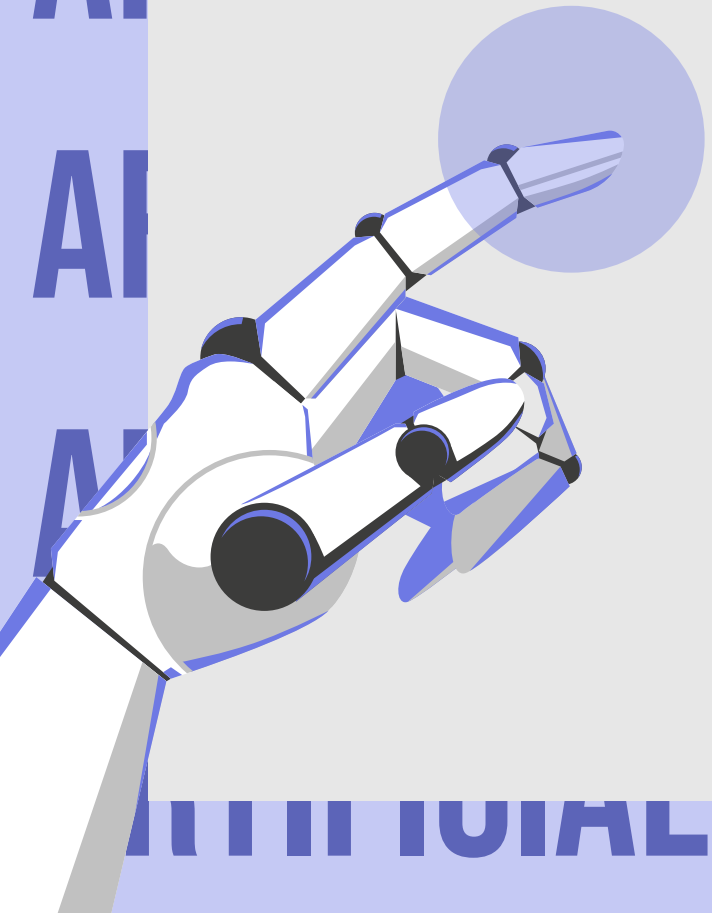


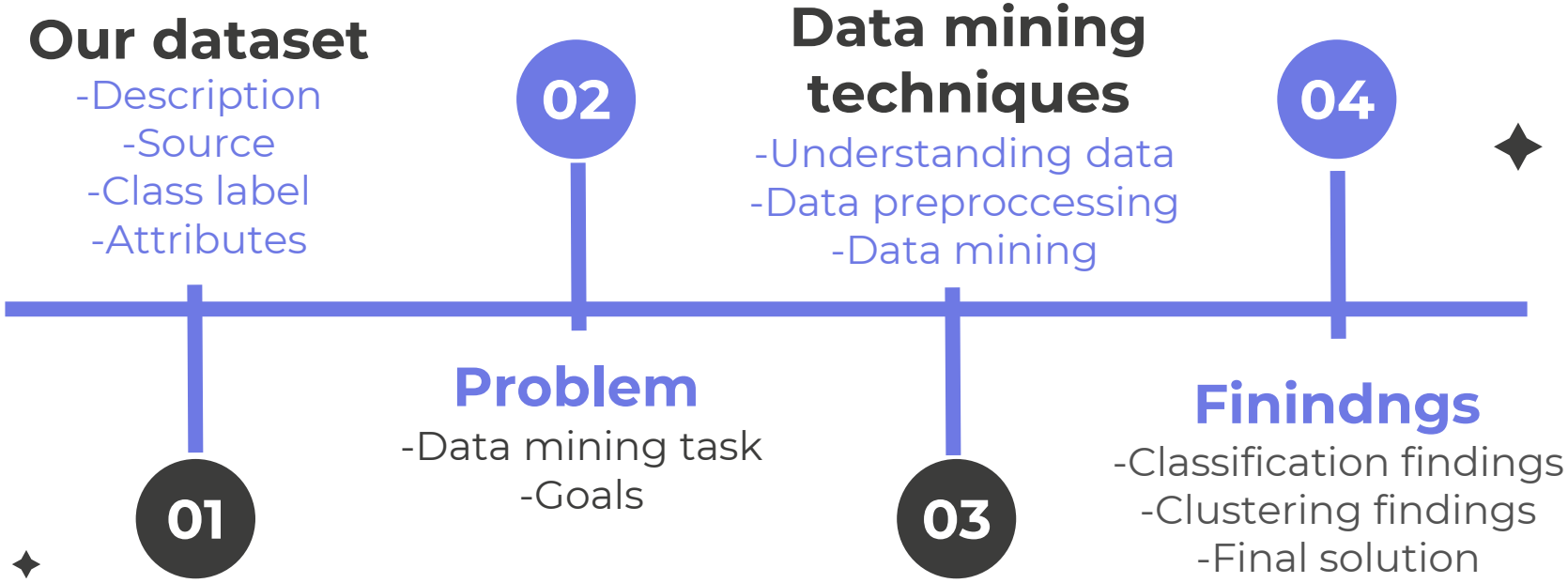
CYBER SECURITY SALARIES

Data Mining Project Presentation

Section #56545

Group #2

OUTLINE


Our dataset
-Description
-Source
-Class label
-Attributes
01
02
Problem
-Data mining task
-Goals
Data mining techniques
-Understanding data
-Data preproccessing
-Data mining
03
04
Finindngs
-Classification findings
-Clustering findings
-Final solution

01. OUR DATASET

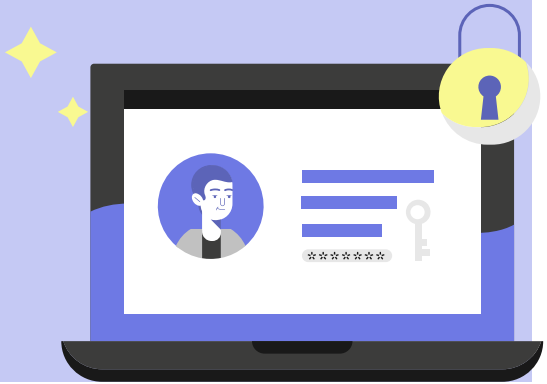
"CYBER SECURITY SALARIES" is:

- A dataset with 1247 instances

- Shows cybersecurity Employees' salaries

- Uses 11 attributes of different data types

Source: Kaggle.com

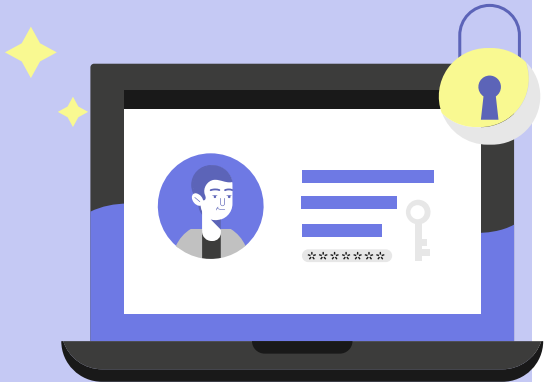


01. OUR DATASET

Class label: salary_in_usd

Attributes:

- work_year
- experience_level
- employment_type
- job_title
- salary
- salary_currency
- salary_in_usd
- employee_residence
- remote_ratio
- company_location
- company_size



02. PROBLEM

Data is becoming a valuable asset

Need to hire CS employees to guard data

Difficulty in deciding the salaries

Competitors can out advantage them!

02. PROBLEM

Data mining task:

- Prediction of the cyber security employees' salary categories (Very Low, Low, , High, Very High) using classification

- Grouping and describing employees based on shared characteristics using clustering

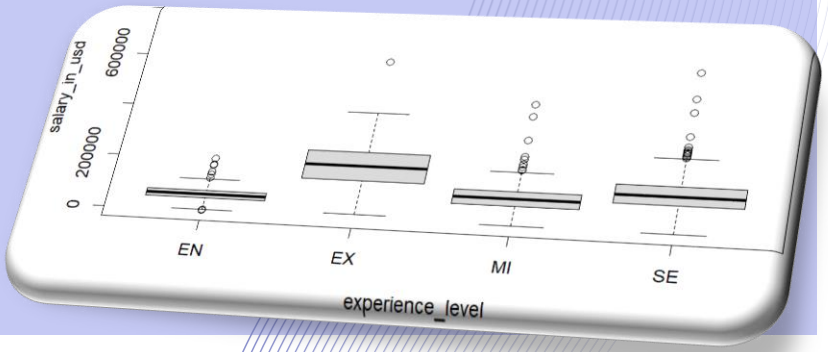
02. PROBLEM
Goals:

- Market segmentation
- Identify trends
- Identify the main cybersecurity employee groups
- Making better decisions
- Increasing loyalty
- Achieving fairness

03. DATA MINING TECHNIQUES

DATA UNDERSATANDING

- No missing values
- Statistical measures show moderate variability
- Graph visualization show positive correlation



03. DATA MINING TECHNIQUES

DATA PREPROCCISSING

Dimensionality Reduction

Removing salary attribute

Removing outliers

From salary_in_usd attribute

Concept hierarchy generation

For company_location, employee_residence attributes into 7 regions

03. DATA MINING TECHNIQUES

DATA PREPROCCISSING

Encoding

Of categorical data

Discretization

of salary_in_usd attribute

Normalization

Of remote_ratio and work_year attributes

03. DATA MINING TECHNIQUES
DATA PREPROCCISSING

Feature selection

Remove employment_type attribute

Balancing data

To avoid biased data mining results

03. DATA MINING TECHNIQUES

DATA MINING TASKS

Classification

-Using Gini index, Gain ratio, and information gain
-Construct a decision tree model, to classify data.
-Using k-fold Cross-validation for partitioning (k=3,5,10)

Clustering

-Using k-means to partition data into clusters (k=2,3,4)
-Using evaluation methods (silhouette, Bcubed precision and recall, WSS)

04. FINDINGS
Classification results

-Gain ratio with k-fold cross validation k=10 provides best performance among all the decision tree models, duo to its favor of unbalanced splits

-"Experience Level" is the first splitting attribute, indicating that it is the strongest predictor in reducing uncertainty

-All methods have similar performance based on evaluation criteria: accuracy, precision, sensitivity and specificity

04. FINDINGS
Clustering results

- k-means clustering with optimal k=2 provides purest clusters with no overlapping, and high intra-cluster similarity with non-sparse clusters

-Evaluation methods favor k=2, such as BCubed Precision and Recall, Average Silhouette Width and graphs

-Other K's , k=3,4 have overlapping and reflect high inter-cluster similarity

04. FINDINGS
Solution:

01 Use classification to predict employees' salaries (using 10-fold gain ratio method)

02 Use clustering to group employees based on their similarties (using k-means with 2 clusters)

That way, we solve current problems and achieve employees' satisfaction which will be reflected on their performance!

REFRENCES

[1] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," 3rd ed., The Morgan Kaufmann Series in Data Management Systems.

[2] Y. Zhao, "R and Data Mining: Examples and Case Studies," 1st ed. Academic Press, 2012. ISBN: 0123969638.

[3] Y. Zhao, "R and Data Mining," RDataMining.com, Available: <https://www.rdatamining.com/>, Accessed on: November 23, 2023.

[4] M. Hahsler, "discretize {arules}R Documentation: Convert a Continuous Variable into a Categorical Variable," R Project, Available: <https://search.r-project.org/CRAN/refmans/arules/html/discretize.html>, Accessed on: November 23, 2023.

THANK YOU FOR LISTENING!

ANY QUESTIONS?

Prepared by:

Sarah Alhindi

Raneem Alyahya

Shahad Bin Makhashen

Shoug Alyahya

Supervised by:

T. Hanan Altamimi