PREDICTIVE ANALYTICS IN R

USING MACHINE LEARNING IN R TO PREDICT MAJOR DEPRESSIVE DISORDER AMONG INDIVIDUALS DIAGNOSED WITH OBSESSIVE COMPULSIVE DISORDER

Jorge Valderrama, PhD

OBJECTIVES

Comprehensive Data Exploration:

- Investigate a dataset comprising individuals diagnosed with Obsessive-Compulsive Disorder (OCD) to enhance our understanding of the factors predicting co-morbid Major Depressive Disorder (MDD).

Advanced Predictive Modeling:

- Employ Machine Learning algorithms in R to predict the occurrence of MDD among individuals with OCD, aiming for a nuanced understanding of the predictive factors.

OBJECTIVES (CONT.)

Algorithmic Evaluation:

- Implement and assess the efficacy of advanced algorithms, including Random Forest, Neural Network, and Xgboost, to predict MDD. Evaluate and compare their accuracy against the model training dataset to ensure robust predictive performance

Model Validation with Test Datasets:

- Rigorously validate the predictive models by comparing their performance against independent test datasets to ensure the reliability and generalizability of the developed models.

DATA

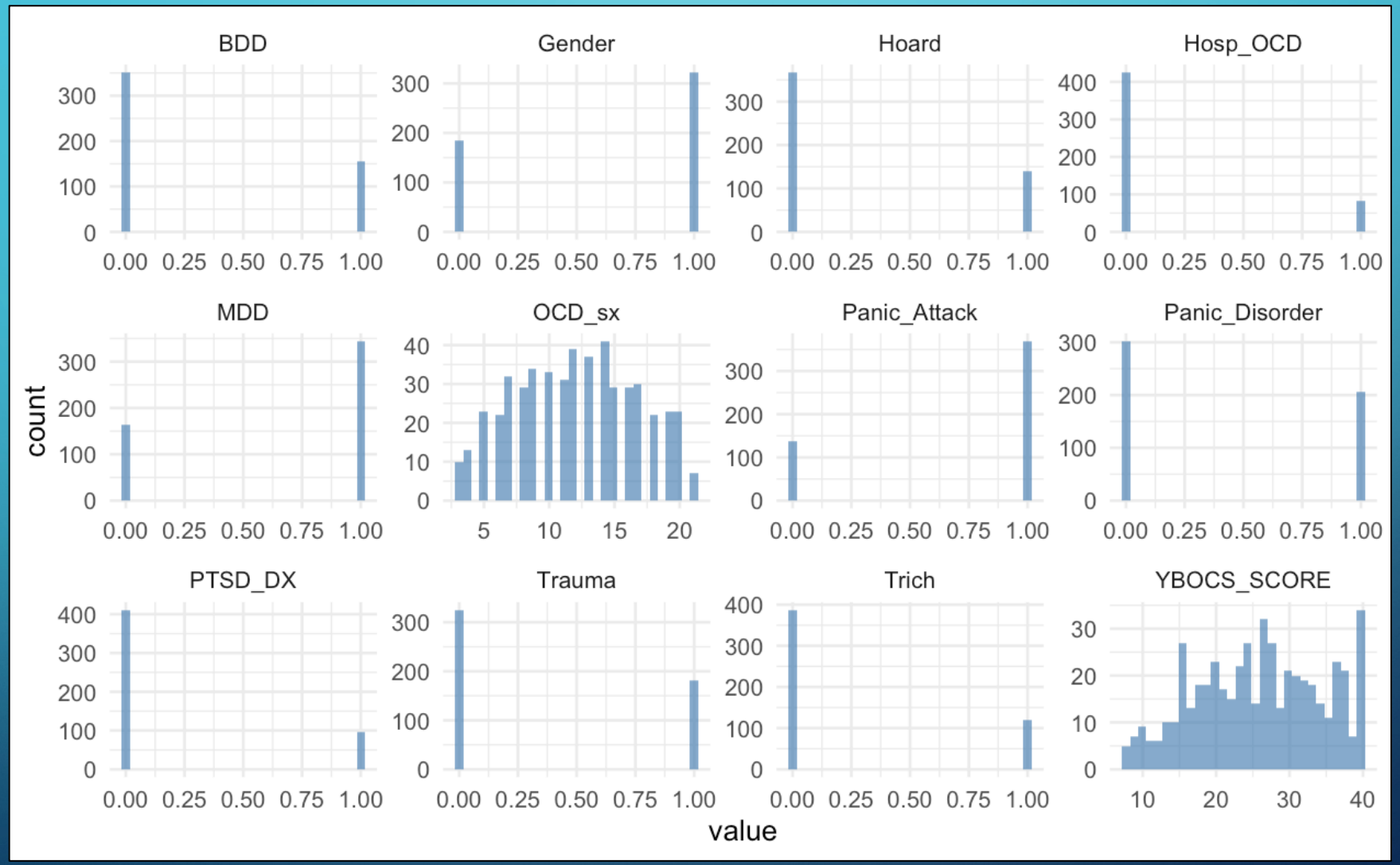
- The data includes a subset of 509 individuals with OCD who were recruited as part of the larger Genomic Psychiatry Cohort OCD research study (GPC-OCD). [1]

[1]Valderrama J., Hansen, S. K., Pato, C., Phillips, K., Knowles, J., & Pato, M. T. (2020). Greater history of traumatic event exposure and PTSD associated with comorbid body dysmorphic disorder in a large OCD cohort. Psychiatry research, 289, 112962.

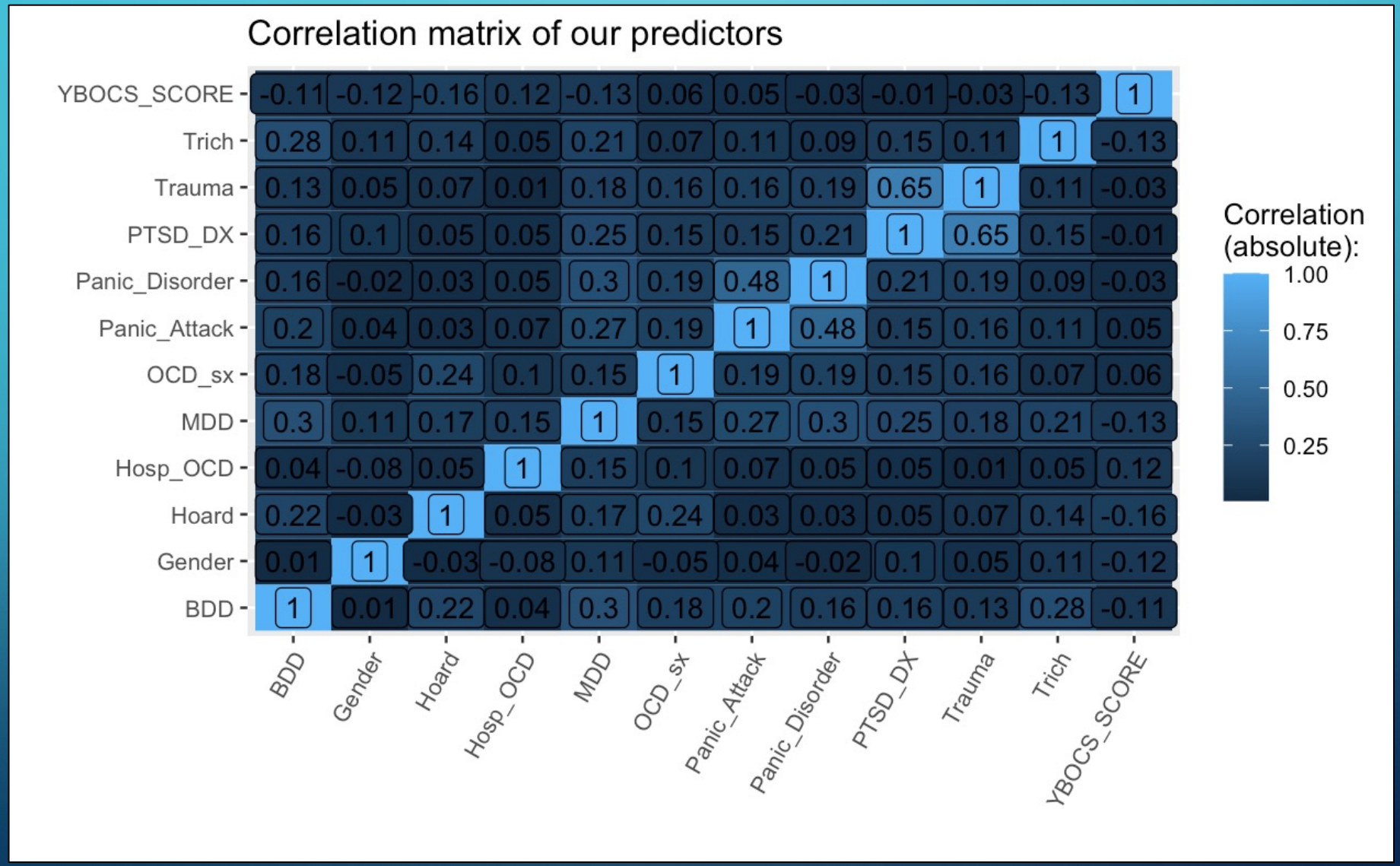
Variable/Feature Name[1]	Definition
BDD	Body Dysmorphic Disorder diagnosis
Hoard	Hoarding Disorder diagnosis
Trich	Trichotillomania diagnosis
Hosp_OCD	Ever hospitalized due to OCD symptoms
YBOCS_SCORE	Score scale that assesses severity of OCD symptoms
Panic_Attack	History of at least one panic attack
Panic_Disorder	Panic Disorder diagnosis
Trauma	History of at least one lifetime traumatic event
PTSD_DX	Presumed PTSD diagnosis
OCD_sx	Number of different types of OCD symptoms

[1]Valderrama J., Hansen, S. K., Pato, C., Phillips, K., Knowles, J., & Pato, M. T. (2020). Greater history of traumatic event exposure and PTSD associated with comorbid body dysmorphic disorder in a large OCD cohort. Psychiatry research, 289, 112962.

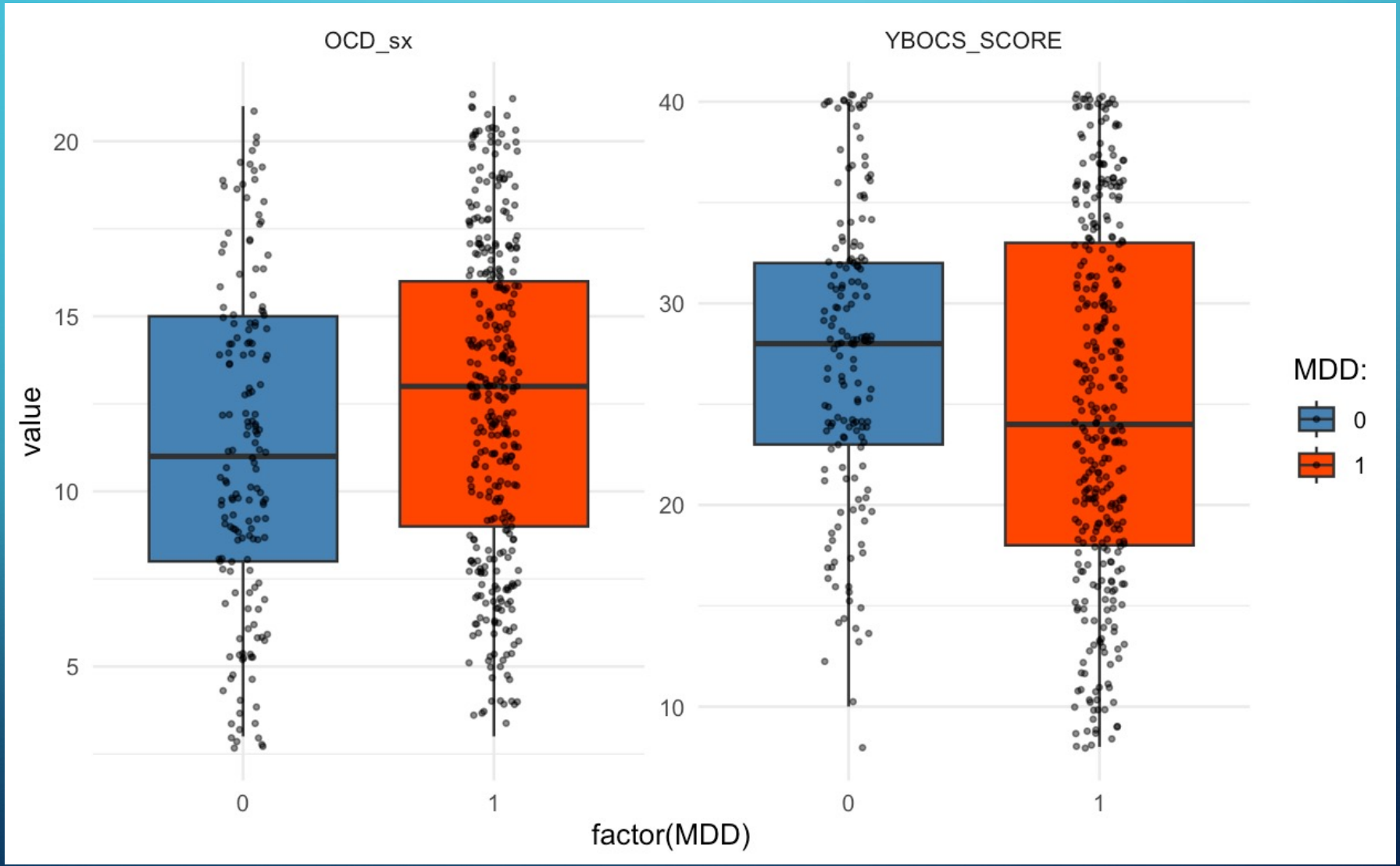
VISUAL INSPECTION OF FEATURES IN DATASET



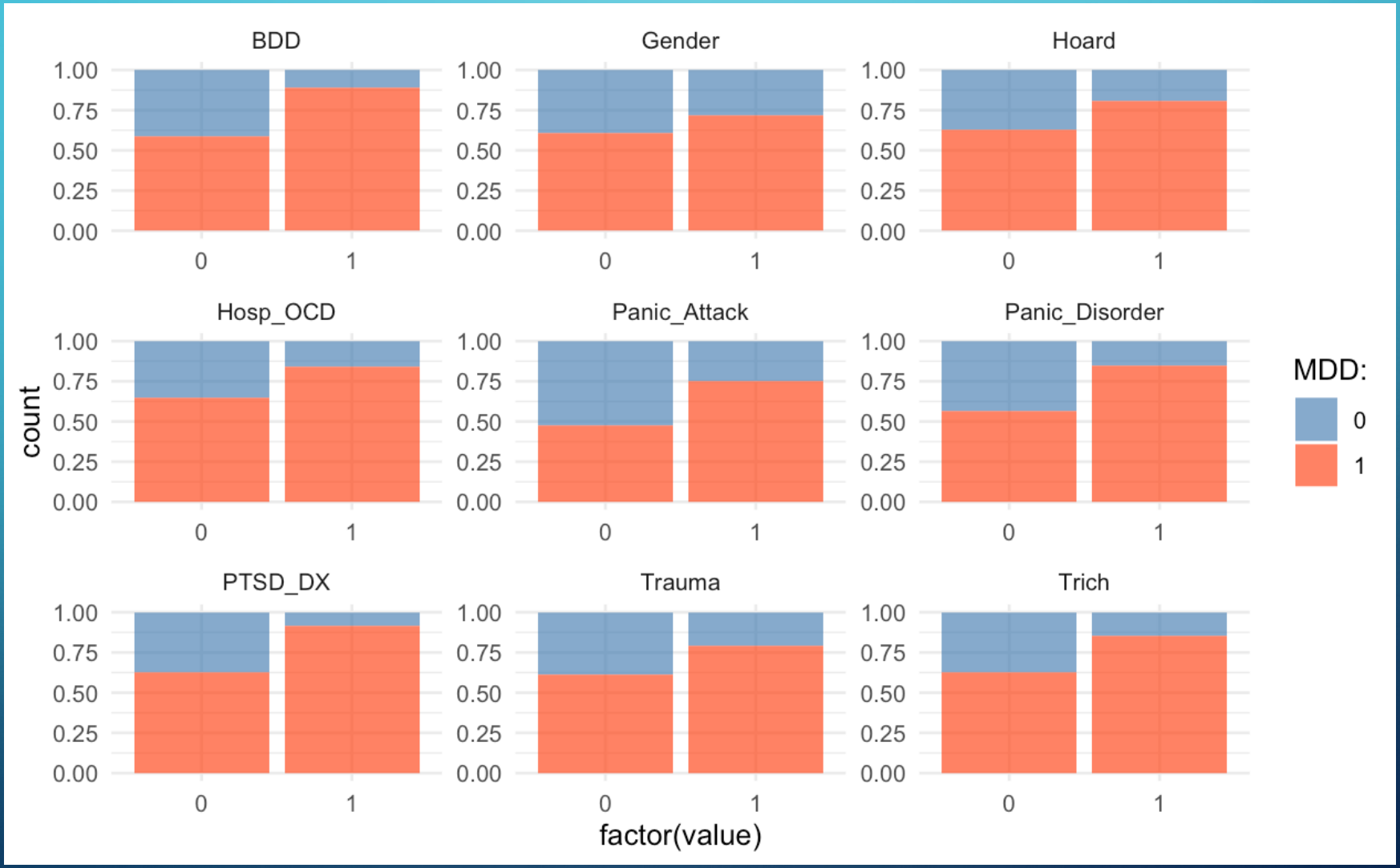
PLOTTING A CORRELATION MATRIX



BIVARIATE RELATIONS BETWEEN CONTINUOUS PREDICTORS AND MDD



BIVARIATE RELATIONS BETWEEN CATEGORICAL PREDICTORS AND MDD



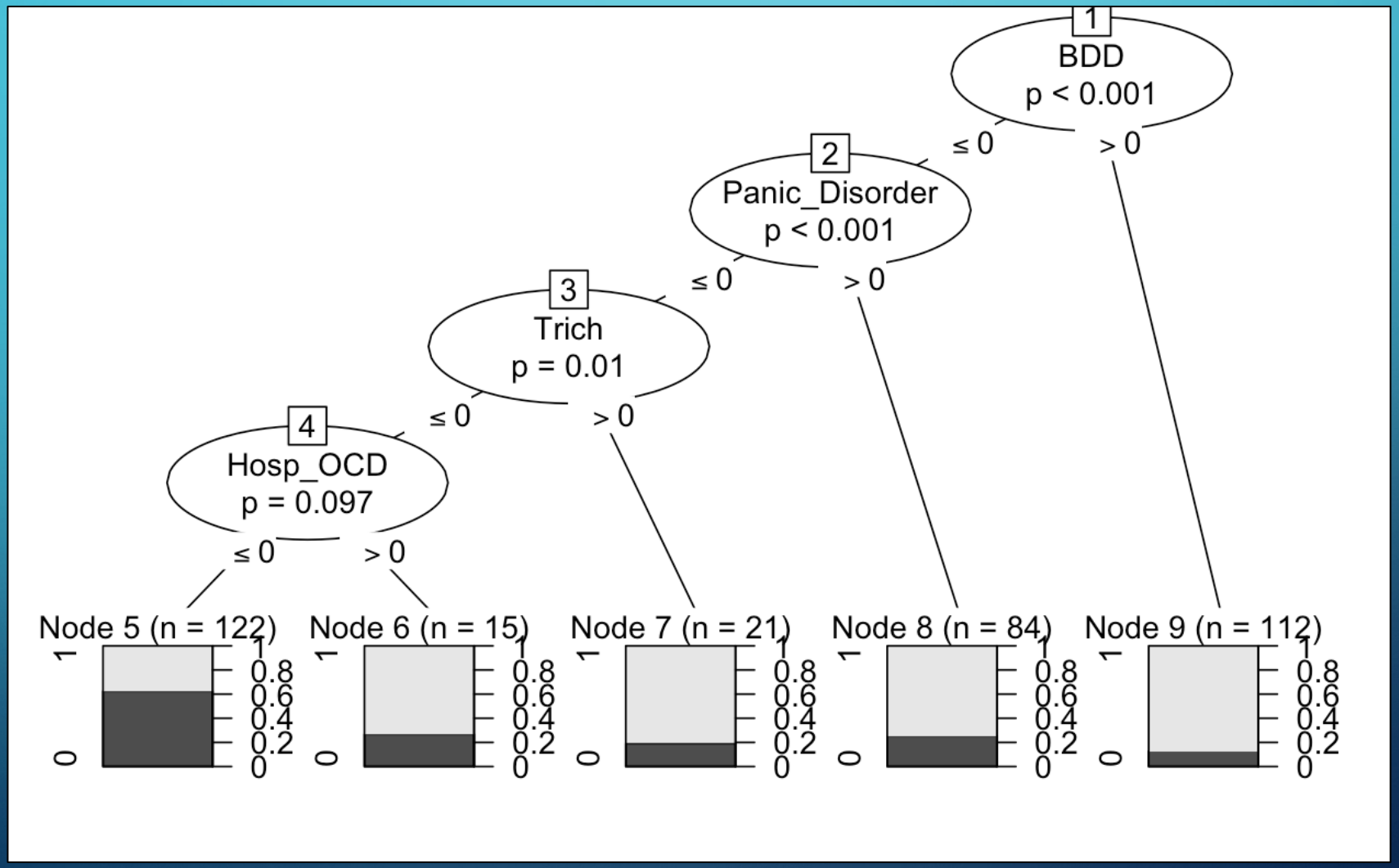
PARTITION DATA INTO TRAINING AND TEST DATASETS

- In R, we want to create training and test datasets to be used for our machine learning algorithms.
- I created a sequence of random numbers which encompass 70% of the dataset, and designated this as “training”, and the rest as a test dataset which will not be touched again until the very end of the analysis.

PRE-PROCESSING

- Since the YBOCS_Score feature was skewed, I created a normalized version of the variable by applying a log transformation
- I kept both variables in the model since the algorithms I used are less sensitive to feature scales (as opposed to a linear regression model)

TRAINING A SIMPLE DECISION TREE ON THE TRAINING DATA

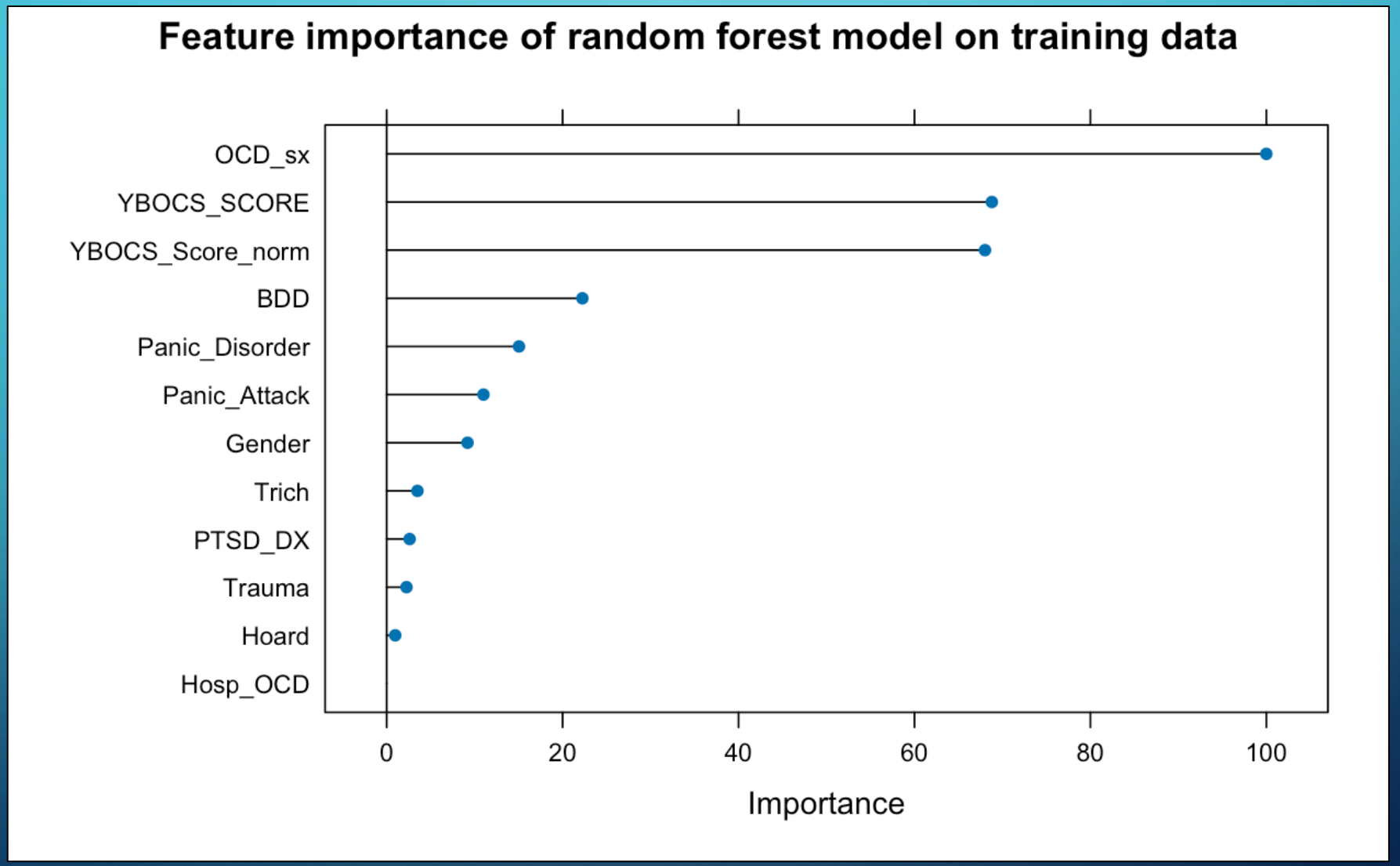


MODEL TRAINING: RANDOM FOREST ALGORITHM

Summary of model training (random forest):

```
354 samples
 12 predictor
  2 classes: '1', '0'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 283, 283, 283, 284, 283
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   5    0.7006036  0.3063659
  10    0.6863984  0.2699655

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 5.
```

MODEL TRAINING: RANDOM FOREST ALGORITHM

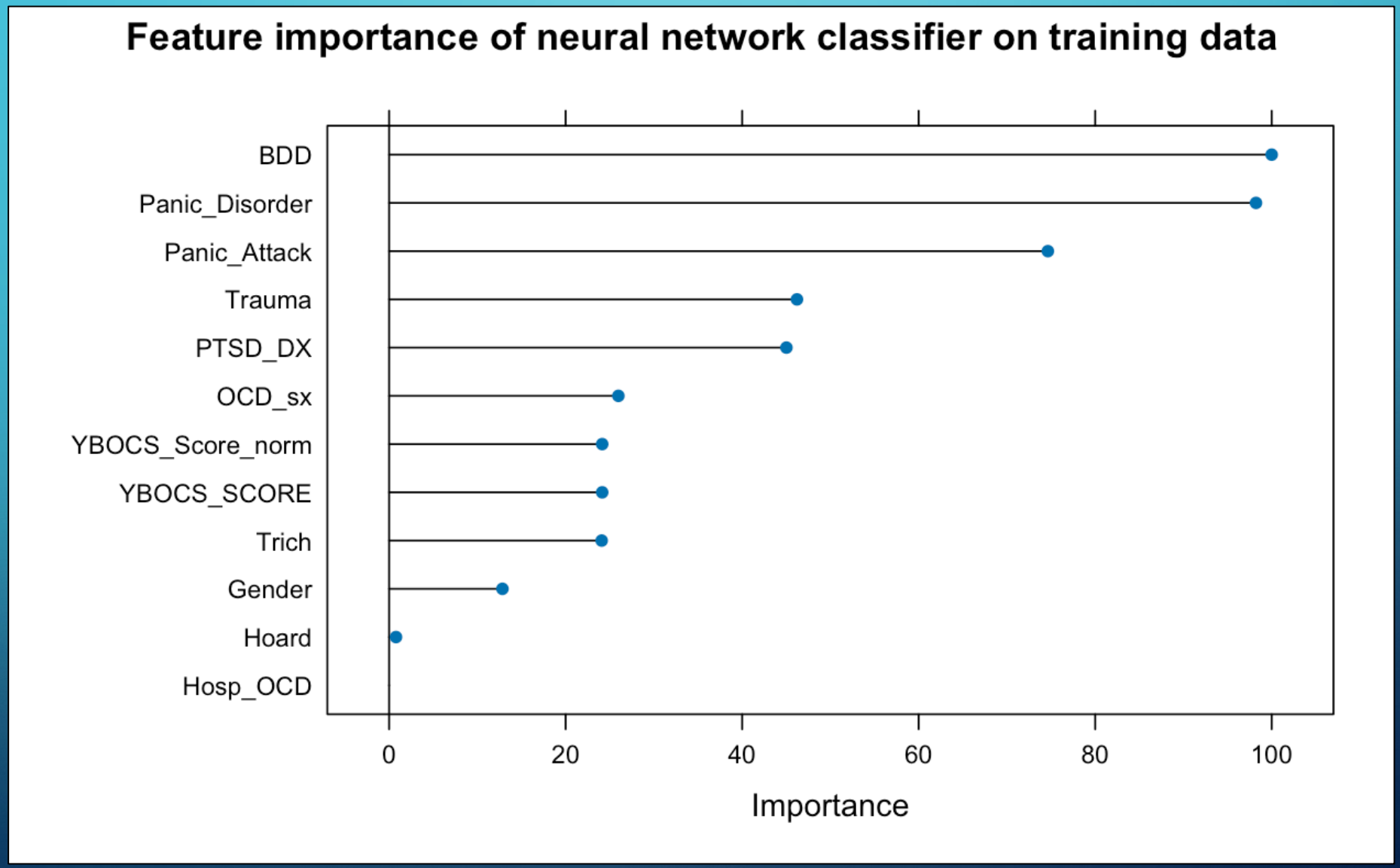


MODEL TRAINING: AVERAGED NEURAL NETWORK

Summary of model training (neural network):

```
Model Averaged Neural Network

354 samples
 12 predictor
  2 classes: '1', '0'

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 283, 283, 283, 284, 283
Resampling results across tuning parameters:

  size  decay  bag    Accuracy   Kappa
   3    0.000  FALSE  0.7005231  0.3175321
   3    0.000   TRUE  0.7006439  0.2903065
   3    0.001  FALSE  0.7032596  0.3432958
   3    0.001   TRUE  0.7315493  0.3835772
   3    0.010  FALSE  0.7173843  0.3598478
   3    0.010   TRUE  0.7118310  0.3452840
   3    0.100  FALSE  0.7286922  0.3781689
   3    0.100   TRUE  0.7343662  0.3748106
   6    0.000  FALSE  0.7030986  0.2978297
   6    0.000   TRUE  0.7033400  0.3059384
```

MODEL TRAINING: AVERAGED NEURAL NETWORK

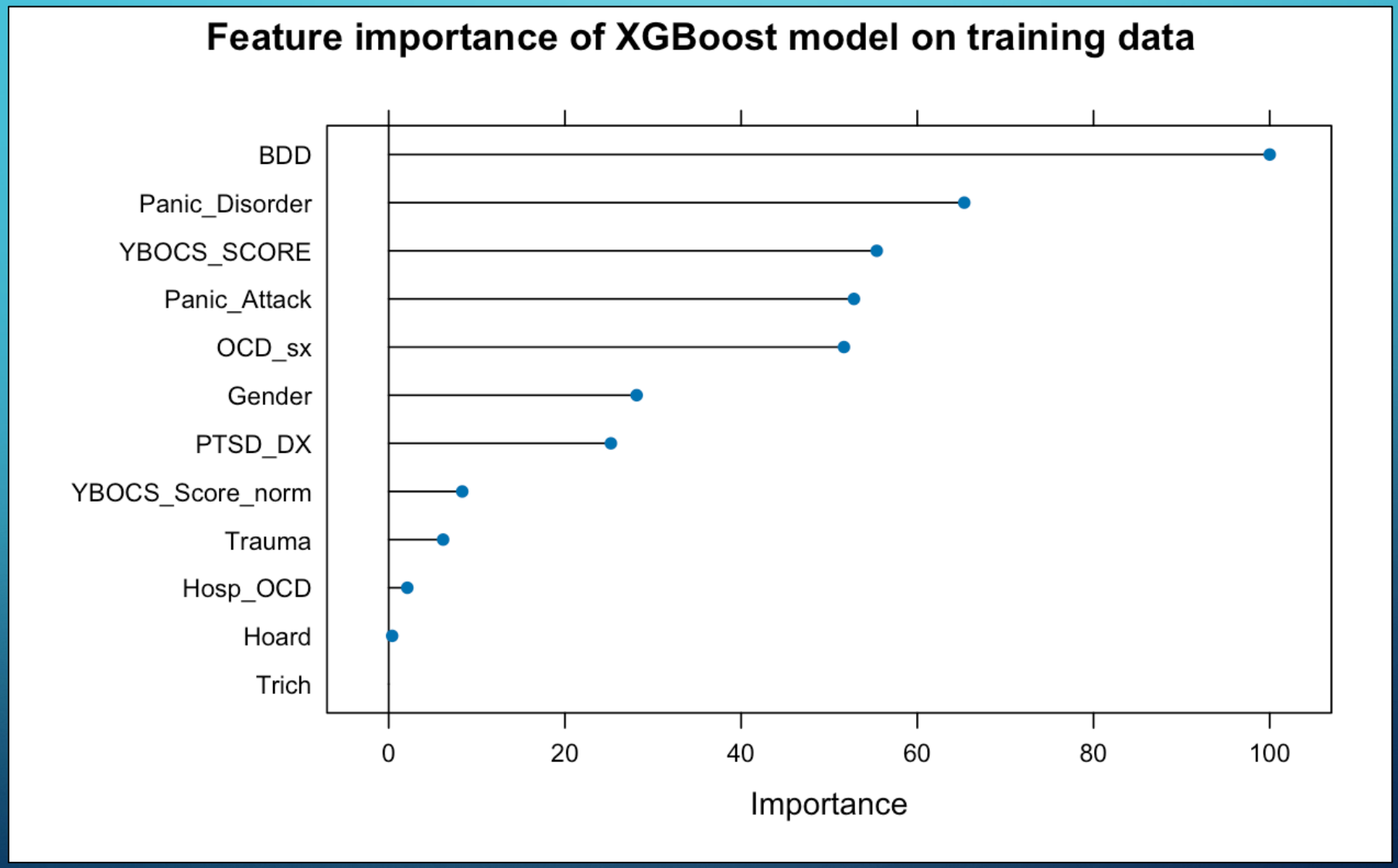


MODEL TRAINING: XGBOOST

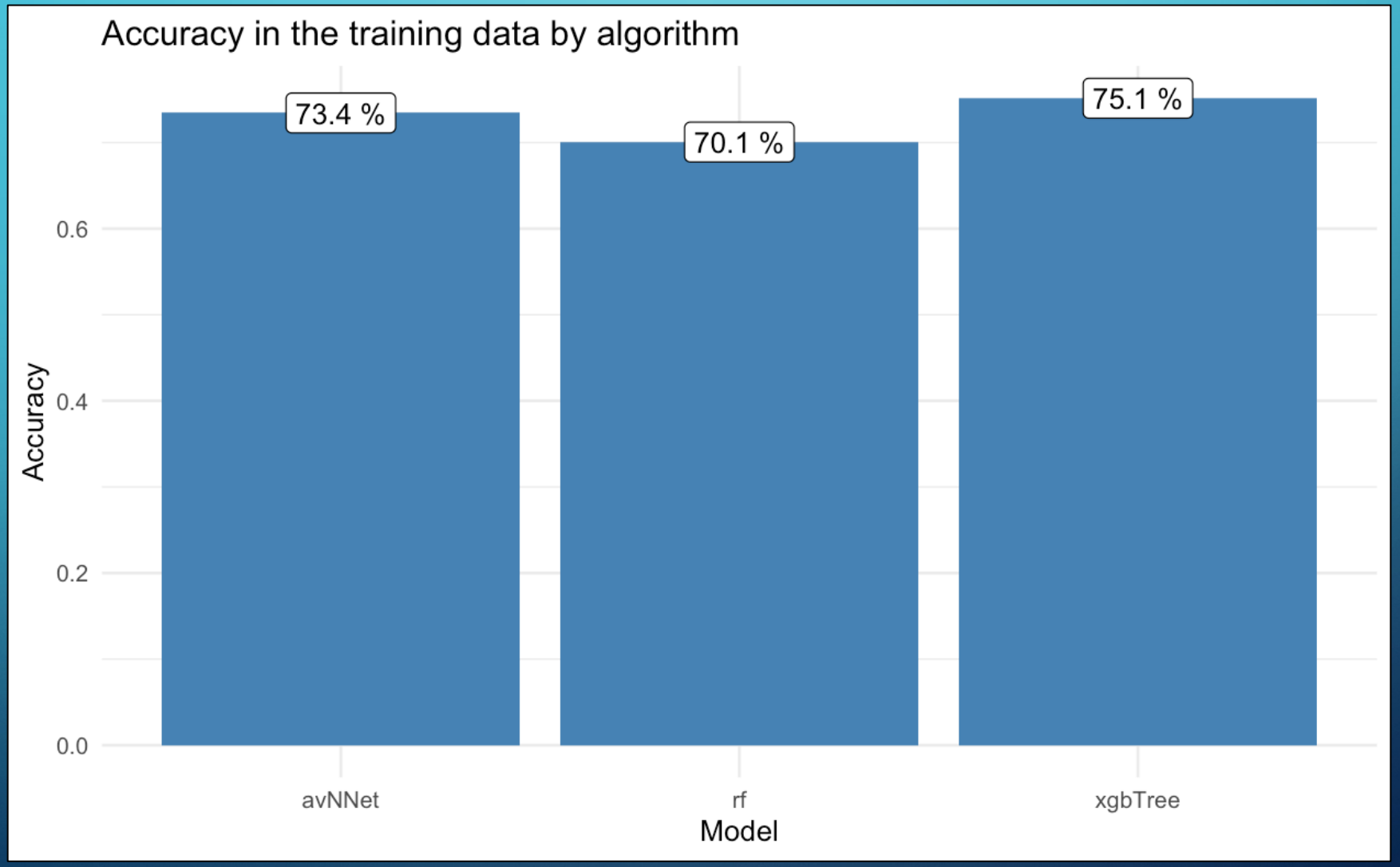
Summary of model training (Xgboost):

```
eXtreme Gradient Boosting

354 samples
 12 predictor
  2 classes: '1', '0'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 283, 283, 283, 284, 283
Resampling results across tuning parameters:

  eta  max_depth  gamma  colsample_bytree  min_child_weight  subsample  nrounds  Accuracy
  0.1  5          0.0    0.8                1                0.8         50      0.7117907
  0.1  5          0.0    0.8                1                0.8        100      0.7060362
  0.1  5          0.0    0.8                1                1.0         50      0.7146479
  0.1  5          0.0    0.8                1                1.0        100      0.7004024
  0.1  5          0.0    0.8                5                0.8         50      0.7288934
  0.1  5          0.0    0.8                5                0.8        100      0.7289336
  0.1  5          0.0    0.8                5                1.0         50      0.7090141
  0.1  5          0.0    0.8                5                1.0        100      0.7147284
```

MODEL TRAINING: XGBOOST



COMPARING THE PERFORMANCE OF THE THREE ALGORITHMS



MODEL EVALUATION AGAINST THE TEST DATA

COMPARE THE RANDOM FOREST MODEL'S PREDICTION AGAINST THE RESERVED TEST DATASET

Prediction	Reference	
	MDD	No MDD
MDD	98	21
No MDD	11	23

Summary Statistics	
Accuracy	79.08%
Sensitivity	83.33%
Specificity	93.02%

Other Metrics for the Random Forest Model	
Precision	82.35%
Recall	89.90%
F1 (Harmonic Mean of Precision and Recall)	85.96%

COMPARE THE NEURAL NETWORK MODEL'S PREDICTION AGAINST THE RESERVED TEST DATASET

Prediction	Reference	
	MDD	No MDD
MDD	90	15
No MDD	19	29

Summary Statistics	
Accuracy	77.78%
Sensitivity	82.57%
Specificity	65.91%

Other Metrics for the Neural Network Model	
Precision	85.71%
Recall	82.56%
F1 (Harmonic Mean of Precision and Recall)	84.11%

COMPARE THE XGBOOST MODEL'S PREDICTION AGAINST THE RESERVED TEST DATASET

Prediction	Reference	
	MDD	No MDD
MDD	97	16
No MDD	12	29

Summary Statistics	
Accuracy	81.70%
Sensitivity	88.99%
Specificity	63.64%

Other Metrics for the Xgboost Model	
Precision	85.84%
Recall	88.99%
F1 (Harmonic Mean of Precision and Recall)	87.38%

SUMMARY

Strategic Machine Algorithms:

- Developed algorithms designed to offer valuable insights for leveraging predictive analytics in identifying Major Depressive Disorder (MDD) among individuals with Obsessive-Compulsive Disorder (OCD).

Significant Predictive Factor:

- Identified co-morbid Body Dysmorphic Disorder as a pivotal feature in effectively predicting MDD during the model training phase.

SUMMARY (CONT.)

Enhanced Modeling Accuracy:

- Demonstrated superior accuracy of Random Forest and Xgboost algorithms in the evaluation of our model against the test data.

Potential for Real-world Application:

- Recognized the potential for deploying one of these advanced algorithms to create a streamlined screening tool. This tool, based on established risk factors for MDD, could be applied in primary care or community settings, facilitating early intervention strategies.