Machine Learning

Quiz 01

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

Question 1

- Your company is providing you with recent sales data and relevant details
- They are wondering: Why did the **sales drop the last 6** months?
- Which area may play a vital role in answering this question?
  - 1) Machine Learning
  - 2) Data Analysis / Data Science

Data analysis. We need to investigate the sales and what happened

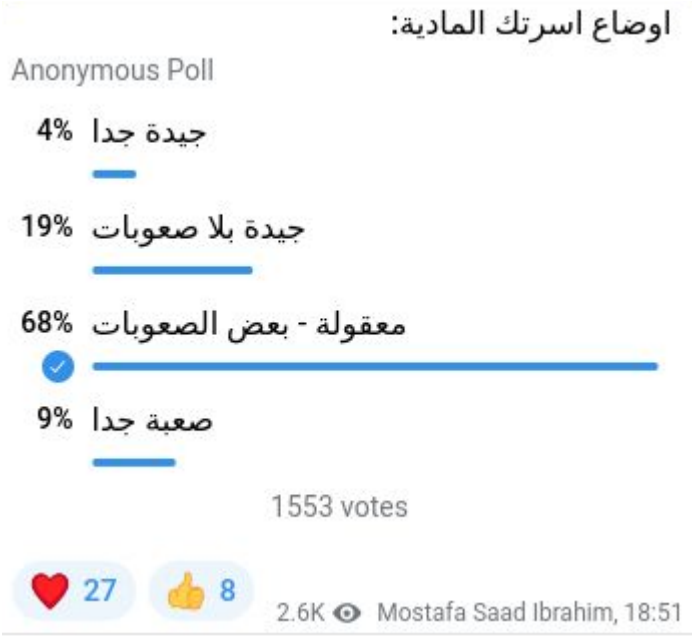
Question 2

- Your company is providing you with recent sales data and relevant details
- They are wondering: How can we **predict our sales in the next 6** months?
- Which area may play a vital role in answering this question?
  - 1) Machine Learning
  - 2) Data Analysis / Data Science

Machine Learning. ML can help in **predictions**

Question 3

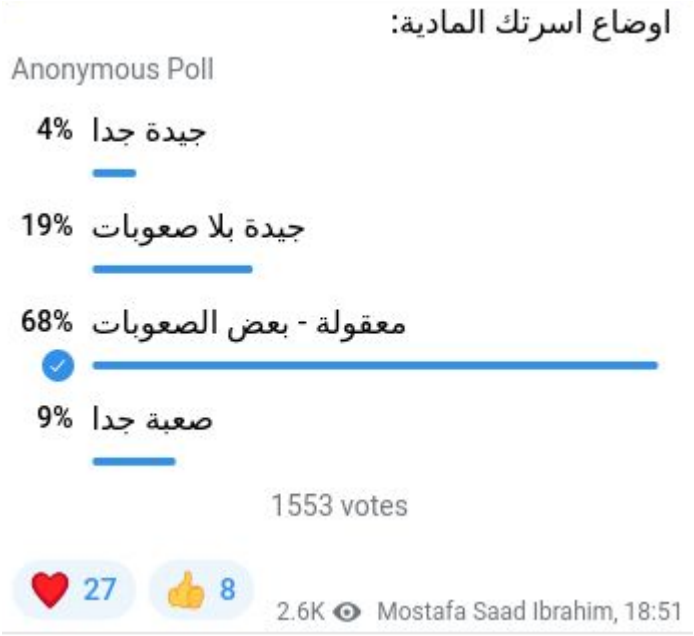
- A pool created in Dr Mostafa Telegram Channel
- Which distribution does the data follow?



Gaussian Distribution

Question 4

- A pool created in Dr Mostafa Telegram Channel
- Is it a representative distribution for Egyptian? Why?

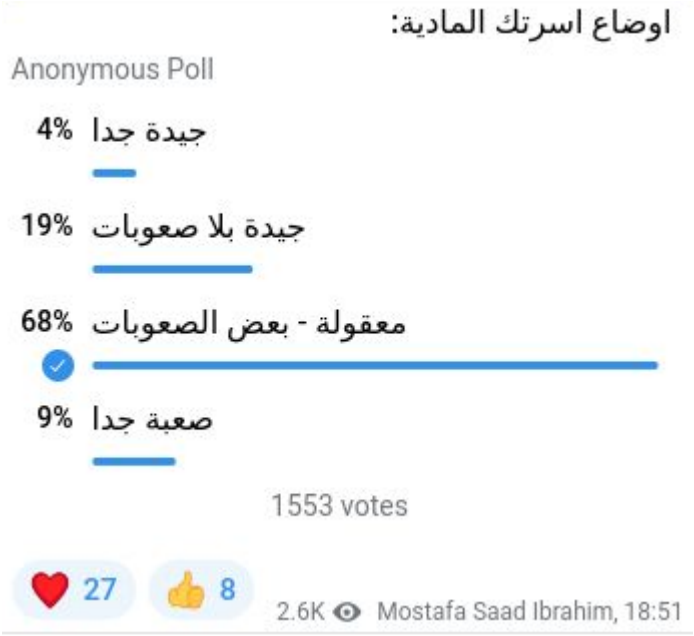


- Biased Sample:
  - Only CS Students
    - Who follows Dr Mostafa!
    - Who use telegram
    - Maybe more males than females
- Small sample size: Egyptians are 100M
- Minor concern
  - Subjectivity: What is the good? Hard life?
    - How can we fix that?

Question 5

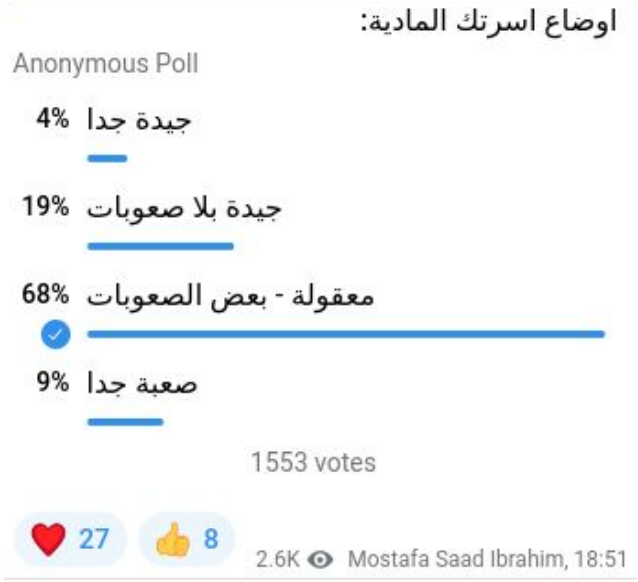
- Imagine we collected a representative sample
- What information might be common with this result? What might not?

- Probably it will also follows gaussian distribution
- However, the ratios might change
  - E.g. more people are poorer / facing difficulties



Question 6

- XXX, a well-known islamic figure in the world, created a survey for his followers asking a yes-or-no question:
  Are you **against same-sex** relationships?
  - What percentages do you expect for the pool results?
  - Figure out an interesting **bias type** that is hidden here?



- **Nonresponse bias** occurs when survey participants are unwilling or unable to respond to a survey question or an entire survey
  - Those who have religions problems with islam or supports LGBT may find the question offensive and **don't answer** it
  - Overall, we will end with a very religious sample that all takes a religious position, even though the sample size is big

Bias

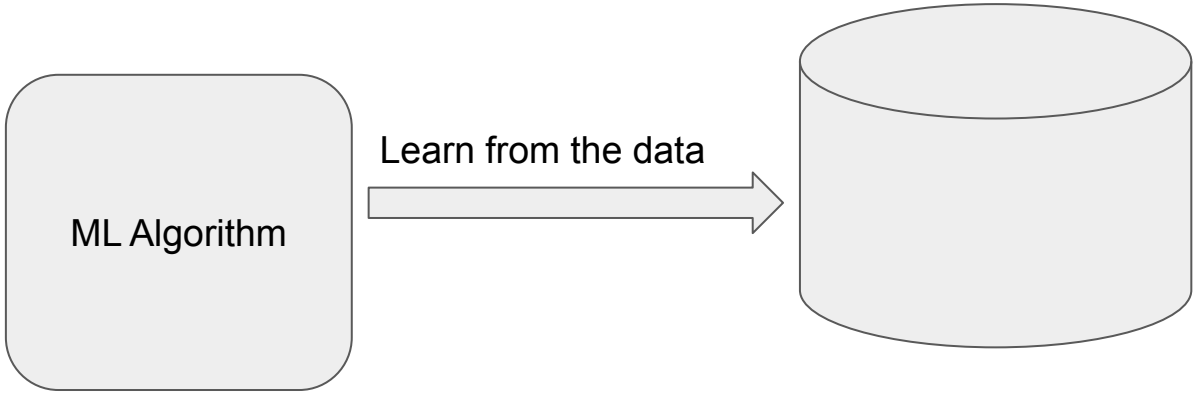
- Bias is a big problem in machine learning
  - See examples from Google course
- Bias in general is a critical concern: wiki
  - Daily logic example: **Confirmation** bias: *the tendency to interpret new evidence as confirmation of one's existing beliefs or theories.*

- **Non-response bias** (or **participation bias**): Data ends up being unrepresentative due to participation gaps in the data-collection process.

EXAMPLE: A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product and with a sample of consumers who bought a competing product. Consumers who bought the competing product were 80% more likely to refuse to complete the survey, and their data was underrepresented in the sample.

Question 7

- A bank collected data and learned a machine learning model to learn **Loan Eligibility**. Most of the data involved men who paid back the loan on time but most of the women failed.
- Any issues in the model?



Similarly

- Most white people paid and black did not!

Question 8

- The machine learning computer vision team is asking the camera team about a setup for their work. The camera team gave them 4 choices
  - A, B, C, D
  - For every choice, they listed: **pros, cons and timeline** to implement the solution
- Logically, how can the vision team make a decision?!

Answer!

- 1) Understand every choice properly
  - One challenge: you may not know the low-level domain
  - Prepare any questions that help you clarify/understand choices
- 2) Verify the mentioned pros and cons
  - Are they valid?
  - More importantly, what about pros/cons from machine learning perspective?
- 3) The missing answers
  - One of the common issues in surveys is missing some answers
  - What if there are other better solutions?
  - Best of 2 worlds: Can we combine solutions A and C?
- 4) Determine what is your criteria (objective) to make a choice?
  - Vision task X must have very high accuracy within 200 ms

Common survey question mistakes

- Building good surveys is not an easy job
- There are many common mistakes
  - Leading Questions
  - Unspecific Questions
  - Answers are Not Mutually Exclusive
  - Missing Answers
  - etc

Question 9

- Part of your work is the following logic
  - Given the 3 vectors of equal length, we compute the euclidean distance between (1st and 2nd) and (1st and 3rd) then compute the sum of the 3 vectors and divide by the multiplication of the 2 distances
- In how many ways, you can present this information?

- Sometimes, it is good to provide the same information in multiple ways
- Sometimes the best is a **visualization** diagram with the **text**
  - Still a bit subjective, but common
- In formal context, you might add **an equation** if possible
  - This ones make it very objective. Good for papers

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."