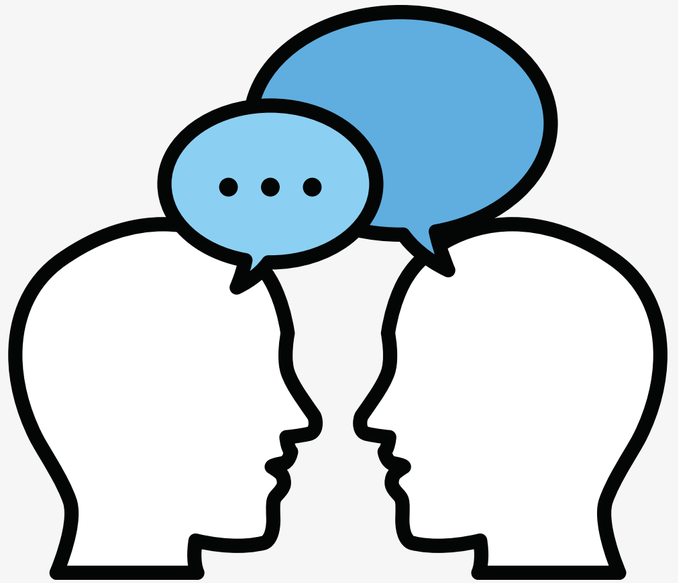
Stackoverflow Dataset MLM

BY HANS WALTER

PURPOSE OF STACKOVERFLOW

QUESTIONS

ANSWERS



COMMUNITY

PROGRAMS

stackoverflow

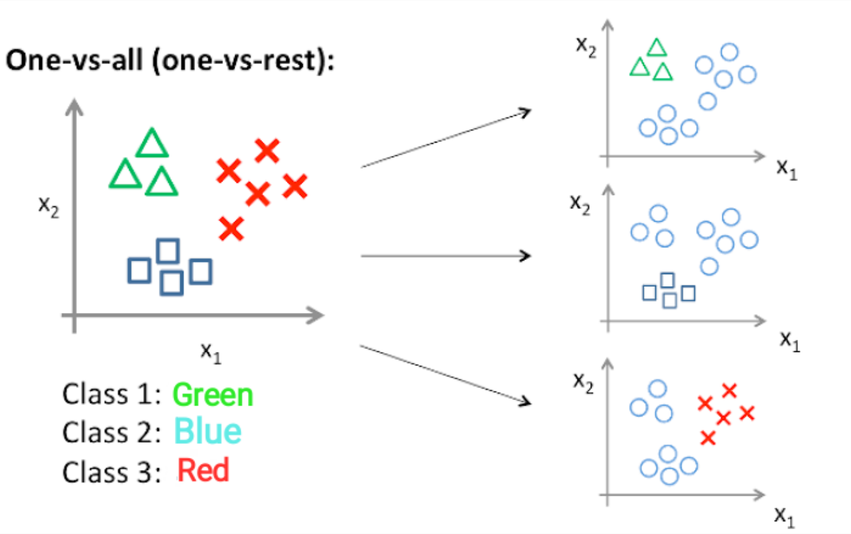
Google BigQuery

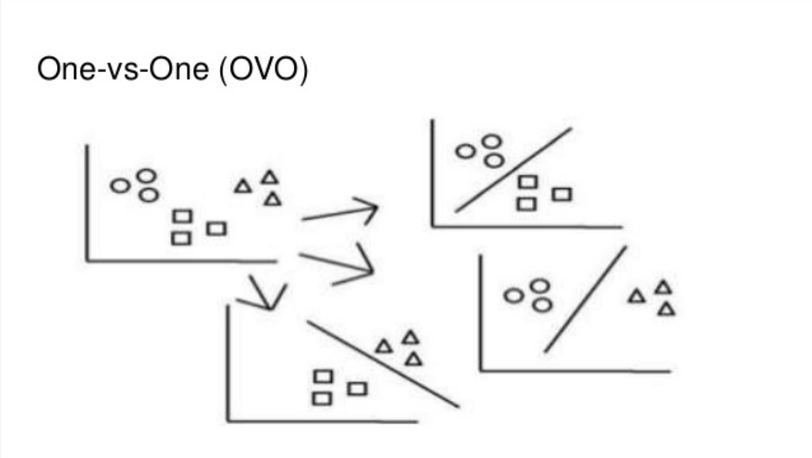
QUESTIONS

- IS THERE A CORRELATION BETWEEN FOR EXAMPLE, THE LENGTH OF THE QUESTION AND THE VARIOUS OTHER PARAMETERS TO A SPECIFIC PROGRAMMING LANGUAGE?
- HOW DOES STACKOVERFLOW ORGANIZE THEIR DATA?
- WHAT SIZE OF DATA ARE WE DEALING WITH?
- IS THERE ROOM FOR IMPROVEMENT, FOR CAPTURING DATA, ORGANIZING IT, ETC.

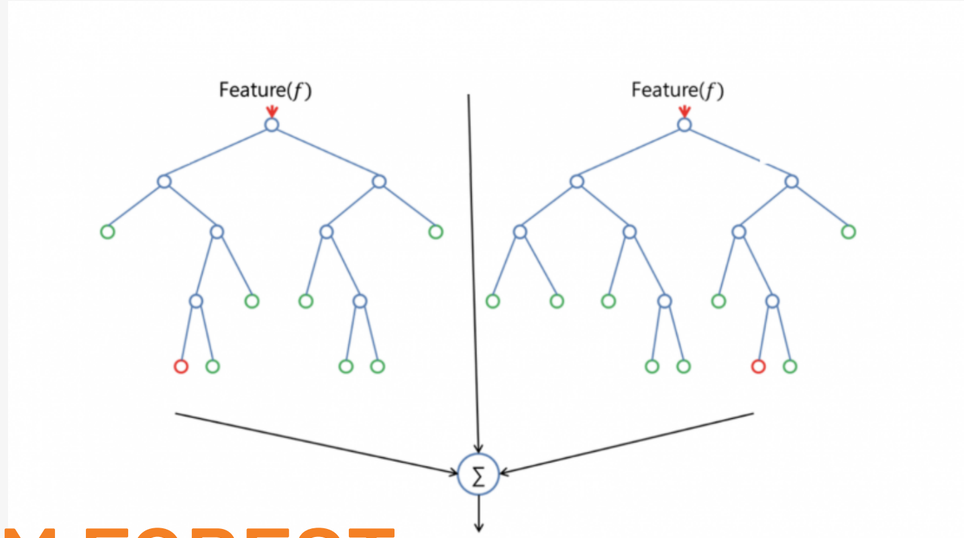
MODELS

OVA

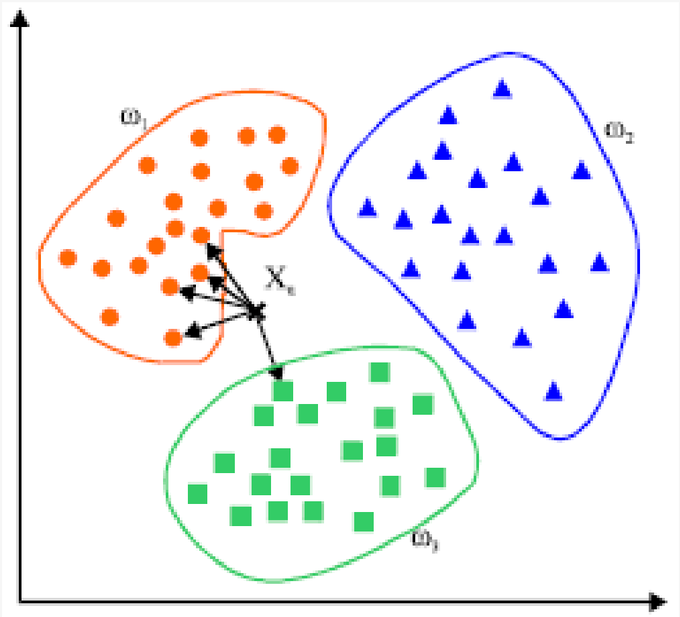




OVO



RANDOM FOREST



KNN
NEIGHBOURS

RESULTS ACCURACY

OVA 28%

OVO 19%

RANDOM FOREST 34%

KNN NEIGHBOURS 26%

CONCLUSIONS AND RECOMMENDATIONS

- LARGE AMOUNTS OF DATA DOESN'T MEAN EITHER GOOD DATA OR GOOD MODEL
- NO RELATIONSHIP BETWEEN THE SET OF FEATURES AND LABEL
- ALGORITHM CAPABLE OF FINDING PATTERNS
- ASK THE RIGHT QUESTIONS
- CONTINUOUS LEARNING AND PRACTICE
- TUNING OF MODELS CAN ENSURE BETTER RESULTS