	[illegible]	F
160	[illegible]	F
163	60	F
163	61	F
160	64	T
163	64	T
165	61	T
165	62	T
165	65	T
168	62	T
168	63	T
168	66	T
170	63	T
170	64	T
170	68	T

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

Our Default Setup: Binary Classification for New Data – What Label?



Height (cm)	Weight (kg)	Large (vs Medium) t-shirt?
158	58	F
158	59	F
158	63	F
160	59	F
160	60	F
163	60	F
163	61	F
160	64	T
163	64	T
165	61	T
165	62	T
165	65	T
168	62	T
168	63	T
168	66	T
170	63	T
170	64	T
170	68	T

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

k-Nearest Neighbors (kNN)

To predict category label y of a new point $\boldsymbol{x}$ (classification):

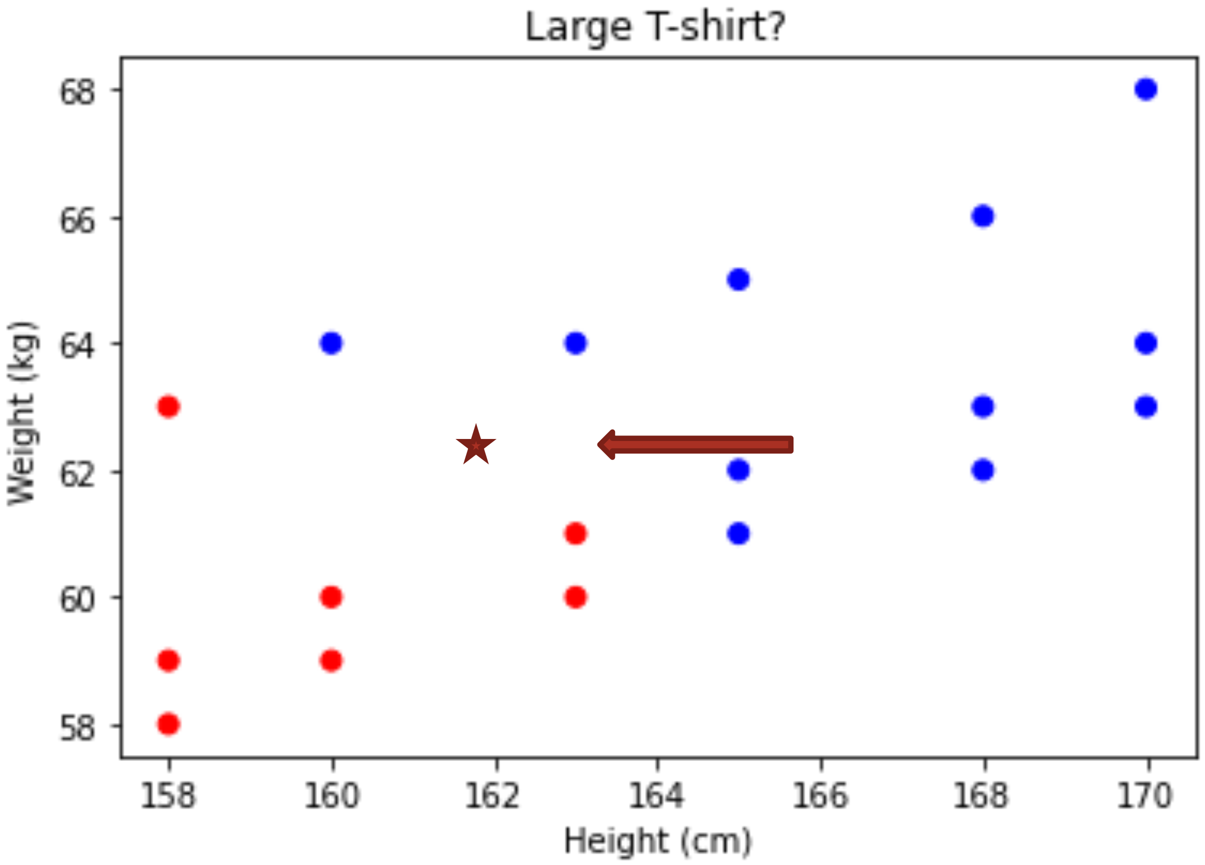
- Find k nearest neighbors (according to some distance metric)
- Assign the **majority label** to the new point

To predict numeric value y of a new point $\boldsymbol{x}$ (regression):

- Find k nearest neighbors
- "Average" the values associated with the neighbors

If we change k we may get a different prediction

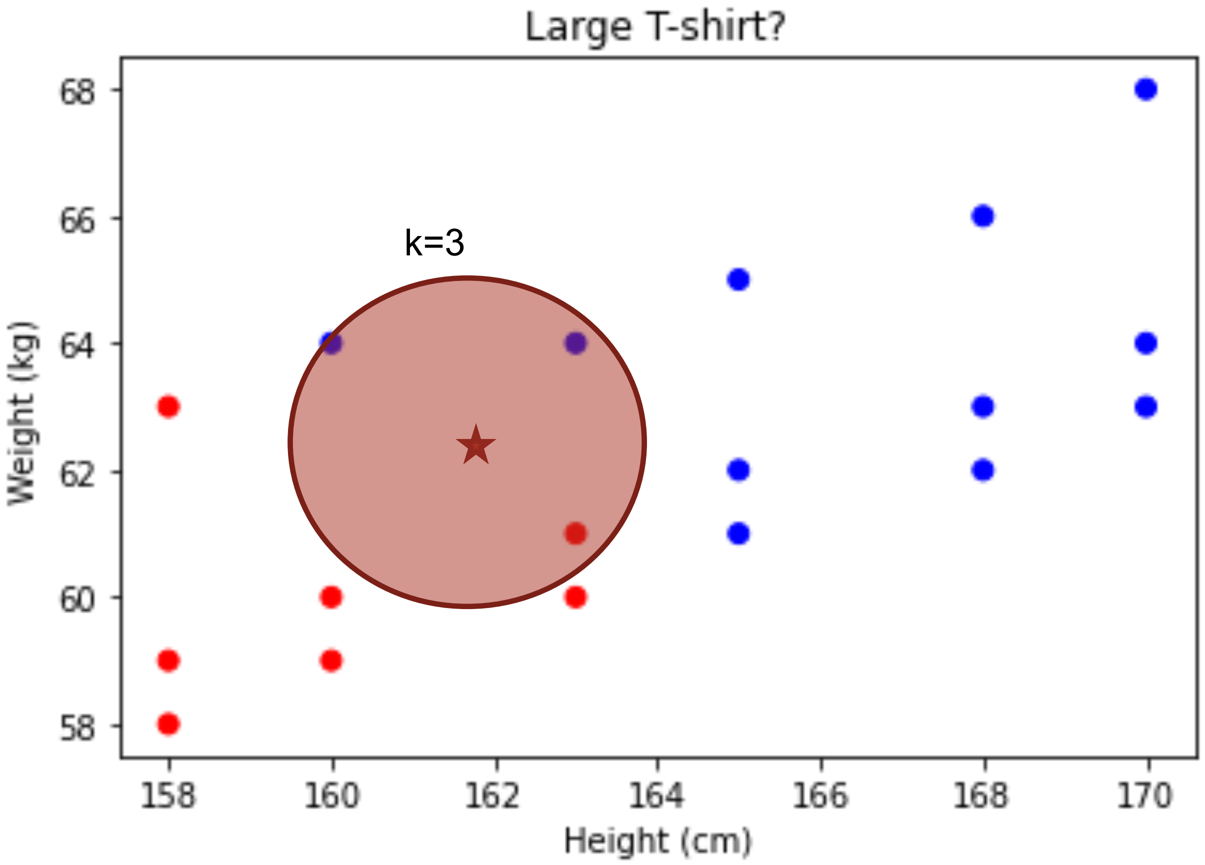
kNN Prediction: What Label?



Height (cm)	Weight (kg)	Large (vs Medium) t-shirt?
158	58	F
158	59	F
158	63	F
160	59	F
160	60	F
163	60	F
163	61	F
160	64	T
163	64	T
165	61	T
165	62	T
165	65	T
168	62	T
168	63	T
168	66	T
170	63	T
170	64	T
170	68	T

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

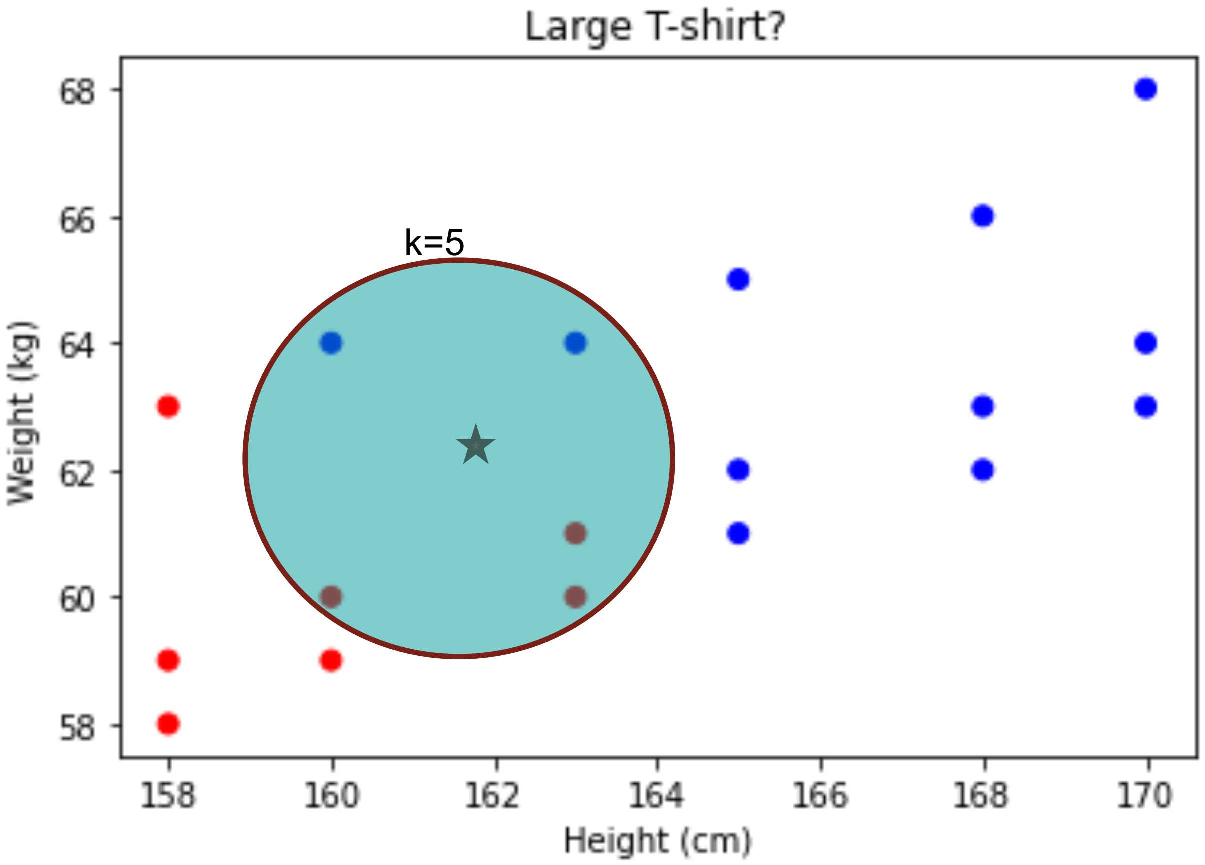
kNN Prediction: What Label?



Height (cm)	Weight (kg)	Large (vs Medium) t-shirt?
158	58	F
158	59	F
158	63	F
160	59	F
160	60	F
163	60	F
163	61	F
160	64	T
163	64	T
165	61	T
165	62	T
165	65	T
168	62	T
168	63	T
168	66	T
170	63	T
170	64	T
170	68	T

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

kNN Prediction: What Label?



Height (cm)	Weight (kg)	Large (vs Medium) t-shirt?
158	58	F
158	59	F
158	63	F
160	59	F
160	60	F
163	60	F
163	61	F
160	64	T
163	64	T
165	61	T
165	62	T
165	65	T
168	62	T
168	63	T
168	66	T
170	63	T
170	64	T
170	68	T

Based on data from https://www.listendata.com/2017/12/k-nearest-neighbor-step-by-step-tutorial.html

What Does "Nearest" Mean?

Must define a "distance function" between any two samples $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$

Note: boldface $\boldsymbol{x}$ denotes a vector in widely used notation. In our case, each of these is a 2D vector: $\boldsymbol{x_i} = [x_{i1}, x_{i2}]$

"Nearest neighbor" = sample with least "distance". Some commonly used distances:

$$\left(\sum_d (x_{1j} - x_{2j})^1\right)^{\frac{1}{1}}$$

ℓ_1 distance

$$\sum_d |x_{1j} - x_{2j}|$$

$$\left(\sum_d (x_{1j} - x_{2j})^2\right)^{\frac{1}{2}}$$

ℓ_2 distance

Also, "Euclidean" distance

$$\left(\sum_d (x_{1j} - x_{2j})^{\to\infty}\right)^{\to 0}$$

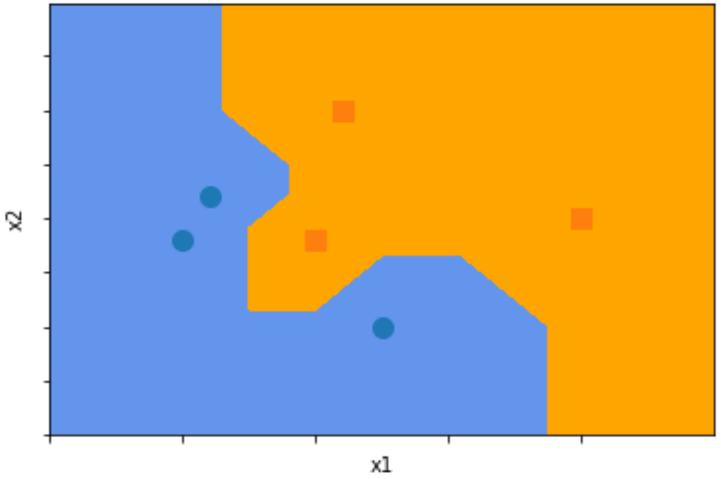
ℓ_∞ distance

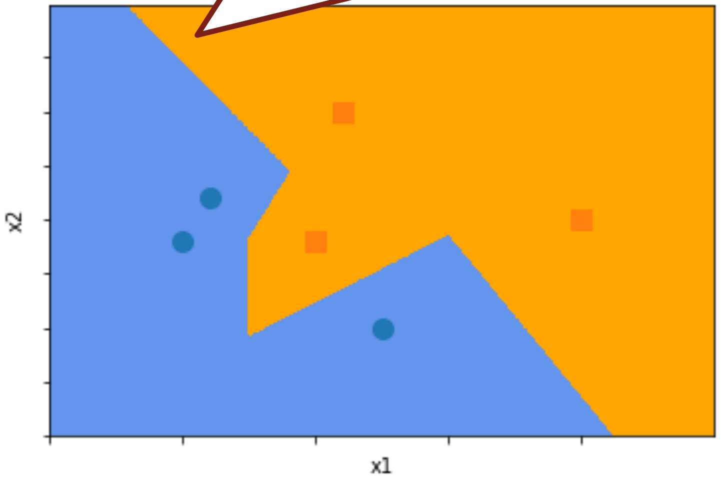
$$\max_d \, (x_{1j} - x_{2j})$$

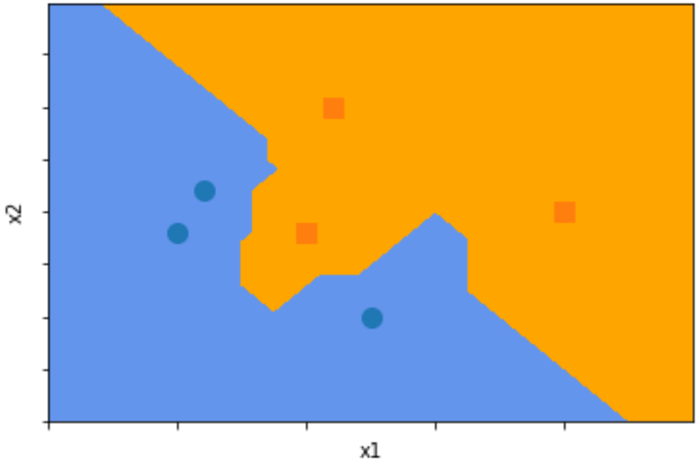
Different Distances Produce Different Outcomes

Fix k =1 neighbors

Classifier "decision boundary" plots show what class would be assigned at *every point* $\boldsymbol{x}$







ℓ_1 distance

$$\sum_d |x_{1j} - x_{2j}|$$

ℓ_2 distance

Also, "Euclidean" distance

ℓ_∞ distance

$$\max_d \, (x_{1j} - x_{2j})$$

What about Distances between Non-numeric Data? Consider Strings…

Hamming distance (number of characters that are different)

ABCDE vs AGDDF → 3

Edit distance (number of character inserts/replacements/deletes to go from one to the other)

ROBOT vs BOT → 2

Jaccard distance between sets $\frac{|A \cap B|}{|A \cup B|}$

between **n-grams** (n-character substrings of the strings, with (n-1) character padding)

$$ROBOT$$ vs $$BOT$$ → |{BOT,OT$,T$$}| / |{$$R,$RO,ROB,OBO,$$B,BO,BOT,OT,T$$}|

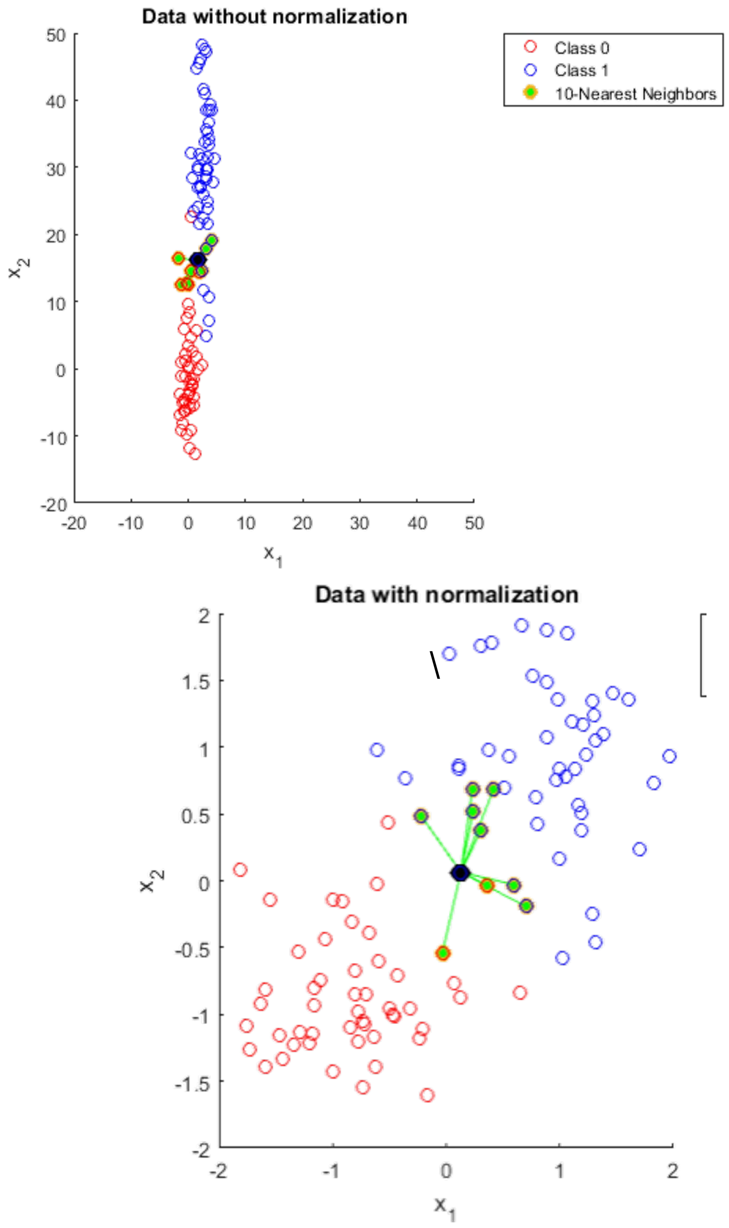
3 9

Beware: Feature Scaling affects Nearest Neighbors

Our previous study of linear / logistic regression:

- OLS regression was *scale-invariant*
- Regularization was affected by the scale of different features

Even more of a concern with kNN: note that we are using a distance measure like L2, which is affected dramatically by feature scales!



https://stats.stackexchange.com/questions/287425/why-do-you-need-to-scale-data-in-knn

What Happens If We Have Many Dimensions?

Predict y = acceleration of an object being pushed by a remote-controlled robot

- What if input features are:
  - x_1 = mass
  - x_2 = Force
  - x_3 = color of object
  - x_4 = temperature
  - x_5 = air pressure
  - x_6 = what the operator ate for breakfast that morning

…

As you add more irrelevant variables, distance functions, which are so critical for k-NN methods, get dominated by irrelevant dimensions in $\boldsymbol{x}$

General Problem: "Curse of Dimensionality"

Adding more dimensions makes lots of things weird and counterintuitive

e.g., the percentage of the volume of a D-dimensional sphere with radius r, that lies beyond $\ell_2$distance $0.99r$ from the center is:

- 3% at $D = 3$
- 63% at $D = 100$
- 99.99% at $D = 1000$

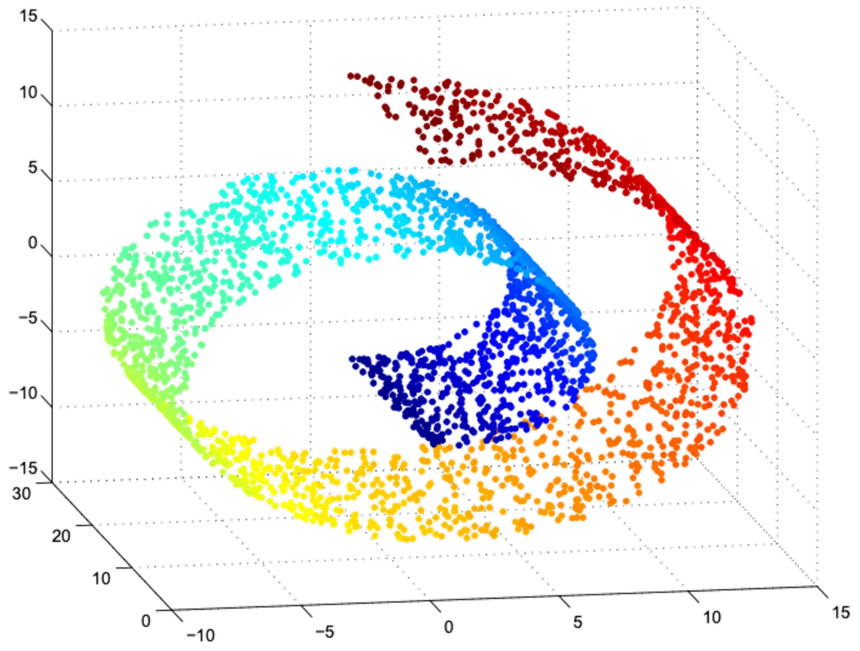
also, with enough dimensions most points are of roughly equal distance!

For k-NN, nearest neighbors become very far apart, and of similar distance – therefore **unreliable predictors**

General Advice …

Always worth trying k-nearest neighbors!

- It's so simple to code up that it's worth it.
- *Often* works surprisingly well, and is very widely used as a simple and reliable baseline, even with for really high-dimensional data



https://axon.cs.byu.edu/~martinez/classes/778/Papers/Manifold_Learning.pd

How Can We Scale kNN?

High D also makes it computationally expensive to compute neighbors.
Naively, must compute N distances between D-dimensional data pairs to compute neighbors before classifying a single new point. O(|training set||data set|)

Indexing

- Use kd-trees and other multidimensional indices to capture the training data
- Each lookup is O(log *n*) but on disk

Parallelism (e.g., PANDA, LBL)

- Use multiple cores / processors, and either compare against in-memory data or kd trees

Approximation

- Compare against a sample, not all of the training data
- See, e.g., https://www.kaggle.com/code/pawanbhandarkar/knn-vs-approximate-knn-what-s-the-difference/notebook

Stepping back…
where are the *parameters* we learn?

Think broadly of the "parameters" as everything required to produce the output, for a given model class. i.e.

Model class + parameters + new input $\boldsymbol{x} \rightarrow$ predicted y

"kNN classifier" ??

A: The full training dataset!

Funnily, methods like these where the parameters are either *the training data* itself, or *grow in size "automatically"* with the training data, are called "**non-parametric**" machine learning approaches.

Summary of k-Nearest Neighbors

A case of "non-parametric" learning

- Uses the full training dataset as parameters
- Requires careful treatment of feature scaling

- Main decisions: the value of k, the distance function

Tends to work well in practice. but beware scalability

Decision Trees

CIS 4190/5190 – Fall 2022

A Motivating Example, with Some Data

Need help modeling diabetes risks!

Over the years, I've collected data from lots of patients, recording their physical information, their demographic information, habits, and done their lab work to diagnose diabetes. I'm wondering now: from all this data, could I model the risk of other people with similar characteristics having diabetes given all this other information about them? And would your applied ML class be able to help? I've attached the data here for you to take a look.

Eventually, we'll want to explain our findings to patients, and point out any behavioral changes that would mitigate their risk for diabetes. Even if the risk factors we find are non-modifiable, insurance companies would be interested in understanding and estimating this risk. Either way, it'd be great to have something that we can understand and interpret well!

Diabetes Data

data matrix X

label y_i

sample x_i

SEQN	RIDAGEYR	BMXWAIST	BMXHT	LBXTC	BMXLEG	BMXWT	BMXBMI	RIDRETH1	BPQ020	ALQ120Q	DMDEDUC2	RIAGENDR	INDFMPIR	LBXGH	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	110.8	161.8	168.0	37.1	93.4	35.7	Non-Hispanic White	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	85.5	152.8	278.0	32.4	61.8	26.5	Non-Hispanic White	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	22.0	Non-Hispanic White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispanic White	no	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	28.0	90.3	175.1	198.0	40.5	92.2	30.1	Other or Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589	35.0	94.6	172.9	192.0	39.1	78.3	26.2	Non-Hispanic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73595	58.0	114.8	175.3	165.0	40.1	96.0	31.2	Other Hispanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73596	57.0	117.8	164.7	151.0	35.3	104.0	38.3	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73600	37.0	122.9	185.1	189.0	48.1	126.2	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	161.0	36.5	111.9	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

Data from NHANES 2013/14 survey

Diabetes Data

ID	AGE	WAIST	HE.	CHOLESTEROL	UPPER LEG LENGTH	WEIGHT	BMI	RACE	HIGH BP	ALCOHOL USE	EDUCATION	GENDER	FAMILY INCOME RATIO	GLYCOHAEM	DIABETIC
SEQN	RIDAGEYR	BMXWAIST	BMXHT	LBXTC	BMXLEG	BMXWT	BMXBMI	RIDRETH1	BPQ020	ALQ120Q	DMDEDUC2	RIAGENDR	INDFMPIR	LBXGH	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	[illegible]	[illegible]	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	[illegible]	no	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577	[illegible]	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Multi-Racial	no	4.0	some college or AA degree	male	[illegible]	5.0	no
73589	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	no	2.0	high school graduate / GED	[illegible]	1.74	5.5	no
73595	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	no	1.0	[illegible]	[illegible]	3.09	7.7	no
73596	57.0	117.8	164.7	151.0	35.3	104.0	38.3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.9	no
73600	37.0	122.9	185.1	189.0	48.1	126.2	36.8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.4	no
73607	75.0	130.5	169.6	161.0	36.5	111.9	38.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

columns X_j denote features

Patient number: should this really be a feature?

The diabetes test outcome: would make our ML pointless …

Data from NHANES 2013/14 survey

Data Dictionary

Data sets are often accompanied by a data dictionary that describes each feature

It is critical to understand the data before analyzing it!

The dictionary for our data: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx

ID (SEQN)	AGE (RIDAGEYR)	WAIST_CIRCUM (BMXWAIST)	HEIGHT (BMXHT)	CHOLESTEROL (LBXTC)	UPPER_LEG_LEN (BMXLEG)	WEIGHT (BMXWT)	BMI (BMXBMI)	RACE (RIDRETH1)	HIGH_BP (BPQ020)	ALCOHOL_USE (ALQ120Q)	EDUCATION (DMDEDUC2)	GENDER (RIAGENDR)	FAMILY_INCOME_RATIO (INDFMPIR)	GLYCOHEMOGLOBIN (LBXGH)	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	[illegible]	[illegible]	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no

777 = refused; 999 = don’t know

Data from NHANES 2013/14 survey

numeric

nominal

ordinal

binary

ID	AGE	WAIST	HE.	CHOLESTEROL	UPPER LEG LENGTH	WEIGHT	BMI	RACE	HIGH BP	ALCOHOL USE	EDUCATION	GENDER	FAMILY INCOME RATIO	GLYCOHAEM	DIABETIC
SEQN	RIDAGEYR	BMXWAIST	BMXHT	LBXTC	BMXLEG	BMXWT	BMXBMI	RIDRETH1	BPQ020	ALQ120Q	DMDEDUC2	RIAGENDR	INDFMPIR	LBXGH	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	110.8	161.8	168.0	37.1	93.4	35.7	Non-Hispanic White	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	85.5	152.8	278.0	32.4	61.8	26.5	Non-Hispanic White	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	22.0	Non-Hispanic White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispanic White	no	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	28.0	90.3	175.1	198.0	40.5	92.2	30.1	Other or Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589	35.0	94.6	172.9	192.0	39.1	78.3	26.2	Non-Hispanic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73595	58.0	114.8	175.3	165.0	40.1	96.0	31.2	Other Hispanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73596	57.0	[illegible]	[illegible]	[illegible]	35.3	[illegible]	[illegible]	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73600	37.0	122.9	185.1	189.0	48.1	[illegible]	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	[illegible]	[illegible]	[illegible]	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

This column seems binary, but also has "refused to answer" and "don't know" categories

Data from NHANES 2013/14 survey

Deciding on a Diagnosis / Prediction

How do we train a human to make a diagnosis?

- Often, a kind of flowchart based on tests! "Decision Tree"

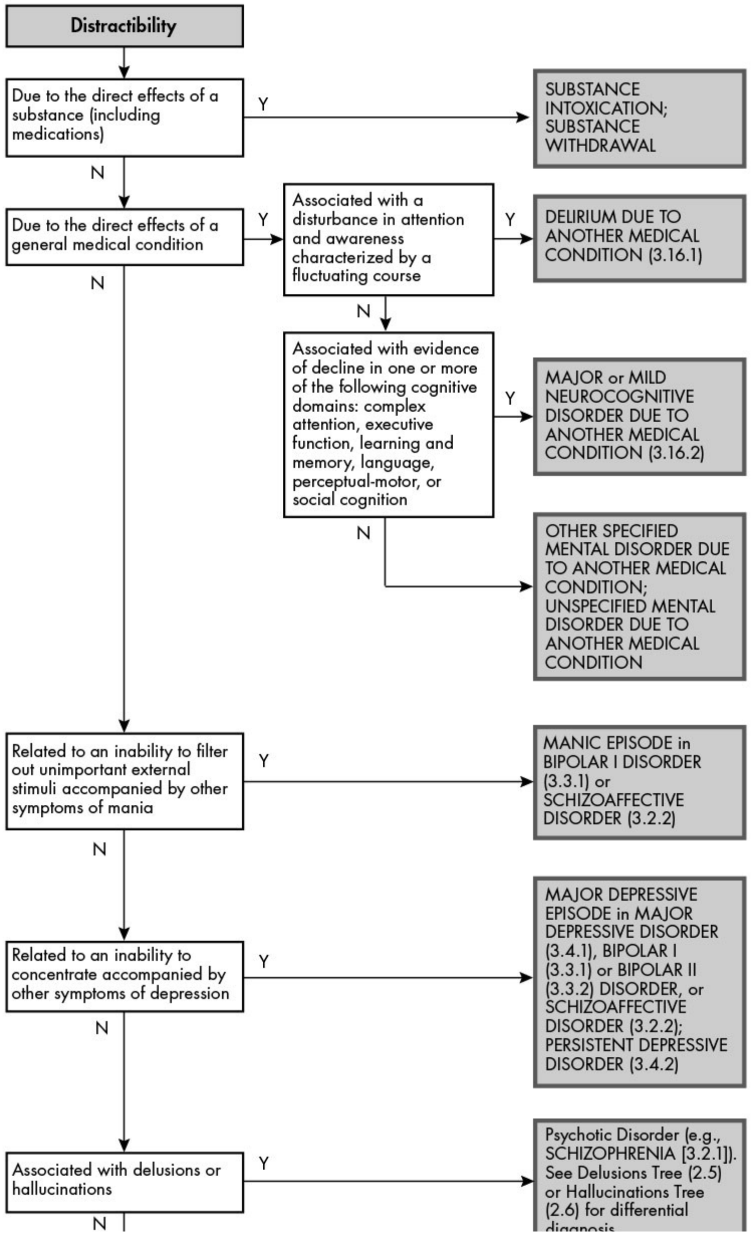
  e.g., how we train psychiatrists to make diagnoses?

- "Explainable" in a clear way, easy to evaluate

Idea: Let's create decision trees computationally! (ie learn them)

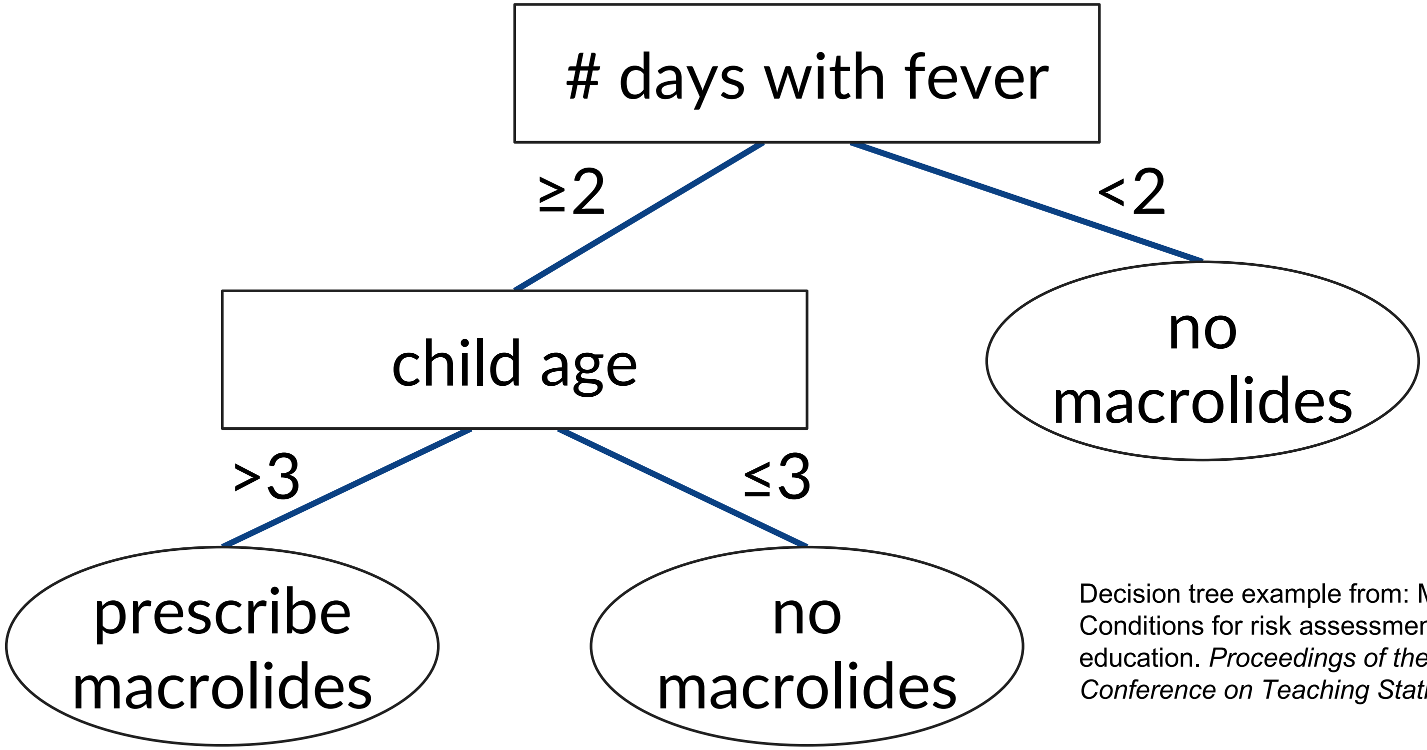
First: let's formalize what we mean by a decision tree…

APA DSM Library



Decision Trees for Humans

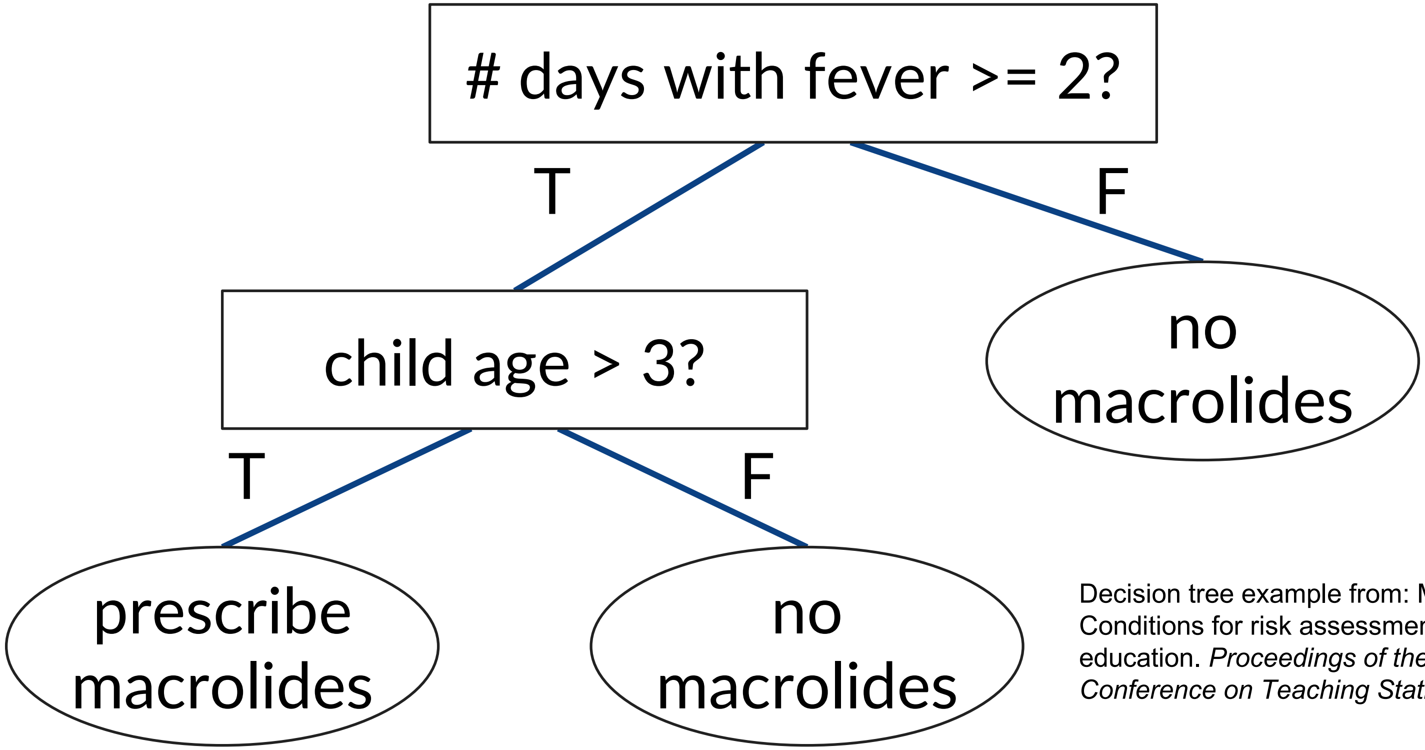
Simple decision tree used in medicine:



Decision tree example from: Martignon and Monti. (2010). Conditions for risk assessment as a topic for probabilistic education. *Proceedings of the Eighth International Conference on Teaching Statistics* (ICOTS8).

A Decision Tree Based on Boolean Tests

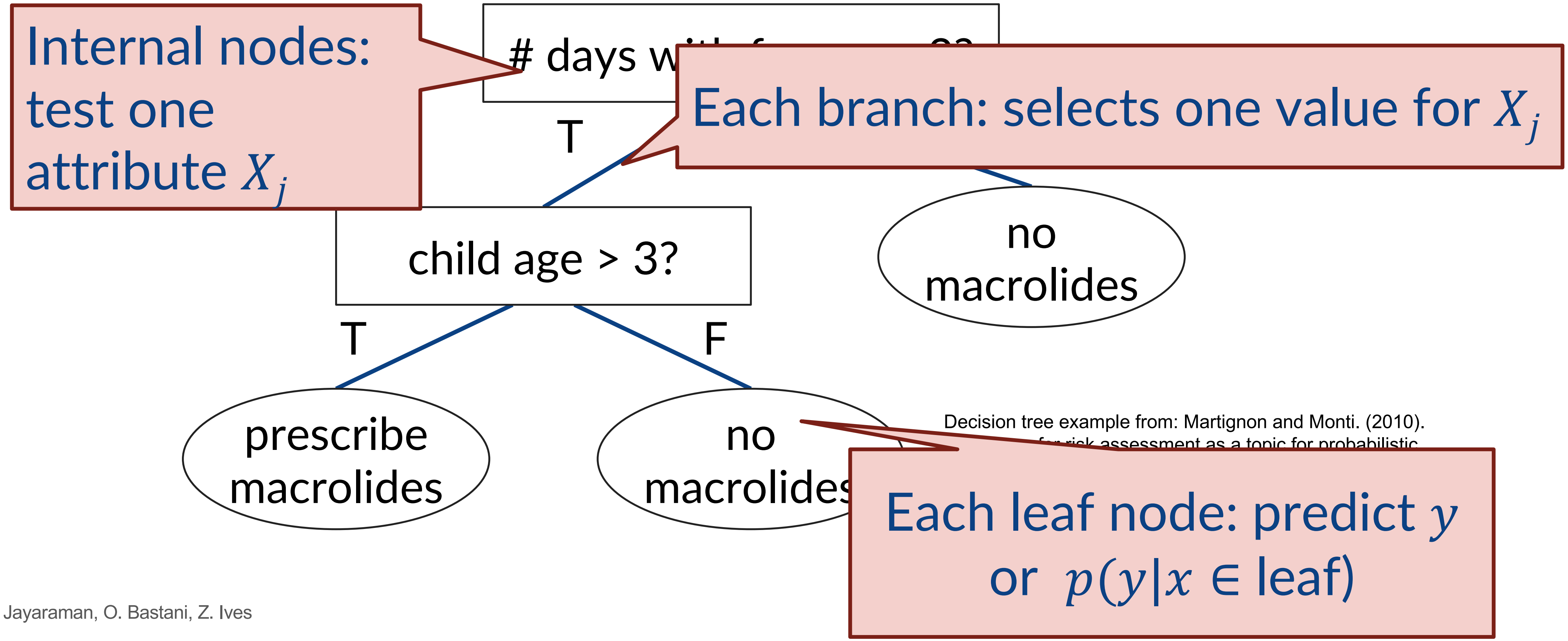
For continuous features, we'll restrict our study to internal nodes that can test the **value of one attribute**. We can generalize to categorical values (binary decision tree).



Decision tree example from: Martignon and Monti. (2010). Conditions for risk assessment as a topic for probabilistic education. *Proceedings of the Eighth International Conference on Teaching Statistics* (ICOTS8).

A Decision Tree Based on Boolean Tests

For continuous features, we'll restrict our study to internal nodes that can test the **value of one attribute**. We can generalize to categorical values (binary decision tree).



A Decision Tree Interior Node “Splits” Training Data

ColorOfCoat	TypeOfHorse
black	thoroughbred
bay	Arabian
black	thoroughbred
chestnut	quarter
black	Arabian

N=5; 3 classes

ColorOfCoat = ‘black’

ColorOfCoat	TypeOfHorse
black	thoroughbred
black	thoroughbred
black	Arabian

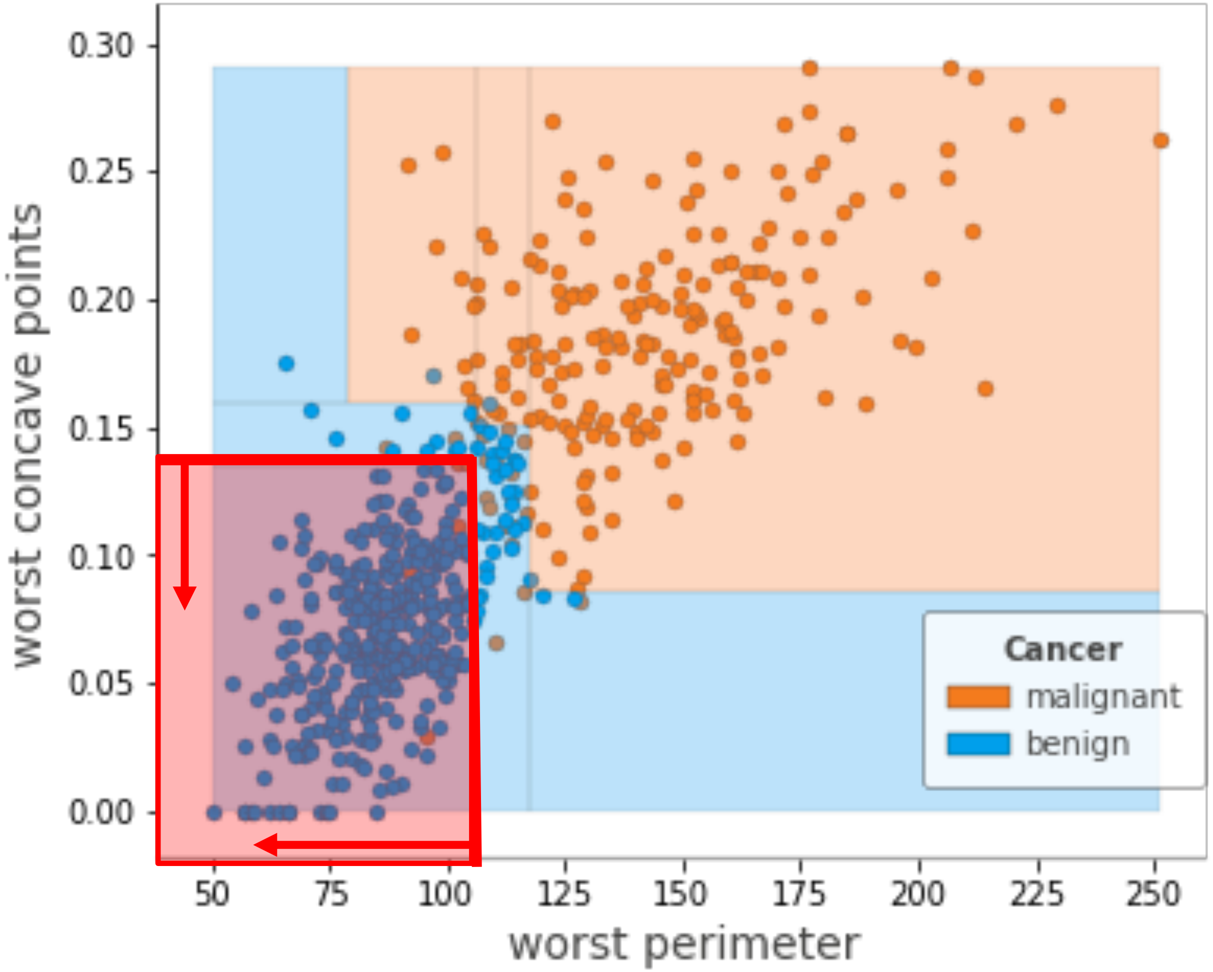
N=3; 2 classes

ColorOfCoat	TypeOfHorse
bay	Arabian
chestnut	quarter

N=2; 2 classes

More Generally: Decision Tree Induces a *Partition*

```
|--- worst perimeter <= 105.95
|  |--- worst concave points <= 0.135
|  |  |--- class: benign
|  |--- worst concave points > 0.135
|  |  |--- worst concave points < 0.16
|  |  |  |--- class: benign
|  |  |--- worst concave points > 0.16
|  |  |  | --- worst perimeter > 80
|  |  |  |  | --- class: malignant
|  |  |  | --- worst perimeter < 80
|  |  |  |  | --- class: benign
                              …
                              …
```



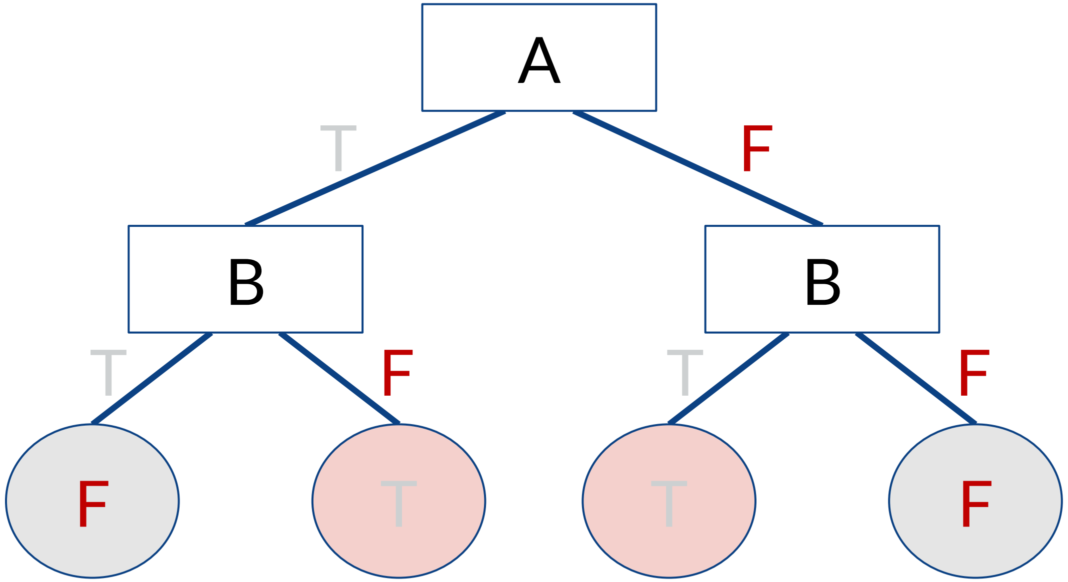
So what is the hypothesis class expressed by a DT?

Decision trees divide the feature space into axis-aligned "hyperrectangles"

Decision Trees and Boolean Tests

Decision trees can represent **any Boolean function** of the features

A	B	A xor B
T	T	F
T	F	T
F	T	T
F	F	F



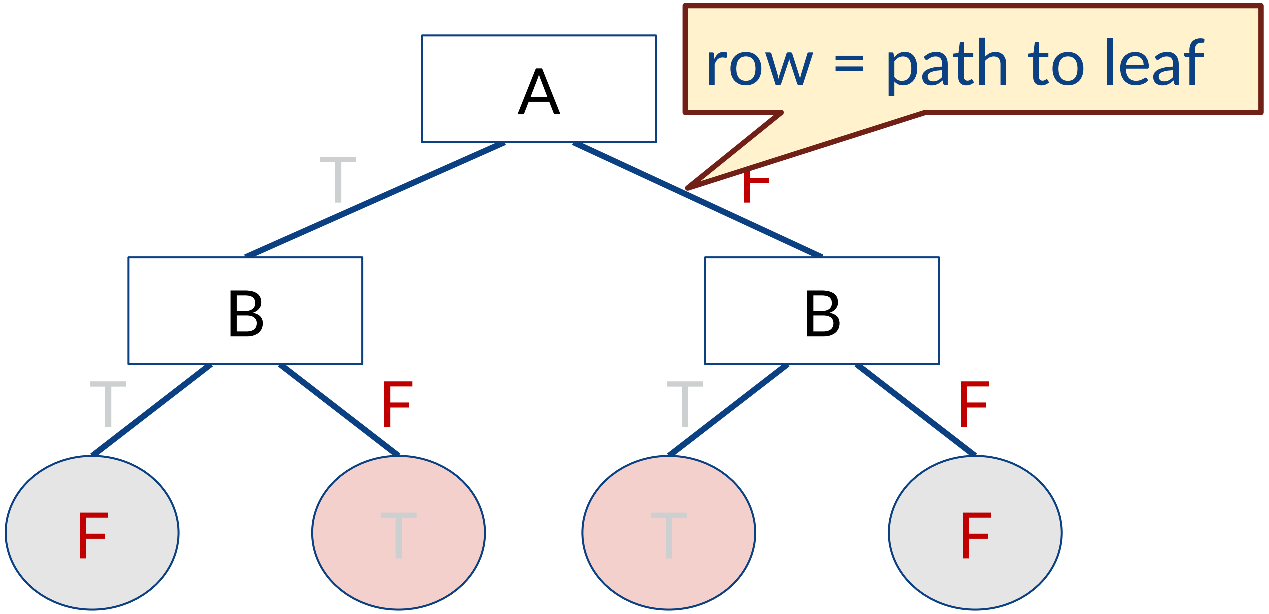
In the worst case, the tree will require **exponentially** many nodes

Decision Trees and Boolean Tests

Decision trees can represent **any Boolean function** of the features

A	B	A xor B
T	T	F
T	F	T
F	T	T
F	F	F

In the worst case, the tree will require **exponentially** many nodes

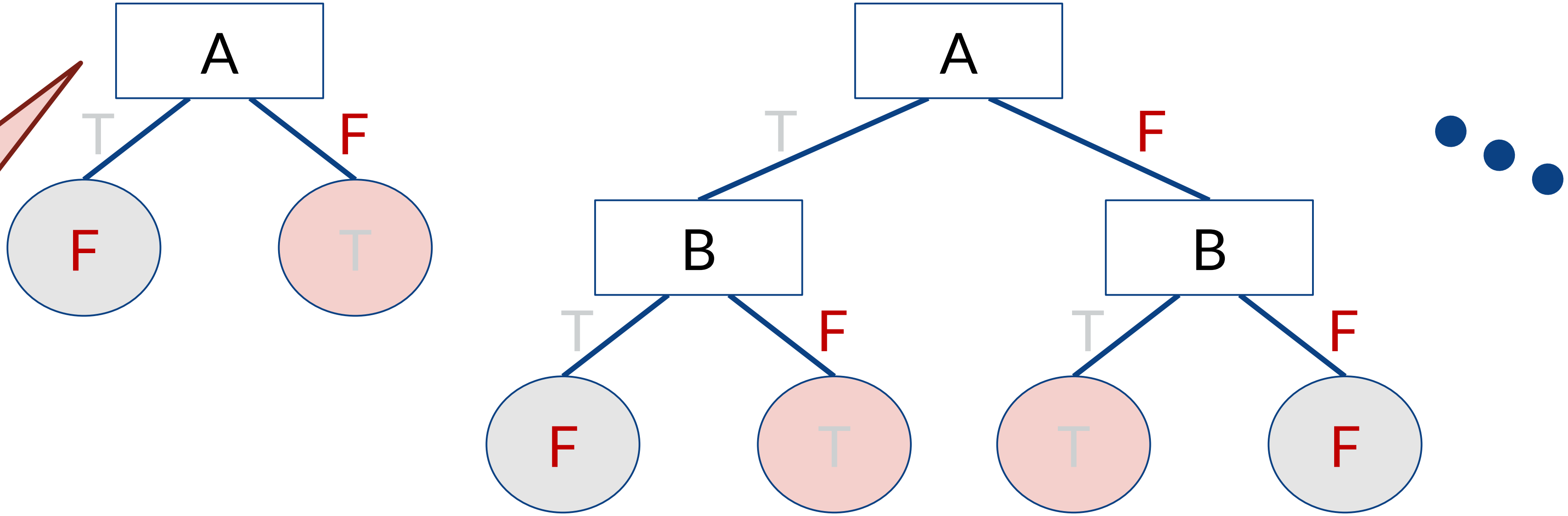


Decision Trees and Boolean Tests

DTs have a **variable-sized hypothesis space** based on their **depth**

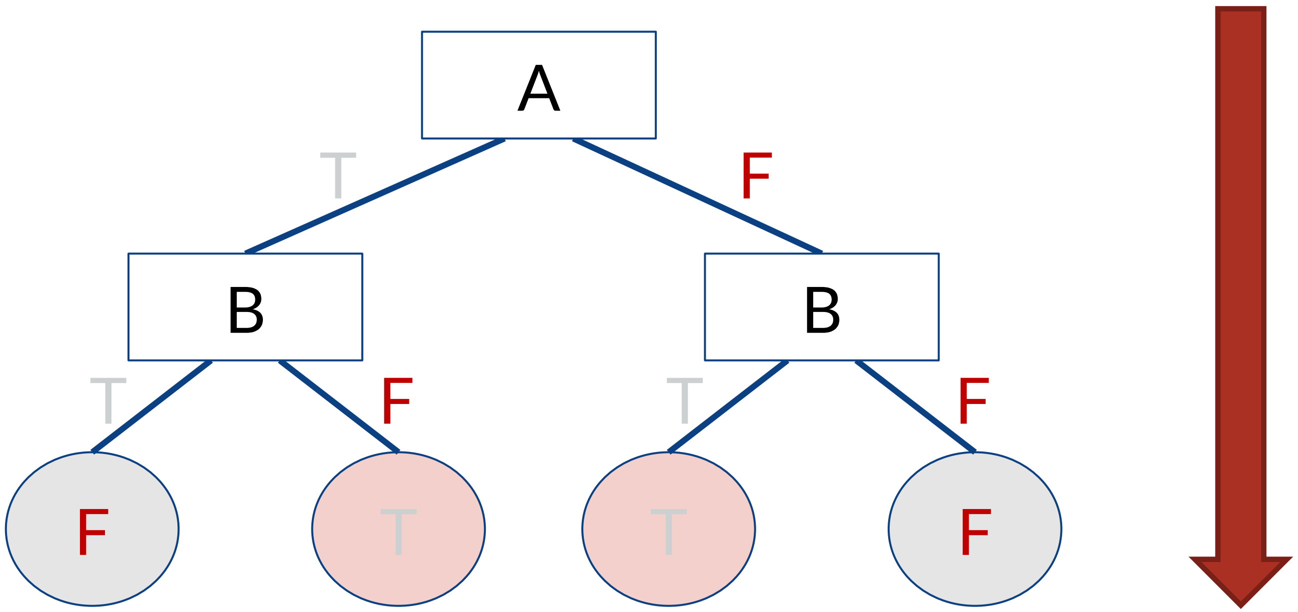
- Depth 1: any Boolean function based on one feature
- Depth 2: any Boolean function based on two features
- . . .

DTs of depth 1 are also called decision stumps



Training Decision Trees

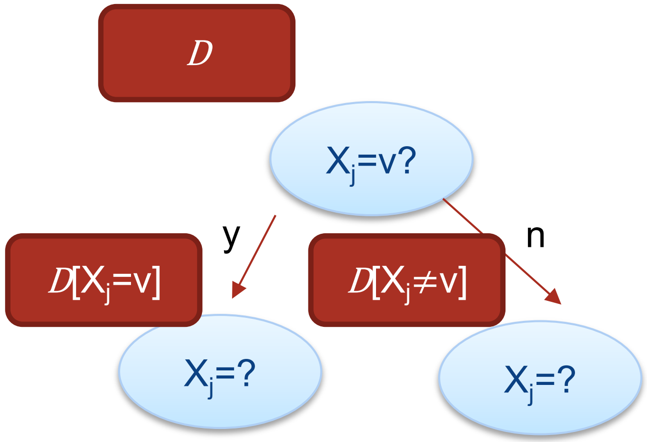
Decision Tree Training – Grow Top-Down



Top-Down Decision Tree Induction
[ID3 (1986), C4.5(1993) by Quinlan]

Let $\mathcal{D}$ be a set of labeled instances; $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} = [X_{N\times D}, \boldsymbol{y}_{N\times 1}]$

Let $\mathcal{D}[X_j = v]$ be the subset of $\mathcal{D}$ where feature X_j has value v

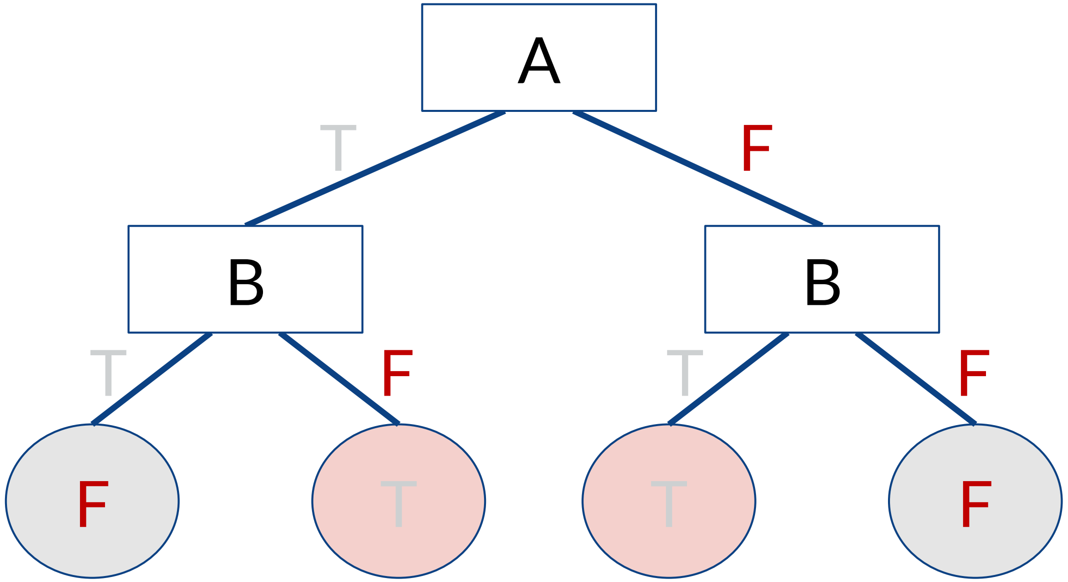


function `train_tree(`$\mathcal{D}$`)`

1. If data $\mathcal{D}$ **all have the same label** y, return new `leaf_node(`y`)`, else:
2. Pick the "best" feature X_j to partition $\mathcal{D}$
3. Set `node = new decision_node(`X_j`)`
4. For each value v that X_j can take

   Recursively create a new child `train_tree(`$\mathcal{D}[X_j = v]$`)` of `node`
5. Return `node`

Top-Down Decision Tree Training



Do we think this is going to be optimal, or greedy?

Top-Down Decision Tree Induction

[ID3 (1986), C4.5(1993) by Quinlan]

Let $\mathcal{D}$ be a set of labeled instances; $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} = [X_{N\times D}, \boldsymbol{y}_{N\times 1}]$

Let $\mathcal{D}[X_j = v]$ be the subset of $\mathcal{D}$ where feature X_j has value v

How do we choose which feature is best?

function `train_tree(`$\mathcal{D}$`)`

1. If data $\mathcal{D}$ **all have the same label** y, return new `leaf_node(`y`)`, else:

2. Pick the "best" feature X_j to partition $\mathcal{D}$
3. Set `node = new decision_node(`X_j`)`
4. For each value v that X_j can take

   Recursively create a new child `train_tree(`$\mathcal{D}[X_j = v]$`)` of `node`
5. Return `node`

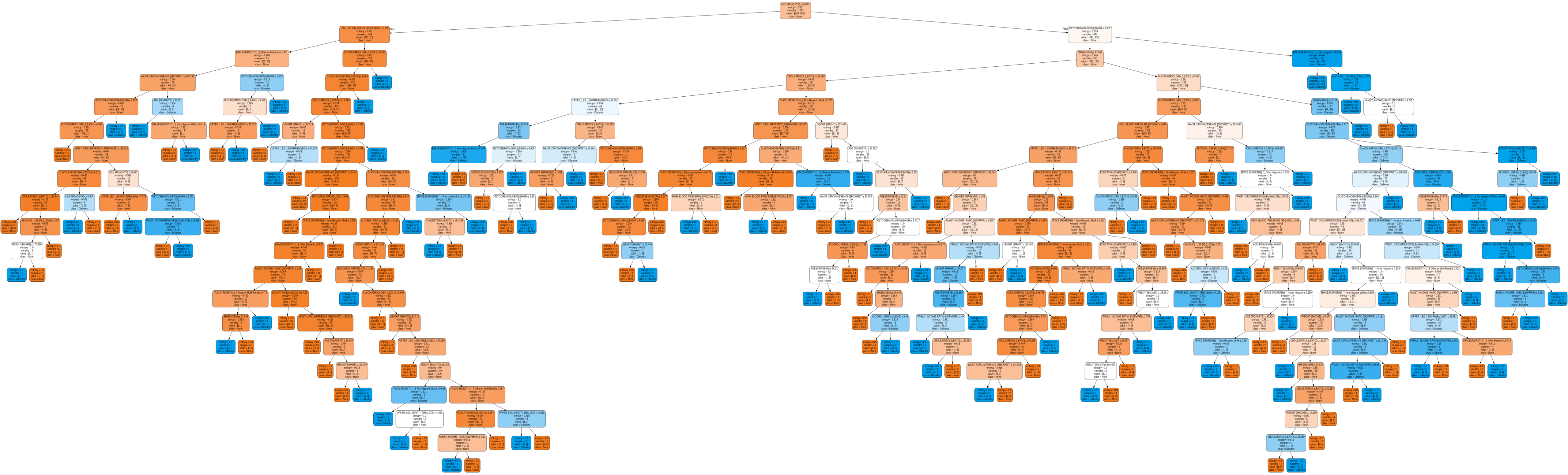
Choosing the Best Feature

Key problem: how should we choose which feature to split the data?

Possibilities:

Random	
Choose any feature at random	?

DT to Predict Diabetes – Random Features



Is this really the best way to choose decision nodes?

What Might be Better?

Learning Bias: Occam’s Razor

Principle stated by William of Ockham (1285-1347)

- “non sunt multiplicanda entia praeter necessitatem” -- ”entities are not to be multiplied beyond necessity”

- also called Ockham’s Razor, Law of Economy, or Law of Parsimony

Key Idea: The simplest consistent explanation is the best

https://en.wikipedia.org/wiki/William_of_Ockham

Choosing the Best Feature

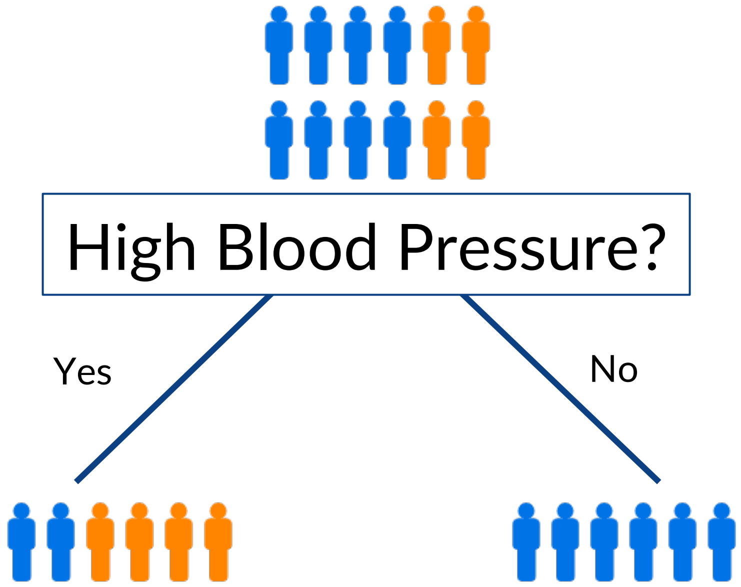
Key problem: how should we choose which feature to split the data?

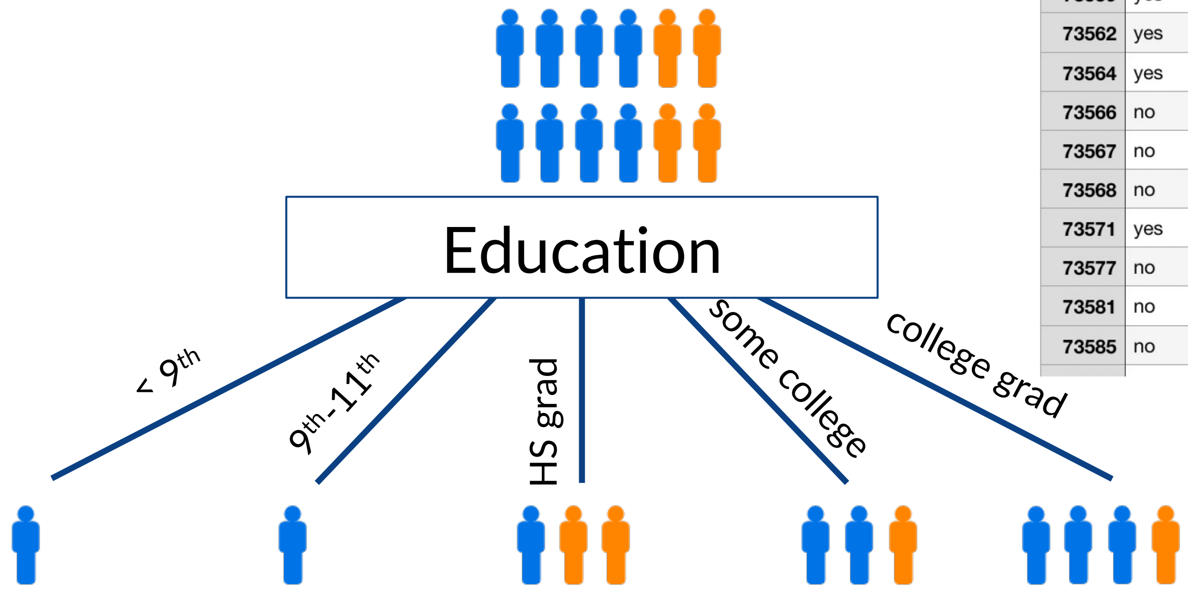
Random	Least-Values	Most-Values	Max-Gain
Choose any feature at random	Choose the feature with the fewest possible values	Choose the feature with the most possible values	Choose the feature with the largest expected *information gain*

i.e., the feature that is expected to result in the shortest subtree

Choosing Features for Short Decision Trees

Key Idea: good features partition the data into subsets that are either "all positive" or "all negative" (ideally)





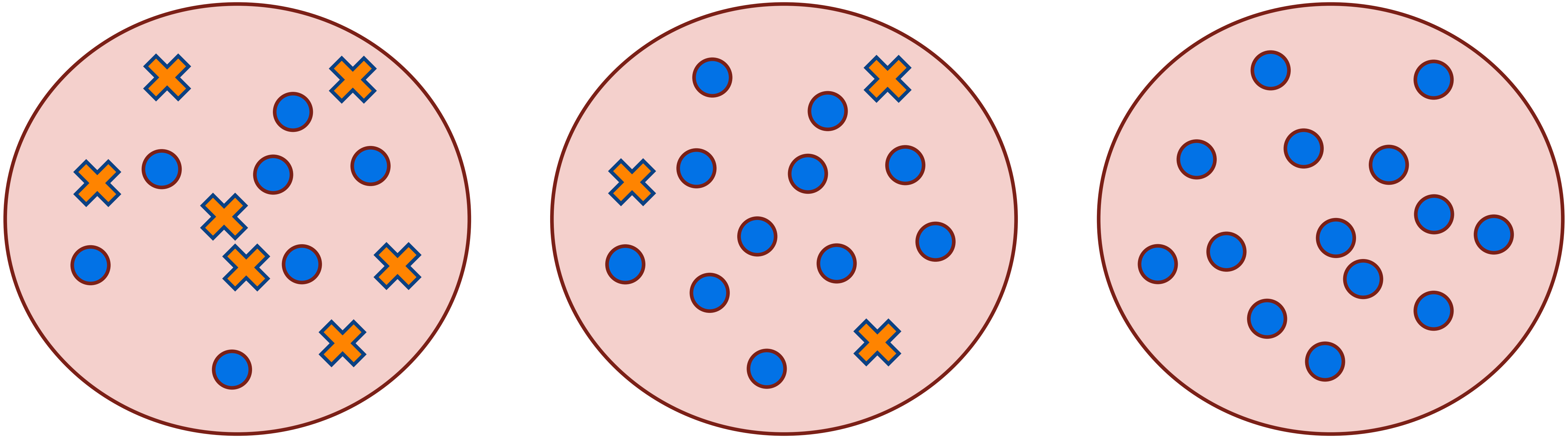
Subset of Data

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no

Which split is more informative?

Formalizing this: Impurity

Could we come up with an "impurity function" of a set of samples?



maximally impure ⟷ minimally impure

Note: All x's is also "pure"

A Candidate For An "Impurity Function": *Entropy*

Let Y be any discrete random variable that can take on *n* values

The entropy of Y is given by

$$H(Y) = -\sum_{i=1}^{n} P(Y=i)\log_2 P(Y=i)$$

Strictly, the entropy $H(Y)$ maps from a probability distribution (over the class label random variable Y) to an impurity score

$\updownarrow$

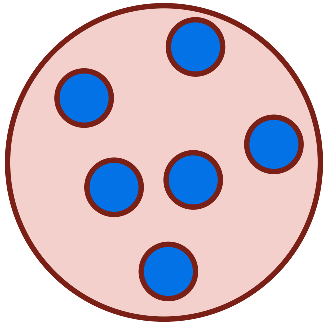
We'll denote $H(\mathcal{D})$ to map from a data subset $\mathcal{D}$ to the impurity score, by setting probability distribution $\approx$ distribution of labels Y in $\mathcal{D}$

Entropy of Binary Classes

Entropy $H(\mathcal{D}) = -\sum_c P(Y=c)\log_2 P(Y=c)$,
where different $c's$ correspond to different class labels

Min Impurity

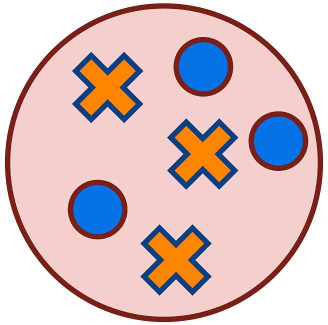
All instances in same class



$$H(\mathcal{D}) = -1\log 1 = 0$$

Max Impurity

Instances split evenly among classes



$$H(\mathcal{D}) = -0.5\log 0.5 - 0.5\log 0.5 = 1$$

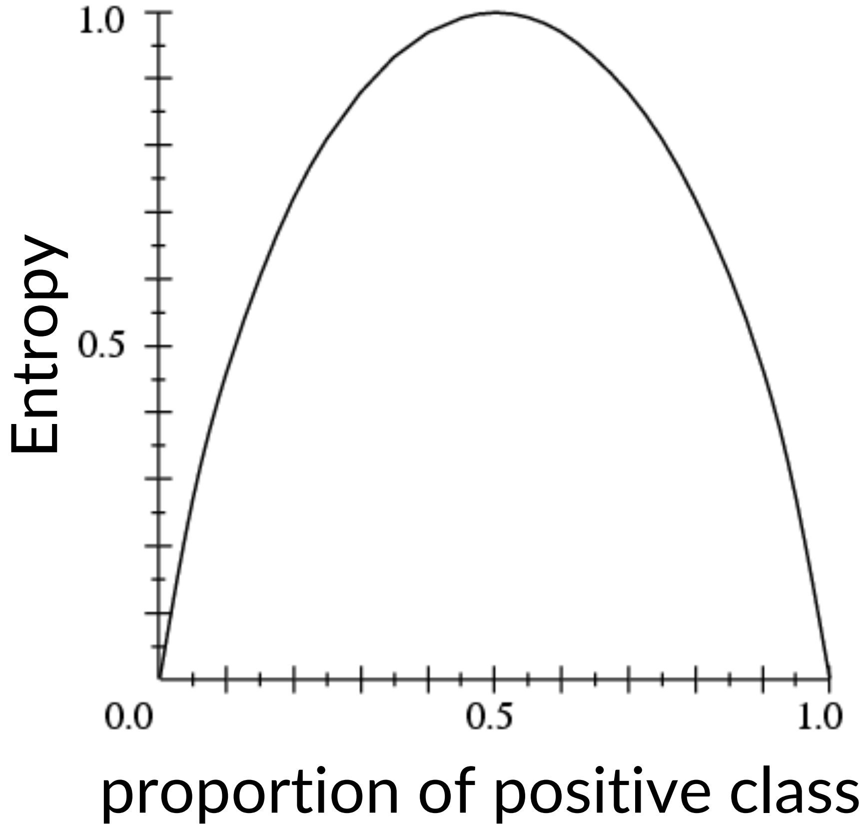
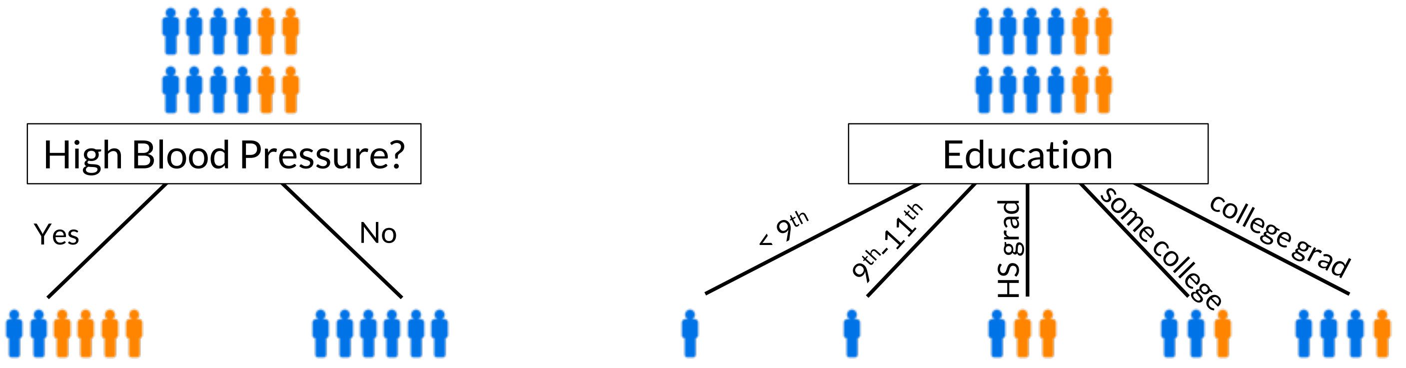


Image: Tom Mitchell

Choosing Features for Short Decision Trees



Recall: Ask questions such that the answers will reduce impurity in child nodes

When considering splitting on attribute / feature X_j,

- Need to estimate the "**expected** drop in impurity" after "getting the answer"/partitioning the data
- "Information Gain" based on our entropy function:

$$\mathrm{IG}(\mathcal{D}, X_j) = H(\mathcal{D}) - \sum_v H(\mathcal{D}[X_j = v])P(X_j = v)$$

Information Gain

Entropy $H(\mathcal{D}) = -\sum_c P(Y = c) \log_2 P(Y = c)$,
where different $c's$ correspond to different class labels

$$\text{IG}(\mathcal{D}, X_j) \;= H(\mathcal{D}) - \textstyle\sum_v H(\mathcal{D}[X_j = v])P(X_j = v)$$

The second term is sometimes called the "conditional entropy":

$$H(\mathcal{D}|X_j) = \sum_v H(\mathcal{D}[X_j = v])P(X_j = v)$$

The information gain may then also be written as:

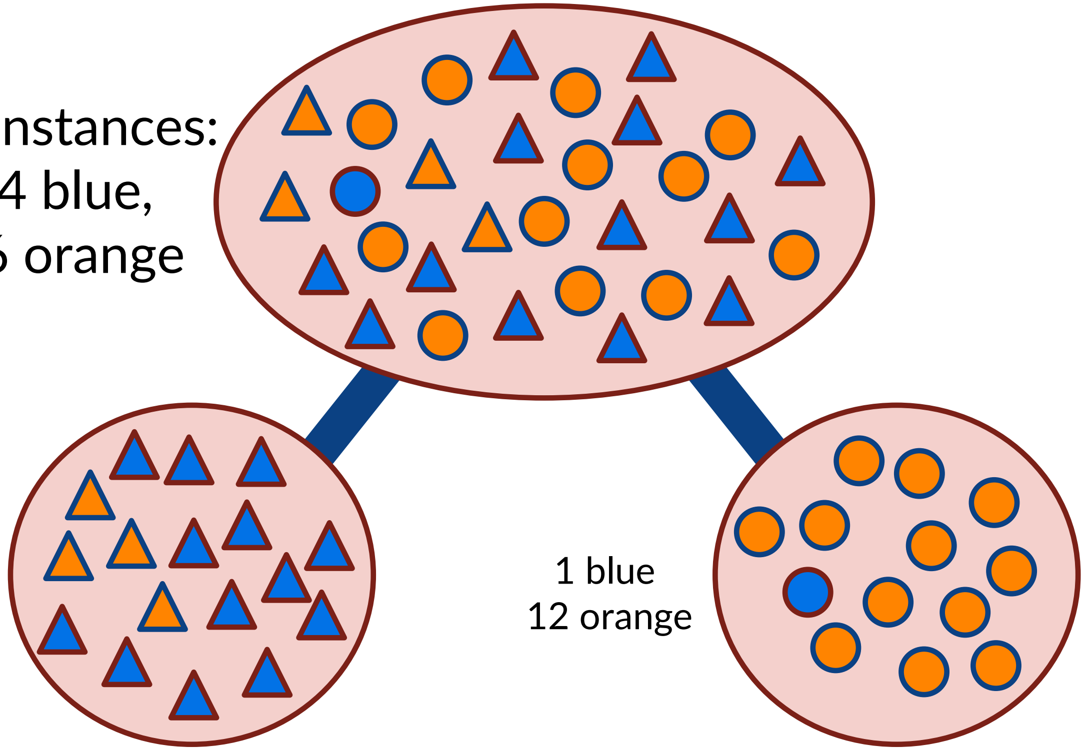
$$IG(\mathcal{D}, X_j) = H(\mathcal{D}) - H(\mathcal{D}|X_j)$$

Example IG Calculation

$$H(\mathcal{D}) = -\sum_c P(Y=c)\log_2 P(Y=c),$$

$$\text{IG}(\mathcal{D}, X_j) = H(\mathcal{D}) - \sum_v H(\mathcal{D}[X_j = v])P(X_j = v)$$

30 instances:
14 blue,
16 orange



13 blue
4 orange

1 blue
12 orange

H(parent) =

$$-\left(\frac{14}{30}\log_2\frac{14}{30}\right) - \left(\frac{16}{30}\log_2\frac{16}{30}\right)$$
$$= 0.996$$

weighted_mean(H(children)) =

$$\frac{17}{30}\cdot 0.787 + \frac{13}{30}\cdot 0.391$$
$$= 0.615$$

H(child)) =

$$-\left(\frac{13}{17}\log_2\frac{13}{17}\right) - \left(\frac{4}{17}\log_2\frac{4}{17}\right)$$
$$= 0.787$$

H(child)) =

$$-\left(\frac{1}{13}\log_2\frac{1}{13}\right) - \left(\frac{12}{13}\log_2\frac{12}{13}\right)$$
$$= 0.391$$

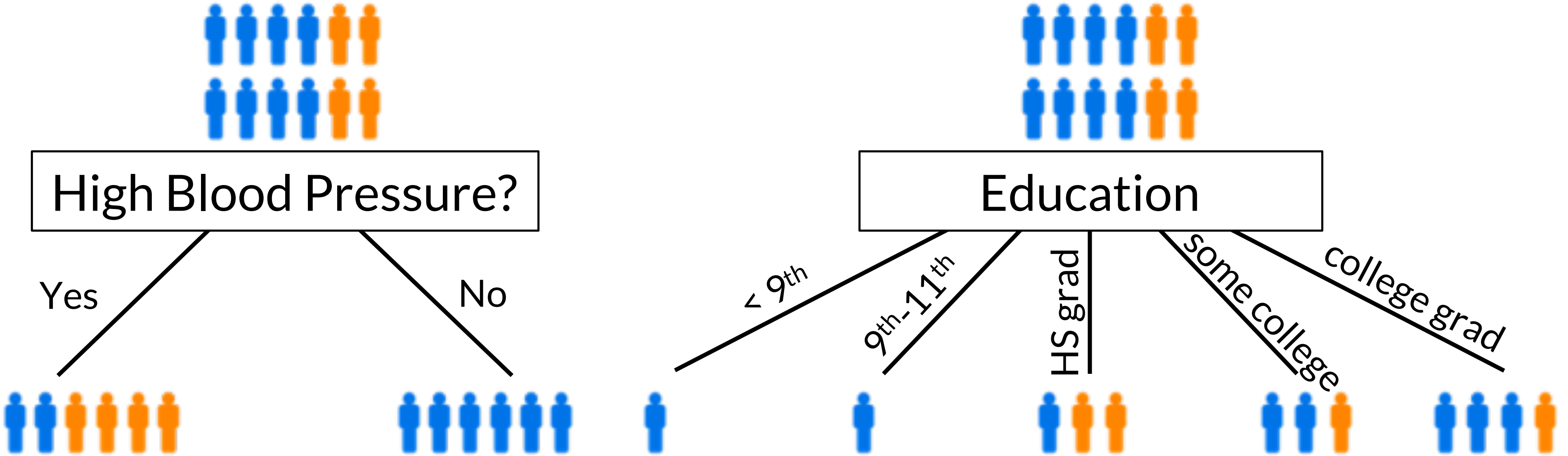
IG = 0.996 – 0.615 = $\boxed{0.381}$

Based on example by Pedro Domingos

Returning to the Diabetes Example Use Case

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no

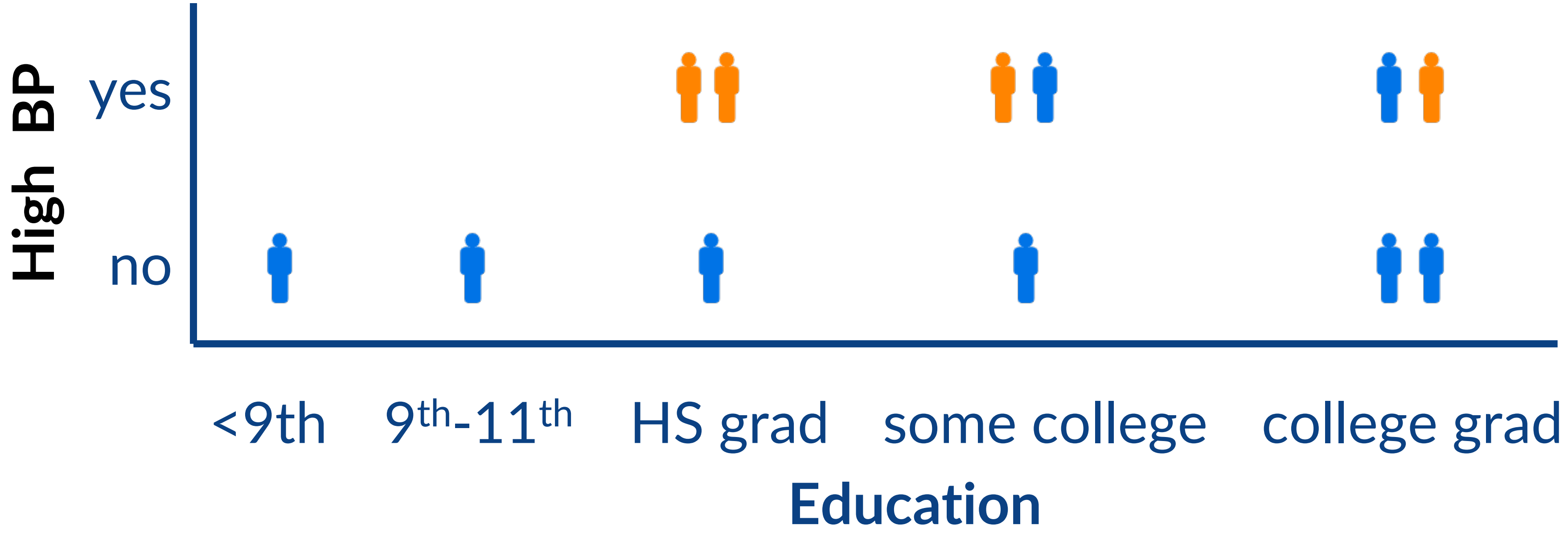
Which split is more informative?



Now we can solve it computationally via information gain

Information Gain Example for Diabetes

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



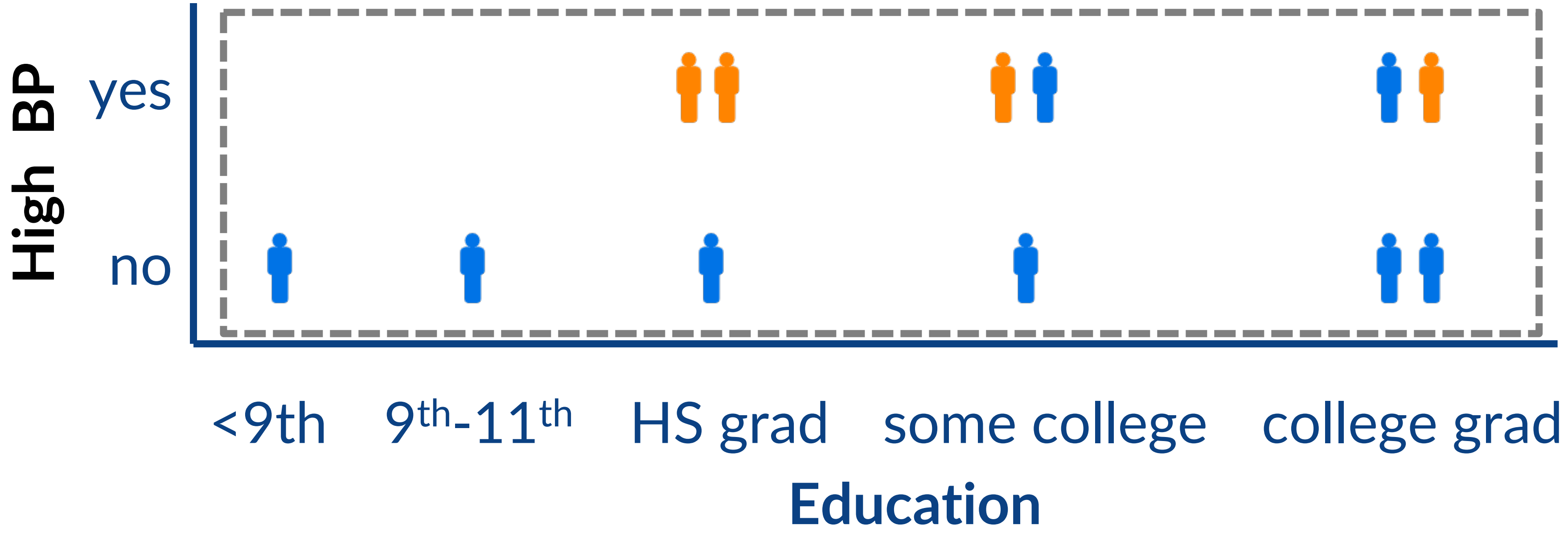
Need to compute:

$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D}\,|High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D}|\,Education)$$

Information Gain Example for Diabetes

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Need to compute:

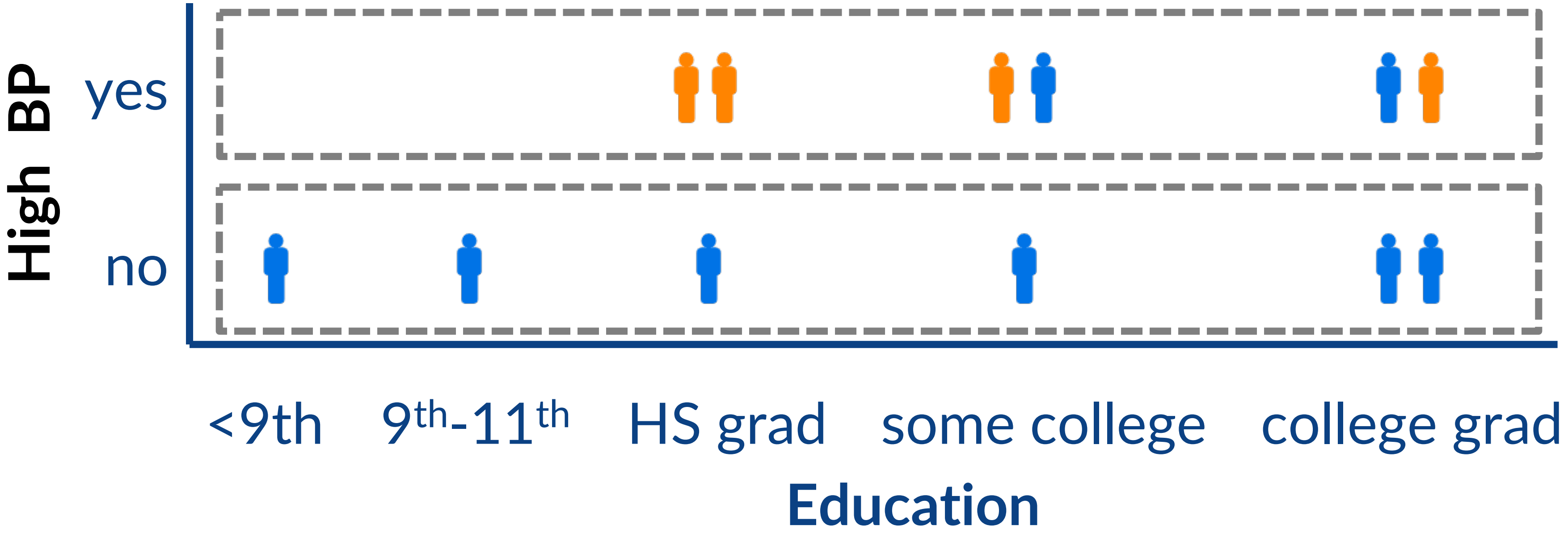
$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D}|High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D}|\ Education)$$

$H(\mathcal{D})$ = – 4/12 lg 4/12
– 8/12 lg 8/12
= 0.918

Information Gain Example for Diabetes

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Need to compute:

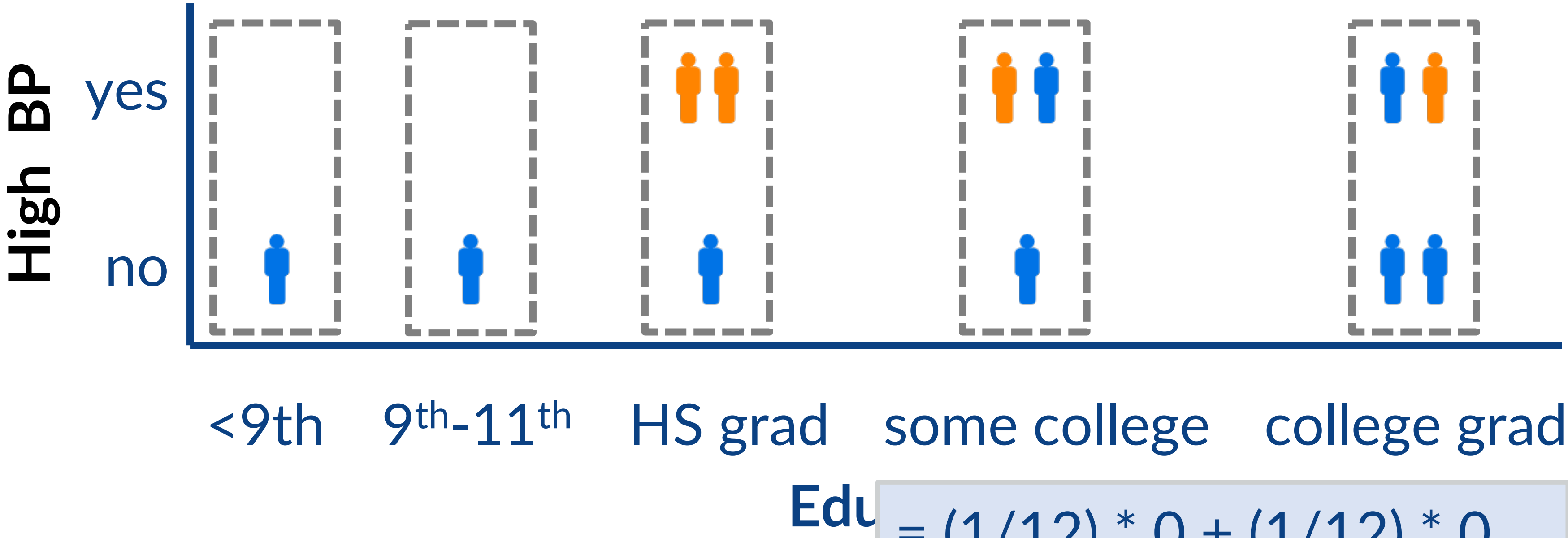
$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D}\,|High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D}|\,Education)$$

= (6/12) * (-2/6 lg 2/6
- 4/6 lg 4/6)
+ (6/12) * (0)
= 0.459

Information Gain Example for Diabetes

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



$$= (1/12) * 0 + (1/12) * 0$$
$$+ (3/12) * (-1/3 \lg 1/3 - 2/3 \lg 2/3)$$
$$+ (3/12) * (-2/3 \lg 2/3 - 1/3 \lg 1/3)$$
$$+ (4/12) * (-3/4 \lg 3/4 - 1/4 \lg 1/4)$$
$$= 0.730$$

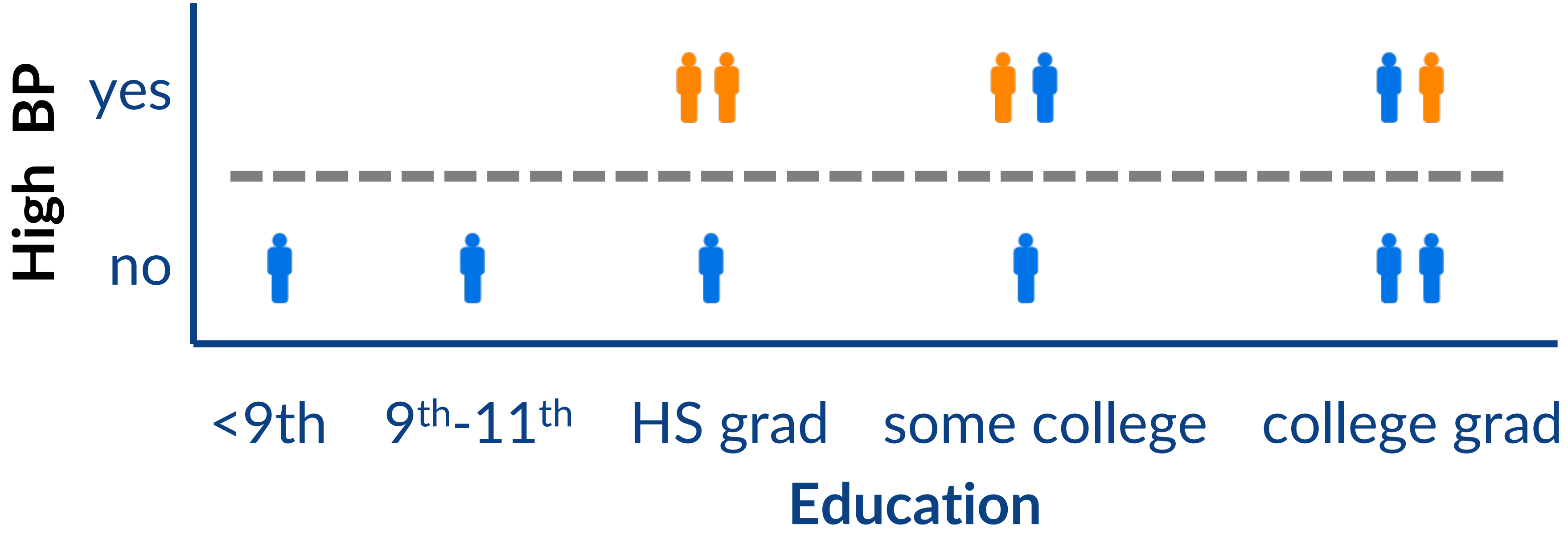
Need to compute:

$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D} \mid High\ BP)$$

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D} \mid Education)$$

Information Gain Example for Diabetes

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Need to compute:

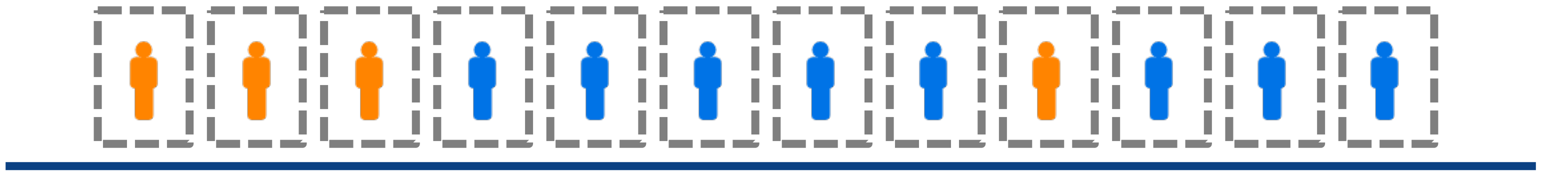
$$IG(\mathcal{D}, High\ BP) = H(\mathcal{D}) - H(\mathcal{D}\,|High\ BP) = 0.918 - 0.459 = 0.459$$

0.459

$$IG(\mathcal{D}, Education) = H(\mathcal{D}) - H(\mathcal{D}|\ Education) = 0.918 - 0.730 = 0.188$$

Information Gain Example for Diabetes

ID (SEQN)	HIGH_BP (BPQ020)	EDUCATION (DMDEDUC2)	DIABETIC
73557	yes	high school graduate / GED	yes
73558	yes	high school graduate / GED	yes
73559	yes	some college or AA degree	yes
73562	yes	some college or AA degree	no
73564	yes	college graduate or above	no
73566	no	high school graduate / GED	no
73567	no	9th-11th grade	no
73568	no	college graduate or above	no
73571	yes	college graduate or above	yes
73577	no	Less than 9th grade	no
73581	no	college graduate or above	no
73585	no	some college or AA degree	no



Patient ID

Need to compute:

$$IG(\mathcal{D}, ID) = H(\mathcal{D}) - H(\mathcal{D} \mid ID)$$

=1/12*0+1/12*0+....
= 0

IG = 0.918 … highest possible!

Compensating for Features with Many Values

IG tends toward selecting features that have many values

- e.g., unique identifiers, dates, etc.
- For deterministic f's, splitting on a unique identifier would immediately maximize the IG!

Gain Ratio can compensate for this:

$$GR\left(\mathcal{D}, X_j\right) = \frac{IG\left(\mathcal{D}, X_j\right)}{SplitInfo\left(\mathcal{D}, X_j\right)}$$

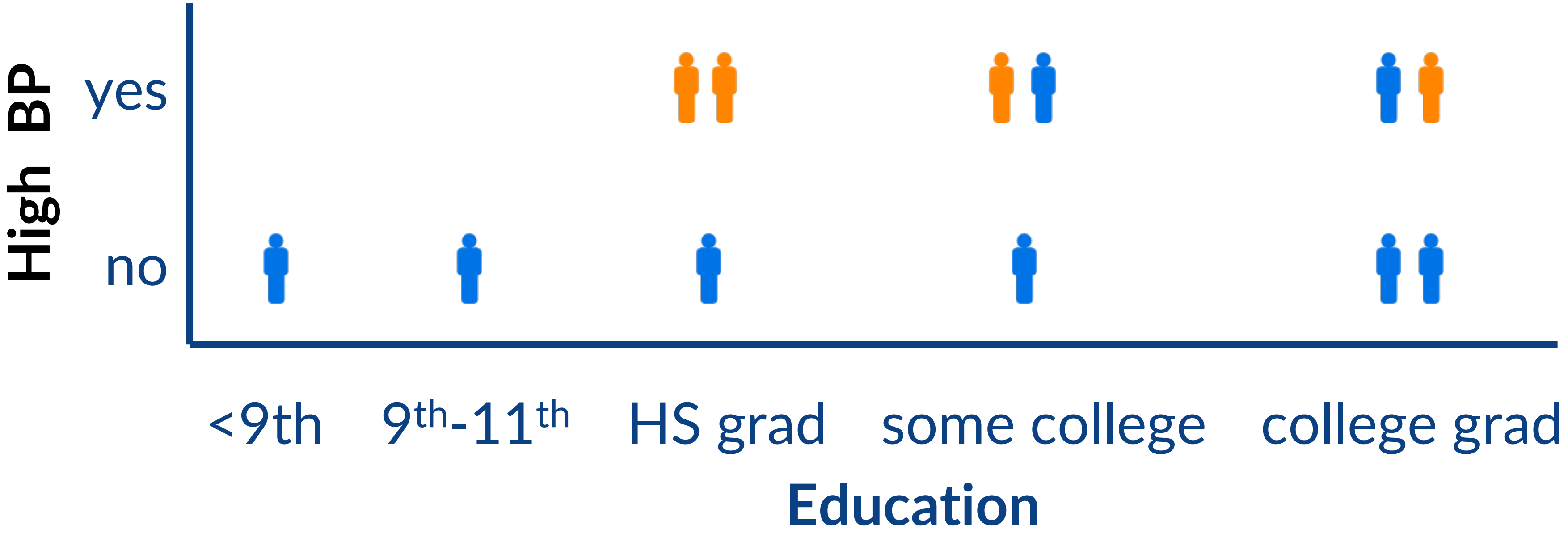
This scales by the entropy of the split, ignoring classes

$$SplitInfo\left(\mathcal{D}, X_j\right) = -\sum_{v} P\left(X_j = v\right) \log_2 P\left(X_j = v\right)$$

$$\frac{\left|\mathcal{D}\left[X_j = v\right]\right|}{|\mathcal{D}|}$$

Gain Ratio Example

Already Computed:
- H($\mathcal{D}$) = 0.918
- H ($\mathcal{D}$ | High BP) = 0.459
- H ($\mathcal{D}$ | Education) = 0.730
- IG($\mathcal{D}$ High BP) = 0.459
- IG($\mathcal{D}$, Education) = 0.188



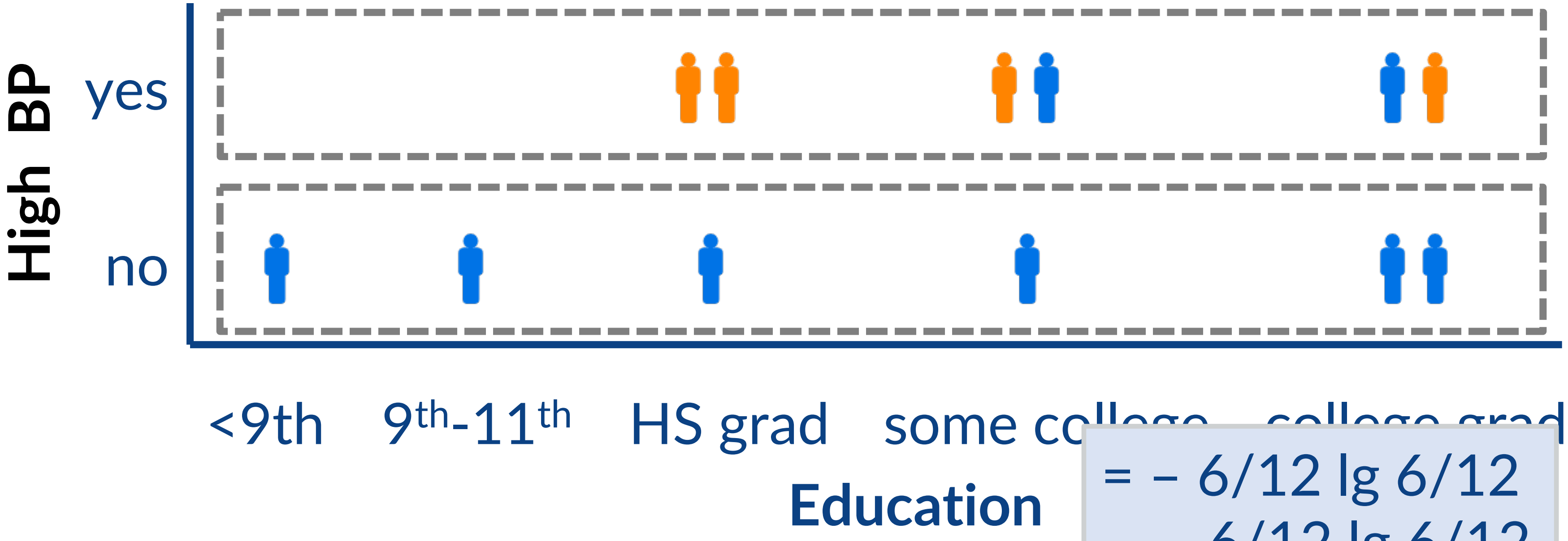
Need to compute:

GainRatio($\mathcal{D}$ High BP) = IG($\mathcal{D}$, High BP) / SplitInfo($\mathcal{D}$, High BP)

GainRatio($\mathcal{D}$, Education) = IG($\mathcal{D}$, Education) / SplitInfo($\mathcal{D}$, Education)

Gain Ratio Example

Already Computed:
- H($\mathcal{D}$) = 0.918
- H ($\mathcal{D}$ | High BP) = 0.459
- H ($\mathcal{D}$ | Education) = 0.730
- IG($\mathcal{D}$ High BP) = 0.459
- IG($\mathcal{D}$, Education) = 0.188



$= -\, 6/12 \lg 6/12$
$\quad -\, 6/12 \lg 6/12$
$= 1$

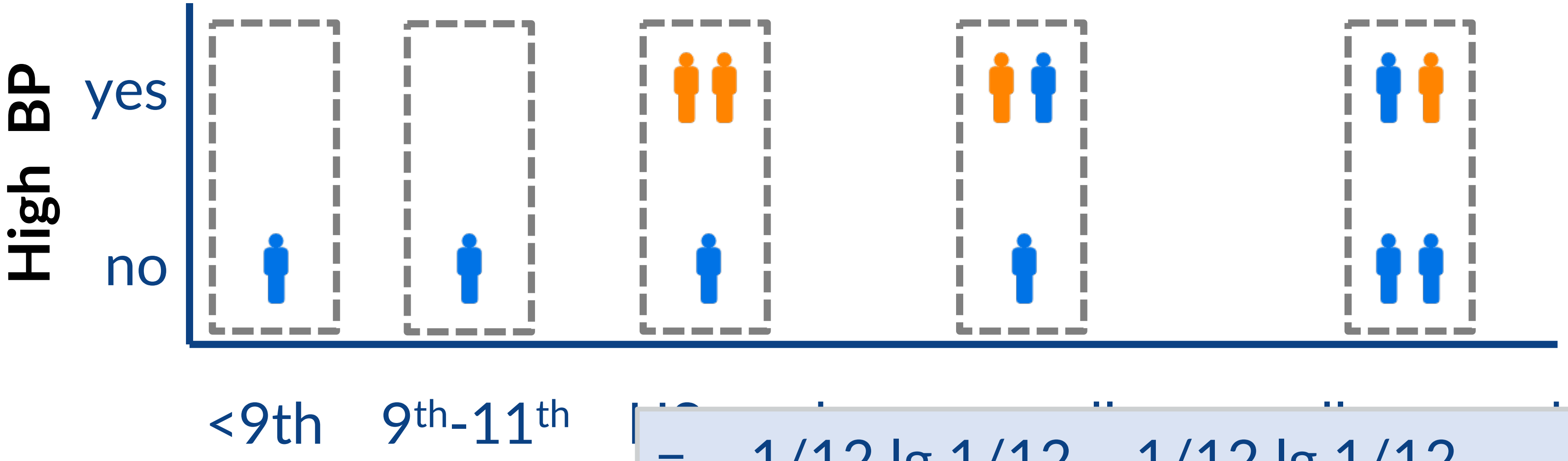
Need to compute:

GainRatio($\mathcal{D}$ High BP) = IG($\mathcal{D}$, High BP) / SplitInfo($\mathcal{D}$, High BP)

GainRatio($\mathcal{D}$, Education) = IG($\mathcal{D}$, Education) / SplitInfo($\mathcal{D}$, Education)

Gain Ratio Example

Already Computed:
- $H(\mathcal{D}) = 0.918$
- $H(\mathcal{D} \mid \text{High BP}) = 0.459$
- $H(\mathcal{D} \mid \text{Education}) = 0.730$
- $IG(\mathcal{D}\ \text{High BP}) = 0.459$
- $IG(\mathcal{D}, \text{Education}) = 0.188$



$= -\,1/12 \lg 1/12 - 1/12 \lg 1/12$
$\quad -\,3/12 \lg 3/12 - 3/12 \lg 3/12$
$\quad -\,4/12 \lg 4/12$
$= 2.1258$

Need to compute:

GainRatio($\mathcal{D}$ High BP) = IG($\mathcal{D}$, High BP) / SplitI

GainRatio($\mathcal{D}$, Education) = IG($\mathcal{D}$, Education) / SplitInfo($\mathcal{D}$, Education)

DT Training via Information Gain

We are Ready to Train the DT for Diabetes!

SEQN	RIDAGEYR	BMXWAIST	BMXHT	LBXTC	BMXLEG	BMXWT	BMXBMI	RIDRETH1	BPQ020	ALQ120Q	DMDEDUC2	RIAGENDR	INDFMPIR	LBXGH	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	110.8	161.8	168.0	37.1	93.4	35.7	Non-Hispanic White	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	85.5	152.8	278.0	32.4	61.8	26.5	Non-Hispanic White	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	22.0	Non-Hispanic White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispanic White	no	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	28.0	90.3	175.1	198.0	40.5	92.2	30.1	Other or Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589	35.0	94.6	172.9	192.0	39.1	78.3	26.2	Non-Hispanic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73595	58.0	114.8	175.3	165.0	40.1	96.0	31.2	Other Hispanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73596	57.0	117.8	164.7	151.0	35.3	104.0	38.3	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73600	37.0	122.9	185.1	189.0	48.1	126.2	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	161.0	36.5	111.9	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes
73616	62.0	95.5	172.8	171.0	38.4	71.8	24.0	Non-Hispanic White	no	2.0	some college or AA degree	female	5.0	5.5	no
73619	36.0	91.1	173.1	162.0	38.9	81.7	27.3	Mexican American	no	2.0	high school graduate / GED	female	0.84	5.0	no
73621	80.0	98.2	176.2	161.0	40.4	76.4	24.6	Non-Hispanic White	no	5.0	college graduate or above	male	5.0	5.6	no
73622	72.0	115.6	185.4	186.0	39.7	99.5	28.9	Non-Hispanic White	no	4.0	college graduate or above	male	5.0	6.0	no

Entropy-Based Greedy DT Construction

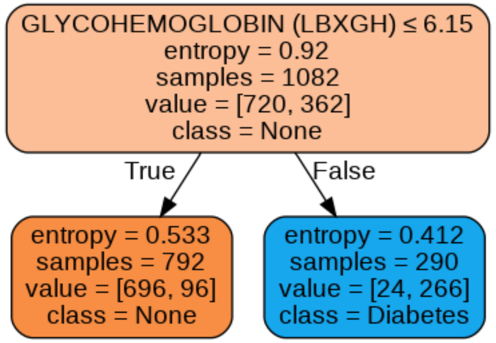
SEQN	RIDAGEYR	BMXWAIST	BMXHT	LBXTC	BMXLEG	BMXWT	BMXBMI	RIDRETH1	BPQ020	ALQ120Q	DMDEDUC2	RIAGENDR	INDFMPIR	LBXGH	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	110.8	161.8	168.0	37.1	93.4	35.7	Non-Hispanic White	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	85.5	152.8	278.0	32.4	61.8	26.5	Non-Hispanic White	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	22.0	Non-Hispanic White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispanic White	no	2.0	college graduate or above	female	5.0	5.2	no
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	28.0	90.3	175.1	198.0	40.5	92.2	30.1	Other or Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589	35.0	94.6	172.9	192.0	39.1	78.3	26.2	Non-Hispanic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73595	58.0	114.8	175.3	165.0	40.1	96.0	31.2	Other Hispanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73596	57.0	117.8	164.7	151.0	35.3	104.0	38.3	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73600	37.0	122.9	185.1	189.0	48.1	126.2	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	161.0	36.5	111.9	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

X_1 X_2 . . . X_{14}

Given dataset $\mathcal{D} = [X, \boldsymbol{y}]$

- Pick feature X_j to split upon with the highest IG (or GainRatio)
- Partition $\mathcal{D}$ via X_j
- Recurse until nodes are homogenous

X_{14} (LBXGH) ≤ 6.15 has the highest IG



Dataset partition $\mathcal{D}$[LBXGH ≤ 6.15]

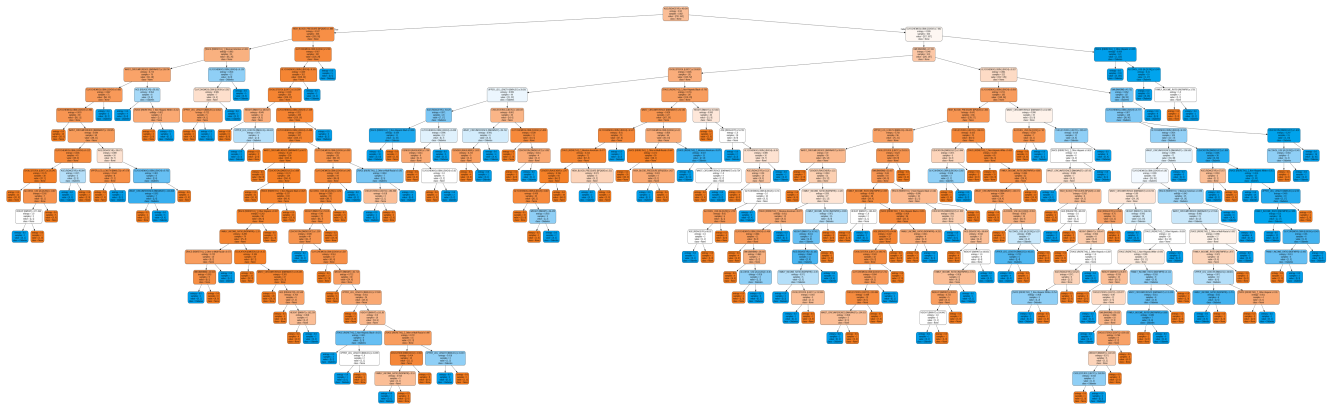
SEQN	RIDAGEYR	BMXWAIST	BMXHT	LBXTC	BMXLEG	BMXWT	BMXBMI	RIDRETH1	BPQ020	ALQ120Q	DMDEDUC2	RIAGENDR	INDFMPIR	LBXGH	DIABETIC
73562	56.0	123.1	158.7	226.0	34.2	105.0	41.7	Mexican American	yes	5.0	some college or AA degree	male	4.79	5.5	no
73564	61.0	110.8	161.8	168.0	37.1	93.4	35.7	Non-Hispanic White	yes	2.0	college graduate or above	female	5.0	5.5	no
73566	56.0	85.5	152.8	278.0	32.4	61.8	26.5	Non-Hispanic White	no	1.0	high school graduate / GED	female	0.48	5.4	no
73567	65.0	93.7	172.4	173.0	40.0	65.3	22.0	Non-Hispanic White	no	4.0	9th-11th grade	male	1.2	5.2	no
73568	26.0	73.7	152.5	168.0	34.4	47.1	20.3	Non-Hispanic White	no	2.0	college graduate or above	female	5.0	5.2	no
73577	32.0	100.0	166.2	182.0	36.5	79.7	28.9	Mexican American	no	20.0	Less than 9th grade	male	0.29	5.3	no
73581	50.0	99.3	185.0	202.0	42.8	80.9	23.6	Other or Multi-Racial	no	0.0	college graduate or above	male	5.0	5.0	no
73585	28.0	90.3	175.1	198.0	40.5	92.2	30.1	Other or Multi-Racial	no	4.0	some college or AA degree	male	2.26	5.0	no
73589	35.0	94.6	172.9	192.0	39.1	78.3	26.2	Non-Hispanic White	no	2.0	high school graduate / GED	male	1.74	5.5	no
73596	57.0	117.8	164.7	151.0	35.3	104.0	38.3	Other or Multi-Racial	yes	1.0	college graduate or above	female	5.0	5.9	no
73604	69.0	96.6	156.9	203.0	37.0	59.5	24.2	Non-Hispanic White	no	1.0	some college or AA degree	female	2.44	5.4	no
73607	75.0	130.5	169.6	161.0	36.5	111.9	38.9	Non-Hispanic White	yes	0.0	high school graduate / GED	male	1.08	5.0	no
73610	43.0	102.6	176.8	200.0	38.8	90.2	28.9	Non-Hispanic White	no	5.0	college graduate or above	male	2.03	4.9	no
73613	60.0	113.6	163.8	203.0	41.6	104.9	39.1	Non-Hispanic Black	yes	2.0	9th-11th grade	female	5.0	6.1	no
73614	55.0	90.9	167.9	256.0	43.5	60.9	21.6	Non-Hispanic White	no	0.0	high school graduate / GED	female	1.29	5.0	no

Dataset partition $\mathcal{D}$[LBXGH > 6.15]

SEQN	RIDAGEYR	BMXWAIST	BMXHT	LBXTC	BMXLEG	BMXWT	BMXBMI	RIDRETH1	BPQ020	ALQ120Q	DMDEDUC2	RIAGENDR	INDFMPIR	LBXGH	DIABETIC
73557	69.0	100.0	171.3	167.0	39.2	78.3	26.7	Non-Hispanic Black	yes	1.0	high school graduate / GED	male	0.84	13.9	yes
73558	54.0	107.6	176.8	170.0	40.0	89.5	28.6	Non-Hispanic White	yes	7.0	high school graduate / GED	male	1.78	9.1	yes
73559	72.0	109.2	175.3	126.0	40.0	88.9	28.9	Non-Hispanic White	yes	0.0	some college or AA degree	male	4.51	8.9	yes
73571	76.0	122.1	172.5	167.0	35.5	102.4	34.4	Non-Hispanic White	yes	2.0	college graduate or above	male	5.0	6.9	yes
73595	58.0	114.8	175.3	165.0	40.1	96.0	31.2	Other Hispanic	no	1.0	some college or AA degree	male	3.09	7.7	no
73600	37.0	122.9	185.1	189.0	48.1	126.2	36.8	Non-Hispanic Black	yes	2.0	high school graduate / GED	male	0.63	6.2	yes
73615	65.0	100.3	145.9	166.0	30.0	55.4	26.0	Other Hispanic	yes	1.0	Less than 9th grade	female	1.22	6.3	yes

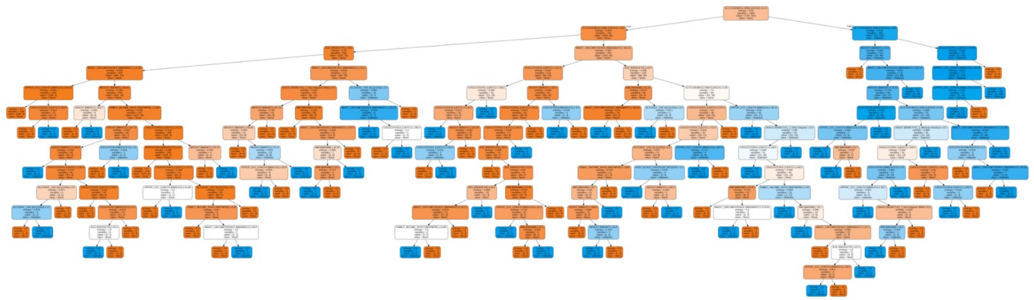
Diabetes DT – Random vs IG Features

DT with random feature splits



Accuracy on diabetes data = 100%

DT via IG



Accuracy on diabetes data = 100%

- Well, it is smaller while retaining 100 % accuracy on our training data
- Still rather complex, though …

Overfitting and Decision Trees

Accuracy – Decision Tree (Version 1)

Original Patient Data: 100.000 % (n = 1082)

New Patient Data: 82.796 % (n = 465)

Avoiding Overfitting

How can we avoid overfitting?

1. Stop growing when data split is not statistically significant
2. Acquire more training data
3. Remove irrelevant attributes (manual process – not always possible)
4. Grow full tree, then post-prune

Try various tree hyperparameters (e.g., tree depth, splitting criterion, termination criterion) and pick the one with the best estimated generalization performance. How to estimate?

- Cross-validation
- Add a complexity penalty to performance measure e.g. training accuracy – average depth of leaf node

Based on Slide by Pedro Domingos

Reduced-Error Pruning

Split the original training data into training and validation sets

Training Stage

Grow the decision tree based on the training set

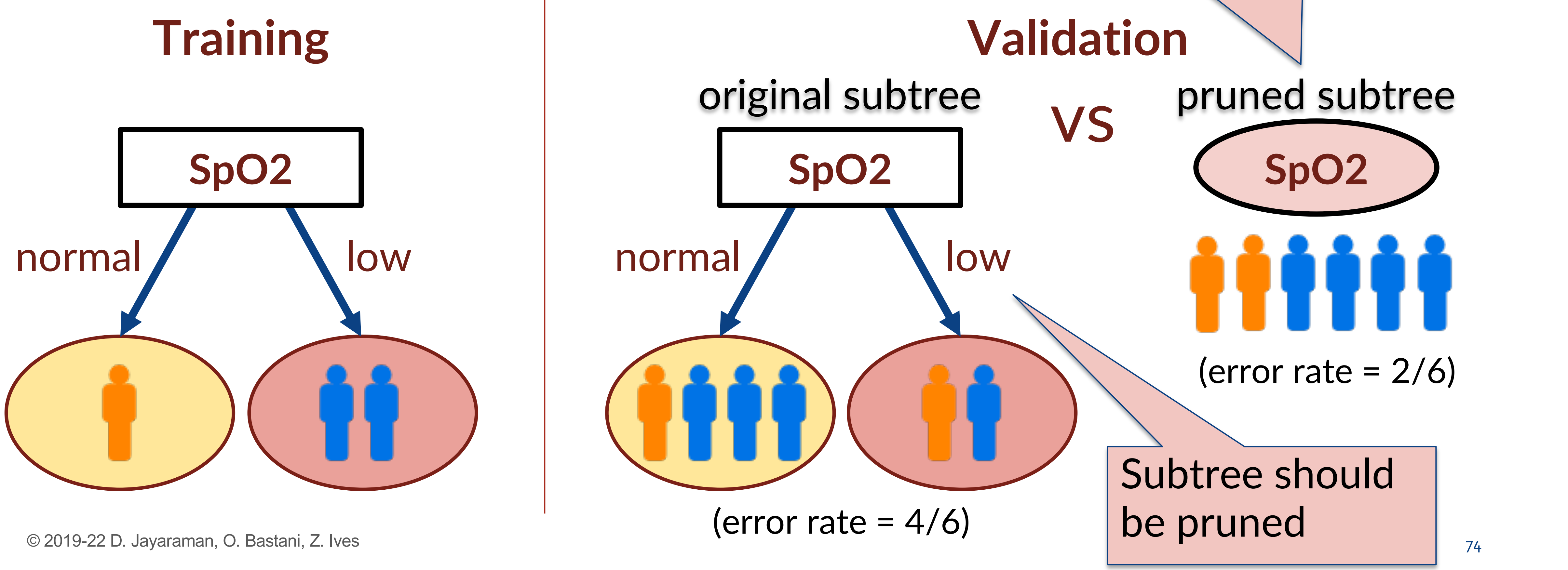
Pruning Stage

Loop until further pruning hurts validation performance:

- Measure the validation performance of pruning each node (and its children)
- Greedily remove the node that most improves validation performance

Reduced-Error Pruning

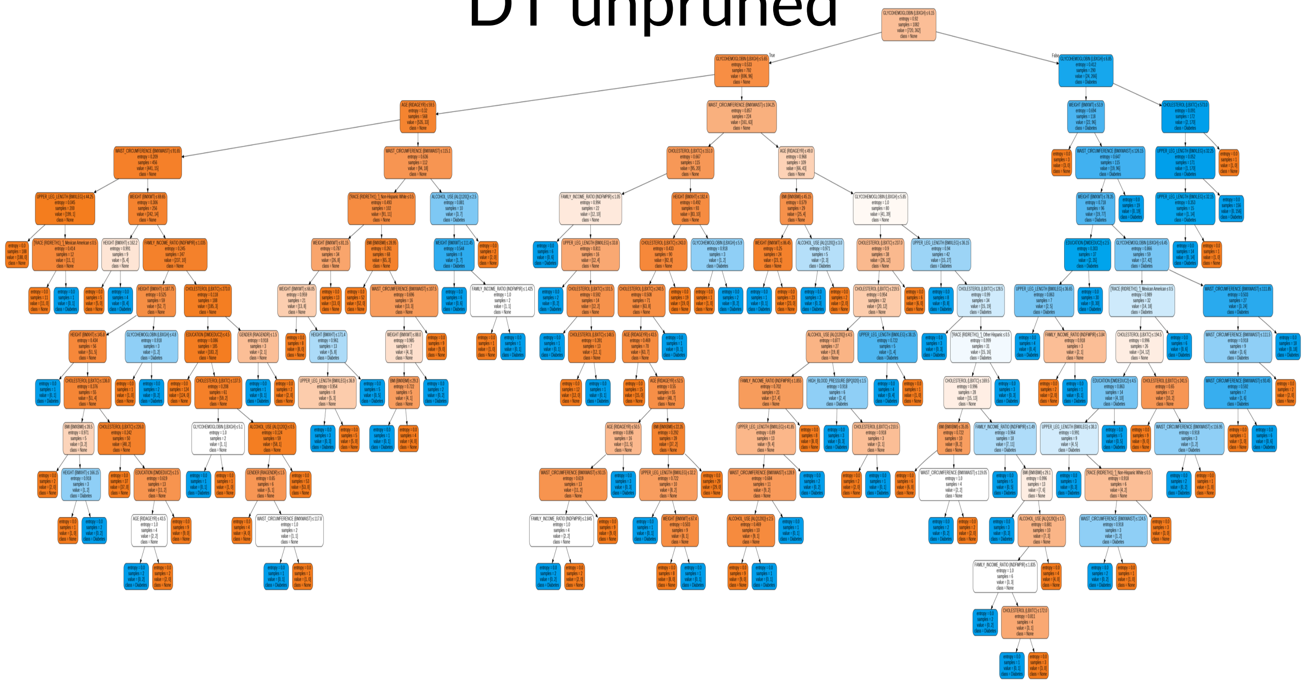
- Pruning replaces a whole subtree with a leaf node
- Replacement occurs if the expected error rate of the subtree is greater than that of the leaf



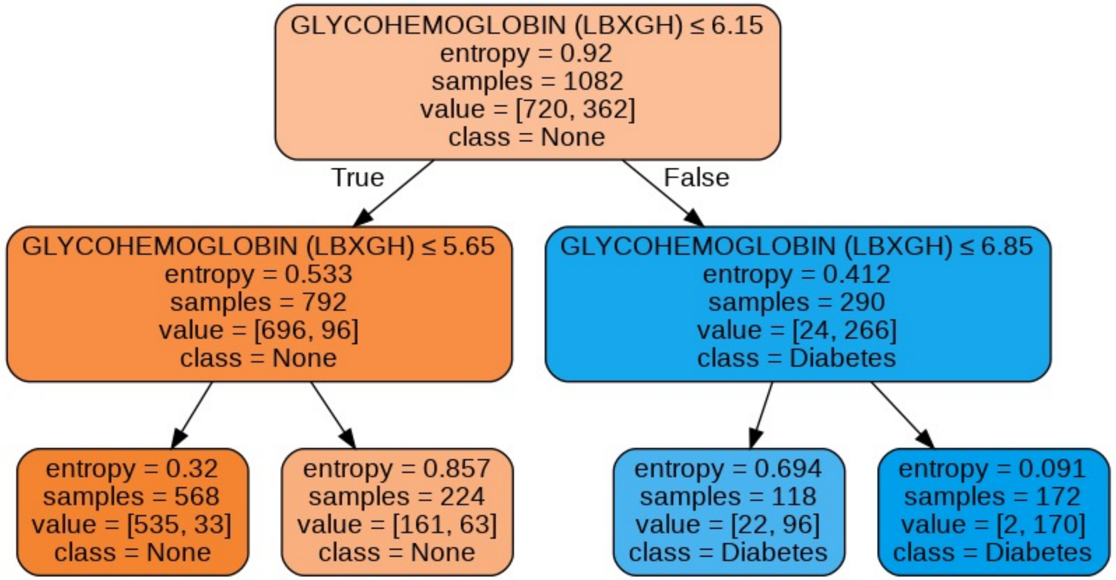
Accuracy – Decision Trees

	DT unpruned	DT pruned	
Original Patient Data:	100.000 %	88.909 %	(n = 1082)
New Patient Data:	82.796 %	85.591 %	(n = 465)

DT unpruned

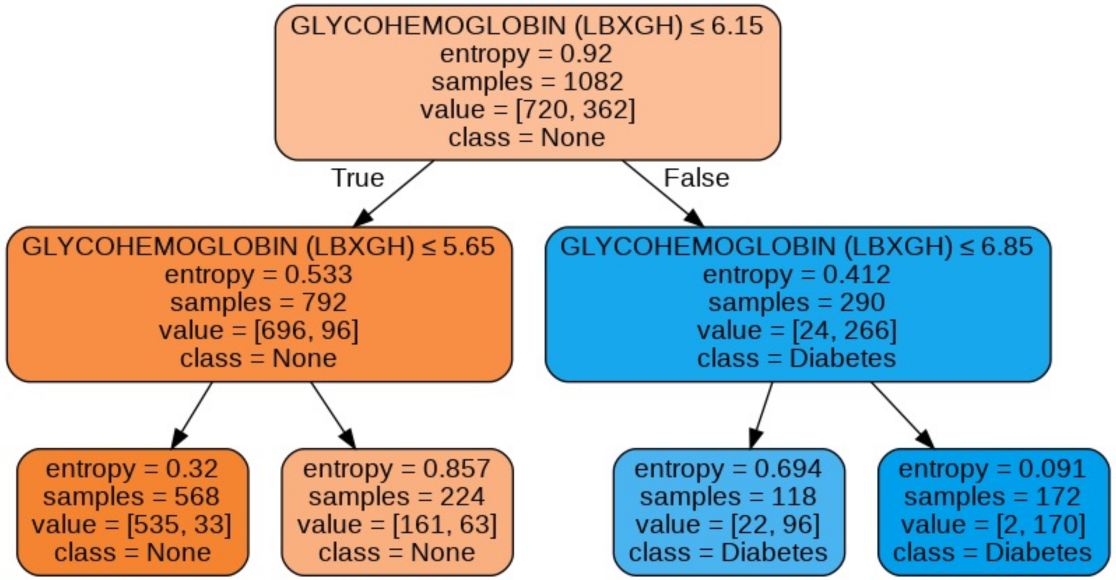


DT pruned



The Final Diabetes DT

Our Pruned Decision Tree



How Diabetes is Actually Diagnosed

- If your A1C level is between 5.7 and less than 6.5%, your levels have been in the prediabetes range.
- If you have an A1C level of 6.5% or higher, your levels were in the diabetes range.

(screenshot from diabetes.org)

Strong similarity to how diabetes is *actually* diagnosed!

Decision Tree Algorithms

ID3

- Information gain on nominal features

C4.5

- Can use info gain or gain ratio
- Nominal or numeric features
- Missing values
- Post-pruning
- Rule generation

CART (Classification and Regression Tree)

- Similar to C4.5
- Can handle continuous target prediction (regression)
- No rule sets
- Sklearn's `DecisionTreeClassifier` is based on CART, but can't handle nominal features (as of version 0.22.1)

Many Other Algorithms …

Strengths and Weaknesses of DTs

Strengths

👍 Widely used in practice
👍 Fast and simple to implement
👍 Small trees are easily interpretable
👍 Handles a variety of feature types
👍 Can convert to rules
👍 Handles noisy / missing data
👍 Insensitive to feature scaling
👍 Handles irrelevant features
👍 Handles large datasets

Weaknesses

👎 Univariate partitions limit potential trees
👎 Limited predictive power
👎 Heuristic-Based Greedy Training

Comparison of Learning Methods

Characteristic	Trees	k-NN, Kernels
Natural handling of data of "mixed" type	▲	▼
Handling of missing values	▲	▲
Robustness to outliers in input space	▲	▲
Insensitive to monotone transformations of inputs	▲	▼
Computational scalability (large N)	▲	▼
Ability to deal with irrelevant inputs	▲	▼
Ability to extract linear combinations of features	▼	◆
Interpretability	◆	▼
Predictive power	▼	▲

[Table 10.3 from Hastie, et al. *Elements of Statistical Learning*, 2nd Edition]