OVERFITTING IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

ENSEMBLE LEARNING

ENSEMBLE LEARNING

INVOLVES THE USE OF MULTIPLE LEARNERS AND COMBINING THEIR RESULTS

IN 2006, NETFLIX HELD AN OPEN COMPETITION FOR A MACHINE LEARNING ALGORITHM TO PREDICT A USER'S RATING OF A MOVIE

THE GRAND PRIZE WAS A COOL MILLION !

THE COMPETITION WENT ON FOR 3 YEARS, BEFORE A GRAND PRIZE WINNER WAS DECLARED

AN INTERESTING THING HAPPENED DURING THIS TIME...

THE CONTESTANTS FOUND THAT, INSTEAD OF USING 1 SINGLE MODEL, COMBINING MULTIPLE MODELS WORKED BETTER

TEAMS STARTED MERGING INTO LARGER TEAMS, THEY WOULD COMBINE THEIR MODELS TO DO BETTER

IN THE END, THE GRAND PRIZE WINNER (AND A VERY CLOSE RUNNER UP) WERE BOTH ENSEMBLES OF MORE THAN A 100 LEARNERS EACH..

AND COMBINING THEM IMPROVED THE RESULTS EVEN FURTHER!

THE IDEA OF ENSEMBLE LEARNING IS SIMPLE..

THE IDEA OF ENSEMBLE LEARNING IS SIMPLE..

MODELS TEND TO OVERFIT

IF YOU TRAIN MULTIPLE MODELS

THE OVERFITTING COMPONENTS OF EACH OF THE MODELS WOULD BE DIFFERENT

WHEN YOU COMBINE THESE MODELS

THE OVERFITTING COMPONENTS OF THE MODELS WOULD CANCEL EACH OTHER OUT

AND YOU ARE LEFT WITH THE COMPONENTS THAT REALLY DESCRIBE YOUR DATA

LET'S TAKE AN EXAMPLE

CLASSIFY A TWEET AS POSITIVE OR NEGATIVE SENTIMENT

(THIS IS A CLASSIFICATION PROBLEM)

METHOD 1. CHOOSE 1 TECHNIQUE

NAIVE BAYES (OR) SUPPORT VECTOR MACHINES (OR) NEURAL NETWORKS

METHOD 2. USE AN ENSEMBLE

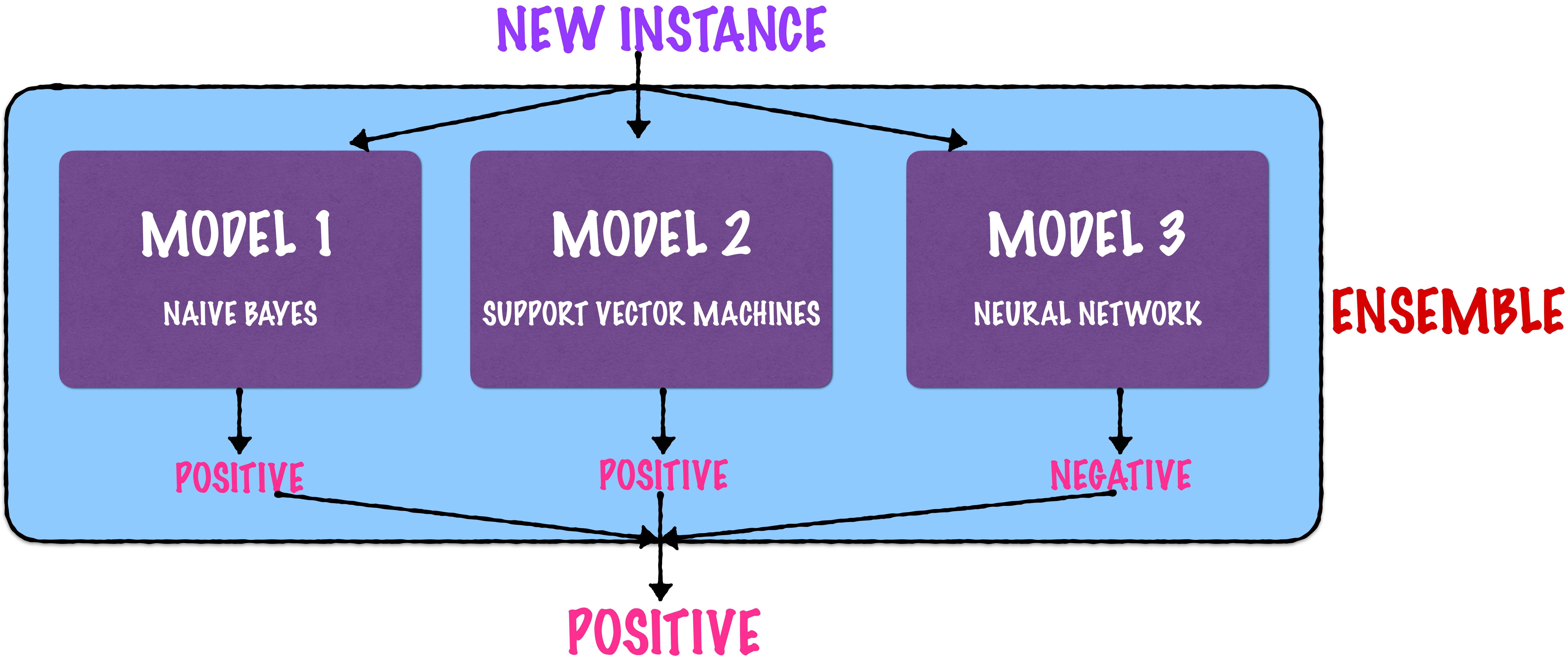
NAIVE BAYES (AND) SUPPORT VECTOR MACHINES (AND) NEURAL NETWORKS

METHOD 2. USE AN ENSEMBLE NAIVE BAYES (AND) SUPPORT VECTOR MACHINES (AND) NEURAL NETWORKS

1. TAKE THE TRAINING SET AND TRAIN EACH OF THE ABOVE CLASSIFIERS ON IT

2. WHEN A NEW INSTANCE (TWEET) COMES IN, GET THE PREDICTIONS FROM EACH OF THE MODELS

3. TAKE THE MAJORITY VOTE OF THE MODELS AND THAT WILL BE THE FINAL PREDICTION



A MACHINE LEARNING ENSEMBLE IS A COLLECTION OF MODELS

THE MODELS IN THE ENSEMBLE CAN BE

- BASED ON DIFFERENT TECHNIQUES
  - A COLLECTION WITH 1 SVM, 1 DECISION TREE, 1 NAIVE BAYES, 1 KNN
- TRAINED ON DIFFERENT TRAINING SETS
  - A COLLECTION OF SVMS, EACH TRAINED ON A DIFFERENT TRAINING SET
- USING DIFFERENT FEATURES
  - A COLLECTION OF DECISION TREES, EACH GIVEN A DIFFERENT SET OF FEATURES
- USING DIFFERENT VALUES OF PARAMETERS
  - A COLLECTION OF K-NEAREST NEIGHBOURS, EACH WITH A DIFFERENT VALUE OF K

AN ENSEMBLE LEARNER COMBINES THE RESULTS FROM INDIVIDUAL MODELS

THE FINAL RESULT CAN BE

A MAJORITY VOTE OF THE INDIVIDUAL MODELS

AVERAGE OF THE RESULT FROM INDIVIDUAL MODELS

A WEIGHTED FUNCTION OF THE RESULT FROM INDIVIDUAL MODELS

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

ENSEMBLE LEARNING

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

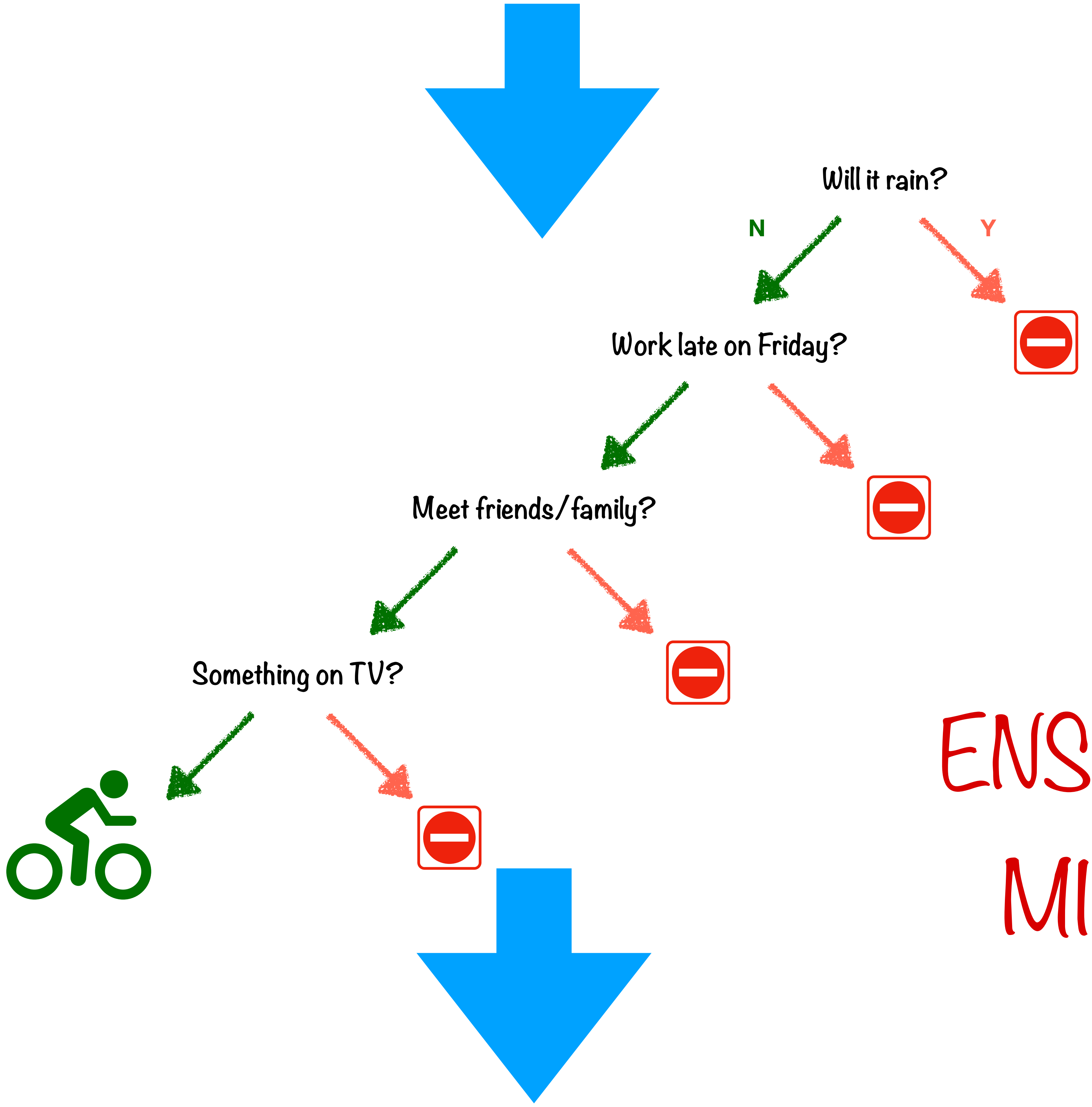
SOME OF THE WAYS TO MITIGATE THIS PROBLEM

ENSEMBLE LEARNING

Decision Tree

DECISION TREES ARE VERY PRONE TO THE RISK OF OVERFITTING

Input Variables/Predictors



Outcome/Output Variables

ENSEMBLE LEARNING CAN MITIGATE THE RISK OF OVERFITTING

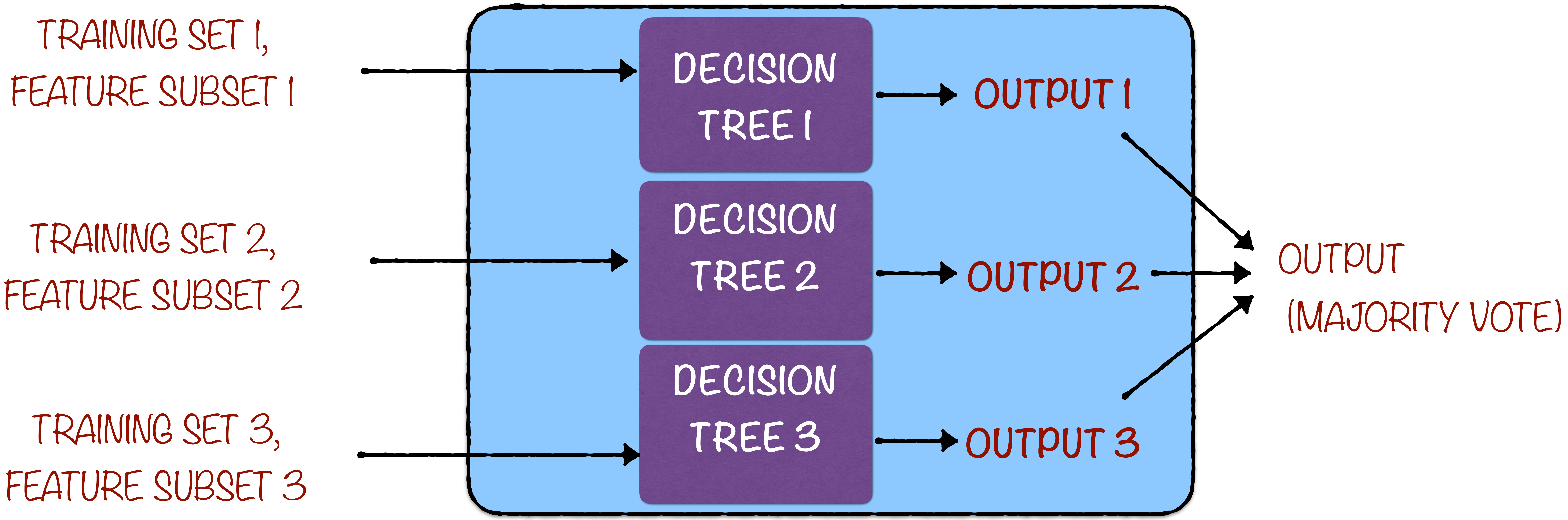
A RANDOM FOREST IS AN ENSEMBLE OF DECISION TREES

EACH DECISION TREE IN THE ENSEMBLE IS

TRAINED ON DIFFERENT TRAINING SETS

USING DIFFERENT FEATURES
(A RANDOMLY SELECTED SUBSET OF FEATURES)

Random Forest



TRAINING SET 1,
FEATURE SUBSET 1

TRAINING SET 2,
FEATURE SUBSET 2

TRAINING SET 3,
FEATURE SUBSET 3

DECISION
TREE 1

DECISION
TREE 2

DECISION
TREE 3

OUTPUT 1

OUTPUT 2

OUTPUT 3

OUTPUT
(MAJORITY VOTE)