t-SNE

t distribution

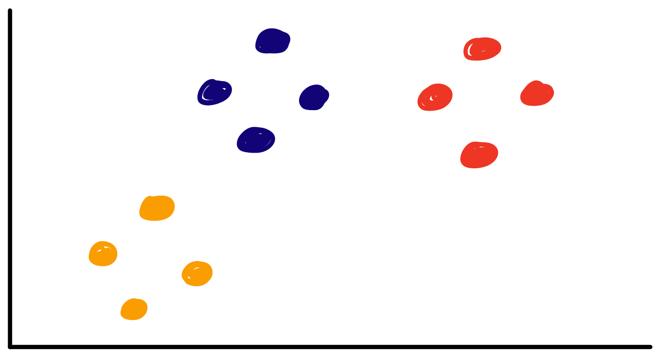
Stochastic Neighbour Embedding

↳ Probabilities

B: 1
R: 0
} hard

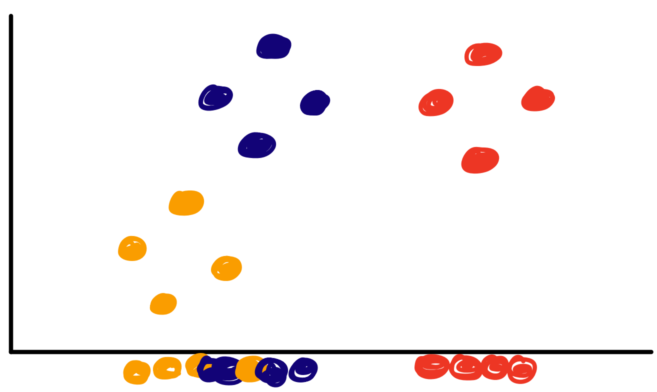
B: 0.8
R: 0.2
} stochastic

2D Scatter Plot

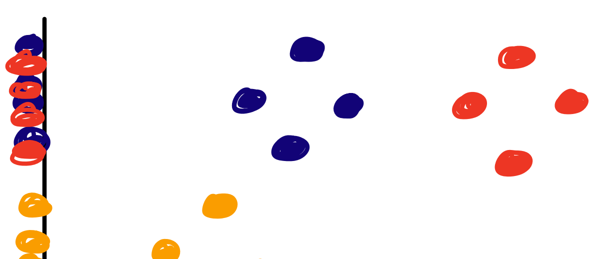


1D number line

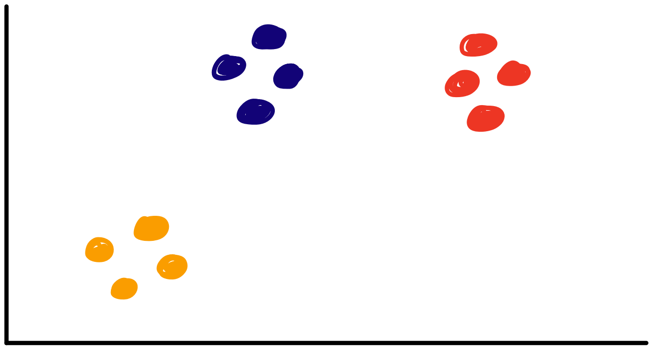
Spread (Projection on x)

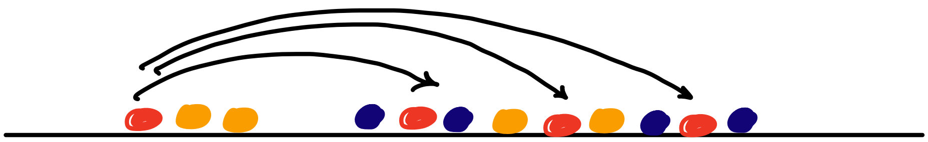


Projection on y



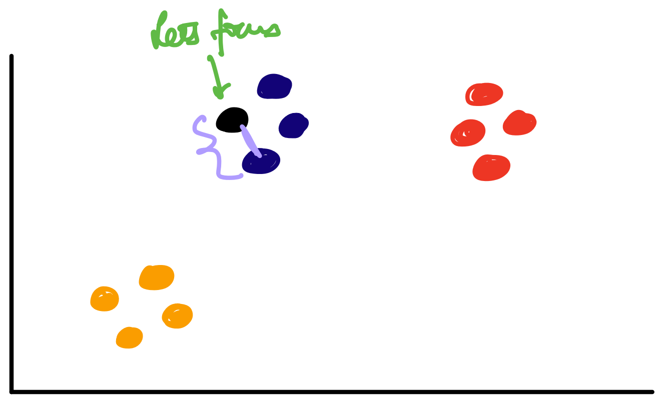
Intuition



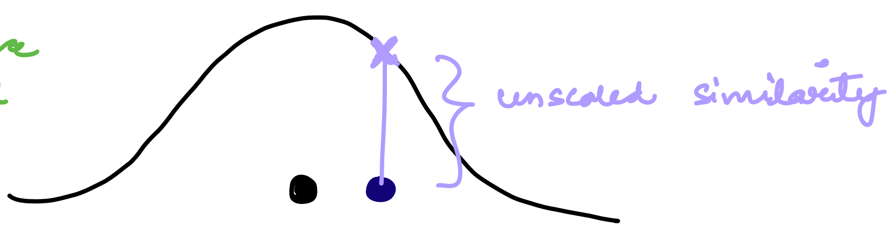


Step 1:

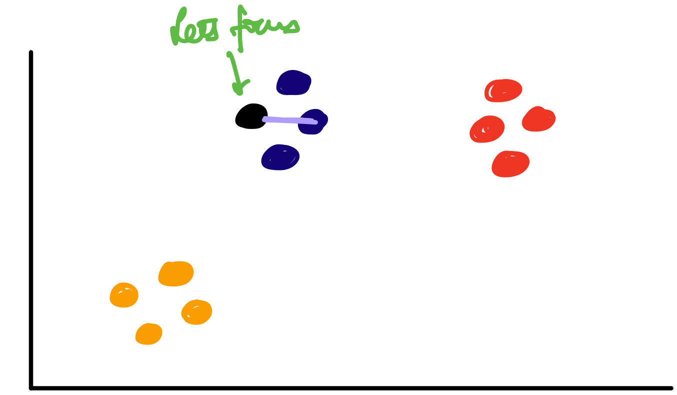
Determine the similarity of all the points in the scatter plot

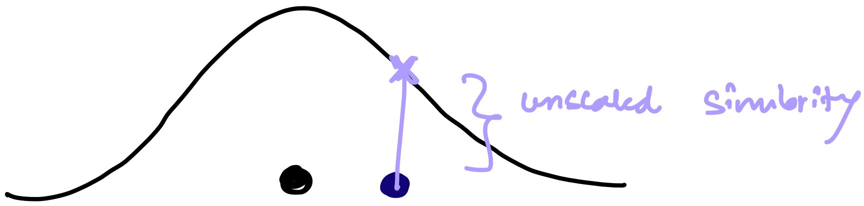


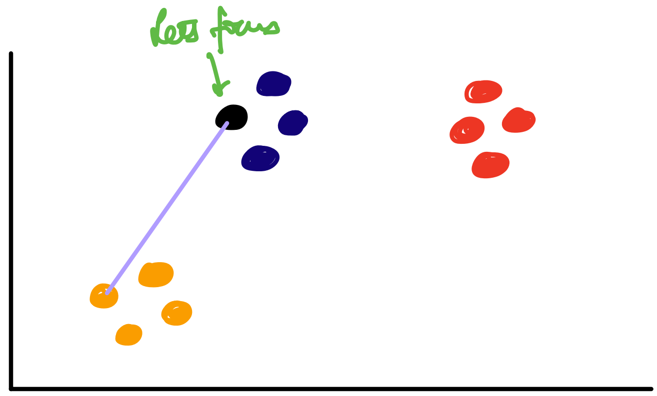
Plot the distance on a normal curve that is centered on point of interest



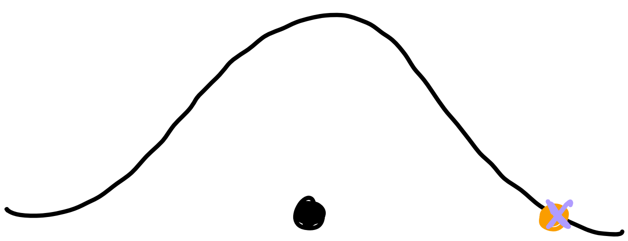
Close points have high similarity



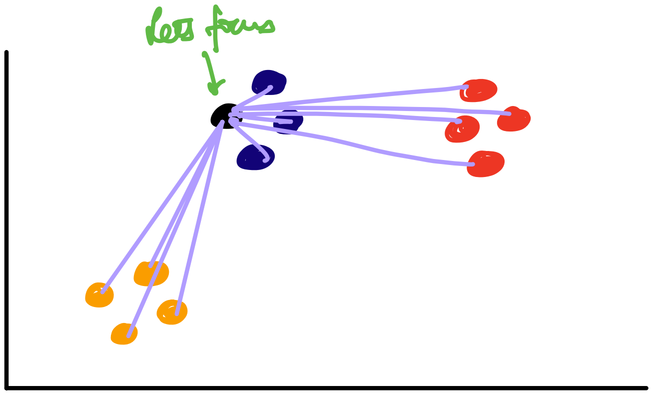


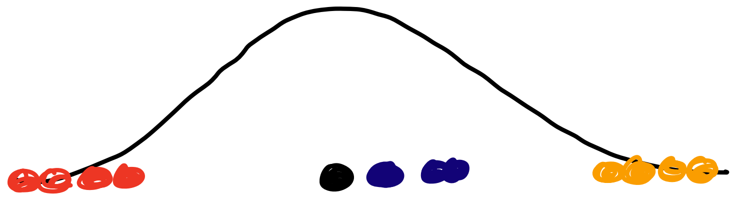


Distant points have very low similarity

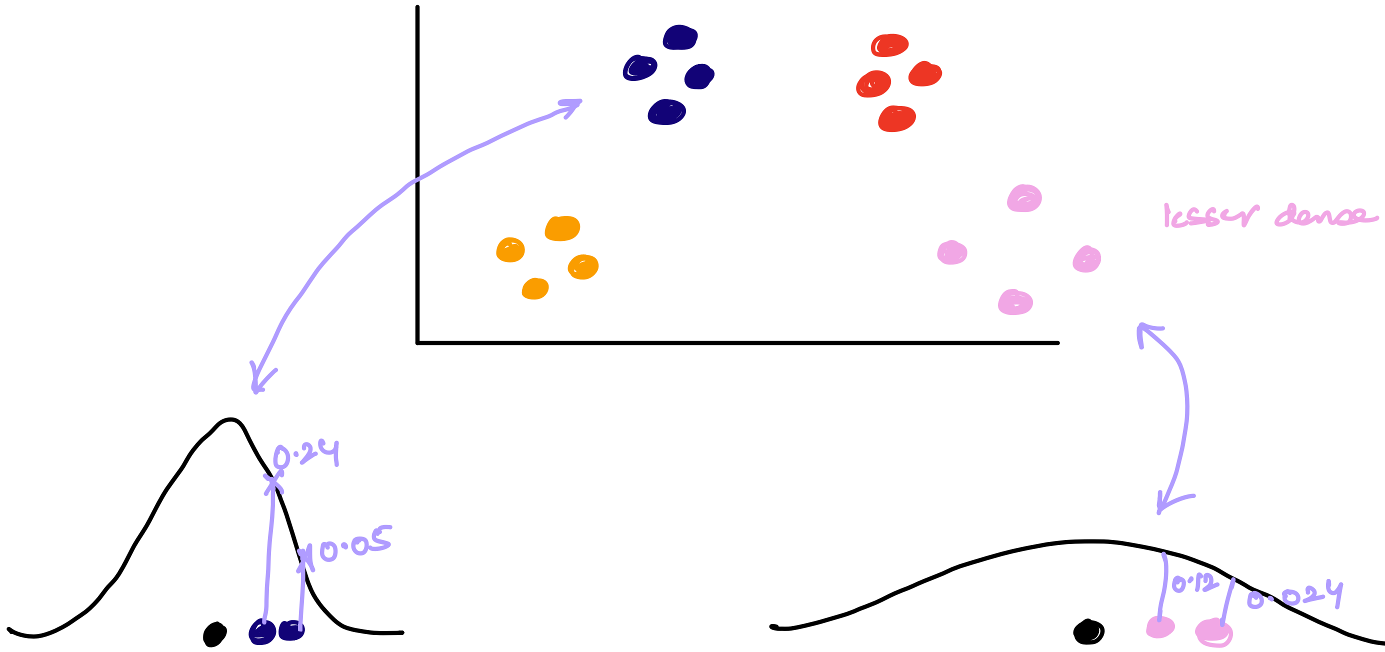


we measure the distances b/w all the points & point of interest.



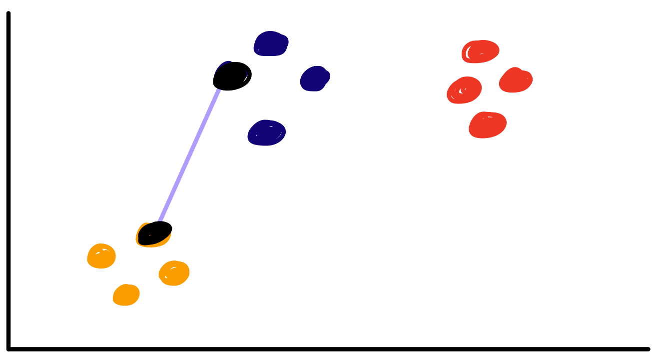


Unscaled → Scaled ?

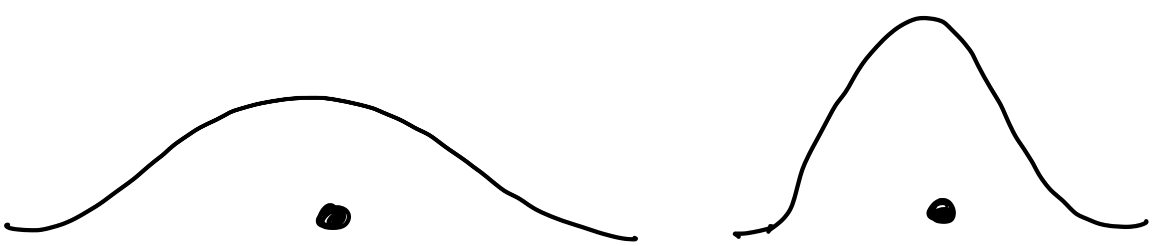


$$\frac{0.24}{0.24+0.05}, \quad \frac{0.05}{0.24+0.05} \qquad\qquad \frac{0.12}{0.12+0.024}, \quad \frac{0.024}{0.12+0.024}$$

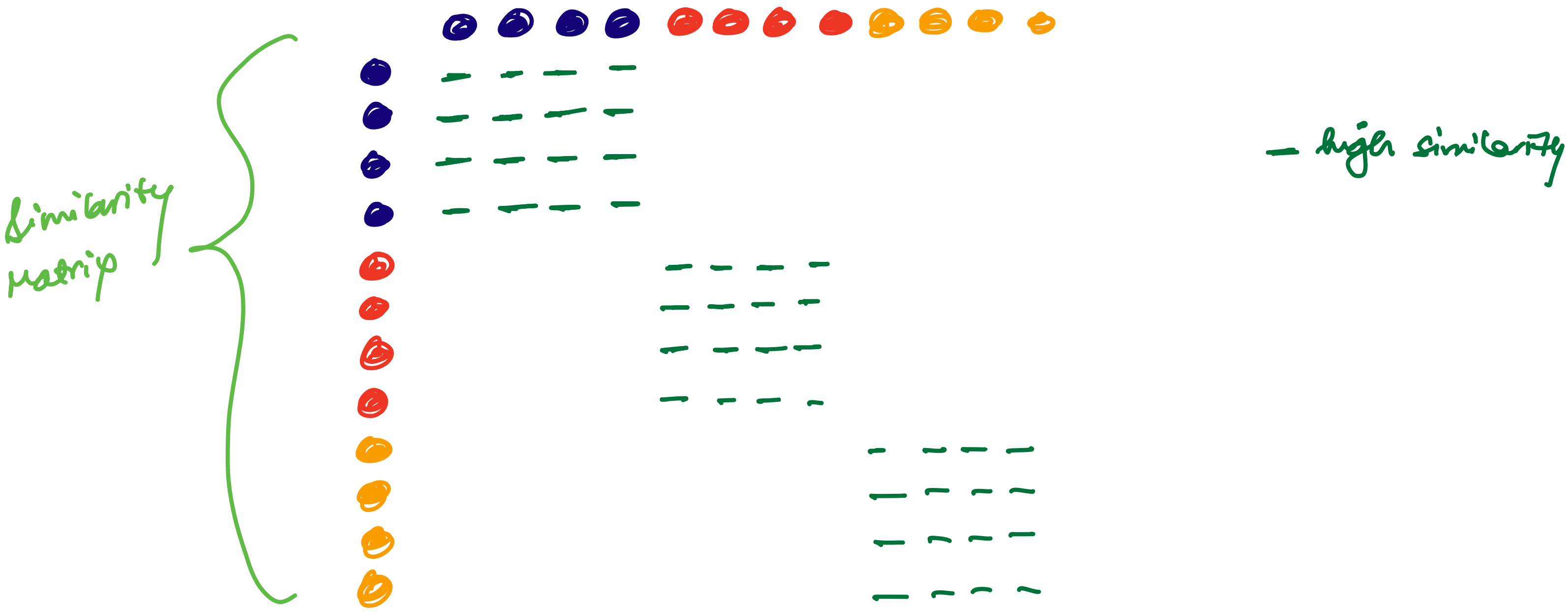
$$0.82, \quad 0.18 \qquad\qquad 0.82, \quad 0.18$$

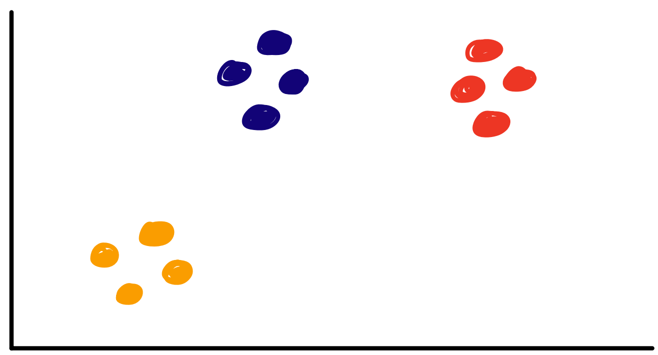


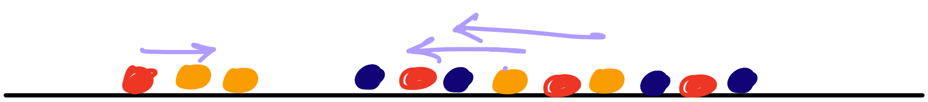
Keep each point in focus & calculate the distance.

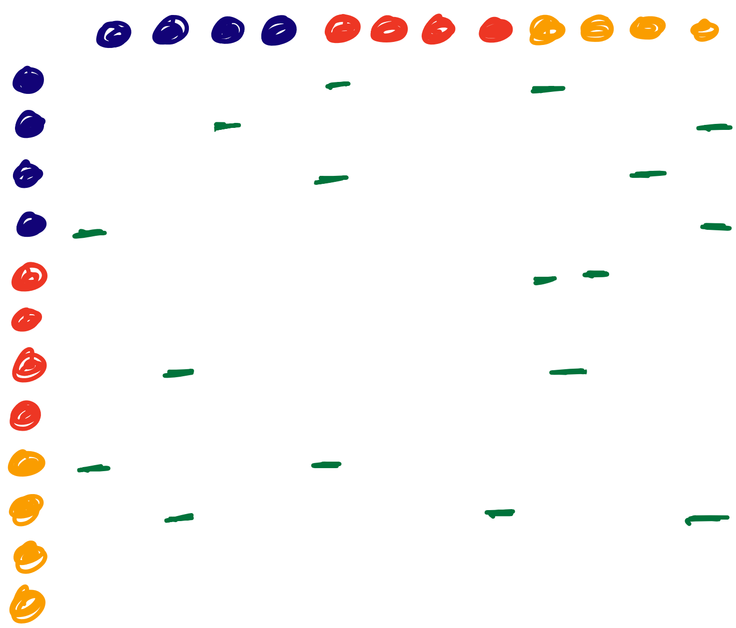


Avg similarity score from both dr's


similarity matrix
— high similarity







Fairness

Sensitive: Race, Age, Gender...

Group fairness

males & females

Individual fairness

2 individuals

$P(A|B)$: Probability of A happening given that B has already happened.

y = actual y / true y

$\hat{y}$ = predicted y

G = groups

Group fairness:

- **Statistical Parity / Demographic Parity:**

$$P(\hat{y}=1 | G=f) = P(\hat{y}=1 | G=m)$$

	group	$\hat{y}$
1.	f	✓
2.	f	
3.	f	✓
4.	f	
5.	m	✓
6.	m	
7.	m	✓
8.	m	
9.	m	✓
10.	m	
11.	m	✓
12.	m	

$$\frac{2}{4} \qquad \frac{4}{8}$$

$$\frac{1}{2} = \frac{1}{2} \quad ✓$$

males:
$P(\hat{y}=1 | G=m)$

0.4 — 0.6 — 0.8

20% 20%

$0.4 \rightarrow 0.8$

thats fine

	group	$\hat{y}$
1.	f	✓
2.	f	
3.	f	
4.	f	
5.	m	✓

$$\frac{1}{4} \qquad \frac{6}{8}$$

not fair

5.	m	✓
6.	m	
7.	m	✓
8.	m	✓
9.	m	✓
10.	m	✓
11.	m	✓
12.	m	

Equal opportunity

(TPR should be same across both groups)

$$P(\hat{y}=1 \mid Y=1, G=f) = P(\hat{y}=1 \mid Y=1, G=m)$$

$$\frac{1}{2} \qquad \frac{2}{6}$$

	group	y	$\hat{y}$
1.	f	✓	✓
2.	f		
3.	f	✓	
4.	f		✓
5.	m	✓	
6.	m		
7.	m	✓	✓
8.	m	✓	
9.	m	✓	✓
10.	m		✓
11.	m	✓	
12.	m	✓	

Equalized odds:

TPR & FPR should be same across groups

$$P(\hat{y}=1 \mid Y=1, G=f) = P(\hat{y}=1 \mid Y=1, G=m)$$

and

$$P(\hat{y}=1 \mid Y=0, G=f) = P(\hat{y}=1 \mid Y=0, G=m)$$

Overall Accuracy Equality:

Acc. should be same across both the groups

$$\frac{TP_f + TN_f}{TP_f + TN_f + FP_f + FN_f} = \frac{TP_m + TN_m}{TP_m + TN_m + FP_m + FN_m}$$

Individual fairness:

- Causal Independence (Causal Discrimination)

CGPA, Branch, Exp., Projects, Gender

m

f

Result should be same for these 2 individuals.

- Fairness through Unawareness:

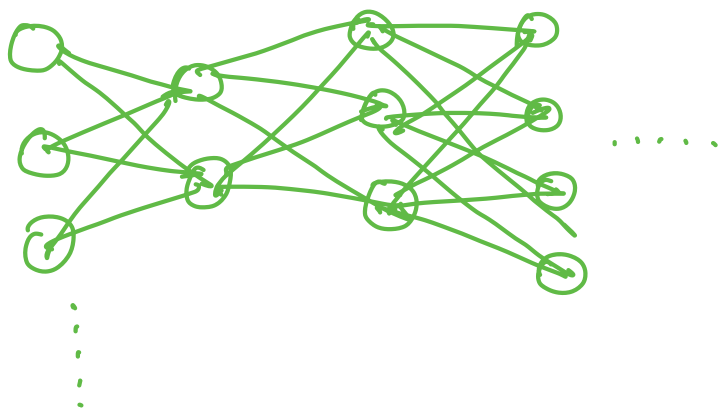
Sensitive attribute are excluded in the decision making process.

Explainable AI

- Transparency: Functioning & Decision making process.

C, D, H

(weights)

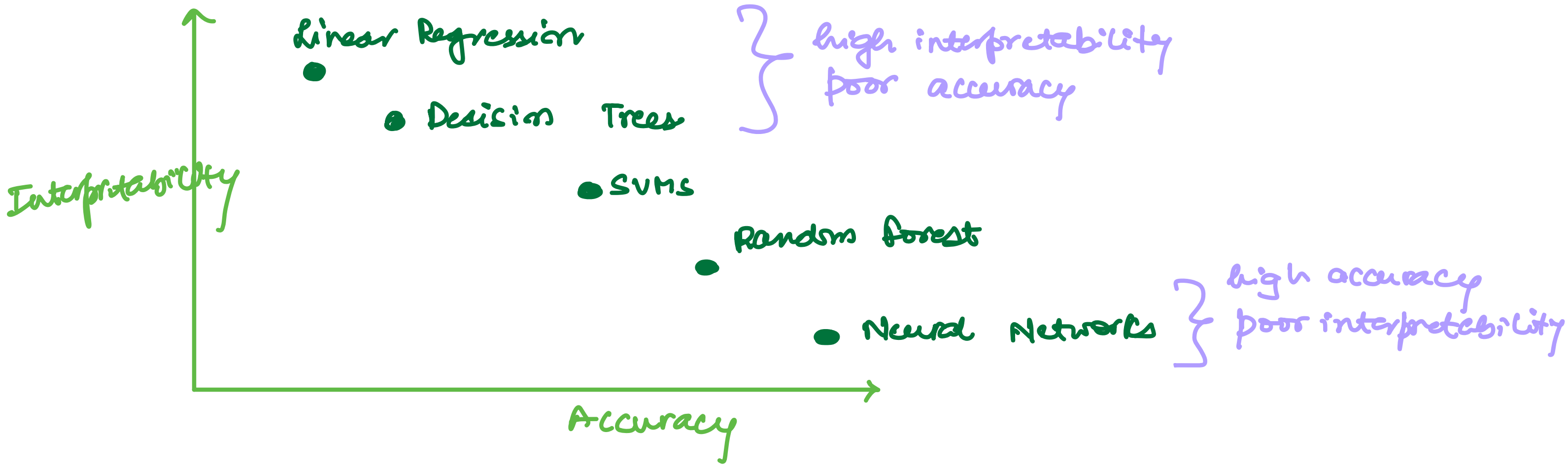


- Interpretability

Human understandable

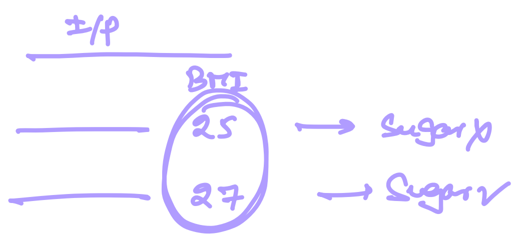
- Accountability

Interpretability v/s Accuracy

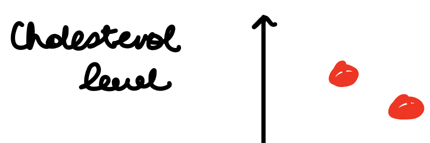


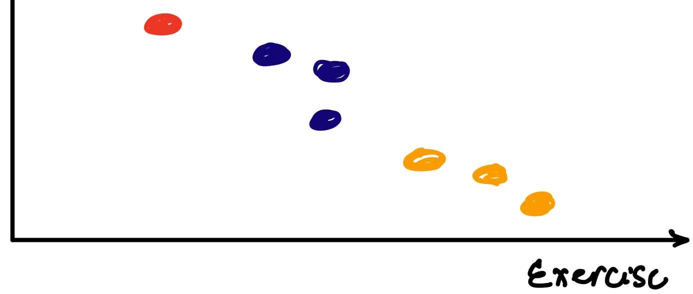
Interpretability

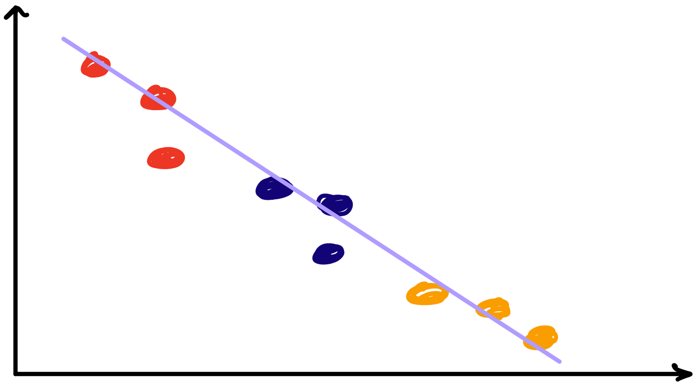
- Build interpretable ML models (Model based interpretable)
- Derive explanations for complex ML models (Post-hoc interpretability)
  - Black-box (Model parameters unknown. Derive explanations using the correlations b/w input & output)
  - White box (Parameters are unknown)

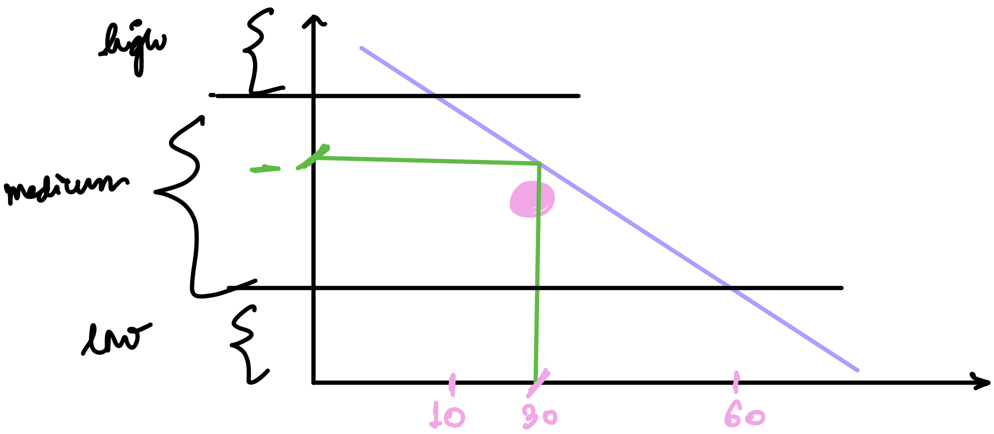


Linear Regression







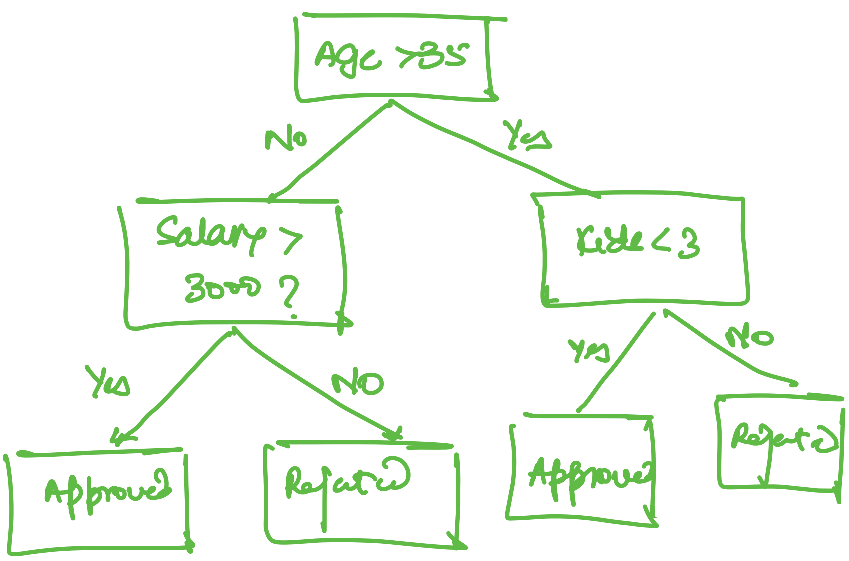


<10 : high
10 – 60 : medium
>60 : low

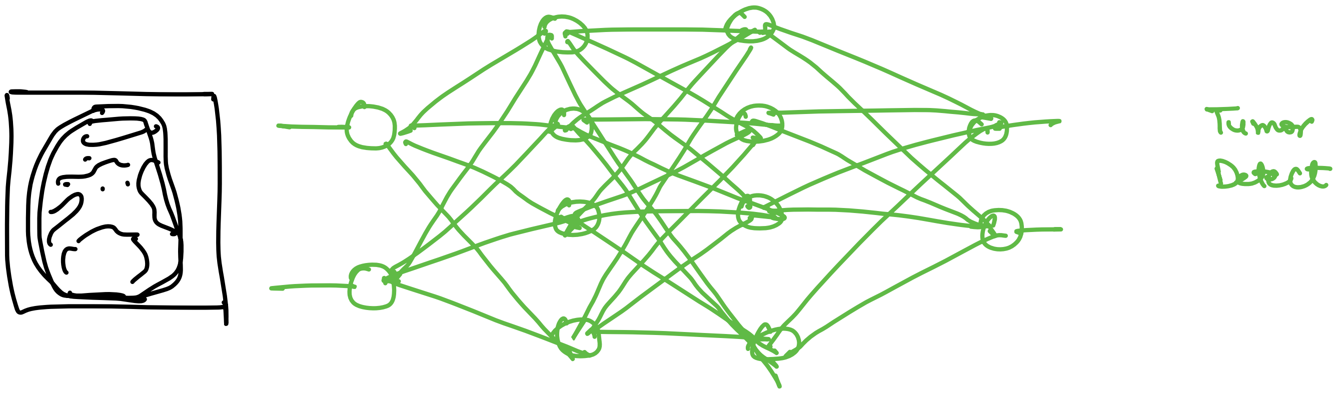
Decision Trees

Age: 32
Salary: 2800
Kids: 2
↓
Reject.

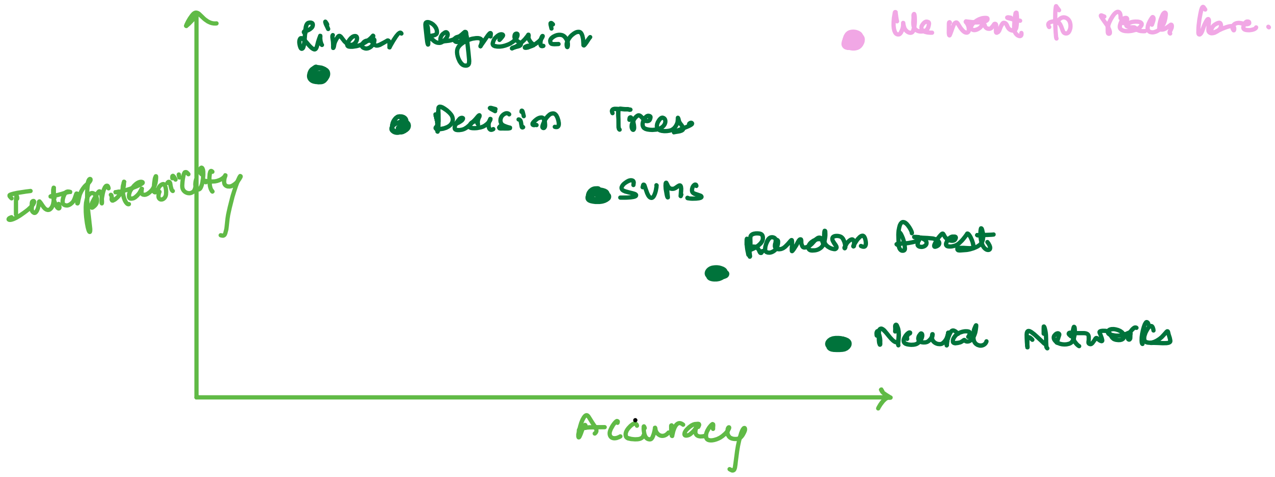
Loan approval?



Neural Networks

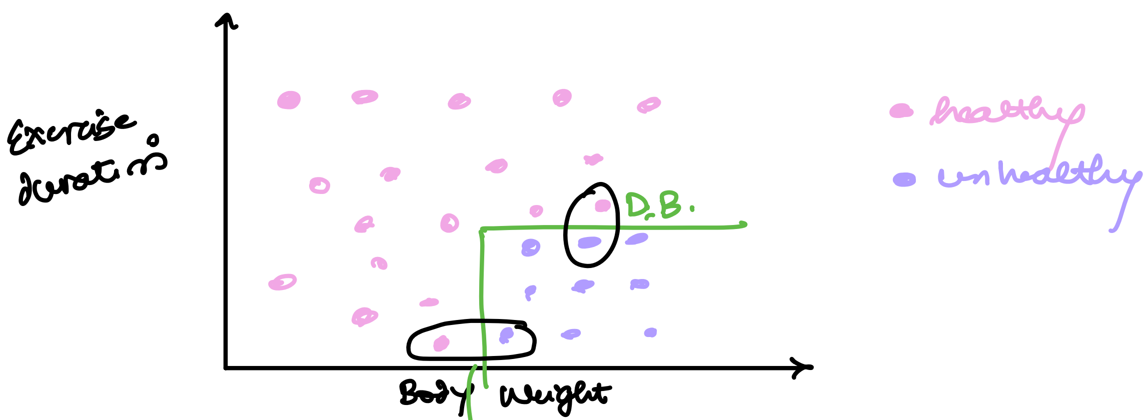


NN are not inherently interpretable. we don't know which feature got maximum importance.



LIME

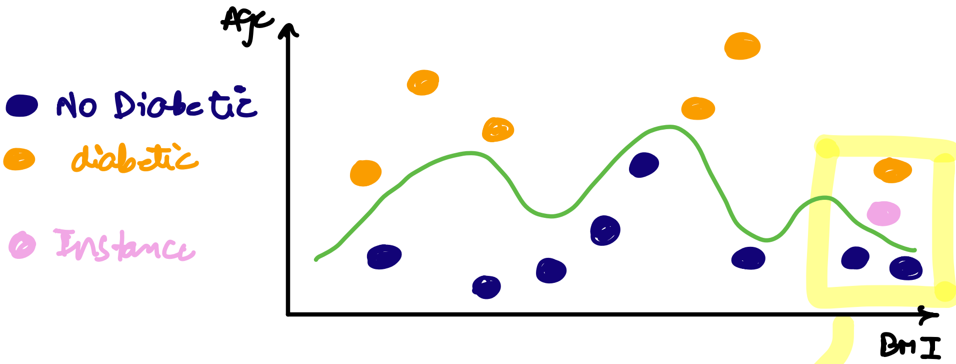
Local ⟶ local neighbourhood of the instance
Interpretable ⟶ A human should be able to interpret
Model-Agnostic ⟶ Applicable to all models
Explanations ⟶ Explanation that help interpretation



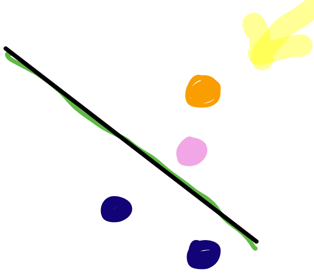
feature differentiating these 2 people ?

LIME Implementation:

1 Global model, local data point

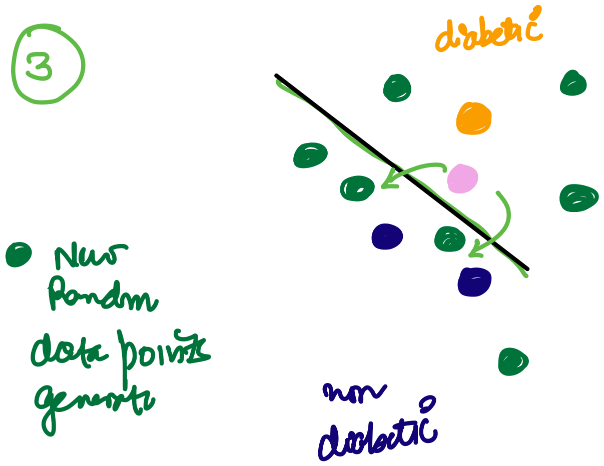


2 local neighbourhood

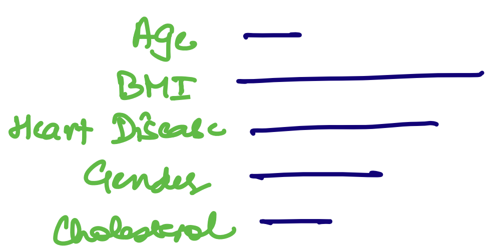


local linear model near the point

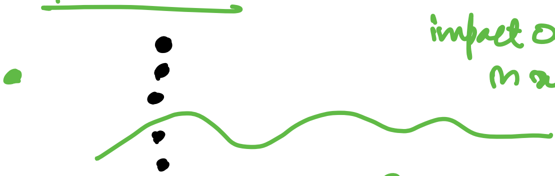
3



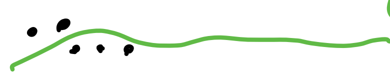
4 features relevant to the data point



Drawback:



impact of (feature on x axis)?



impact of (feature on y axis)?

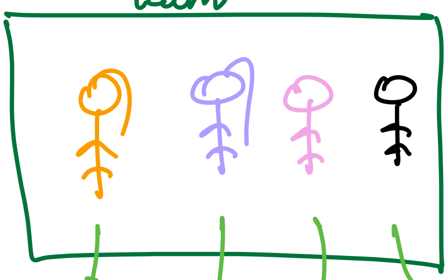
• Linear Assumption

SHAP

Shapley Additive exPlanations

Lloyd Shapley → Game theory

team



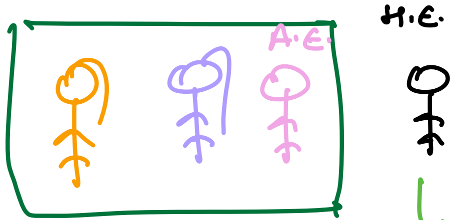
1000 Rs.

Distribution ?

Option 1:
Equal
Distribution:

250 250 250 250

Option 2:



A.E.

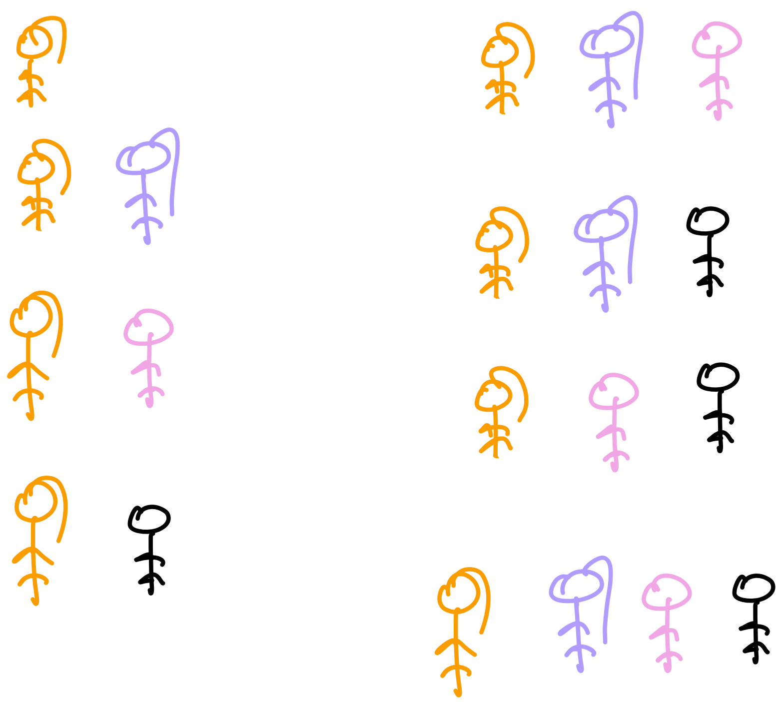
H.E. 2nd pos

600 Rs

400 Rs ?

we don't account for interactions ?

Contribution ?



- Consider different subsets.
- Calculate individual contribution of  in each subset
- weighted avg of these individual contribution will give you the contribution of each player.

① Age   ② Exercise Duration   ③ BMI   ④ Diabetic parents   ⑤ Calorie Intake

BMI → Diabetic   BMI → Not Diabetic

Subset: Age   Random Values   BMI   Random Values   Calorie Intake

Subset: Random Values   Random Values   BMI   Random Values   Calorie Intake

ML model

Data point

weight

M = total # features

z' = # features in subset

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \boxed{\frac{|z'|!\,(M - |z'| - 1)!}{M!}} \left[\underbrace{f_x(z')}_{60\%} - \underbrace{f_x(z' \setminus i)}_{15\%} \right]$$

Shapley value for feature i

Subsets

if features are:

Age | BMI | Calorie Intake

then the result from ML model

BMI | Calorie Intake

result from ML model

Drawback:

Computational Complexity is high

subsets: 2^n