Data Analysis in Software Engineering

MESW
FEUP

30th May 2025

Mental Health Dataset

Final Presentation

André Butuc, 202400040
José Mendes, 202402636
Matilde Faro, 202108853
Rui Soares, 202103631
WG-14

OUR DATASET

17 FEATURES

291 364 ENTRIES

SOME UNBALANCED

OUR DATASET

16 FEATURES

284 858 ENTRIES

SOME UNBALANCED

FEATURES

GENDER

CHANGE IN HABITS

GROWING STRESS

COUNTRY

MENTAL HEALTH HISTORY

OCCUPATION

SOCIAL WEAKNESS

SELF EMPLOYMENT

MOOD SWINGS

MENTAL HEALTH INTERVIEW

FAMILY HISTORY

COPING STRUGGLES

TREATMENT

CARE OPTIONS

DAYS INDOOR

WORK INTEREST

VARIATIONS OF DATASET

UNTOUCHED

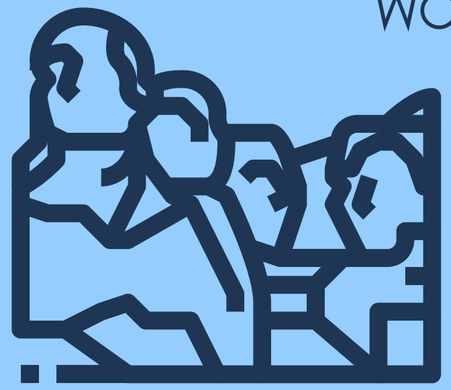
DECIDED

ANGLOPHONE COUNTRIES

REST OF THE WORLD

MALE

FEMALE



USA

DATA ANALYSIS

ONE-HOT ENCODING

CORRELATION MATRIX

ELBOW METHOD

SILHOUETTE SCORE

OPTIMAL NUMBER OF CLUSTERS

PRINCIPAL COMPONENT ANALYSIS

VISUALLY PLOTED THE CLUSTERS

KEY TAKEAWAYS

FROM DATASET VARIATION ANALYSIS

CULTURAL INSIGHTS

USA AND OTHER ANGLOPHONE COUNTRIES SHOW VERY SIMILAR MENTAL HEALTH PATTERNS

THE REST OF THE WORLD DATASET DISPLAYED GREATER DIVERSITY

KEY TAKEAWAYS

FROM DATASET VARIATION ANALYSIS

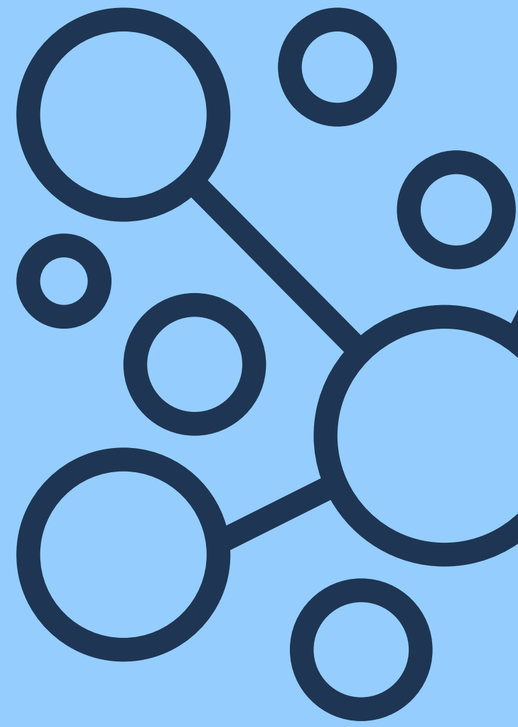
ELBOW METHOD AND SILHOUETTE SCORES OFTEN DISAGREED ON OPTIMAL CLUSTER COUNT

SILHOUETTE SCORES WERE GENERALLY LOW

ONE GENDER DATASETS FORMED RIGID, BOX-LIKE CLUSTERS

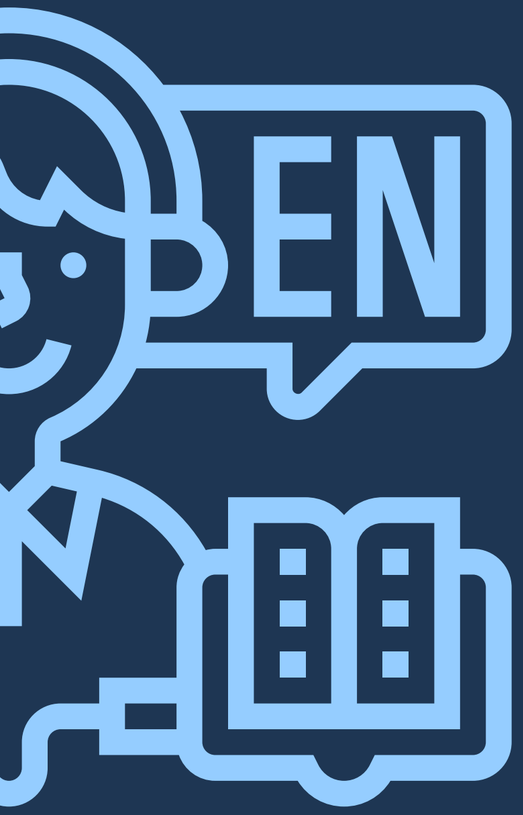
THE DECIDED DATASET HAD SMOOTHER, WELL-SEPARATED CLUSTERS

CLUSTERING OBSERVATIONS



KEY TAKEAWAYS

FROM DATASET VARIATION ANALYSIS



ANGLOPHONE COUNTRIES

FEMALE CORRELATES WITH FAMILY HISTORY, LESS TREATMENT, AND MORE UNCERTAINTY ABOUT CARE OPTIONS

MALE IS THE OPPOSITE

MOOD SWINGS HIGH - STUDENTS
MOOD SWINGS MEDIUM- CORPORATE WORKERS
MOOD SWINGS LOW- HOUSEWIFES

GOING OUT DAILY CORRELATES WITH SOCIAL WEAKNESS

KEY TAKEAWAYS

FROM DATASET VARIATION ANALYSIS

HIGHER CORRELATIONS THAN EARLIER DATASETS

MALE ↔ LOWER MOOD SWINGS
FEMALE ↔ HIGHER CARE AWARENESS

HOUSEWIVES: LINKED TO LOW MOOD SWINGS AND SOCIAL WEAKNESS.

STUDENTS: LINKED TO BEING INDOORS 31-60 DAYS

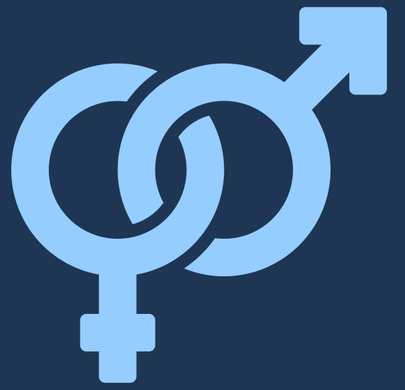
DECIDED DATASET NEW CORRELATIONS

CONCLUSIONS

FROM DATASET VARIATION ANALYSIS

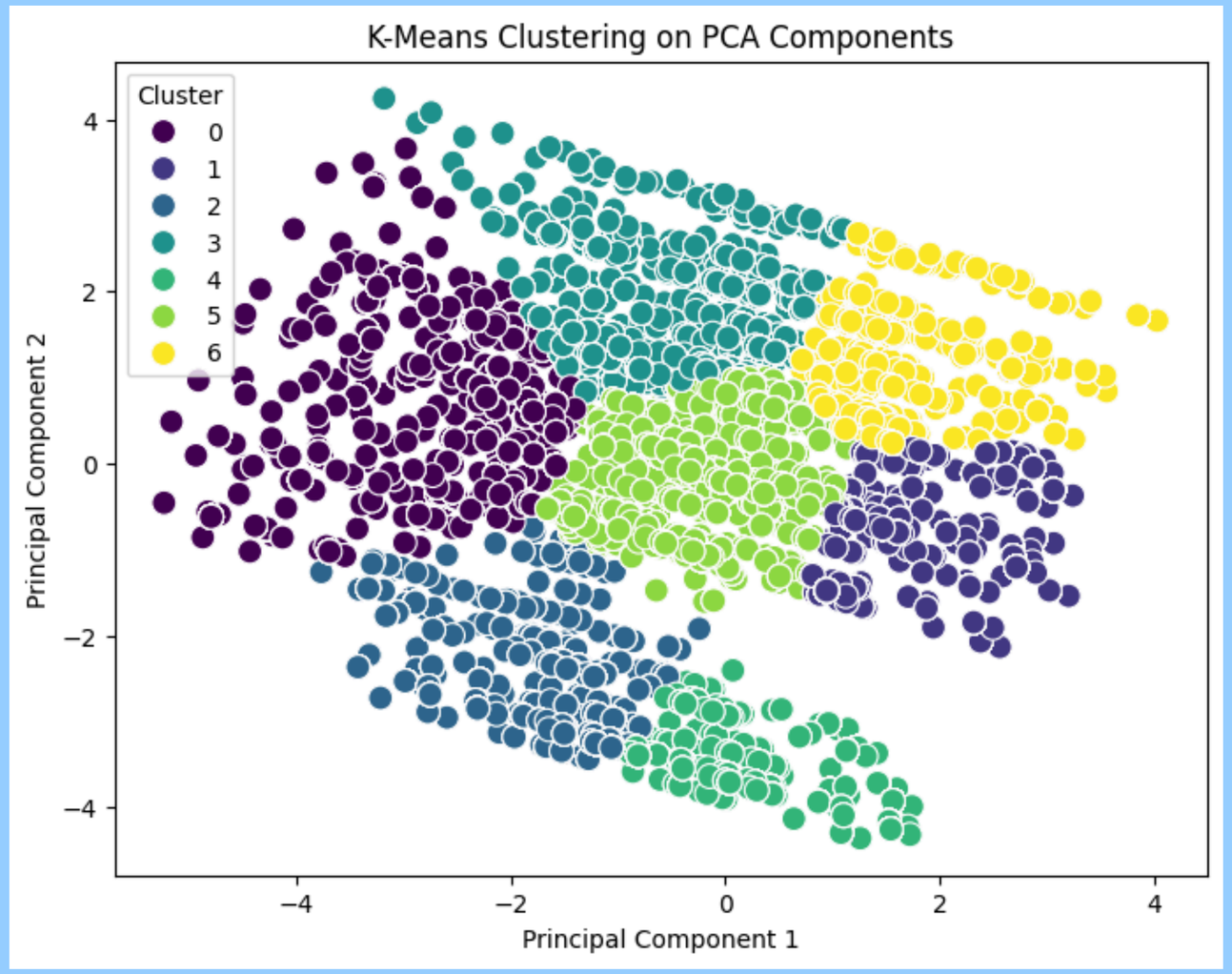
DECIDED

BEST CORRELATION AND CLUSTERING RESULTS



POSITIVE INFLUENCE

DECIDED DATASET

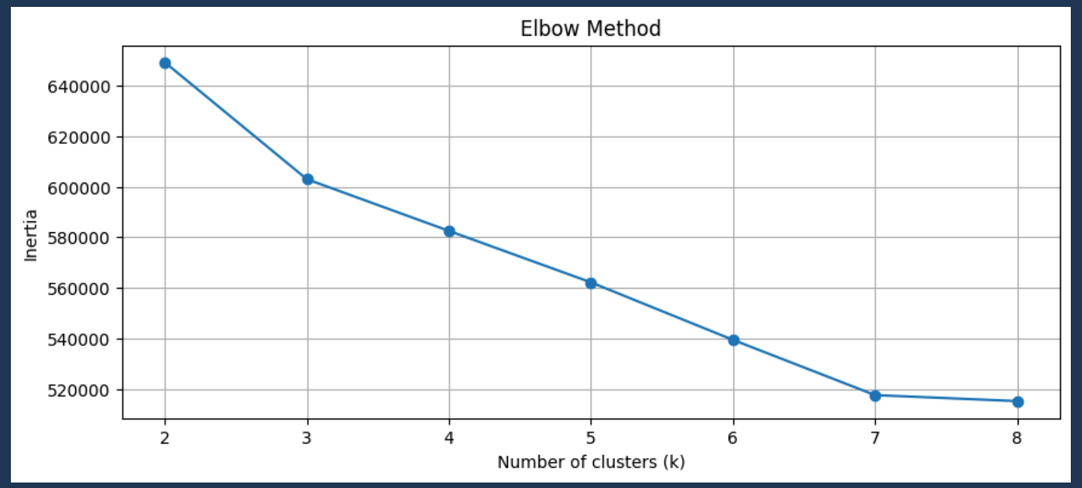


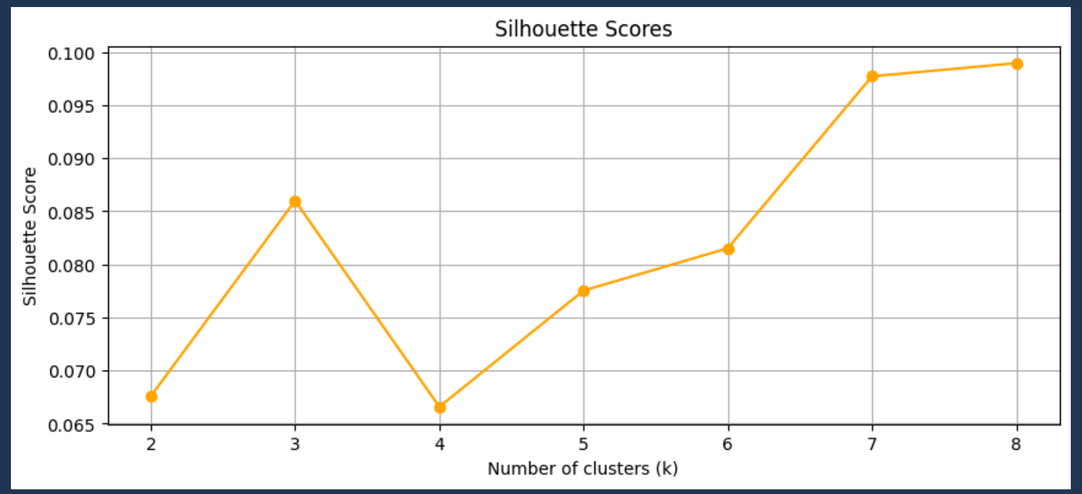
NATURAL SHAPE AND SPREAD

LESS OVERLAP

BALANCED DENSITY

DECIDED DATASET





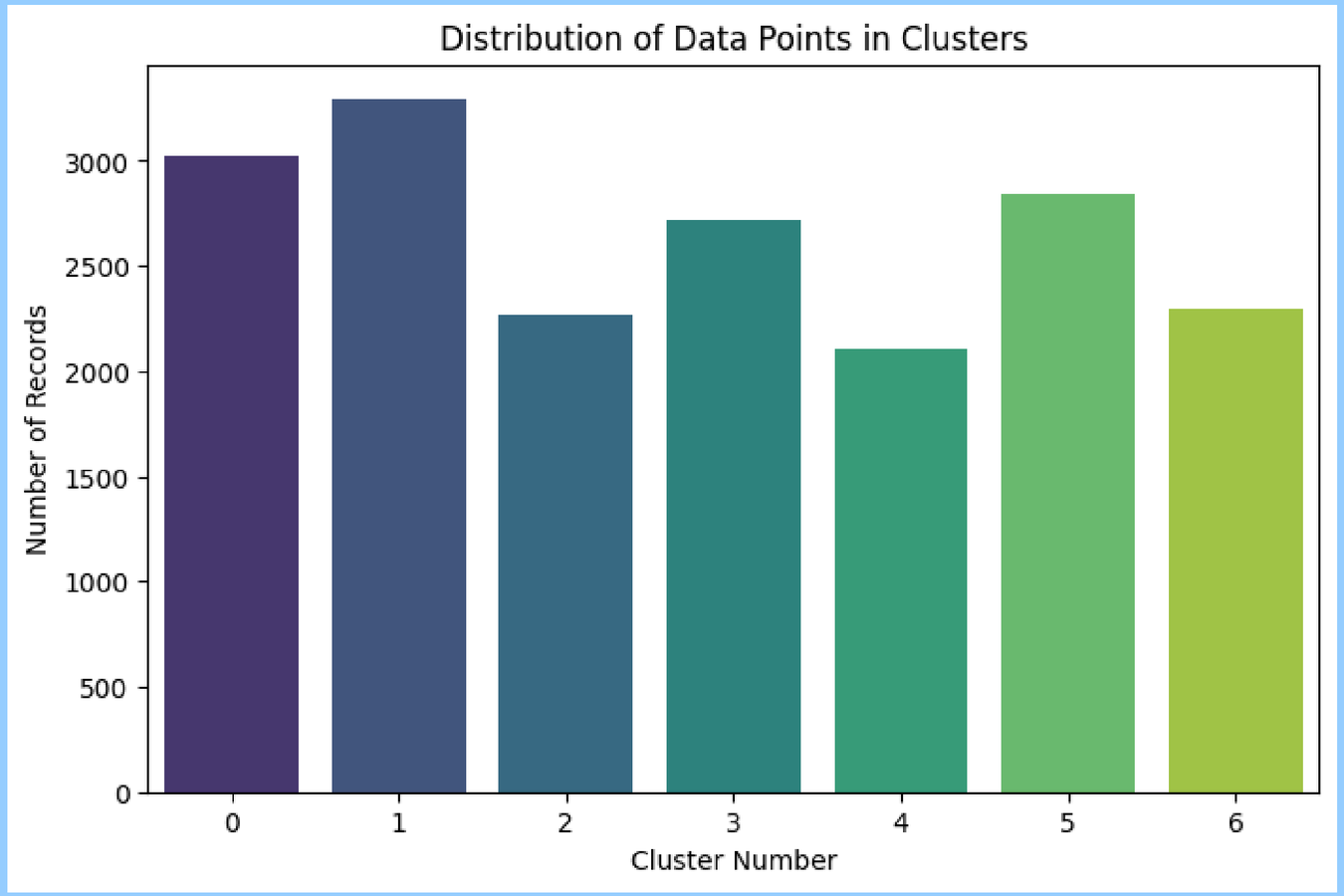
ELBOW METHOD AND SILHOUETTE SCORES WITH THE EXPECTED BEHAVIOUR

POINTED TO OPTIMAL 7 CLUSTERS

DECIDED DATASET

EVENLY DISTRIBUTED

SMALLER DATASET



CLUSTER PATTERNS

THE DECIDED

CLUSTER	LABEL	CHARACTERISTICS
CLUSTER 0	HIGH-RISK TREATED WOMEN	WOMEN WITH PREVIOUS TREATMENT AND HISTORY OF MENTAL HEALTH
CLUSTER 1	STABLE MALE WORKERS	MEN, CORPORATE AND HOUSEWIVES, WITHOUT STRESS OR HABIT CHANGES
CLUSTER 2	LOW-RISK, HIGH-FUNCTIONING MALES	ACTIVE MEN WITH LOW STRESS LEVEL, LOW TREATMENT, LOW FAMILY HISTORY AND STABLE HABITS
CLUSTER 3	SEVERELY AFFECTED MEN WITH HISTORY	MEN IN ACUTE MENTAL HEALTH CRISIS, ALREADY SEEKING OR NEEDING CARE

CLUSTER PATTERNS

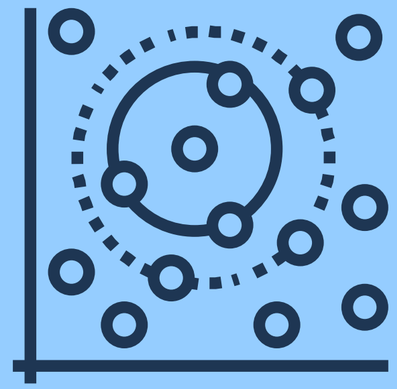
THE DECIDED

CLUSTER	LABEL	CHARACTERISTICS
CLUSTER 4	EMOTIONALLY RESILIENT MALES	MEN WORKING IN NON-TRADITIONAL FIELDS, EMOTIONALLY RESILIENT OR IN DENIAL
CLUSTER 5	UNSTABLE MEN	MEN UNDERGOING LIFESTYLE OR JOB TRANSITIONS.
CLUSTER 6	YOUNG MEN FACING NEW STRESS	YOUNG HIGH-FUNCTIONING STUDENTS WITH RISING STRESS WITH NO CURRENT TREATMENT

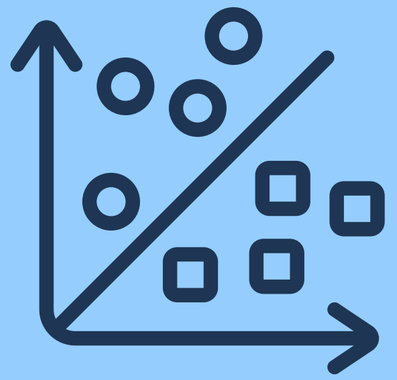
CLASSIFICATION

LOGISTIC REGRESSION

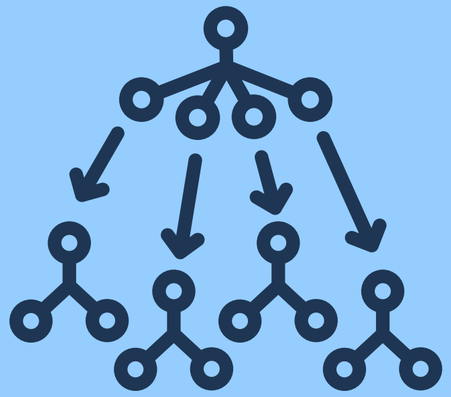
GAUSSIAN NAIVE BAYES



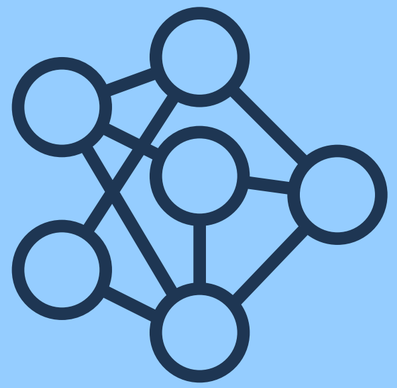
K NEAREST NEIGHBOUR



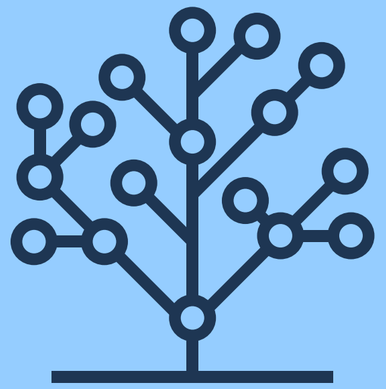
SUPPORT VECTOR MACHINE



RANDOM FOREST



NEURAL NETWORK

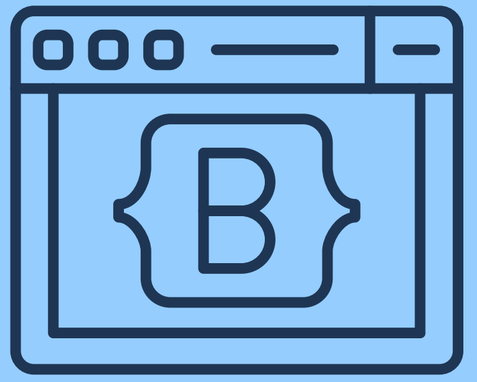


DECISION TREE

DATA SPLIT

HOLD-OUT METHOD

K-FOLD CROSS VALIDATION

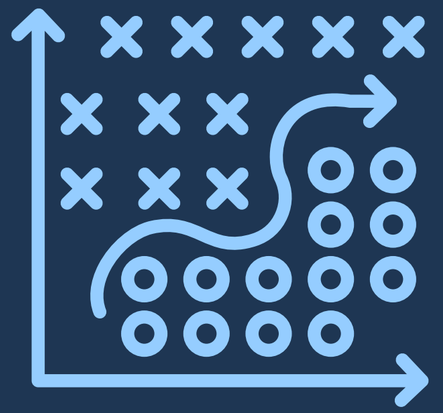


BOOTSTRAPPING

RESULTS

SMALL DATASET

SCORES ON DETECTING
GROWING STRESS



OVERFIT?

CONCLUSIONS

OVER 5 FOLD CV ON UNTOUCHED (285K) AND ANGLOPHONE COUNTRIES (234K):

- RF, DT & KNN ACHIEVED 99% ON ALL WEIGHTED AVERAGE SCORES
- NN ACHIEVED 100% ON ALL WEIGHTED AVERAGE SCORES

POSSIBLE EXPLANATION: CHOSEN CLASSIFICATION TASK IS TOO EASY FOR THE DATASET

THANK YOU

ADES 2025