BORCELLE

CLASSIFICATION OF DIABETIC PATIENTS

SUBTITLE: USING VARIOUS CLASSIFICATION ALGORITHMS

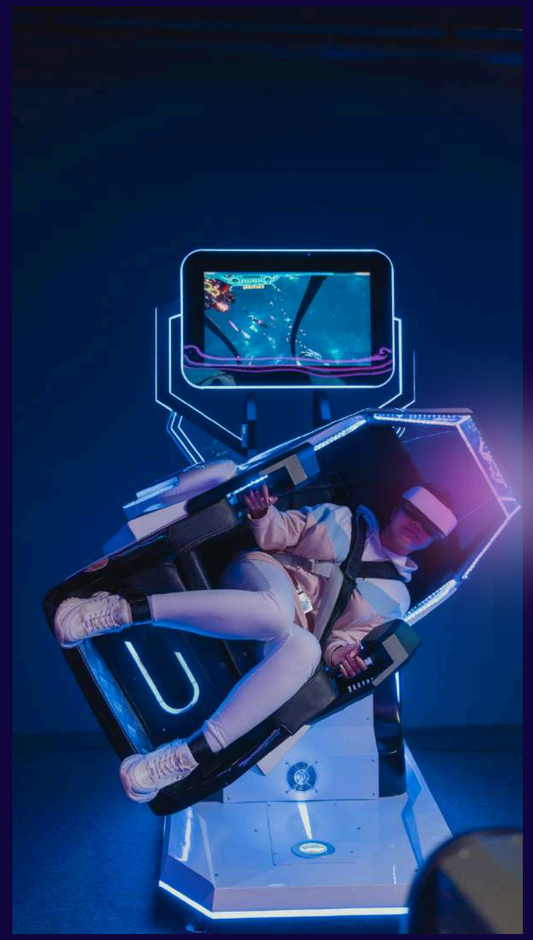
PRESENTED BY: MANAV BARIK

STUDENT ID: 21070122092

- AIM:
- TO CLASSIFY PATIENTS AS DIABETIC OR NON-DIABETIC USING VARIOUS CLASSIFICATION ALGORITHMS AND EVALUATE THEIR PERFORMANCE.
- OBJECTIVE:
  - INSTALL AND EXPLORE CLASSIFICATION PACKAGES IN R.
  - CHOOSE AN APPROPRIATE CLASSIFICATION ALGORITHM.
  - APPLY THE ALGORITHM TO THE DIABETES DATASET.
  - EVALUATE THE MODEL'S PERFORMANCE USING ACCURACY, PRECISION, RECALL, AND F1-SCORE.
  - VISUALIZE THE CLASSIFIER'S DECISION BOUNDARIES.

STEP-BY-STEP GUIDE

1. INSTALL AND LOAD NECESSARY PACKAGES
   - INSTALL PACKAGES: CARET, E1071, GGPLOT2.
   - LOAD PACKAGES IN R.
   -
2. LOAD THE DIABETES DATASET
   - USE THE PIMA INDIANS DIABETES DATASET FROM THE UCI MACHINE LEARNING REPOSITORY.



DATA PREPROCESSING

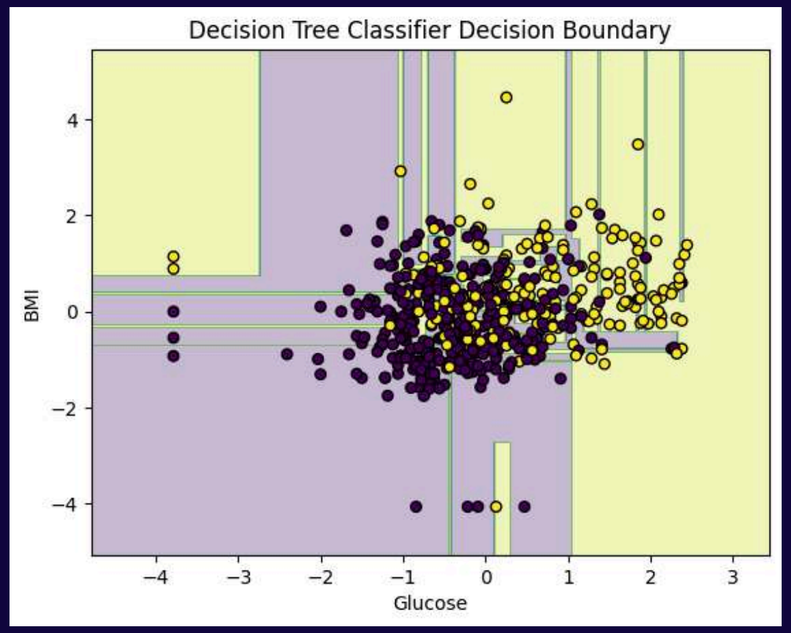
- CHECK FOR MISSING VALUES:
- USE SUM(IS.NA(DATA)) TO IDENTIFY ANY MISSING DATA.
- SPLIT THE DATA:
  - CREATE TRAINING AND TESTING SETS USING CREATEDATAPARTITION.

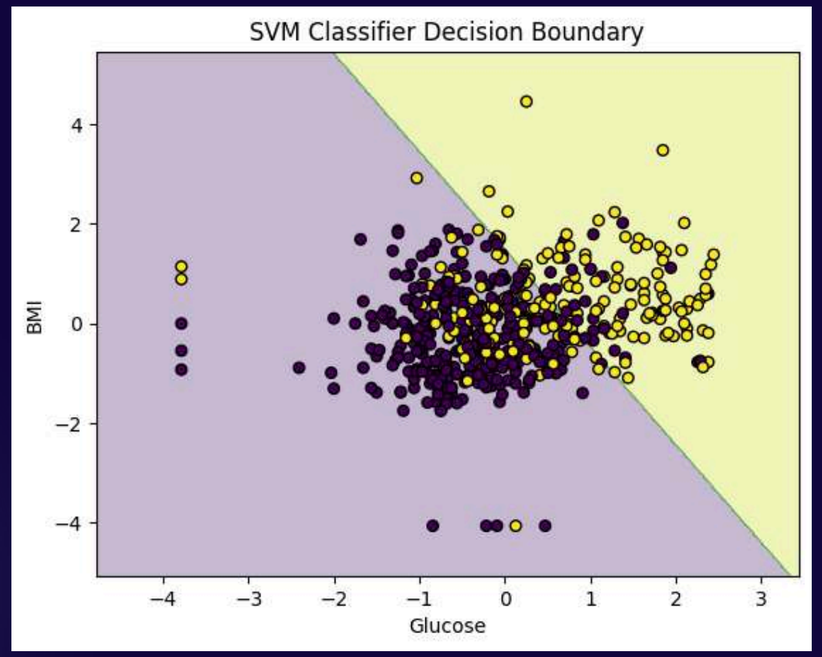
TRAIN A CLASSIFICATION MODEL

- MODEL SELECTION:
- WE'LL USE A DECISION TREE FOR THIS EXAMPLE.
-
- TRAINING THE MODEL:
  - CODE: MODEL <- TRAIN(OUTCOME ~ ., DATA = TRAINDATA, METHOD = "RPART")

MODEL EVALUATION

- MAKE PREDICTIONS:
- USE THE TRAINED MODEL TO PREDICT OUTCOMES ON THE TEST DATA.
- PERFORMANCE METRICS:
  - ACCURACY, PRECISION, RECALL, F1-SCORE.
  - EXAMPLE METRICS FOR DECISION TREE:
    - ACCURACY: 0.675
    - PRECISION: 0.528
    - RECALL: 0.588
    - F1-SCORE: 0.556





DECISION TREE CLASSIFIER PERFORMANCE:

ACCURACY: 0.6753246753246753

PRECISION: 0.5280898876404494

RECALL: 0.5875

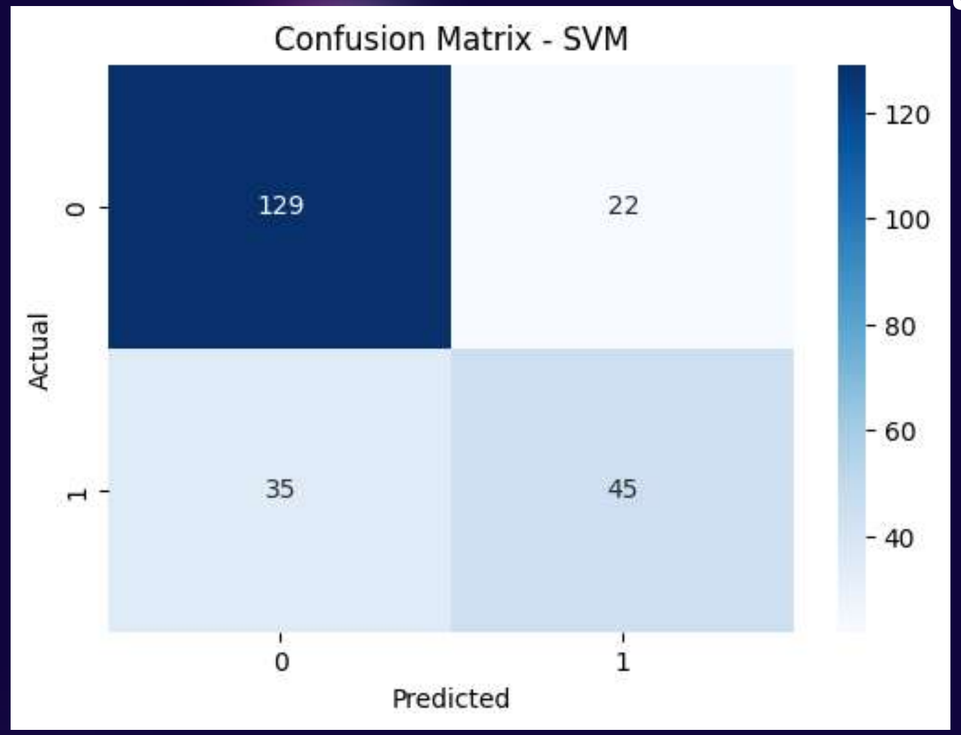
F1-SCORE: 0.5562130177514792

SUPPORT VECTOR MACHINE PERFORMANCE

- SVM CLASSIFIER PERFORMANCE:
  - ACCURACY: 0.753
  - PRECISION: 0.672
  - RECALL: 0.563
  - F1-SCORE: 0.612

VISUALIZING DECISION BOUNDARIES

- VISUALIZATION:
  - USE GGPLOT TO PLOT DECISION BOUNDARIES BASED ON GLUCOSE AND BMI.
  - CODE SNIPPET FOR VISUALIZATION
  -
  -
  - GGPLOT(TRAINDATA, AES(X = GLUCOSE, Y = BMI, COLOR = AS.FACTOR(OUTCOME))) +
  - GEOM_POINT() +
  - STAT_CONTOUR(DATA = AS.DATA.FRAME(PREDICT(MODEL, TRAINDATA, TYPE = "PROB")), AES(Z = ..LEVEL..), BINS = 1) +
  - LABS(TITLE = "DECISION BOUNDARIES", X = "GLUCOSE", Y = "BMI")



CONCLUSION

- SUMMARY:
  - A DECISION TREE WAS USED TO CLASSIFY PATIENTS AS DIABETIC OR NON-DIABETIC.
  -
  - THE MODEL'S PERFORMANCE WAS EVALUATED USING VARIOUS METRICS.
  -
  - VISUALIZATION OF DECISION BOUNDARIES AIDS IN UNDERSTANDING MODEL PREDICTIONS.
  -
  - THIS APPROACH CAN BE EXTENDED TO OTHER CLASSIFICATION ALGORITHMS AND DATASETS FOR COMPREHENSIVE ANALYSIS

THANK YOU FOR YOUR ATTENTION!