Breast Cancer Project Part 4

Question 14: K-nearest neighbors model

2.0/2.0 points (graded)

Set the seed to 7, then train a k-nearest neighbors model on the training set using the `caret` package. Try odd values of k from 3 to 21. Use the final model to generate predictions on the test set.

What is the final value of k used in the model?

21 ✔ **Answer:** 9 **or** 21

21

Explanation

The value of k can be determined using the following code:

```
set.seed(7)
# set.seed(7, sample.kind = "Rounding")    # simulate R 3.5
tuning <- data.frame(k = seq(3, 21, 2))
train_knn <- train(train_x, train_y,
      method = "knn",
      tuneGrid = tuning)
train_knn$bestTune
```

What is the accuracy of the kNN model on the test set?

0.948 ✔ **Answer:** 0.974 **or** 0.948

0.948

Explanation

The accuracy can be determined using the following code:

```
knn_preds <- predict(train_knn, test_x)
mean(knn_preds == test_y)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 15a: Random forest model

3.0/3.0 points (graded)

Set the seed to 9, then train a random forest model on the training set using the `caret` package. Test `mtry` values of 3, 5, 7 and 9. Use the argument `importance=TRUE` so that feature importance can be extracted. Generate predictions on the test set.

What value of `mtry` gives the highest accuracy?

3 ✔ **Answer:** 3

3

Explanation

The value can be found using the following code:

```
set.seed(9)
# set.seed(9, sample.kind = "Rounding")    # simulate R 3.5
tuning <- data.frame(mtry = c(3, 5, 7, 9))    # can expand to seq(3, 21, 2), same
train_rf <- train(train_x, train_y,
                  method = "rf",
                  tuneGrid = tuning,
                  importance = TRUE)
train_rf$bestTune
```

What is the accuracy of the random forest model on the test set?

0.974 ✔ **Answer:** 0.948 **or** 0.974

0.974

Explanation

The accuracy can be found using the following code:

```
rf_preds <- predict(train_rf, test_x)
mean(rf_preds == test_y)
```

What is the most important variable in the random forest model?

Be sure to enter the variable name exactly as it appears in the dataset.

area_worst ✔ **Answer:** area_worst

Explanation
The most important variable can be found using the following code:

```
varImp(train_rf)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 15b: Random forest model

1.0/1.0 point (graded)
Consider the top 10 most important variables in the random forest model.

Which set of features is most important for determining tumor type?

- mean values
- standard errors
- worst values ✔

Explanation
`varImp(train_rf)` gives the importance of the various variables. When looking at the top 10 most important features, 6 of the top 10 (including the top 4!) are worst values.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 16a: Creating an ensemble

1.0/1.0 point (graded)
Create an ensemble using the predictions from the 7 models created in the previous exercises: k-means, logistic regression, LDA, QDA, loess, k-nearest neighbors, and random forest. Use the ensemble to generate a majority prediction of the tumor type (if most models suggest the tumor is malignant, predict malignant).

What is the accuracy of the ensemble prediction?

0.983

✔ **Answer:** 0.965 **or** 0.983

0.983

Explanation
The ensemble accuracy can be calculated using the following code:

```
ensemble <- cbind(glm = glm_preds == "B", lda = lda_preds == "B", qda = qda_preds == "B", loess = loess_preds == "B", rf = rf_preds == "B", knn = knn_preds == "B", kmeans = kmeans_preds == "B")

ensemble_preds <- ifelse(rowMeans(ensemble) > 0.5, "B", "M")
mean(ensemble_preds == test_y)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

Question 16b: Creating an ensemble

1.0/1.0 point (graded)
Make a table of the accuracies of the 7 models and the accuracy of the ensemble model.

Which of these models has the highest accuracy?

- Logistic regression
- (●) LDA
- Loess
- Random forest
- Ensemble

✔

Explanation
The table can be generated using the following code:

```
models <- c("K means", "Logistic regression", "LDA", "QDA", "Loess", "K nearest
neighbors", "Random forest", "Ensemble")
accuracy <- c(mean(kmeans_preds == test_y),
              mean(glm_preds == test_y),
              mean(lda_preds == test_y),
              mean(qda_preds == test_y),
              mean(loess_preds == test_y),
              mean(knn_preds == test_y),
              mean(rf_preds == test_y),
              mean(ensemble_preds == test_y))
data.frame(Model = models, Accuracy = accuracy)
```

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem