EEP 596: AI and Health Care || Lecture 3

Dr. Karthik Mohan

Univ. of Washington, Seattle

Apr 4, 2022

Logistics

- **Lectures:** Monday in person, Wednesday online

Logistics

- **Lectures:** Monday in person, Wednesday online
- **Grading Office Hours:** Saturday, 5-6 pm (Mathew)

Logistics

- **Lectures:** Monday in person, Wednesday online
- **Grading Office Hours:** Saturday, 5-6 pm (Mathew)
- **TA office hours:** Sunday, ~~5-6 pm~~ (Ayush)
  12-1 PM

Logistics

- **Lectures:** Monday in person, Wednesday online
- **Grading Office Hours:** Saturday, 5-6 pm (Mathew)
- **TA office hours:** Sunday, ~~5-6 pm~~ 12 - 1 PM (Ayush)
- **Quiz Section:** Sunday, ~~12 - 1 pm~~ 5 - 6 PM (Ayush)

Logistics

- **Lectures:** Monday in person, Wednesday online
- **Grading Office Hours:** Saturday, 5-6 pm (Mathew)
- **TA office hours:** Sunday, 5-6 pm (Ayush)
- **Quiz Section:** Sunday, 12 - 1 pm (Ayush)
- **Karthik Office hours:** Wednesday, 6-6:30 pm (Karthik)

Logistics

- **Lectures:** Monday in person, Wednesday online
- **Grading Office Hours:** Saturday, 5-6 pm (Mathew)
- **TA office hours:** Sunday, 5-6 pm (Ayush)
- **Quiz Section:** Sunday, 12 - 1 pm (Ayush)
- **Karthik Office hours:** Wednesday, 6-6:30 pm (Karthik)
- **Programming Assignment 1:** Due Sunday night

Logistics

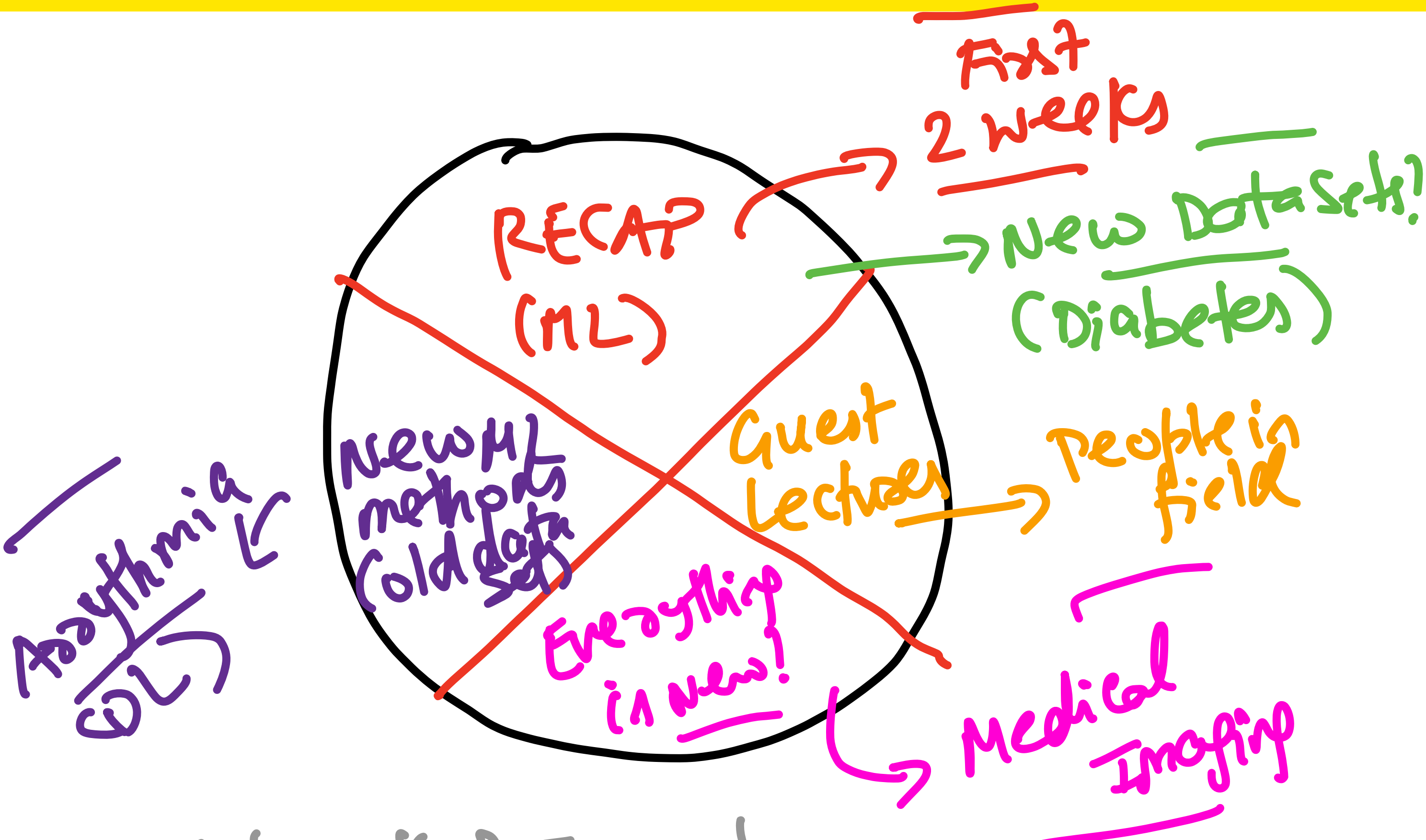
- **Lectures:** Monday in person, Wednesday online
- **Grading Office Hours:** Saturday, 5-6 pm (Mathew)
- **TA office hours:** Sunday, 5-6 pm (Ayush)
- **Quiz Section:** Sunday, 12 - 1 pm (Ayush)
- **Karthik Office hours:** Wednesday, 6-6:30 pm (Karthik)
- **Programming Assignment 1:** Due Sunday night
- **Surveys:** Fill it out - So we can be aligned on hours, topics, etc and have a good feedback loop!

Last lecture

- Classification - Logistic Regression
- Overfitting in Machine Learning
- Methods to overcome overfitting

1. Feature selection
2. Get More Data
3. Regularize
4. DL hacks
   - Dropouts

Lectures Makeup (Pie Chart)



Sit back & Enjoy!

Today

ML good to know!

1. Data pre-processing
2. Diabetes data set
3. Decision Trees for Diabetes classification

Programming 1

Data Transformations

1. **Data-cleaning:** What does this refer to? → Removing / filtering bad entries

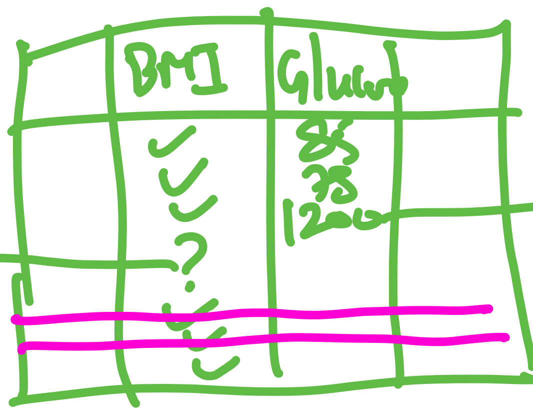
Data Transformations

1. **Data-cleaning:** What does this refer to?
2. **Data-cleaning:** Removing outliers/extreme feature values, handling missing data

1 2

3 Duplicates

2 missing data



outliers! 1

Same!! 3 duplicates

Data Transformations

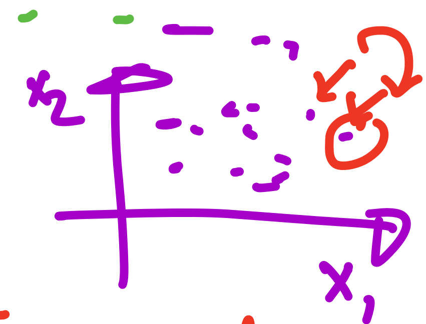
1. **Data-cleaning:** What does this refer to?
2. **Data-cleaning:** Removing outliers/extreme feature values, handling missing data
3. **Missing Data:** Let's say a few columns have missing data - How does one handle it?

4) Matrix completion

1) Interpolate

2) Impute

3) Throw away rows!!



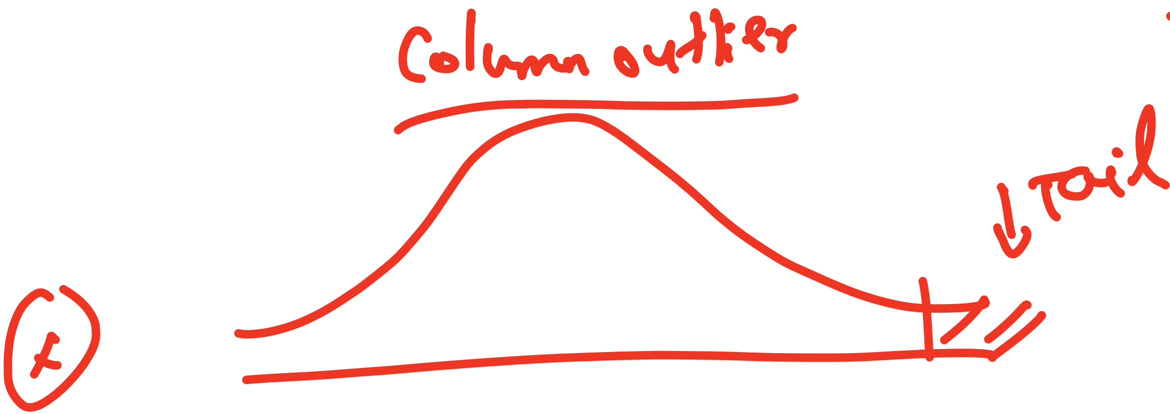
? → NaN

Data Transformations

1. **Data-cleaning:** What does this refer to?
2. **Data-cleaning:** Removing outliers/extreme feature values, handling missing data
3. **Missing Data:** Let's say a few columns have missing data - How does one handle it?
4. **Outlier removal:** How do you remove outliers from data?

Anomaly /

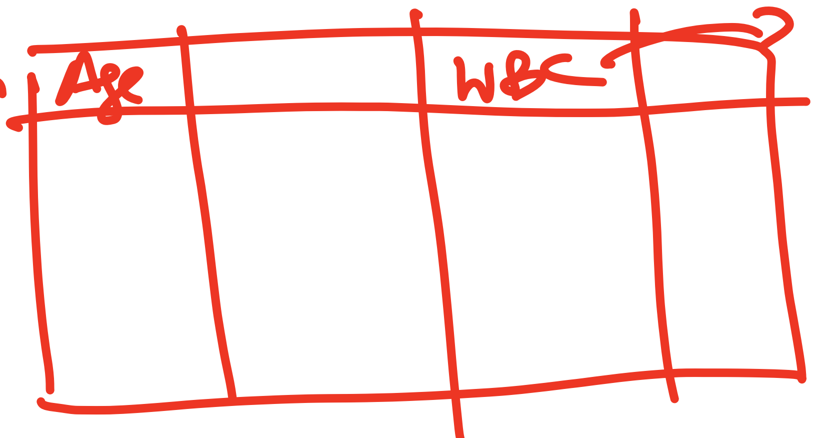
→ outlier detection methods



Data Transformations

1. **Data-cleaning:** What does this refer to?
2. **Data-cleaning:** Removing outliers/extreme feature values, handling missing data
3. **Missing Data:** Let's say a few columns have missing data - How does one handle it?
4. **Outlier removal:** How do you remove outliers from data?
5. **Data Normalization:** What is data normalization and why do we normalize?

0-100



0 - 15000

After norm

0-100

Helps with avoiding gradient issues

Data Transformations

1. **Data-cleaning:** What does this refer to?
2. **Data-cleaning:** Removing outliers/extreme feature values, handling missing data
3. **Missing Data:** Let's say a few columns have missing data - How does one handle it?
4. **Outlier removal:** How do you remove outliers from data?
5. **Data Normalization:** What is data normalization and why do we normalize?
6. **Data pre-processing:** How is data pre-~~por~~cessing different from data cleaning? What else does data pre-processing entail?

pro

Aug hyperparams

Imaging!

Data pre-processing

Data Augmentation

Data cleaning

Feature Eng.

Data Norm.

Data Transformations

1. **Data-cleaning:** What does this refer to?
2. **Data-cleaning:** Removing outliers/extreme feature values, handling missing data
3. **Missing Data:** Let's say a few columns have missing data - How does one handle it?
4. **Outlier removal:** How do you remove outliers from data?
5. **Data Normalization:** What is data normalization and why do we normalize?
6. **Data pre-processing:** How is data pre-porcessing different from data cleaning? What else does data pre-processing entail?
7. **Feature engineering:** Feature engineering is one of the data pre-processing steps - Where we transform raw data into useful features. What are some examples?

ICE #0

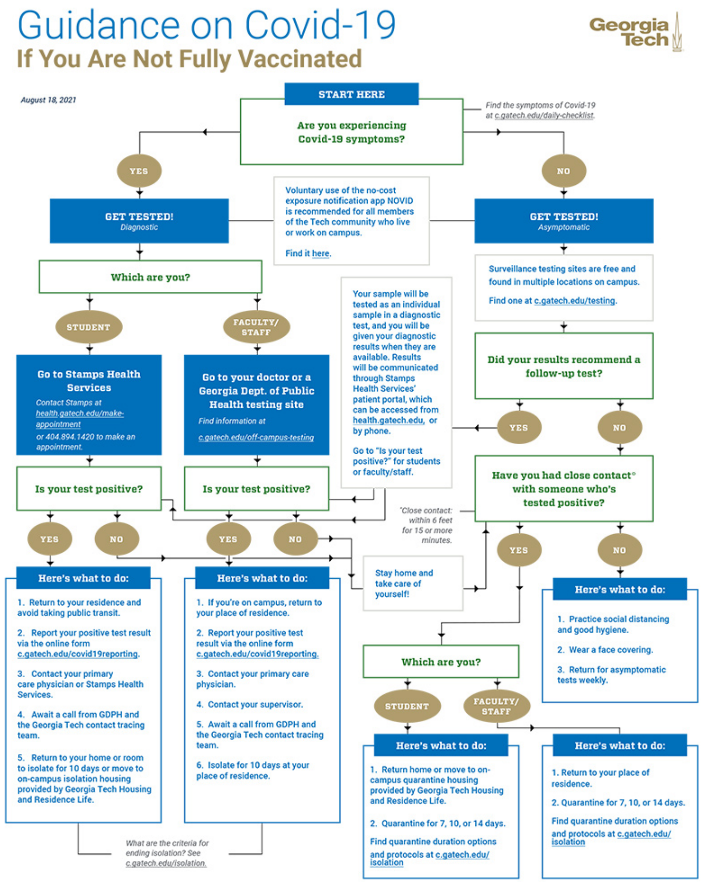
You are tasked with detecting risk of diabetes for a segment of the university population who volunteer to take part in a study. Post data collection, you notice that some of the patients with low risk diabetes have gluocse levels that don't add up. What step in data pre-processing will help you make better predictions?

- Data cleaning
- Outlier removal
- Data normalization
- Feature engineering

Decision Trees Motivation

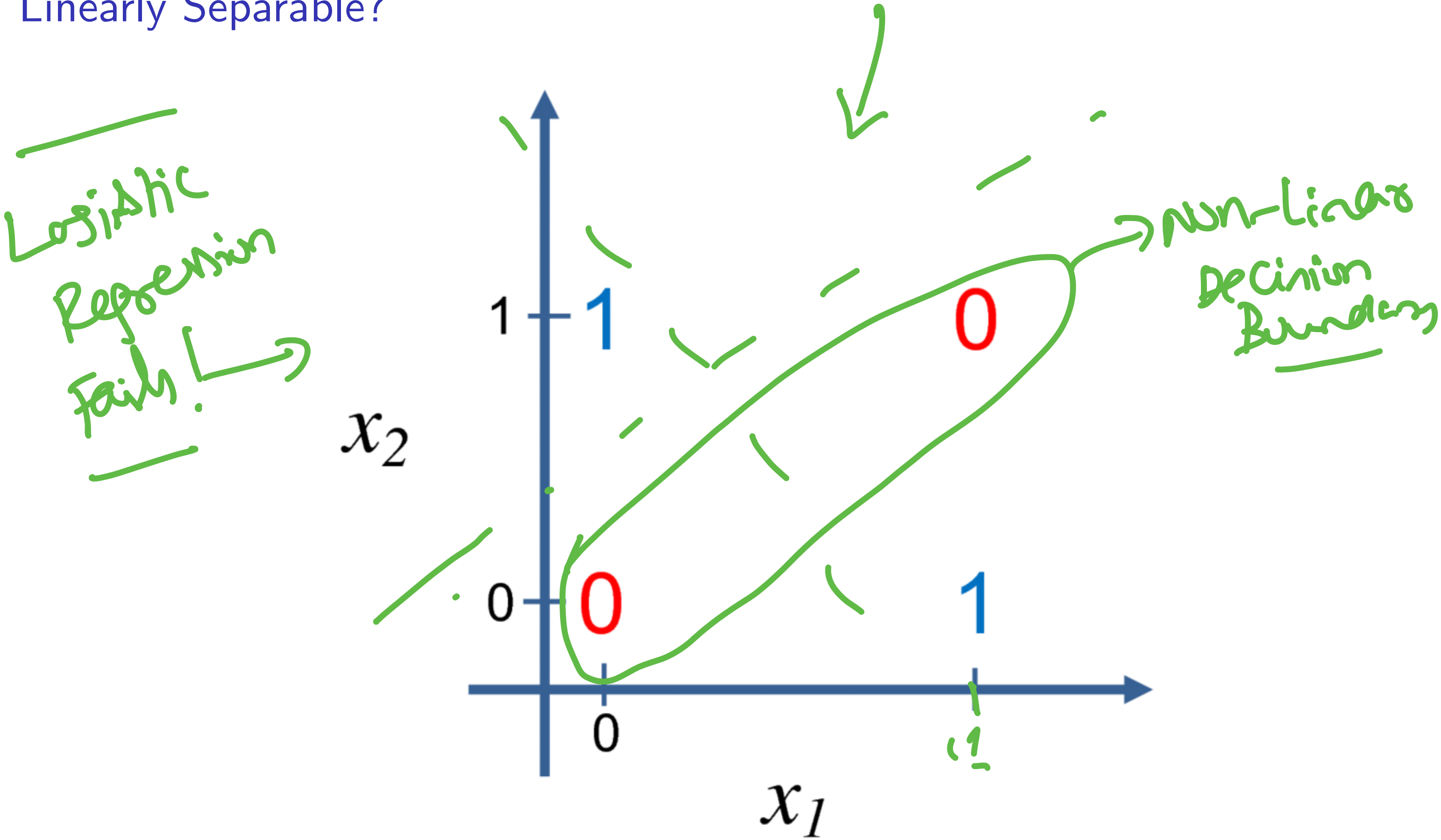
Health care!

Interpretable!



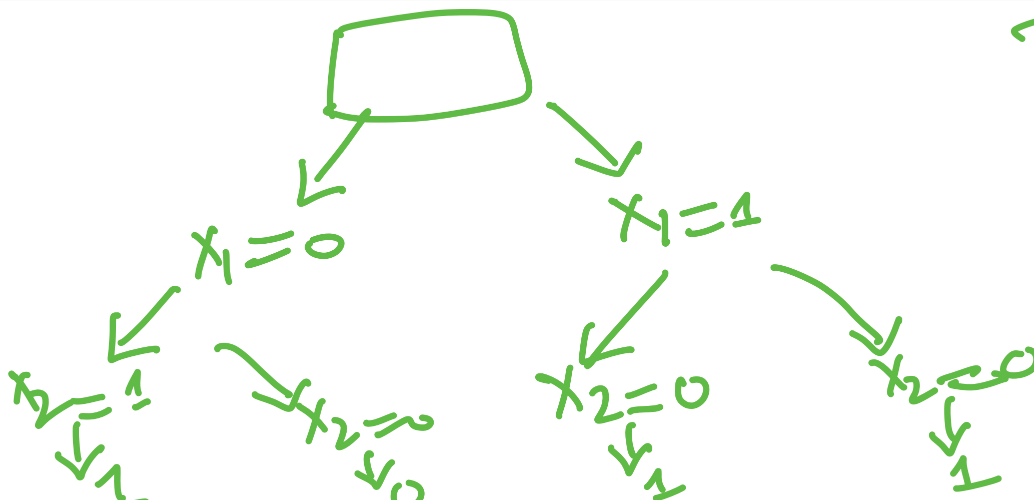
XOR Function

Linearly Separable?



XOR Function

Can XOR be modeled by Decision Tree? ✓



D.T. can learn XOR!!

Learning Decision Trees

Learning

The `learning` for Decision Trees boils down to how to build the tree. Which feature to split on first? Second? And so on... Also, when to stop building the tree

Learning Decision Trees

Learning

The `learning` for Decision Trees boils down to how to build the tree. Which feature to split on first? Second? And so on... Also, when to stop building the tree

Intuition behind building Decision Trees

Start splitting on features that give the maximum information gain or reduce the uncertainty in prediction/reduce the classification error. This is done iteratively and hence can be thought of as a greedy procedure.

Case Study: What factors increase risk of chronic diabetes

RBS

Case Study: What factors increase risk of chronic diabetes

Features

History

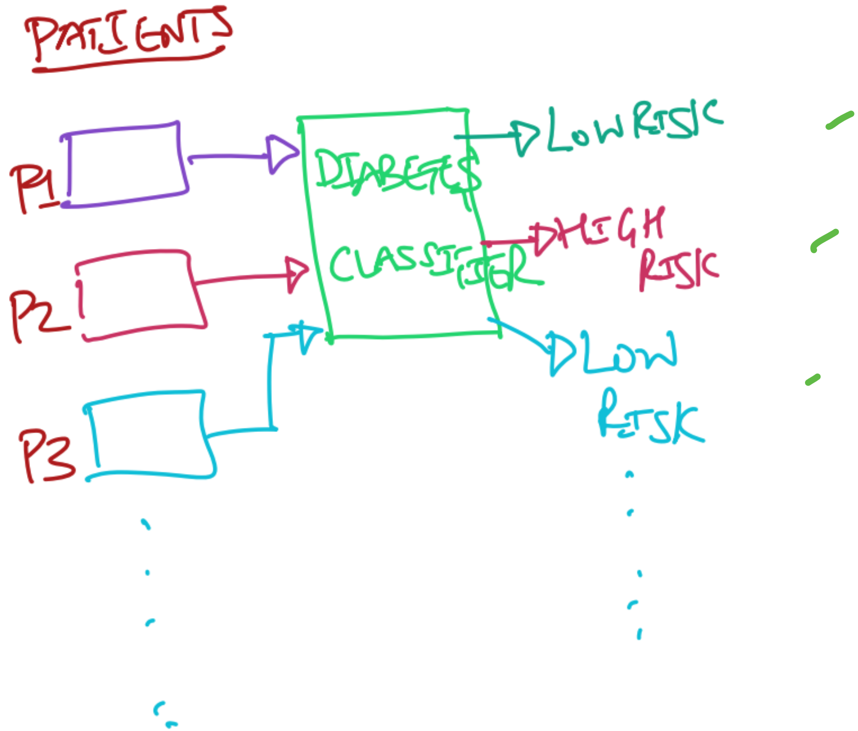
TARGET

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0

→ High Risk

→ Low Risk

Intelligent Diabetes Risk Detection



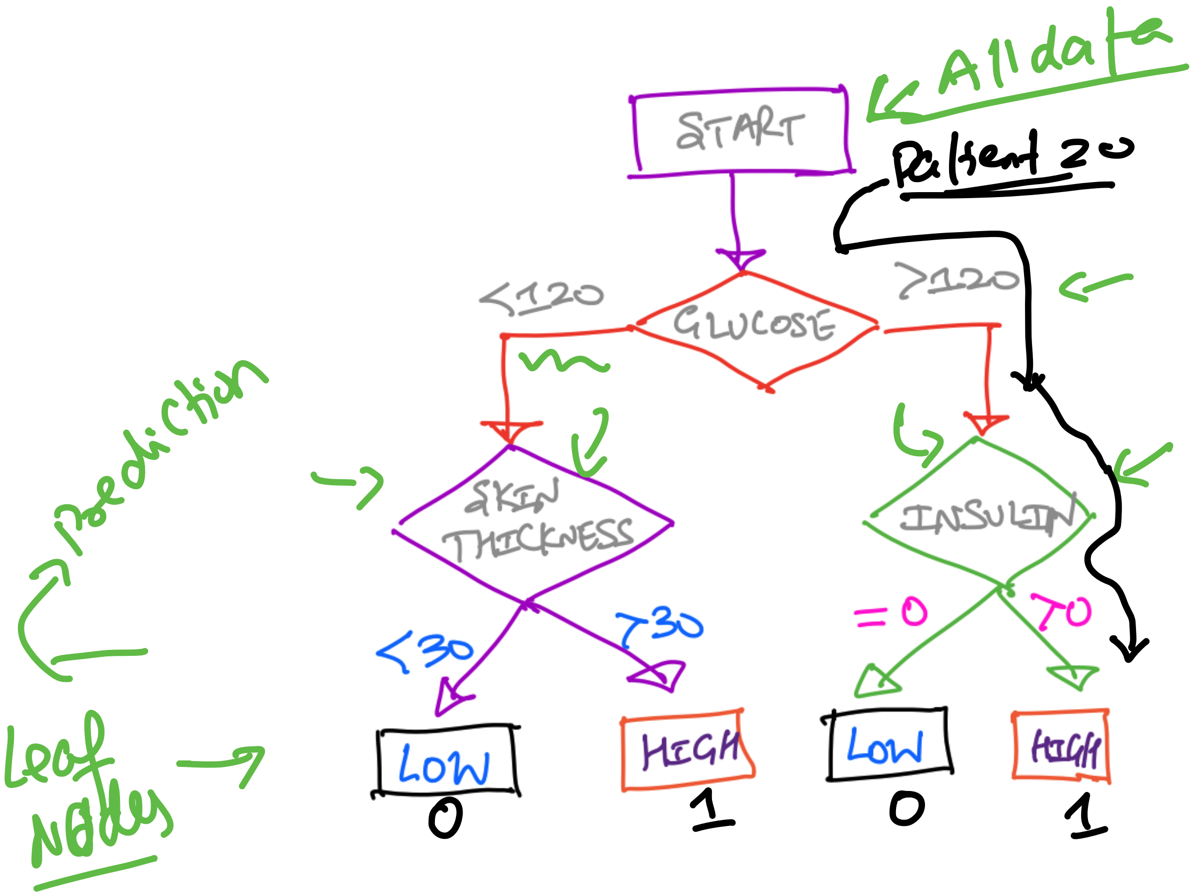
Sample Data

18-22 Body Mass Index

Triceps fold skin

Glucose	Insulin	BMI	SkinThickness	Outcome
148	0	33.6	35	1
85	0	26.6	29	0
183	0	23.3	0	1
89	94	28.1	23	0
137	168	43.1	35	1
116	0	25.6	0	0
78	88	31	32	1
115	0	35.3	0	0
197	543	30.5	45	1
125	0	0	0	1
110	0	37.6	0	0
168	0	38	0	1
139	0	27.1	0	0

Decision Trees

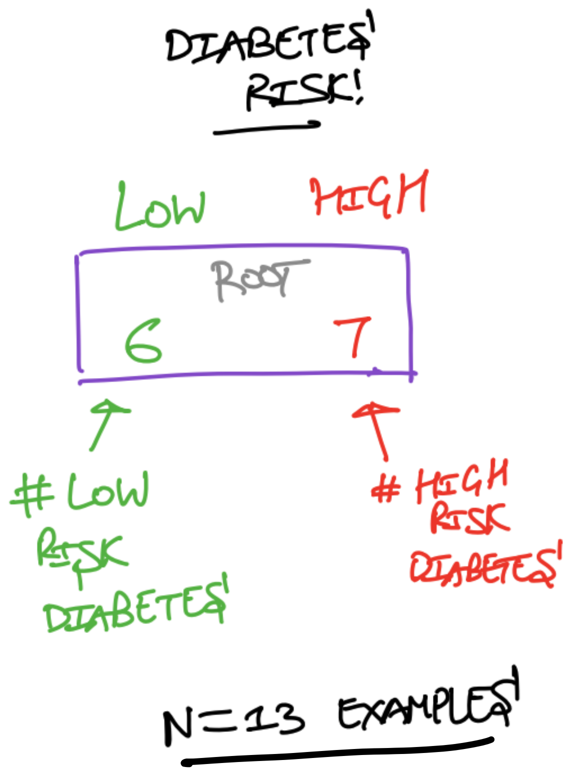


Growing Trees

Questions

- Which features are "good"?
- When to stop growing a tree?

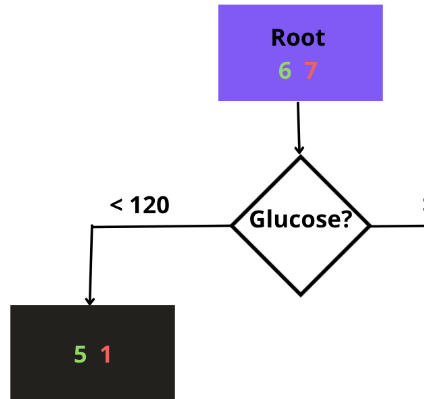
Visual Notation



Decision stump 1

Glucose	Insulin	BMI	SkinThickness	Outcome
148	0	33.6	35	1
85	0	26.6	29	0
183	0	23.3	0	1
89	94	28.1	23	0
137	168	43.1	35	1
116	0	25.6	0	0
78	88	31	32	1
115	0	35.3	0	0
197	543	30.5	45	1
125	0	0	0	1
110	0	37.6	0	0
168	0	38	0	1
139	0	27.1	0	0

N=13



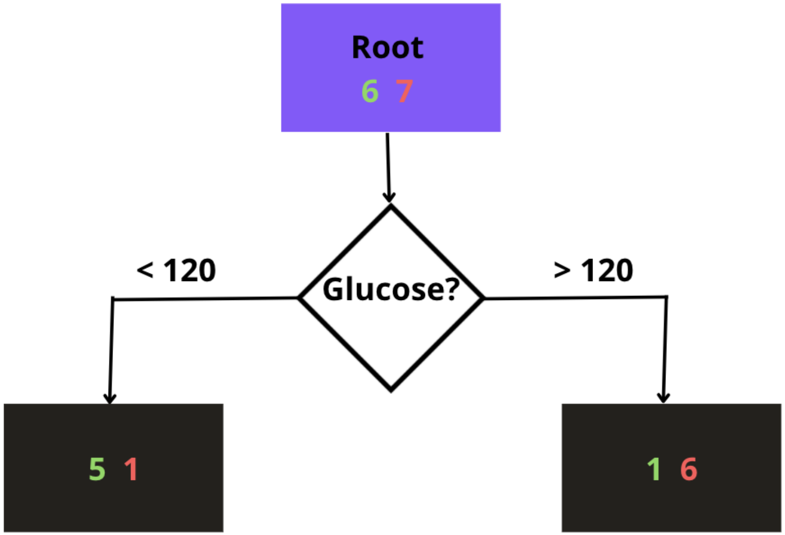
6+7=13

N=6

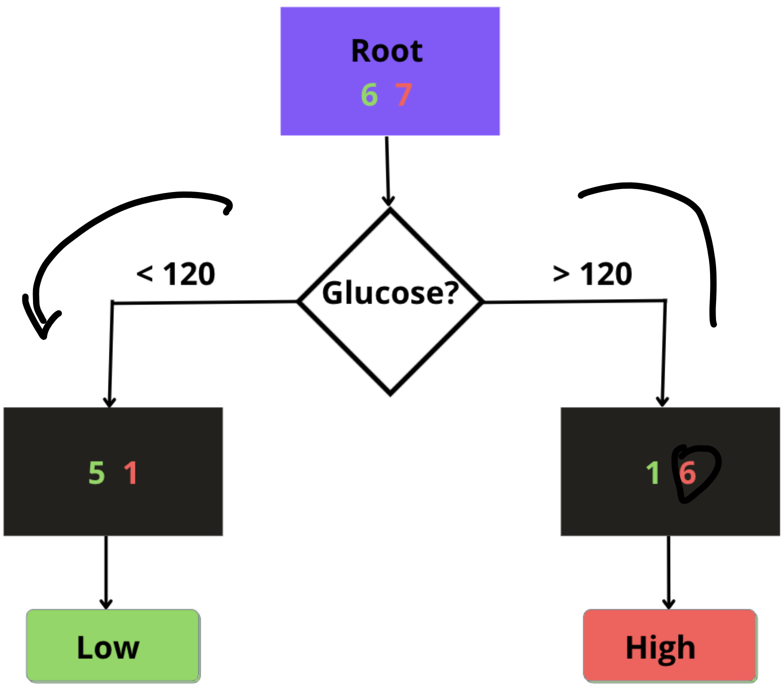
N=7

5+1+1+6=13

Making predictions



Making predictions

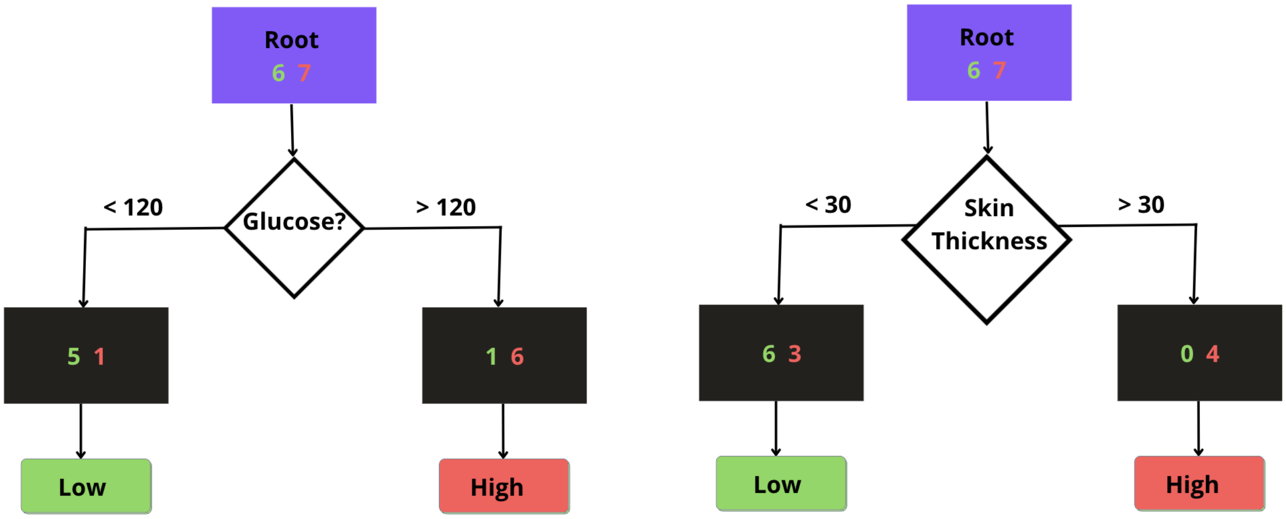


Split only on Glucose!

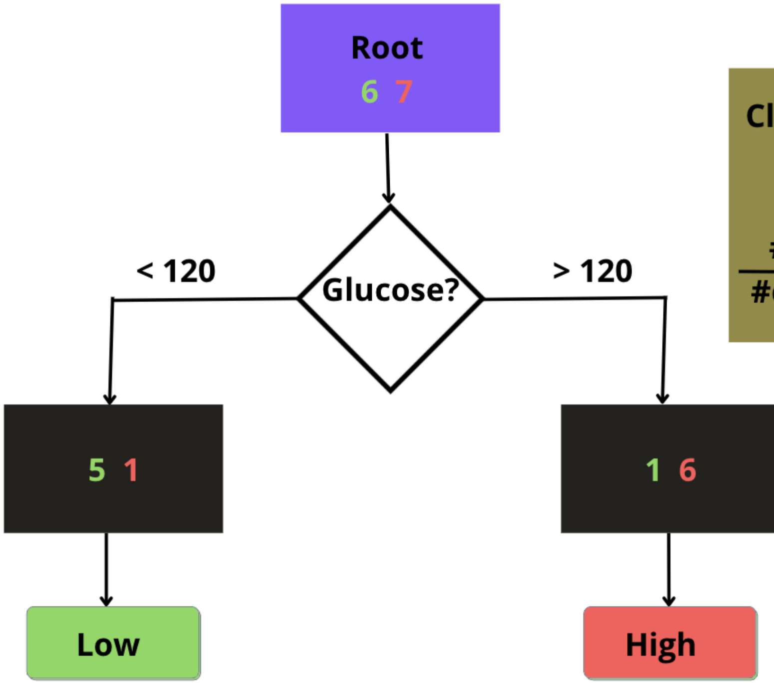
Majority class = "Low risk"

Majority class = "High risk"

Split selection



Split Effectiveness



Classification Error $= \dfrac{\#\text{mistakes}}{\#\text{data points}}$

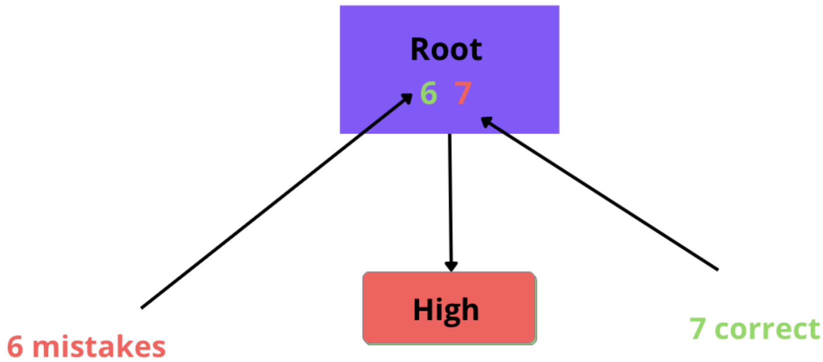
1

1

#mistakes = 2

2/13

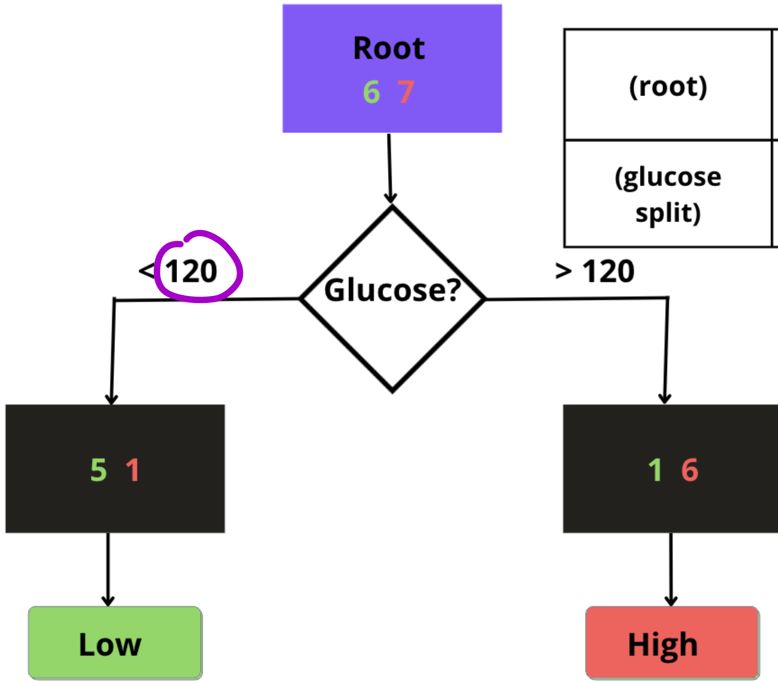
Calculate Classification Error



Tree	Classification Error
(root)	0.46

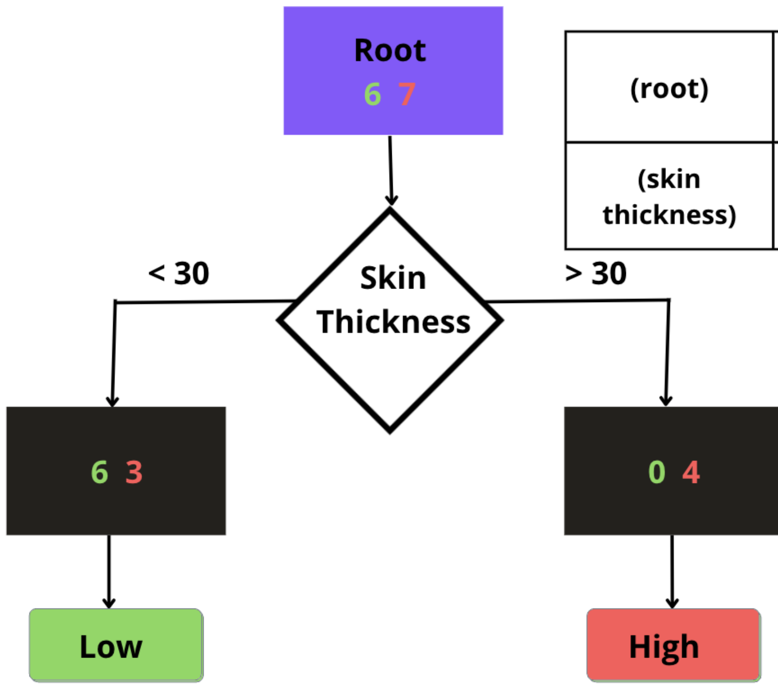
$\rightarrow$ 6/13

Split on Glucose



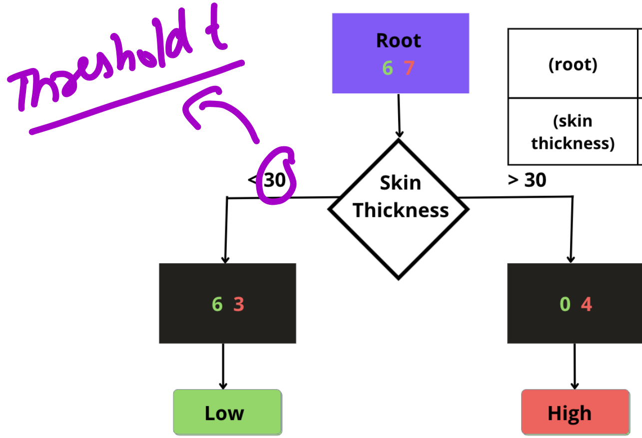
(root)	0.46
(glucose split)	$2/13 = 0.154$

Split on Skin Thickness (ICE #1)



(root)	0.46
(skin thickness)	?

Split on Skin Thickness (ICE #1)

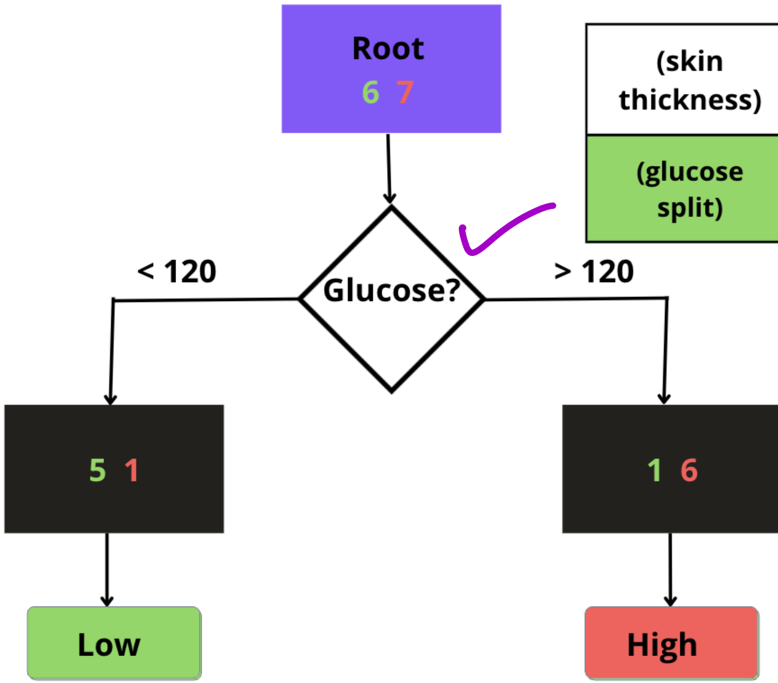


(root)	0.46
(skin thickness)	

Whats the misclassification error of splitting on **skin thickness**?

a. 0.07
b. 0.154
c. 0.23
d. 0.31

Split Winner



(skin thickness)	3/13
(glucose split)	2/13

Split Thickness

Relationship of Skin Thickness to Duration of Diabetes, Glycemic Control, and Diabetic Complications in Male IDDM Patients

Andrew Collier, MRCP
Alan W. Patrick, MRCP
Derek Bell, MRCP
David M. Matthews, MRCP
Cecilia C.A. MacIntyre, MSc
David J. Ewing, FRCP
Basil F. Clarke, FRCP

Skin thickness is primarily determined by collagen content and is increased in insulin-dependent diabetes mellitus (IDDM). We measured skin thickness in 66 IDDM patients aged 24–38 yr and investigated whether it correlated with long-term glycemic control and the presence of certain diabetic complications. With univariate analysis, skin thickness was increased and significantly related to duration of diabetes ($P < .001$), previous glycemic control ($P < .001$), retinopathy ($P < .001$), cheiroarthropathy ($P < .001$), and vibration-perception threshold ($P < .05$). There was a negative correlation between forced expiratory volume at 1 s ($P < .05$) and vital capacity ($P < .05$) with duration of diabetes. Neither skin thickness nor ankle arteriomedial wall calcification correlated with abnormal autonomic function tests. When corrected for duration of diabetes, there was a weak correlation between skin thickness and glycemic control ($P < .05$) but no correlation with

tinely used as indices of glycemic control (3–5). Collagen is the most studied protein regarding advanced NEG, because of the ease with which it can be examined in skin biopsies, and because of its importance as a protein that is present in several tissues subject to complications in diabetes, e.g., vascular basement membrane, arterial wall, and lung (6–10).

Skin thickness (epidermal surface to dermal fat interface), which is primarily determined by collagen content, is greater in insulin-dependent diabetes mellitus (IDDM) patients who have been diabetic for >10 yr (11,12). This possibly reflects increased collagen cross-linkage and reduced collagen turnover (2,3).

The aims of this study were to investigate whether the increase in skin thickness related to long-term glycemic control and correlated with microangiopathic compli-

BMI

South Med J. Author manuscript; available in PMC 2016 Jan 1.
Published in final edited form as:
South Med J. 2015 Jan; 108(1): 29–36.
doi: 10.14423/SMJ.0000000000000214

PMCID: PMC4457375
NIHMSID: NIHMS691426
PMID: 25580754

OTHER FORMATS

PubReader | PDF (438K)

ACTIONS

Cite

Favorites

SHARE

RESOURCES

Similar articles in PubMed

The Relationship between BMI and Onset of Diabetes Mellitus and its Complications

Natallia Gray, Ph.D.,[1] Gabriel Picone, Ph.D., Frank Sloan, Ph.D., and Arseniy Yashkin, Ph.D.

▸ Author information ▸ Copyright and License information Disclaimer

The publisher's final edited version of this article is available at South Med J
See other articles in PMC that cite the published article.

Abstract

Go to: ▸

Objective

This study determines effects of elevated body mass index (BMI) on type 2 diabetes mellitus (DM) onset and its complications among U.S. elderly.

Design and Methods

Data came from the Medicare Current Beneficiary Survey (MCBS), 1991-2010. A Cox proportional hazard model was used to assess effects of elevated BMI at the first MCBS interview on DM diagnosis

Author Manuscript

Case Study: What factors increase risk of chronic diabetes

Glucose	Insulin	BMI	SkinThickness	Outcome
148	0	33.6	35	1
85	0	26.6	29	0
183	0	23.3	0	1
89	94	28.1	23	0
137	168	43.1	35	1
116	0	25.6	0	0
78	88	31	32	1
115	0	35.3	0	0
197	543	30.5	45	1
125	0	0	0	1
110	0	37.6	0	0
168	0	38	0	1
139	0	27.1	0	0

Case Study: What factors increase risk of chronic diabetes

SkinThickness > 30

SkinThickness	Outcome	XOR
35	1	FALSE
29	0	FALSE
0	1	TRUE
23	0	FALSE
35	1	FALSE
0	0	FALSE
32	1	FALSE
0	0	FALSE
45	1	FALSE
0	1	TRUE
0	0	FALSE
0	1	TRUE
0	0	FALSE

#mistakes = 3

Case Study: What factors increase risk of chronic diabetes

Glucose	Outcome	XOR
148	1	FALSE
85	0	FALSE
183	1	FALSE
89	0	FALSE
137	1	FALSE
116	0	FALSE
78	1	TRUE
115	0	FALSE
197	1	FALSE
125	1	FALSE
110	0	FALSE
168	1	FALSE
139	0	TRUE

2/13

Case Study: What factors increase risk of chronic diabetes

Insulin	Outcome	XOR
0	1	TRUE
0	0	FALSE
0	1	TRUE
94	0	TRUE
168	1	FALSE
0	0	FALSE
88	1	FALSE
0	0	FALSE
543	1	FALSE
0	1	TRUE
0	0	FALSE
0	1	TRUE
0	0	FALSE

5/13

Case Study: What factors increase risk of chronic diabetes

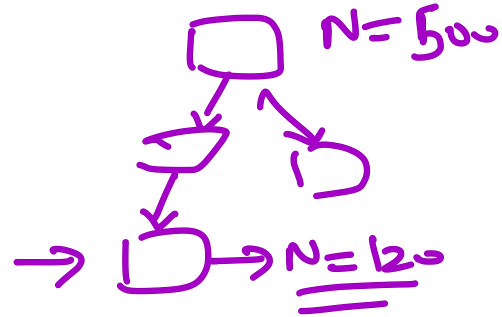
BMI	Outcome	XOR
33.6	1	FALSE
26.6	0	FALSE
23.3	1	TRUE
28.1	0	FALSE
43.1	1	FALSE
25.6	0	FALSE
31	1	FALSE
35.3	0	TRUE
30.5	1	FALSE
0	1	TRUE
37.6	0	TRUE
38	1	FALSE
27.1	0	FALSE

4/13

Case Study: What factors increase risk of chronic diabetes

	Glucose	SkinThickness	BMI	Insulin
Mis-Classification Error	2	3	4	5

Split selection



Split selection procedure

- Given a subset of data set, M at a node
- For each remaining feature $h_i(x)$, split M by feature $h_i(x)$ and compute classification error
- Pick the feature i to split with minimum classification error

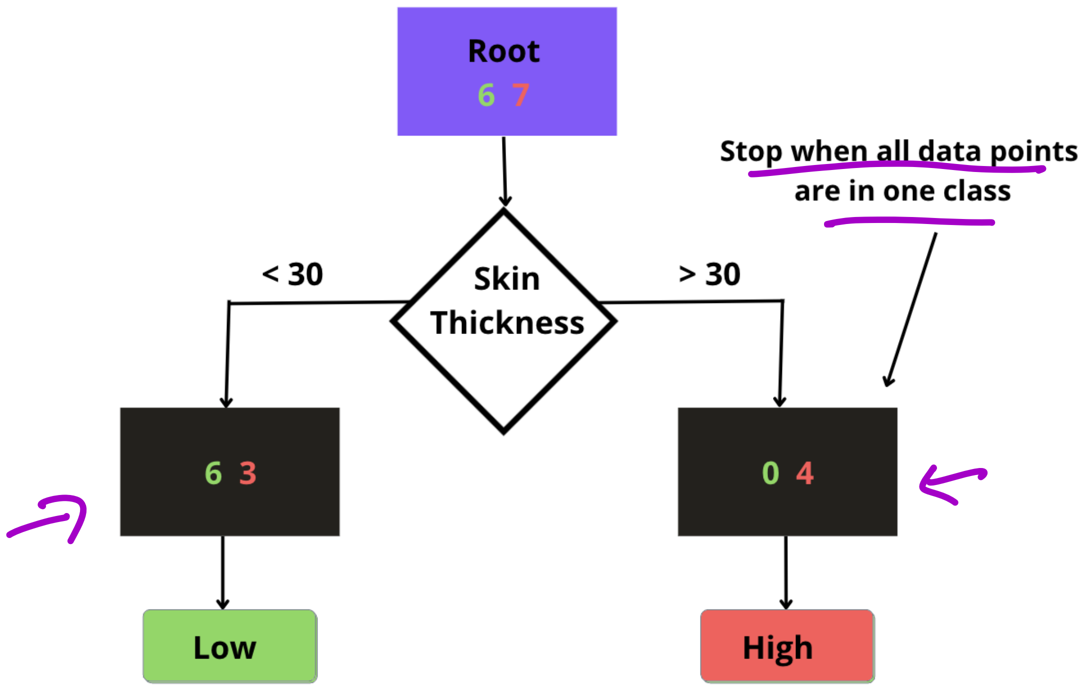
Decision Tree Classification as a Greedy Procedure

DT Classifier Training procedure

If classification splits satisfy criteria (e.g. low classification error), stop,
Else, split further using split selection procedure.

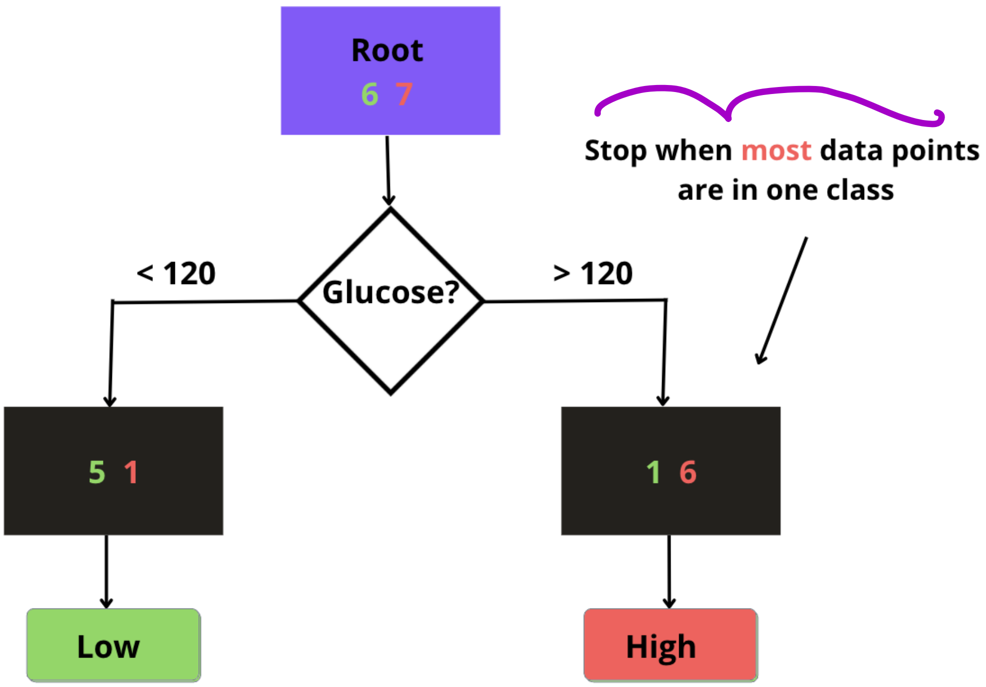
1[illegible] 3

Stopping



Root
6 7
Stop when all data points
are in one class
< 30
> 30
Skin
Thickness
6 3
0 4
Low
High

Stopping



Root
6 7
Stop when most data points
are in one class
Glucose?
< 120
> 120
5 1
1 6
Low
High

Stopping criteria in practice

- Splits with few data points can lead to over-fitting. Example

Stopping criteria in practice

1. Splits with few data points can lead to over-fitting. Example
2. Max tree depth can be a stopping criteria to prevent over-fitting.

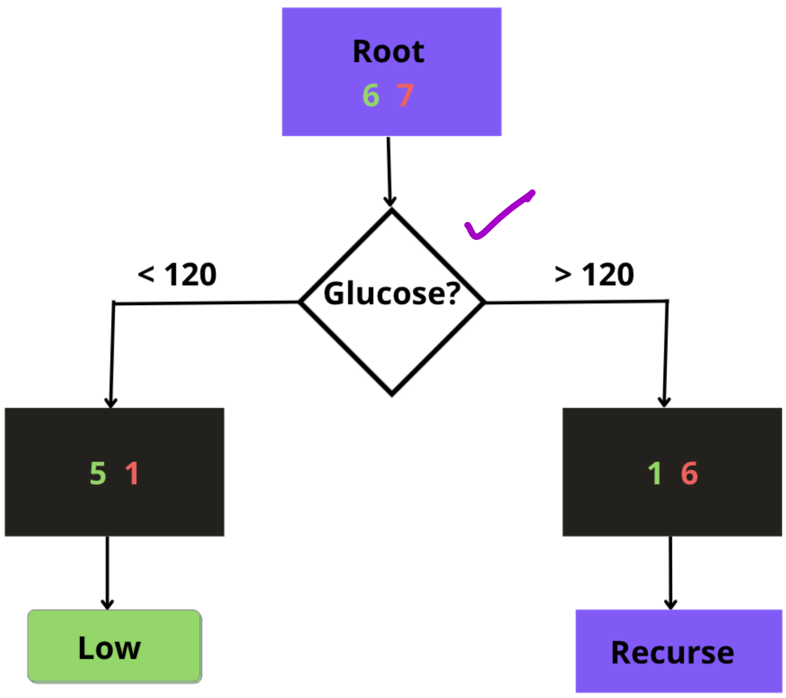
Stopping criteria in practice

1. Splits with few data points can lead to over-fitting. Example
2. Max tree depth can be a stopping criteria to prevent over-fitting.
3. Although theoretically, can aim for 0 classification error - This would lead to over-fitting. Use above 2 to stop earlier.

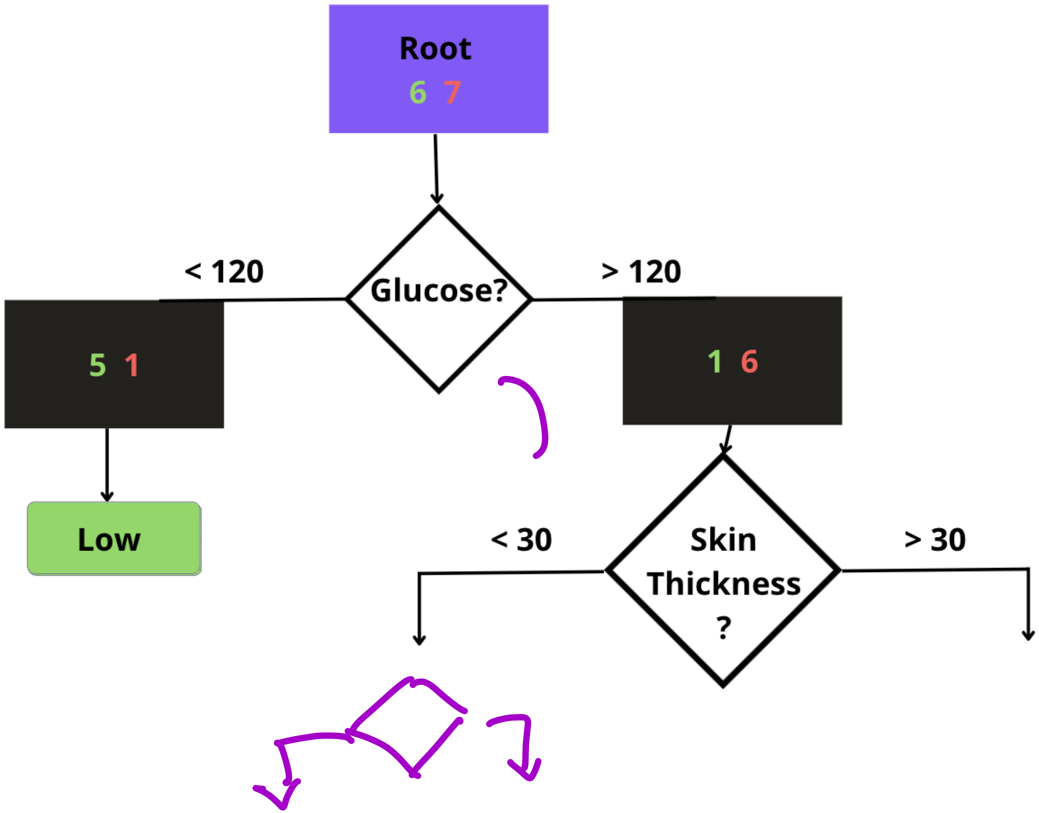
Stopping criteria in practice

1. Splits with few data points can lead to over-fitting. Example
2. Max tree depth can be a stopping criteria to prevent over-fitting.
3. Although theoretically, can aim for 0 classification error - This would lead to over-fitting. Use above 2 to stop earlier.
4. No standard 'regularization' for DTs like for Logistic Regression.

Recursive Splits

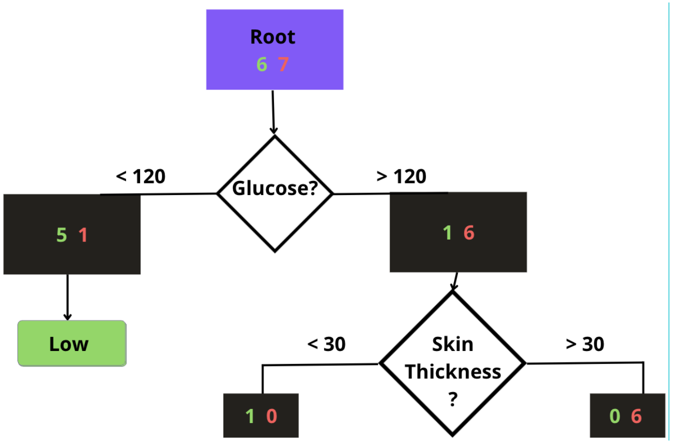


Second level DT



ICE #2

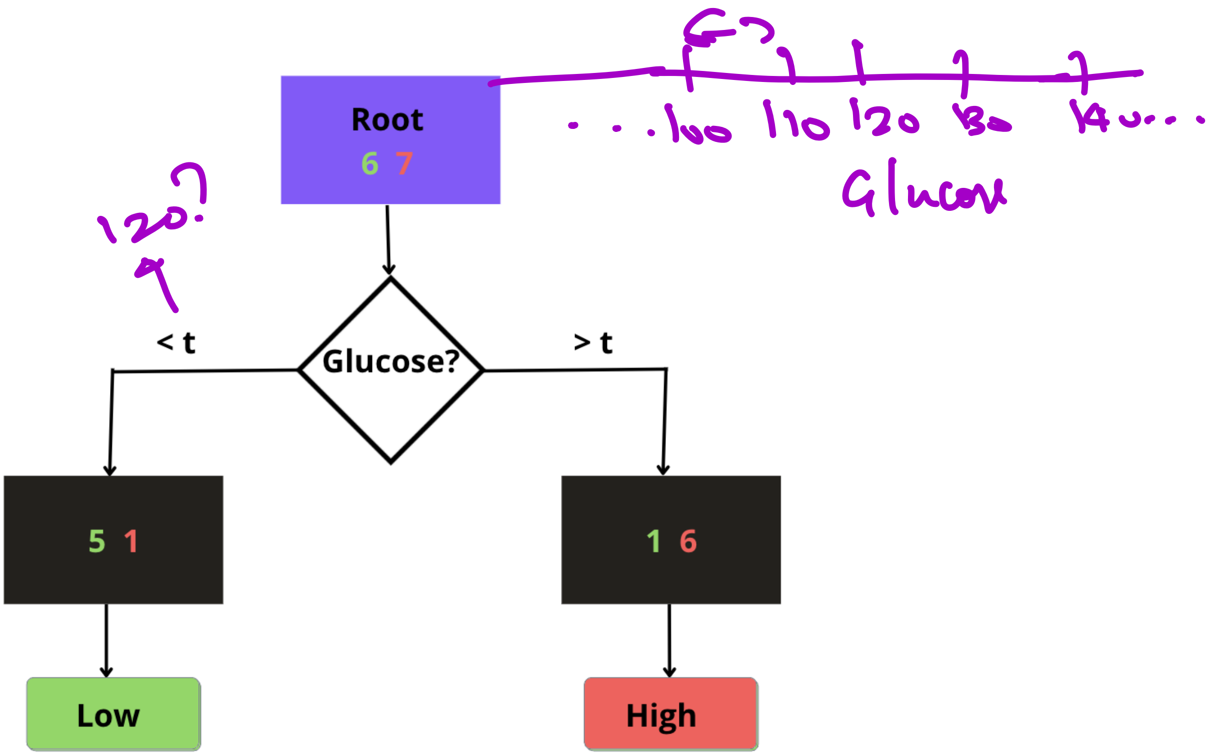
Classification error



The classification error for the DT above is:

a. 0.07
b. 0.154
c. 0.23
d. 0.31

Threshold splits for real valued features



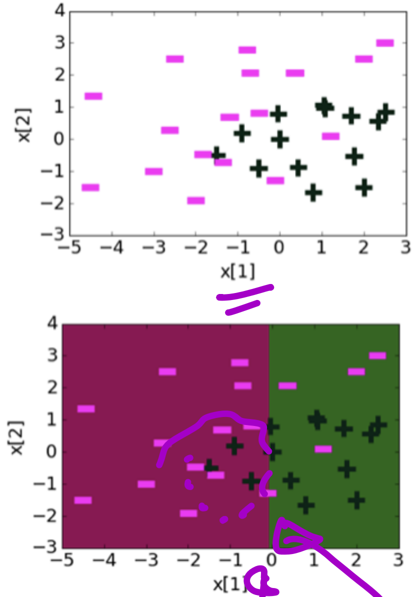
Choosing Split Threshold for Numeric Features

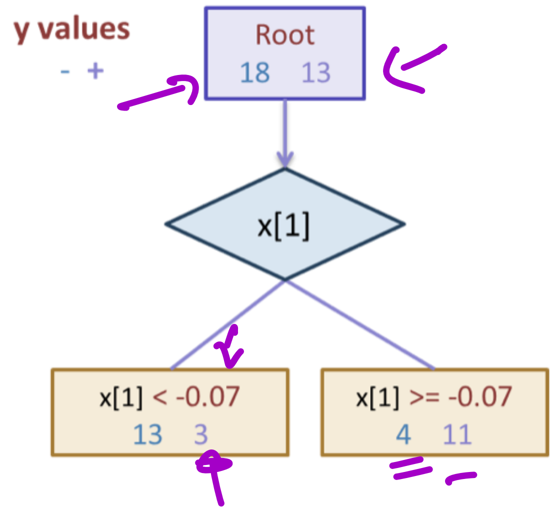
1. Grid search?
2. Numeric vs Categorical Features: Can recurse more than once on a numeric feature. Can't do the same for categorical feature. Why?

Decision Boundary level 1 || Numeric Features

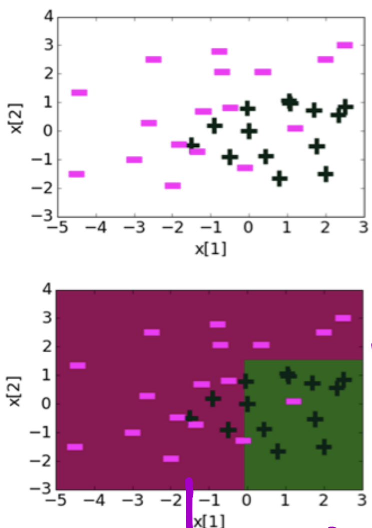
Conceptual level

$\in \mathbb{R}^2$





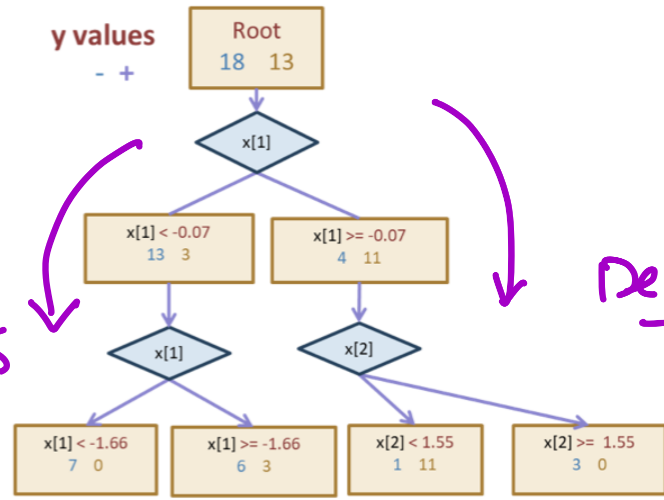
Decision Boundary level 2 || Numeric Features



1.55

-1.66

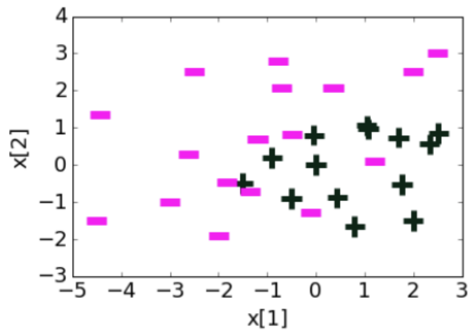
Non-Linear DB



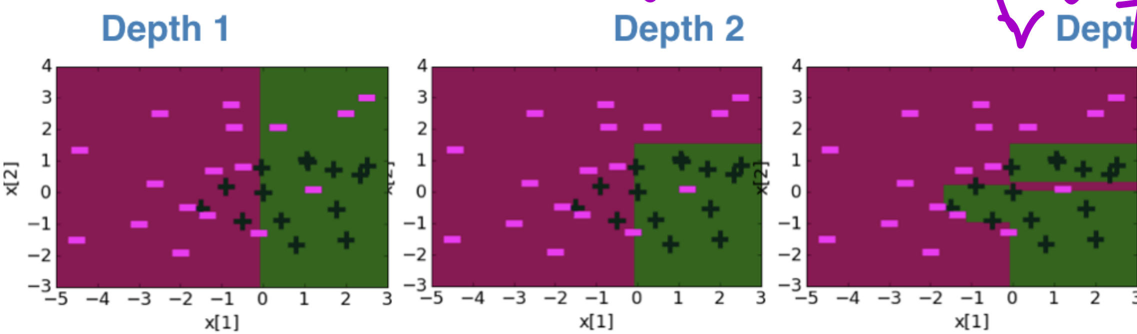
Depth=2

Decision Boundary level 3 || Numeric Features

- Decision boundaries can be complex!



overfit!



Decision Trees Summary

Summary

- Intuitive way to classify by making decsisions by walking down the tree
- Can learn complex non-linear decision boundaries (unlike logistic regression)
- Prone to overfit as tree depth increases (unlike logistic regression)
- Splitting at nodes with few data points can lead to overfitting
- Over-fitting can be avoided by early stopping (depth or error)
- Improve Decision Trees - Random Forests - Next Lecture!

overfitting issue

Decision Trees vs Logistic Regression

1. Both are interpretable in different ways

Decision Trees vs Logistic Regression

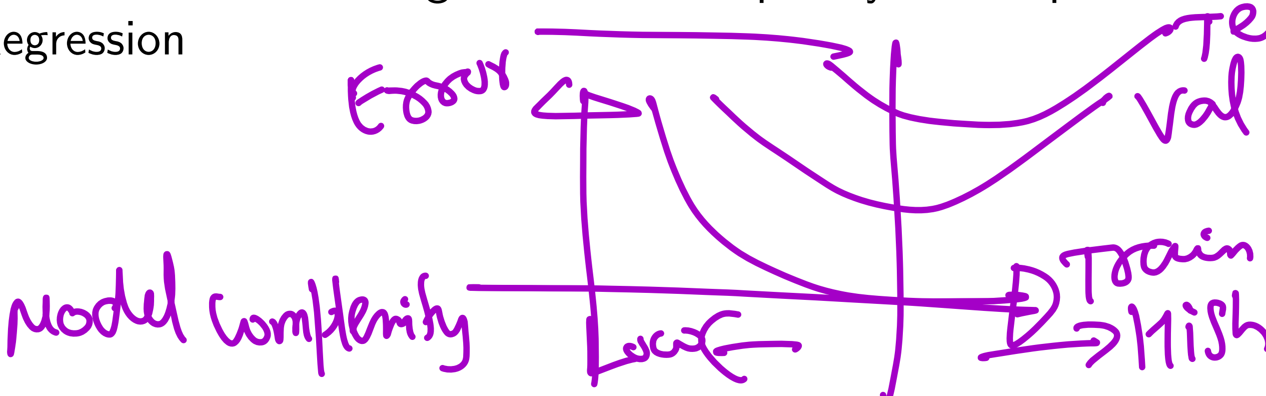
1. Both are interpretable in different ways
2. Decision trees mimick how humans make decisions and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large

Decision Trees vs Logistic Regression

1. Both are interpretable in different ways
2. Decision trees mimick how humans make decisions and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large
3. Decision Trees can easily learn non-linear decision boundaries while Logistic Regression learns linear decision boundary

Decision Trees vs Logistic Regression

1. Both are interpretable in different ways
2. Decision trees mimick how humans make decisions and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large
3. Decision Trees can easily learn non-linear decision boundaries while Logistic Regression learns linear decision boundary
4. Decision Tree has a higher model complexity as compared to Logistic Regression



Decision Trees vs Logistic Regression

1. Both are interpretable in different ways
2. Decision trees mimick how humans make decisions and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large
3. Decision Trees can easily learn non-linear decision boundaries while Logistic Regression learns linear decision boundary
4. Decision Tree has a higher model complexity as compared to Logistic Regression
5. Logistic Regerssion is less prone to over-fitting than Decision Trees with large number of features