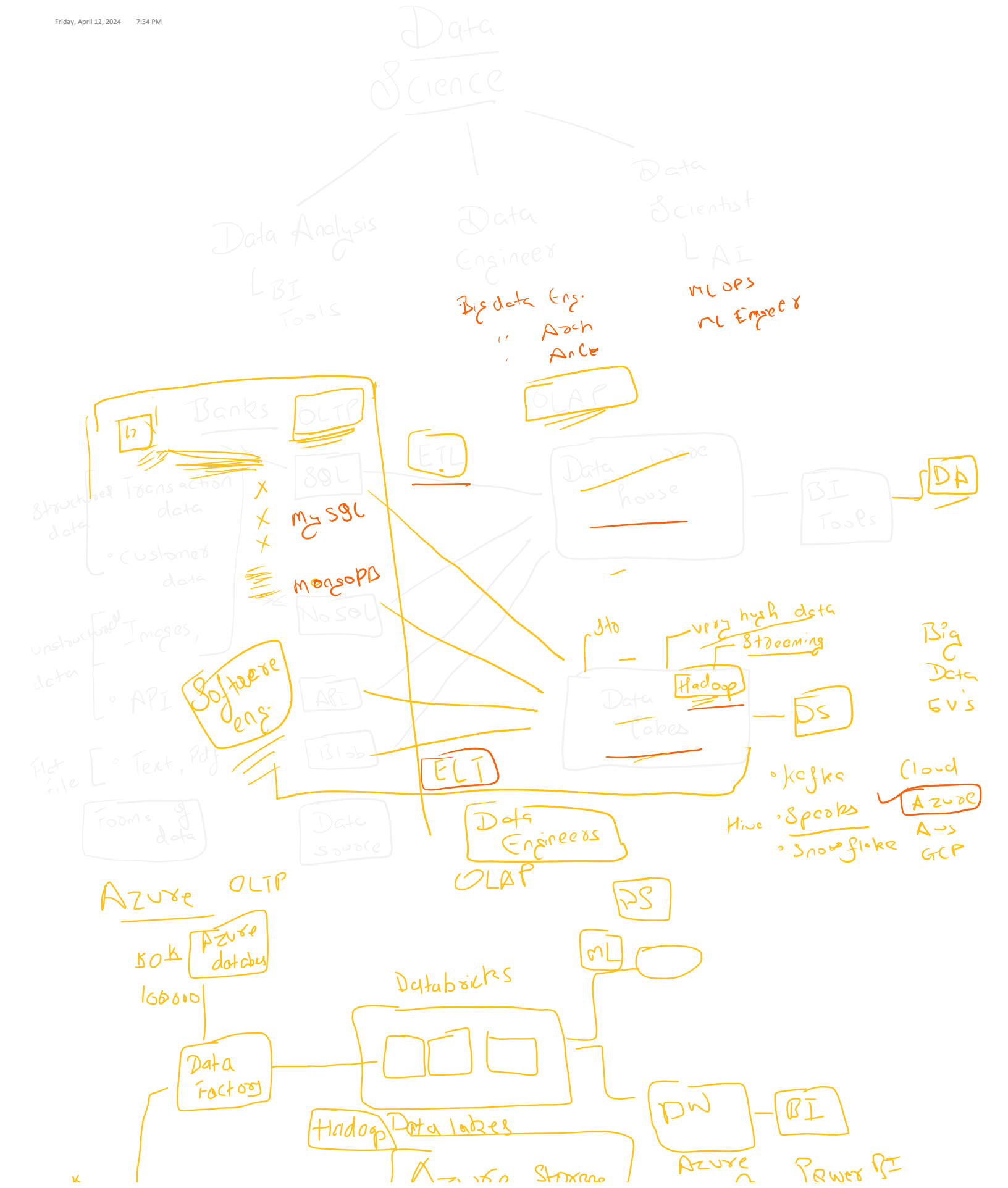
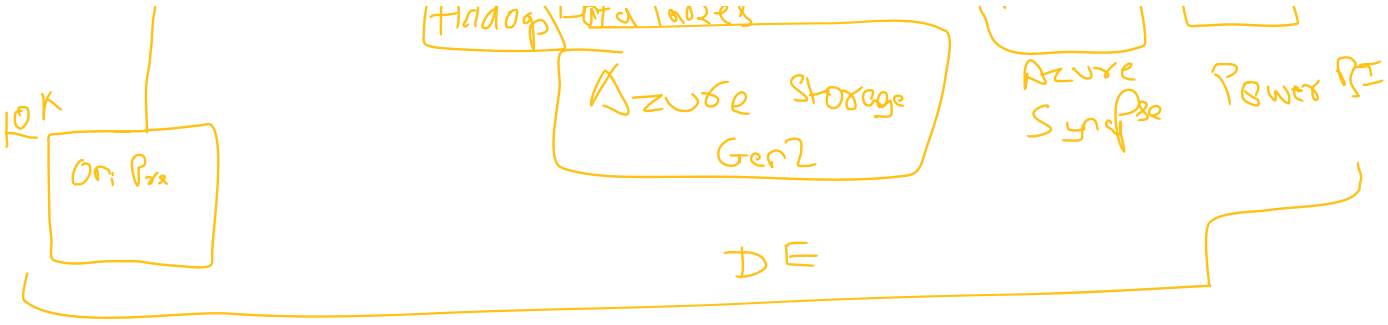

Data Science
Data Analysis
BI Tools
Data Engineer
Data Scientist
AI
Big data Eng.
Arch
Arch
MLOps
ML Engg
Banks
OLTP
SQL
mySQL
mongoDB
NoSQL
API
Blob
Transaction data
customer data
Images
Text, Pdf
Rooms of data
Data source
Software eng
ETL
OLAP
Data warehouse
BI Tools
DA
Data lakes
Hadoop
DS
very high data
Streaming
Kafka
Spooks
Snowflake
Hive
Big Data 6V's
Cloud
Azure
AWS
GCP
Data Engineers
Azure
OLTP
SQL
Azure database
100000
Data factory
Databricks
ML
DS
Hadoop
Data lakes
DW
BI
Azure
Power BI



→ Data

→ Python

→ Database

my SQL

MongoDB

→ Big data technologies

Hadoop

Spark -

- Power BI

→ Cloud services

- Azure → DE → DP-203

→ AWS

→ Kafka

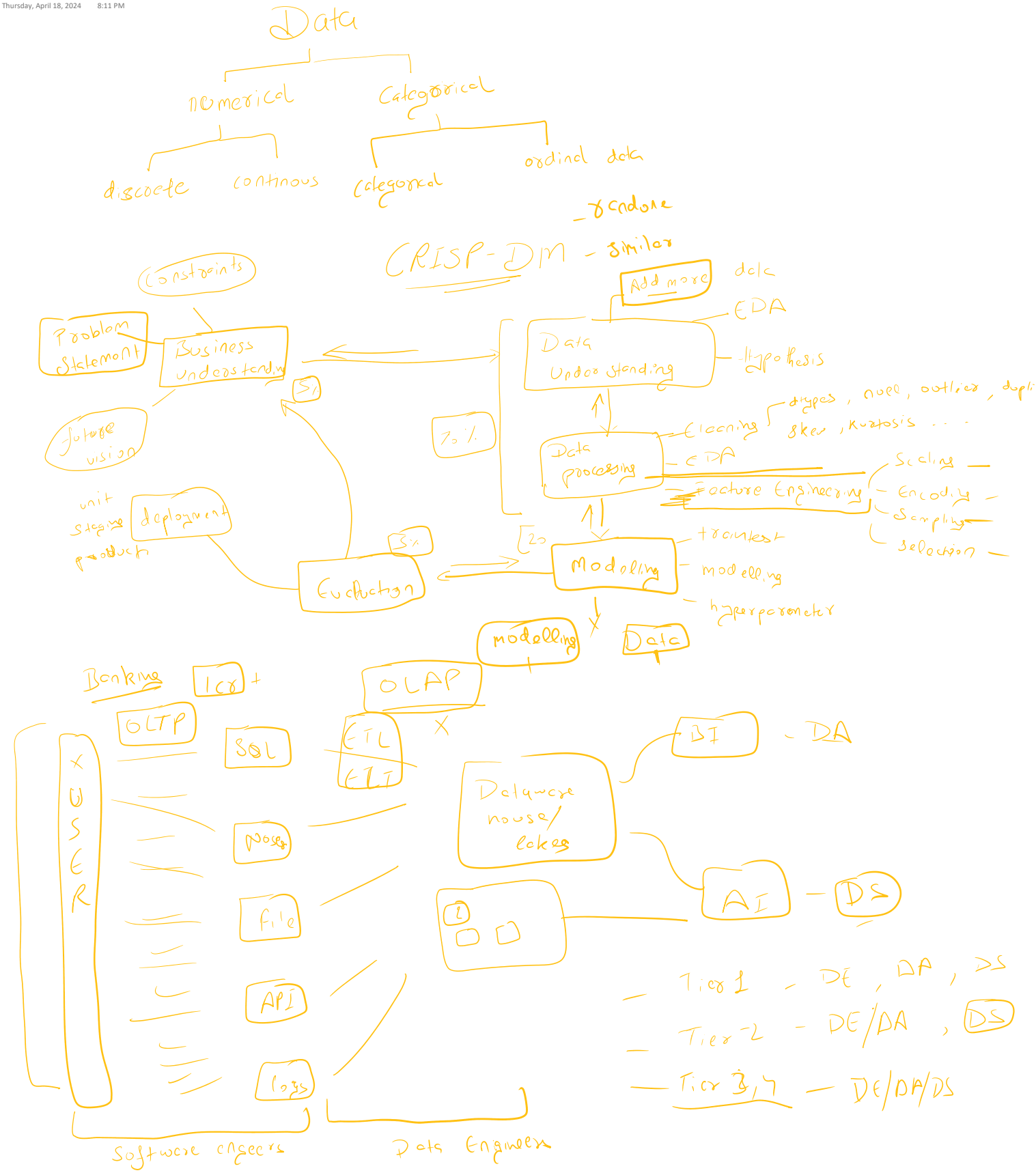
→ Airflow

→ Snowflake

→ informatic

→ Hive

- ML fundamentals


Data
numerical
Categorical
discrete
continous
categorical
ordinal data
CRISP-DM - similar
Constraints
Problem Statement
Business understanding
Future vision
Add more
data
EDA
Data Undoo standing
Hypothesis
Data processing
Cleaning
dtypes , nuel , outlier , dupli
skew , kurtosis . . .
EDA
Feature Engineering
Scaling
Encoding
Sampling
Selection
Modeling
train test
modelling
hyperparameter
deployment
unit
staging
product
Evaluation
modelling
Data
Banking
OLAP
OLTP
SQL
ETL
ELT
Noses
File
API
logs
USER
Software engeers
Data Engineers
Datawore house/ lakes
BI
DA
AI
DS
Tier 1 - DE , DA , DS
Tier-2 - DE/DA , DS
Tier 3,4 - DE/DA/DS
Python

-> Data Preprocessing

-> Maths - Stats Hypothesis testing
- Proof

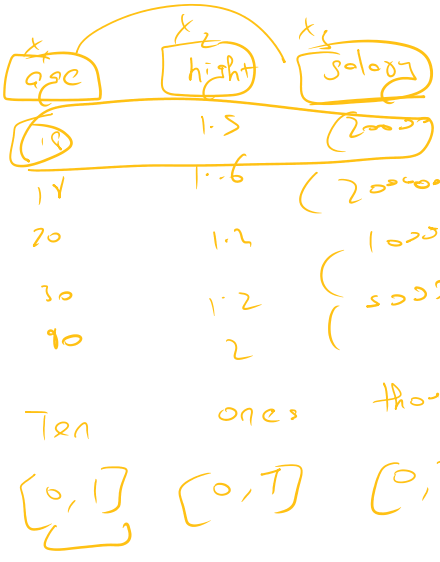
-> Modelling
- ML - Supervised - Classific
  - Regression
- Unsupervised - Clustering
  - Association
  - Dimension Reduct
  - Anomaly
- NLP - text data
- CV - image
- Deep learning - ANN
  - CNN
  - RNN, LSSM
  - GAN
- Generative AI - Attention model
  - transformer
  - transfer learning

-> Deployment

Feature Engineering

$y = m_1x_1 + m_2x_2 + m_3x_3 + c$

- Encoding
- Scaling
- Sampling
- Selection



70:30

Naive bayes model
180
100

5:95

90:10

80:20

e

Encoding :- Label Encoder

workclass
Private - 0
Gov - 1

one Hot Encoder

Private	Gov	SE	NW
1	0	0	0
0	1	0	0
0	0	0	0
1	0		

How — 1 —

SE — 2 —

NO — 3 —

→ Hi How are you → Hi How are you i am good doy

→ Hi, i am good How doy you

- Scaling

- Standardization → [Z-score] $Z = \frac{x - \mu}{\sigma}$ $\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

- Mean of col-0 $\frac{18+20+25+35}{4}$

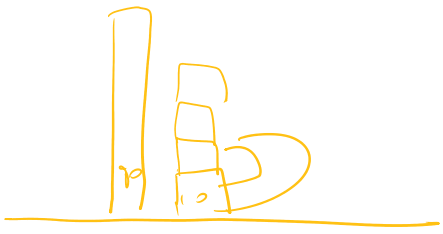
age: 18, 20, 25, 35 → $\mu = 24$, $\sigma = 12.7$ → -.47, -.31, .07, .87 ; $\mu = 0$

- Min Max Scaler → $\frac{Max - x}{max - min}$ $\frac{35-18}{35-18} = 1$ $\frac{35-20}{35-18} =$ — $\frac{35-35}{35-18} = 0$ [0, 1]

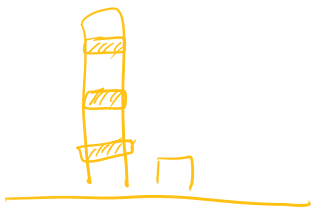
Sampling / Balancing

① Random sampler

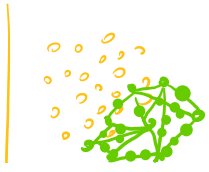
- oversampling



undersampling



② SMOTE

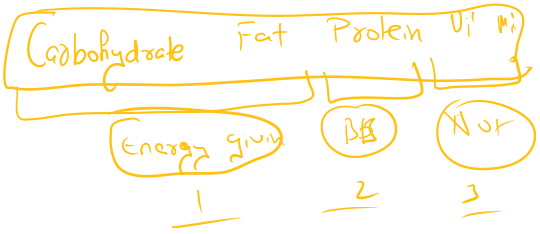


ENN

Feature Selection

- Removing features
  - Correlation
  - Mutual_info_classif
    - Decision
- Data Compression
  - unsupervised
    - PCA - Principle Component Analysis
    - LDA

"curse of dimensionality"



X

Friday, April 19, 2024 4:11 PM

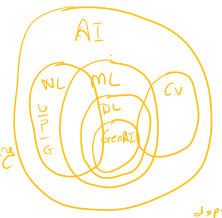
Machine learning

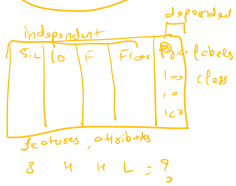
80%

20%

1972

- Predictive Supervised learning
  - Classification (14+)
    - Naive bayes
    - Logistic regression
    - KNN
    - Decision tree → D & R
    - SVM
  - Regression (12+)
    - Linear regression
    - Polynomial reg.
- Descriptive unsupervised learning
  - Clustering
    - Partition (K-means)
    - Hierarchical (Agglomerative)
    - Density (DBSCAN) (17+)
  - Association
    - FP Growth
    - Apriori
    - Eclat
- Reinforcement learning
- Semi Supervised learning





S H H L = ?

Naive Bayes Model

1 Toss = H, T → H = 1/2

2 Toss = HH, HT, TH, TT

- Prob. = occurrence of event / total event
- Conditionally Prob.

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

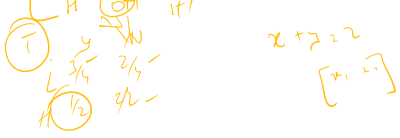
Training:

P(Y) = 2/5, P(N) = 3/5

D	Y	N
S	1/3	2/3
R	1/2	1/2

H	Y	N
L	1/3	2/3
M	1/2	1/2
W	1/2	1/2

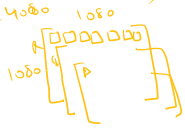
Temp	Y	N
L	2/3	1/2
H	1/3	2/3

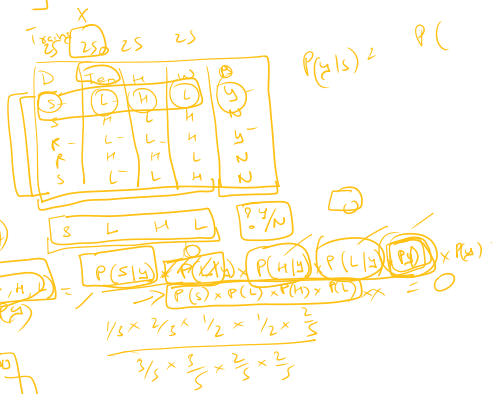


x + y = 2

[x, y]

UHD - 4080 1080





P(S) = 0

- Assumption:
  - All feature put equal contribution to label
  - independent feature

NB → Naive

2/5 2/5 2/5 2/5

- Use Case
  - Text classification
  - Recommendation
  - Medical dignosis
- when to Apply to
  - limited data
  - Simple fast implementation
  - less dependency
- Limitation
  - Assumption of other
  - density of
  - missing data
  - Zero case sensitive
    - Smoothening
      - Laplace

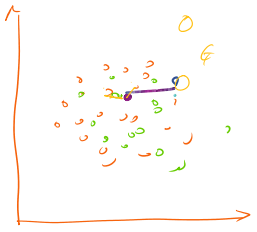
Types of NB

- Gaussian NB — Continuous Datatype

Types of NB

→ ◦ Gaussian NB — Continuous Data
◦ Multinomial NB / Categorical NB
◦ Bernoulli NB — Binary
◦ One vs all
◦ out of core

K-nearest Neighbours = K = 3



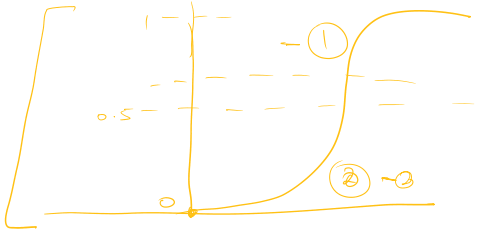
Logistic Regression:

[0, 1]

$y = \frac{1}{1+e^{-x}}$

A B C

◦ Types
◦ Binary classification
◦ multinomial classifica
→ ordinal → Hypothesis testing



1 : 100

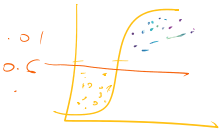
Use Cases:
◦ Customer Churn
◦ Credit scoring
◦ medical dig.

Assumption
◦ Binary target
◦ Independent obs.
◦ data linearly separable

when to Apply
◦ More amount of data
◦ Assumption

◦ Limitation
no outlier.
◦ non linear
◦ Binary classification

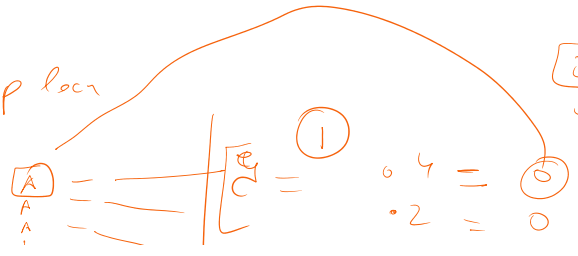
y =

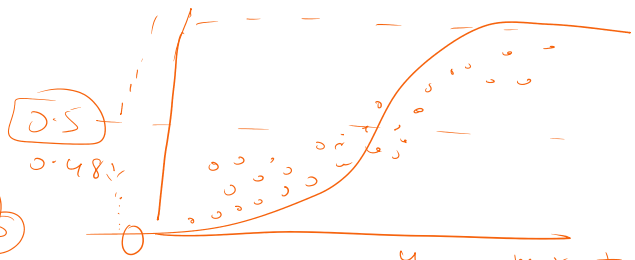


test —

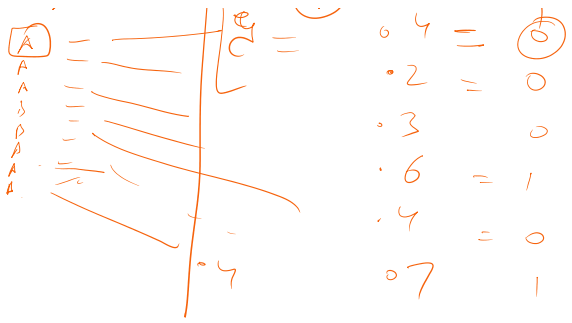
train —

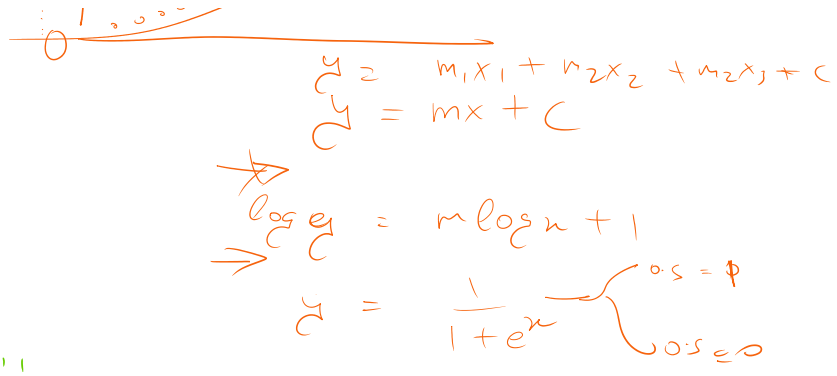
Deep learn





$y = m_1x_1 + m_2x_2 + m_3x_3 + c$



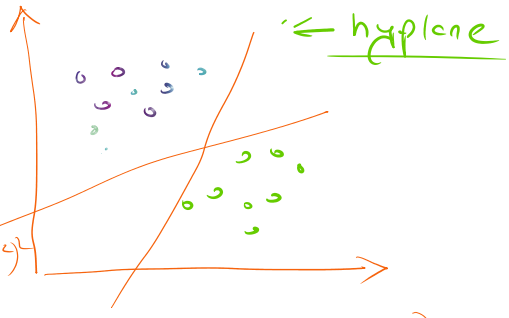


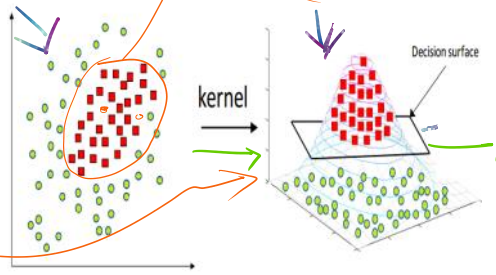
SVM

Support Vector Machines

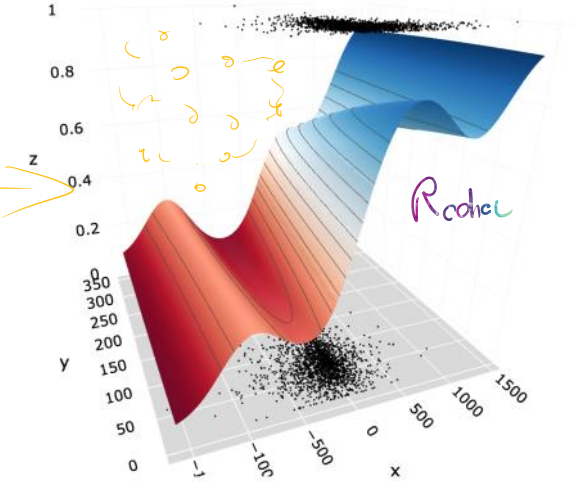
Kernel

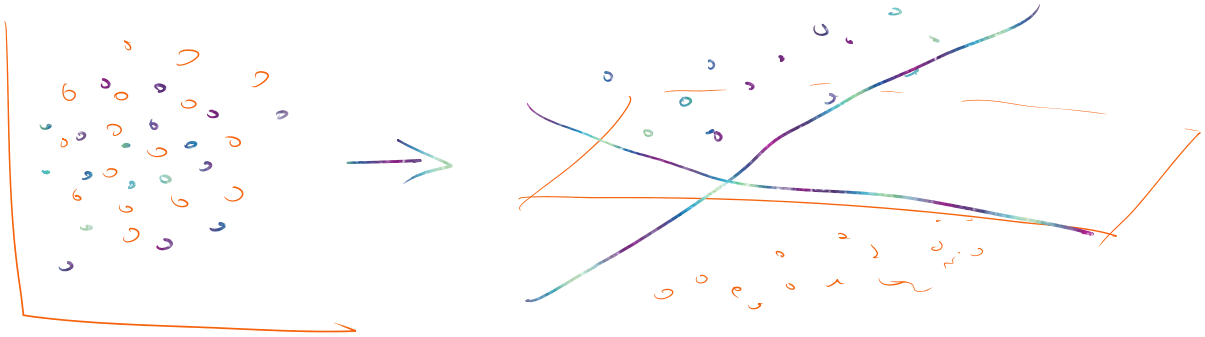
- linear
- polynomial
- Radial





Deeps





$$y = w^T x + b$$

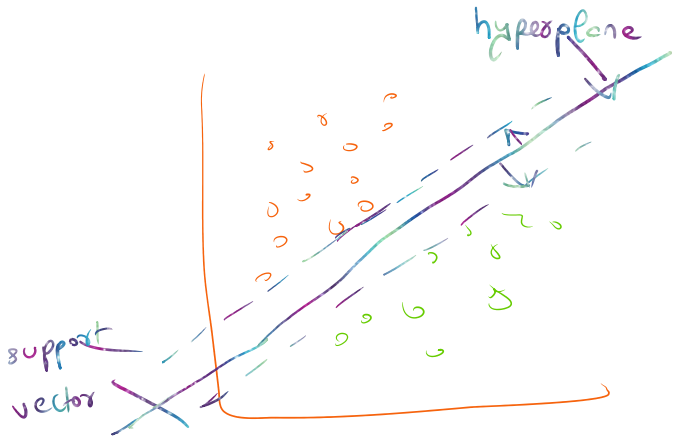
Use Cases

Image classifi

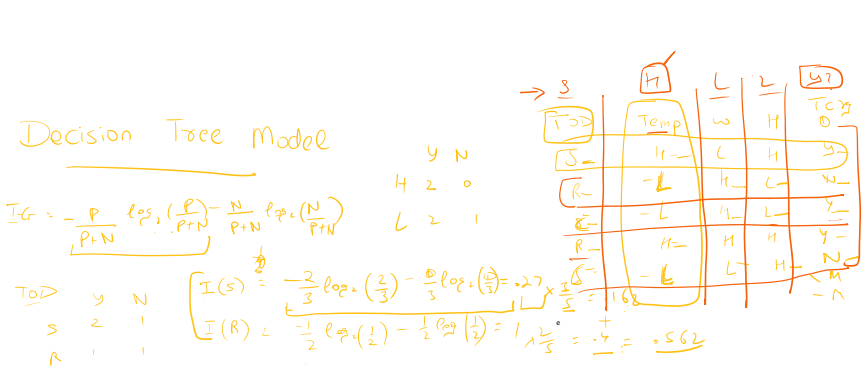
Audio classif

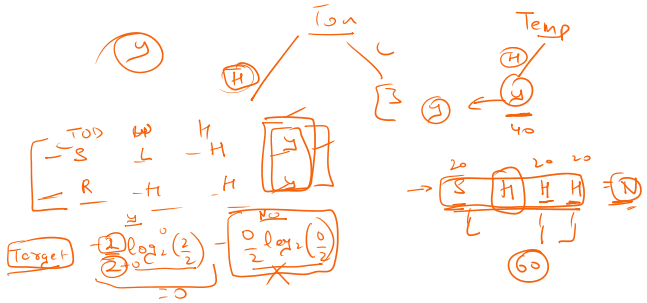
Text data

When to apply



Decision Tree Model



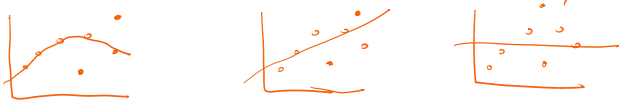


Use cases:

Financial / Bank

Limitation

- Overfitting
- Bias
- Instability



overfitting | good fit | underfitting

when to Apply case

Gini index

$$1 - P(i)$$

$$I_G = 1 - \sum_{i=1}^{n} P_i^2$$

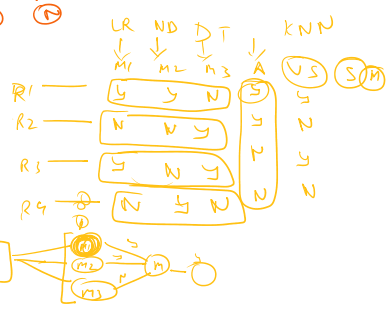
$$I_E = -\sum P_i \log(P_i)$$

Fundamental

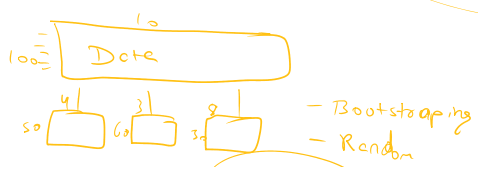
- NB
- KNN
- SVM
- LR
- DT

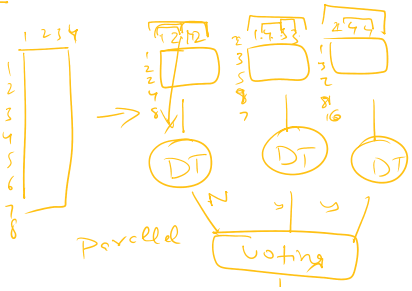
Ensemble models

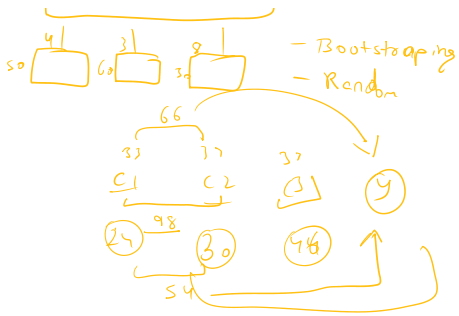
- Voting
- Stacking
- Bagging — Random Forest
- Boosting

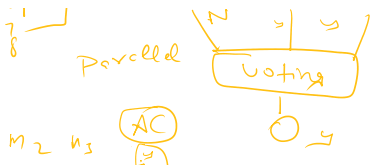


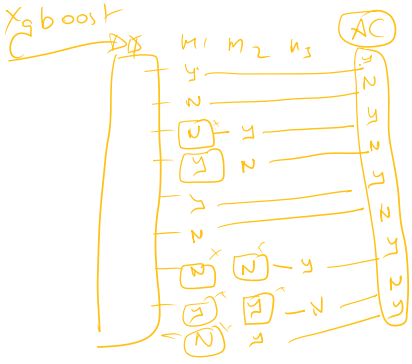
Random Forest











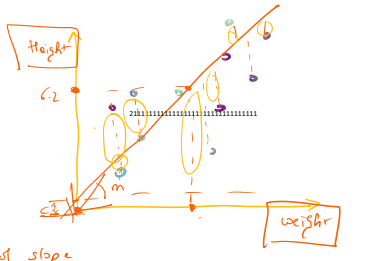
→ Regression Model

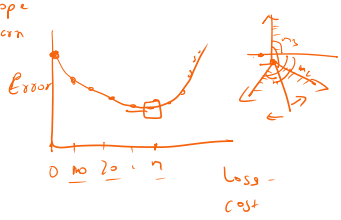
- Linear Regression
  - Residual
  - Simple linear R = $y = mx + c$

    m = Slope of linear – Significance of slope in linear

    c = constant
  - Multiple linear regression

    $y = m_1x_1 + m_2x_2 + m_3x_3 + \dots + c$

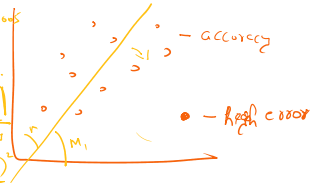




- Loss function – total loss = $\frac{1}{2}\sum_{i}^{n}(\bar{y} - y_i)^2$
- cost function = (m)

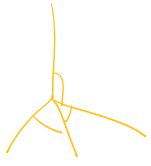
Regularization techniques

- L1 – Lasso Regression = $\frac{1}{2}\sum_{i}^{n}(\bar{y} - y_i)^2 + \lambda \sum |m_i|$ (penality)
- L2 – Ridge Regression = $\frac{1}{2}\sum(\bar{y} - y)^2 + \lambda \sum (m_i)^2$



$y = m_1x_1 + m_2x_2 + m_3x_3 + \dots + c$

Feature selection



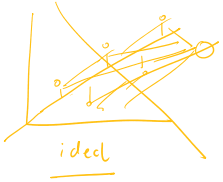
Use Cases

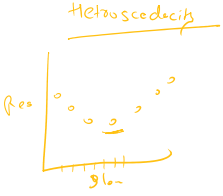
Naive

Assumption
- Linearity
- Homoscedasicity
- Normality of Error

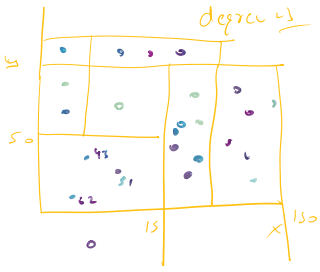
Limitation
- Assumption
- Overfitting
- Categorical data

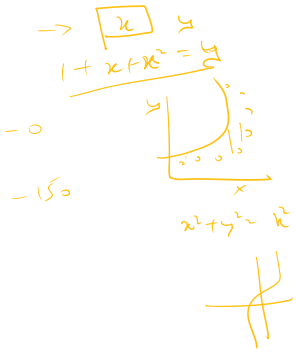
Apply





Polynomial Feature = degree = 2



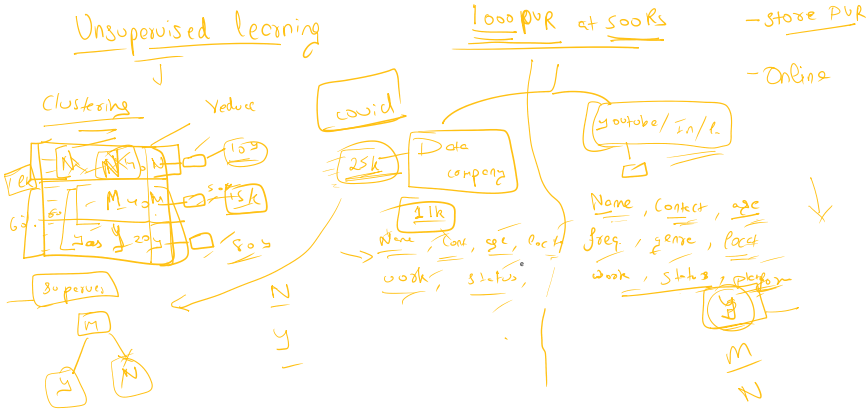


x = 12

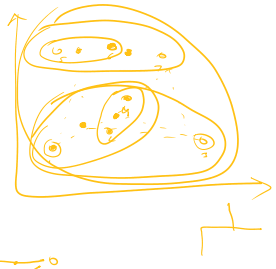
y = 43

8

Unsupervised learning



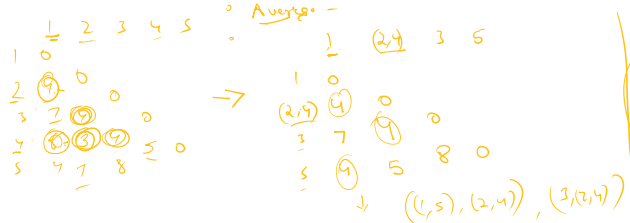
Unsupervised learning

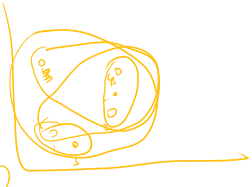


Clustering

- Hierarchical Agglomerative
  - distance matrix
  - linkage
    - Single (min)
    - Complete - max
    - Average -
- Partition Kmeans
- Density DB Scan

Use Case
- Biology
- Anomaly detection
- Market Segmentation





Limitation
- Computationally expensive
- Sensitive to Noisy
- Memory usage
- linkage

Kmeans Clustering → K = 3

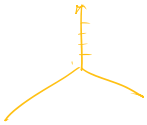
Assumption

Assumption :

- Cluster Shapes - Spherical
- Homogenious Dansity - Almost Equal datapoint
- Single Variance -

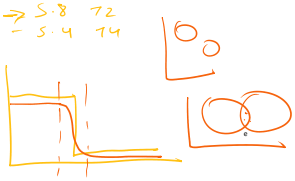
→ 5.7 67 — C1
→ 5.4 68 — C2
6.0 75 — C3
→ 5.8 72
5.4 74

K =



Types of Kmeans

- Fuzzy C means
- kmeans++ — default
- Mini Batch Kn —
- Spectral Clustering -



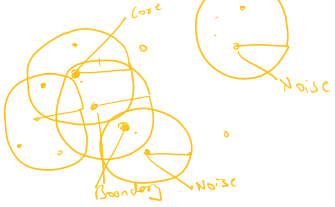
Limitation

- Sensitivity to initial Centroid
- Sensitivity to outlier
- Number of clusters

DB Scan

epsilon - 1 support = 3

- Core
- Boundary
- Noise



Association Technique

- Apriori - 1,2,3
- FP growth

support - 2

step - 1
A - 3
B - 3
C - 4
D - 4
E - 1

1	A B C
2	B C D
3	D A B
4	C D A
5	C D C

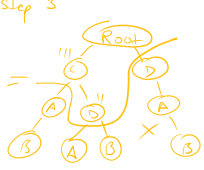
step - 2

1	C A B
2	C D B
3	D A B
4	C D A

C D

Step 3

C → D =



Use Cases

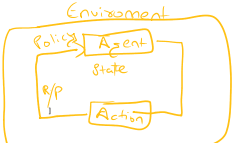
Assumptions

Limitations

- memory usage
- Non Specific

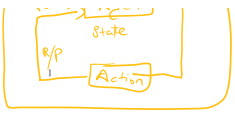
Reinforcement Learning

- Interaction based
- Goal based



- Environment
- State
- Action
- Reward / Penalty
- Agent

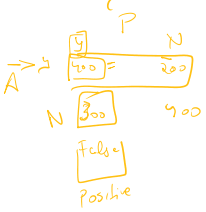
- Interaction based
- Goal based



- Action
- Reward/Penalty
- Agent
- Policy

Accy

Precision :



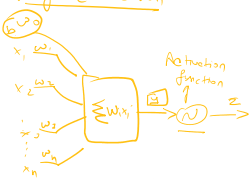
Deep learning

eee

BB-A Neuron

$y = mx + c$

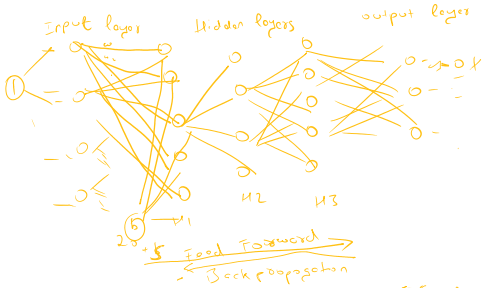
Artificial Neuron



$b w_0 + x_1 w_1 + x_2 w_2 + x_n w_n = f(y) = z$

- Activation function
  - Sigmoid - [0, 1]
  - Step = 0, 1
  - tanh = (-1, 1)
  - relu = [0, ∞)

Feed Forward / Fully connected Neural Network



20 20 20 15 95-w

5 4 5 3 17-b

120×ω 70×ω₂ = 19 21.0

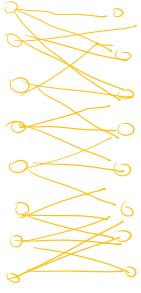
120×1.2 70×1 = 31 21

120×0.1 70×1.5 = 225 21

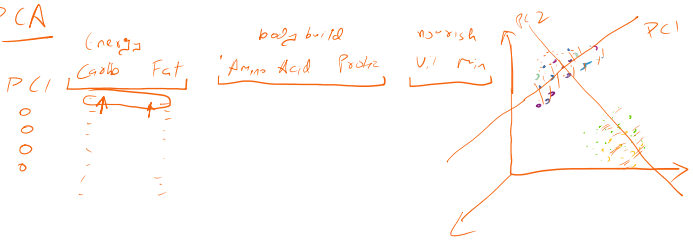
120

-> Convolutional NN - Images

-> Recurrent NN — Text/sequential



PCA



Natural language : - Processing

- Understanding

- Understanding

- Generation

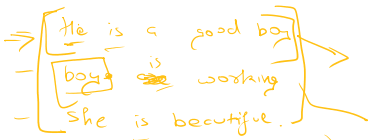
Processing

- Lowercase
- Tokenisation
- Normalization
- Stop word removal
- Stemming & lemmatization
- Cleaning & correction

Understanding

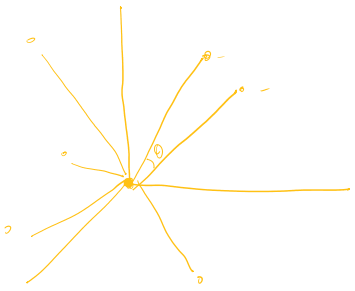
Vectorization

- Bag of words → nL
- TF-IDF → importance of word
- Word2vec → DL



- Tensor
- textblob
- nltk
- spacy

235 - Bank - river bank - 2
789 - Bank - financial - 289

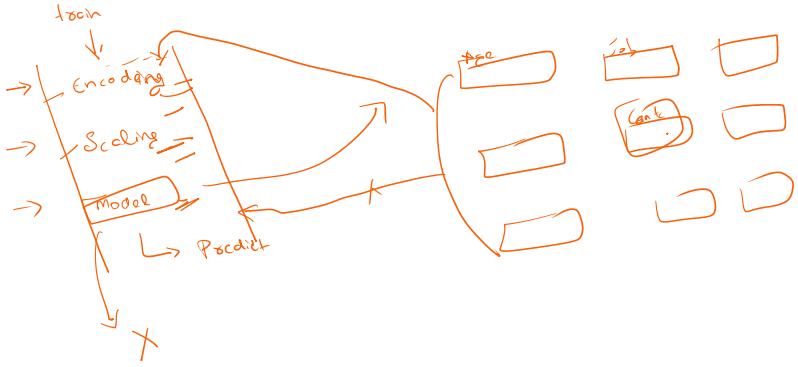


Project Submission

- Sunday, 23/June

- 3 project
  - code
  - Document :
    - Business Problem — 2pages doc
    - Data Description
    - Basic step in data processing & major challenges
    - Choosen model, why, pros & cons

- Bank marketing dataset project



Window functions

select sum(salary) from emp = Groupby

Window functions

select sum(salary) from emp = Groupby

Emp id	dep	salary	Total
1	1	50600	21 -
2	1	60002	23000
3	1	65000	23
4	1	55000	23
5	2	110000	21
6	2	100000	21
7	3	20000	50
8	3	30000	50
9			
		490000	

select sum(salary) over() from emp.

→ Give the total cost of each dep. emp.

select sum (salary) over (partition by dept)

from ~

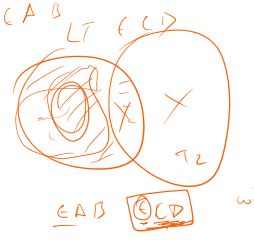
Types of window function

- Aggregate
  - sum()
  - count()
  - avg
- Rank
  - Rank()
  - Dense Rank
  - Row number()
- Value function
  - Lead
  - Lag
  - First value.

Lead(2)

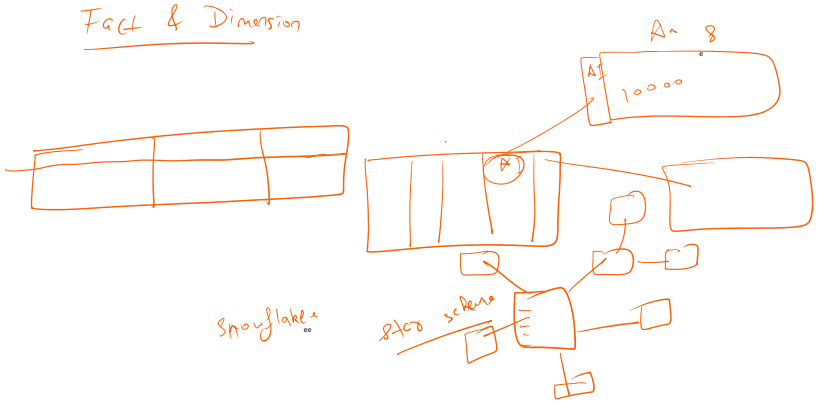
Emp id	Dept id	Salary		Rank	Dense Rank	Rn	Lead
1	1	50000	50000	1	1	1	60000
2	-1	60000	50800	1	1	2	
3	-2	70000	6000	3	2	3	70000
4	-2	65000	650000	4	3	4	65000
5	-1	50000	66000	5	4	1	
6	3	66000	70000	6	5	2	1
7	-1	70000	70000	6	5	3	Null
			80001	8	6		

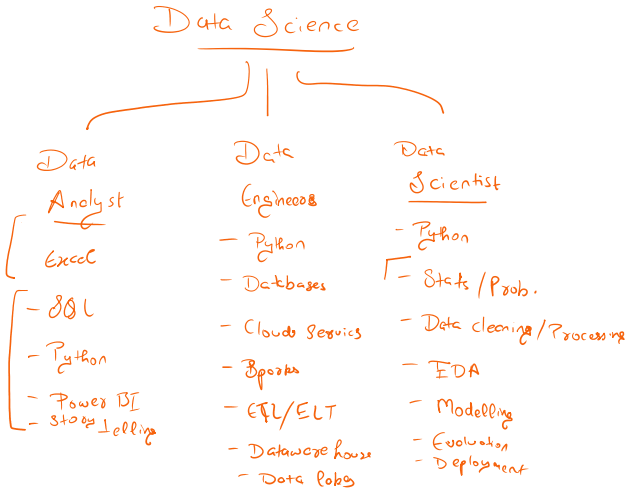
Select custname from customer where title in (select title from file where ctid = 20)

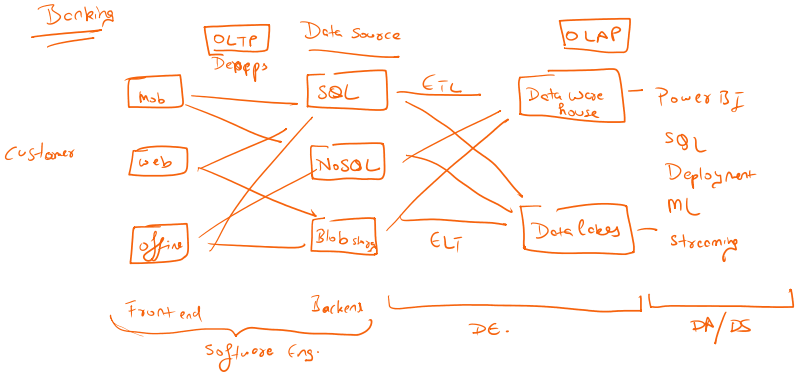


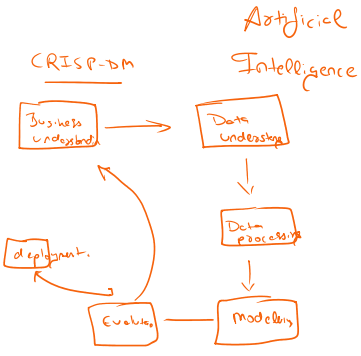
∈ A B ∩ C D where T2.C is Null

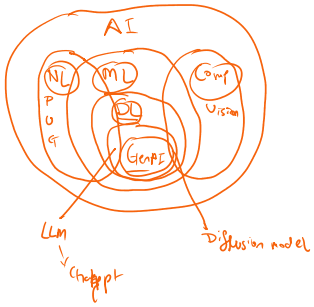
Fact & Dimension




Data Science
Data Analyst
Excel
- SQL
- Python
- Power BI
- Storytelling
Data Engineers
- Python
- Databases
- Cloud Services
- Bigdata
- ETL/ELT
- Datawarehouse
- Data Lakes
Data Scientist
- Python
- Stats / Prob.
- Data cleaning / Processing
- EDA
- Modelling
- Evaluation
- Deployment


Banking
OLTP
Deepp
Data source
OLAP
Mob
Web
Offline
SQL
NoSQL
Blob storage
ETL
ELT
Data Ware House
Data Lakes
Power BI
SQL
Deployment
ML
Streaming
Customers
Frontend
Backend
Software Eng.
DE.
DA/DS


Artificial Intelligence
CRISP-DM
Business understand
Data understanding
Data processing
Modelling
Evaluate
deployment.


AI
NL
ML
DL
GenAI
Comp Vision
LLM
Chatgpt
Diffusion model