Digit Recognizer

Digit Recognition of Handwritten Digits (Kaggle Competition)

The goal of this project is to correctly identify the digits from a dataset of thousands of handwritten images. I have used a Random Forest classifier to predict the handwritten digits. Accuracy achieved against the cross validation data is: 95.4%

The dataset used from MNIST (Modified National Institute of Standards and Technology) is a classic dataset for handwritten digits

```
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

Load the training and test dataset.

The train and test dataset contains the grayscale images of handwritten digits from 0 to 9.

The train dataset has 785 columns. One column is “label”, which specifies the digit drawn by the user. The rest of the 784 columns are pixel values of the associated image.

Each image is a 28 x 28 grayscale image with 784 pixels in total. Each pixel value is associated with it indicate the lightness or the darkness of the pixel - higher the pixel value is, darker will be the pixel.

The test dataset has 784 columns, which represent the 784 pixels of the handwritten image. There is no “label” column in the test dataset (Objective is to create a label column for the test dataset using the predicted values.)

```
# Load the train and test dataset
train.df <- read.csv(file="data/train.csv",header = T)
test.df <- read.csv(file="data/test.csv",header = T)

# "Label" in the train dataset need to be a factor
train.df$label <- as.factor(train.df$label)

# Train dataset is split in 80:20 ratio for cross-validation
div <- sample(1:nrow(train.df), 0.8*nrow(train.df))
train <- train.df[div,]
test <- train.df[-div,]
```

One of the classifiers used for this classification is Random Forest classifier. A random forest can be considered as an ensemble technique. This technique fits N number of decision trees on various sub-samples (with replacement) of the dataset and use voting majority (in case of categorical variables) to improve upon the predictive accuracy and control over-fitting.

```
set.seed(1111)
# Use the 80% of the training dataset to train the model.
# Have set ntree to 500, which is the number of trees to grow for this model
# Have set mtry to 5, which is the number of variables randomly sampled as candidates
at each split.
model.rf <- randomForest(label~., data=train, ntree=500, mtry=5)

# Use the model generated to predict the label of hand-written digits in the cross-va
lidation data.
pred.rf.cv <- predict(model.rf, test)

# Find the accuracy of the prediction on cross-validation data
mean(test$label == pred.rf.cv)
```

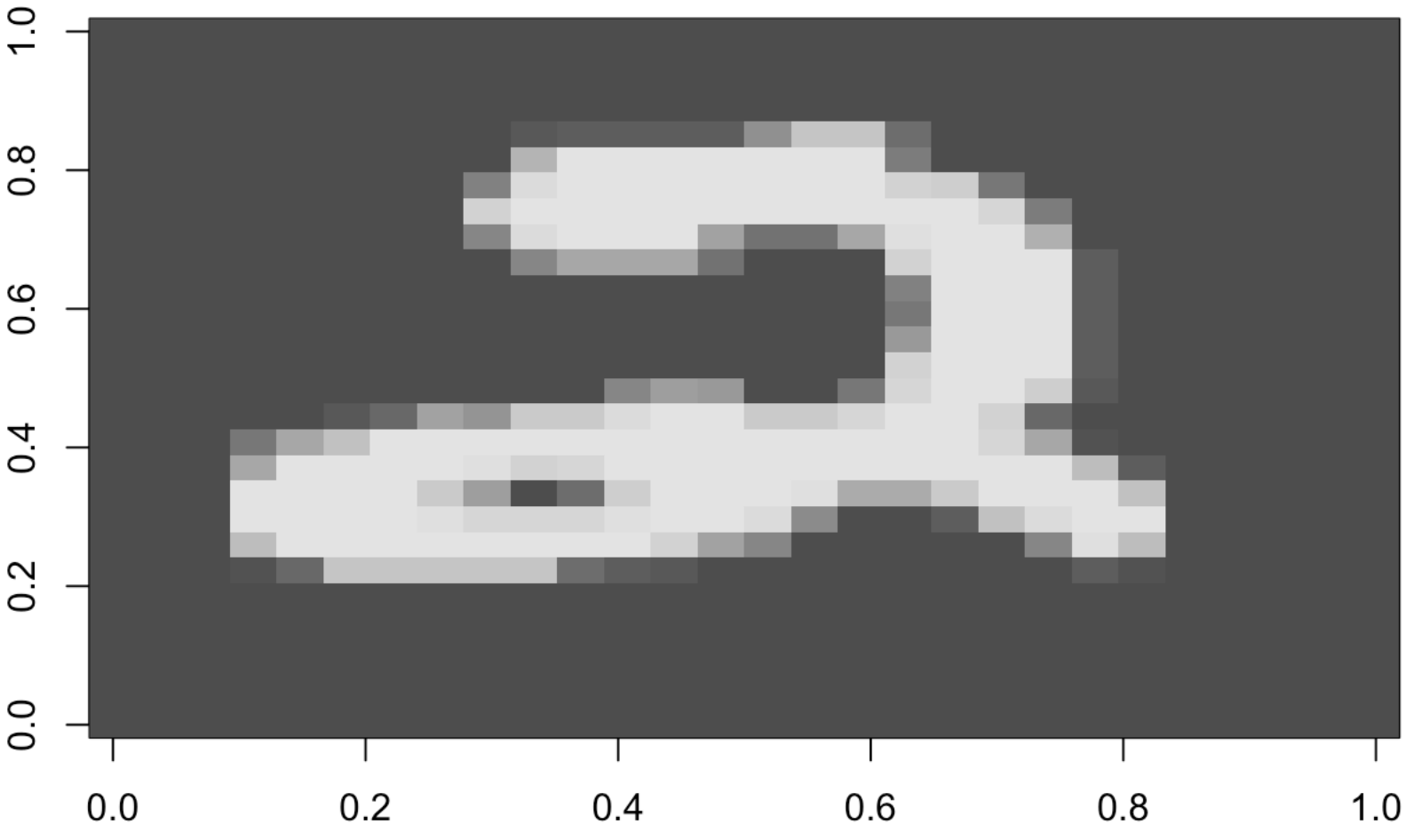
```
## [1] 0.9575
```

Accuracy of the random forest classifier against the cross validation data stands at 95.5%, which is pretty good. Now let’s test it against the test data.

```
# Use the model generated to predict the label of hand-written digits in the test dat
aset.
pred.test <- predict(model.rf, test.df)

# Append the labels predicted to the test dataset.
test.df$label <- pred.test

# Write the result into a csv file.
write.csv(test.df$label, "pred.csv")
```

Test data doesn't include a label field, and hence we will have to manually verify the correctness of the prediction. We can visualize the grayscale images on the test dataset and match them with the predicted values. I have tried to visualize 10 samples from the test dataset as below:

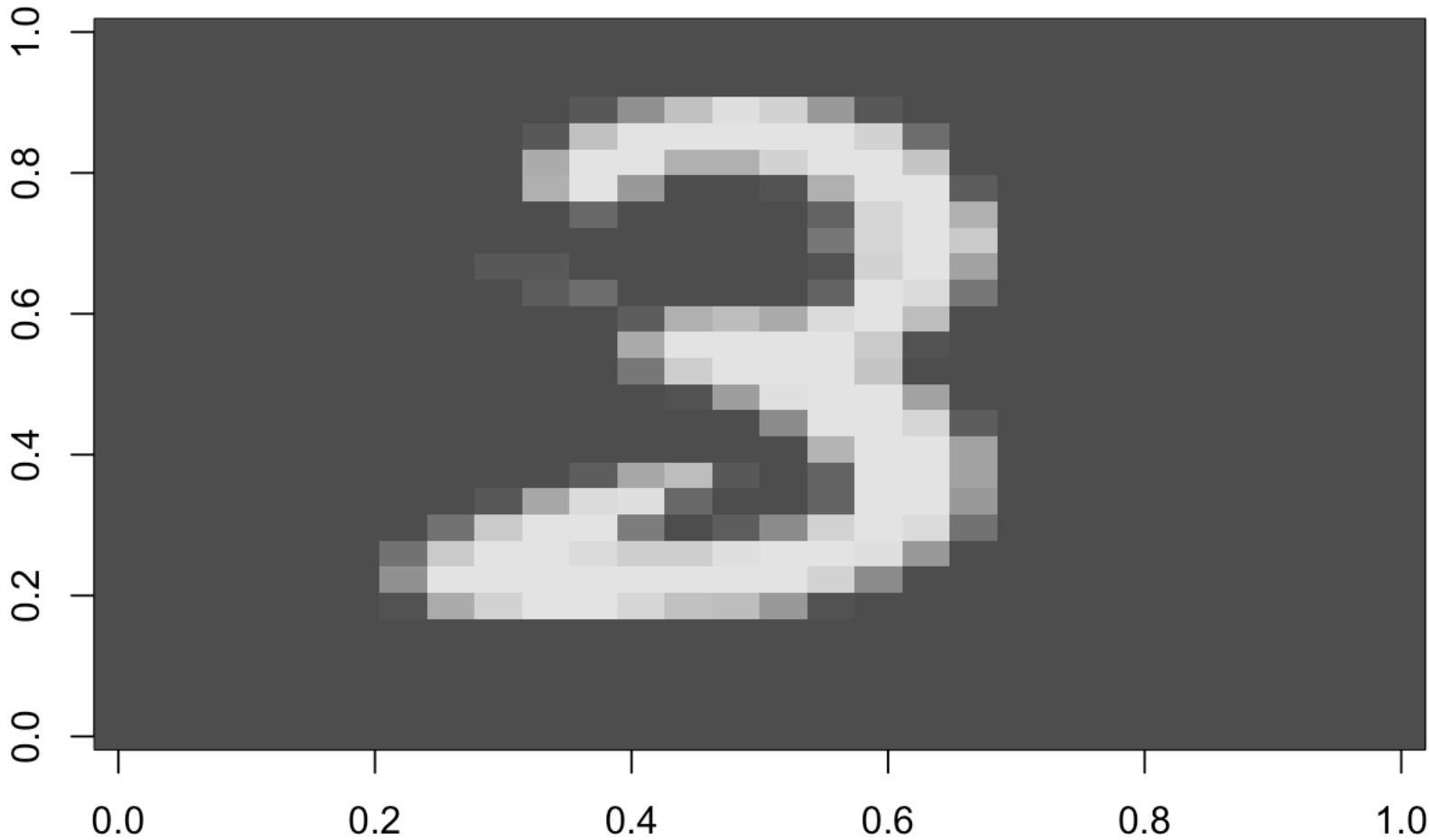
```
rotate <- function(x) t(apply(x, 2, rev))

# Image 1 in test dataset
m1 = rotate(matrix(unlist(test.df[1, -1]), nrow = 28, byrow = T))
image(m1, col = grey.colors(255))
```



```
# Prediction for image 1 using the random forest classifier
print(pred.test[1])
```

```
## 1
## 2
## Levels: 0 1 2 3 4 5 6 7 8 9
```

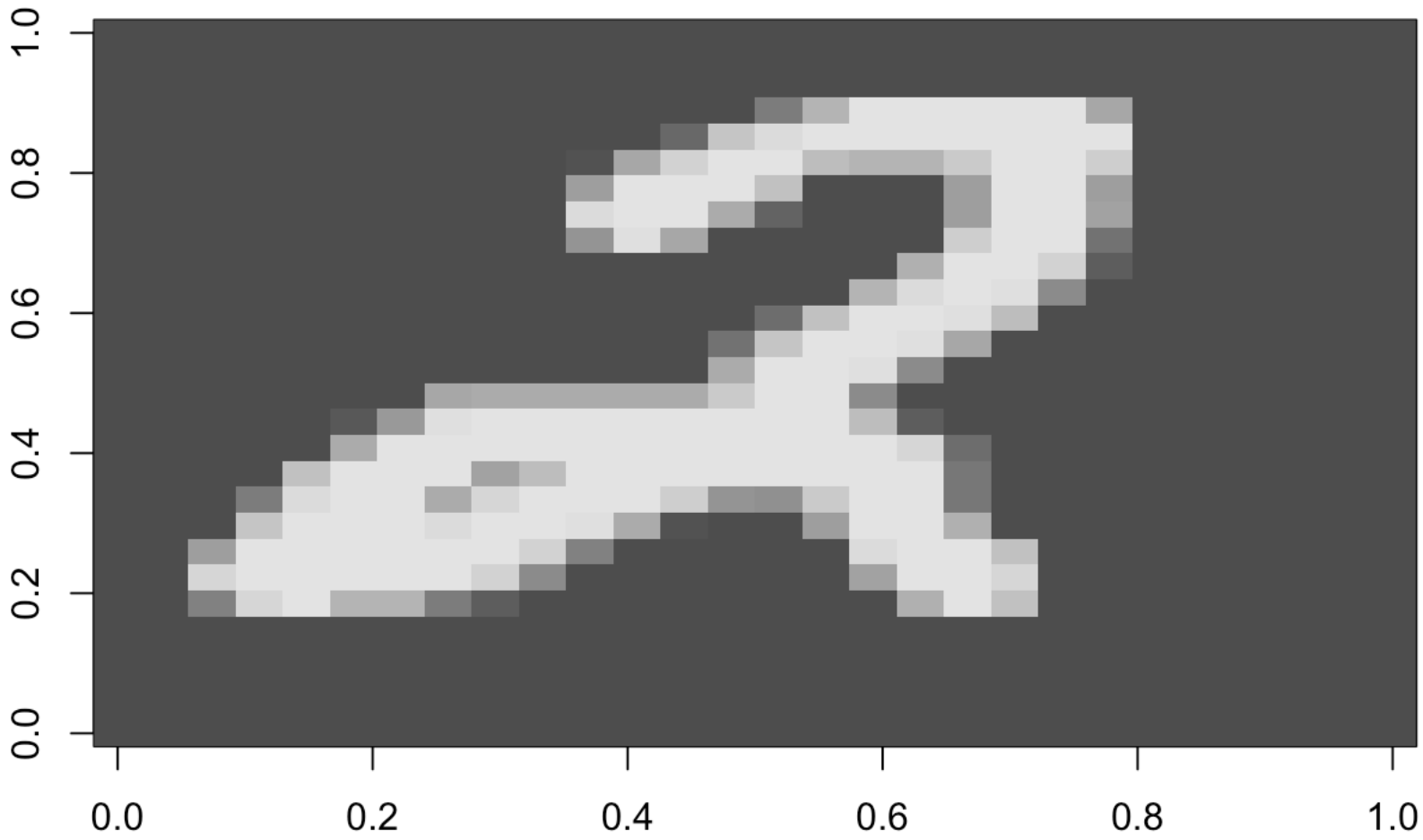
```
# Image 5 in test dataset
m5 = rotate(matrix(unlist(test.df[5, -1]), nrow = 28, byrow = T))
image(m5, col = grey.colors(255))
```




```
# Prediction for image 5 using the random forest classifier
print(pred.test[5])
```

```
## 5
## 3
## Levels: 0 1 2 3 4 5 6 7 8 9
```

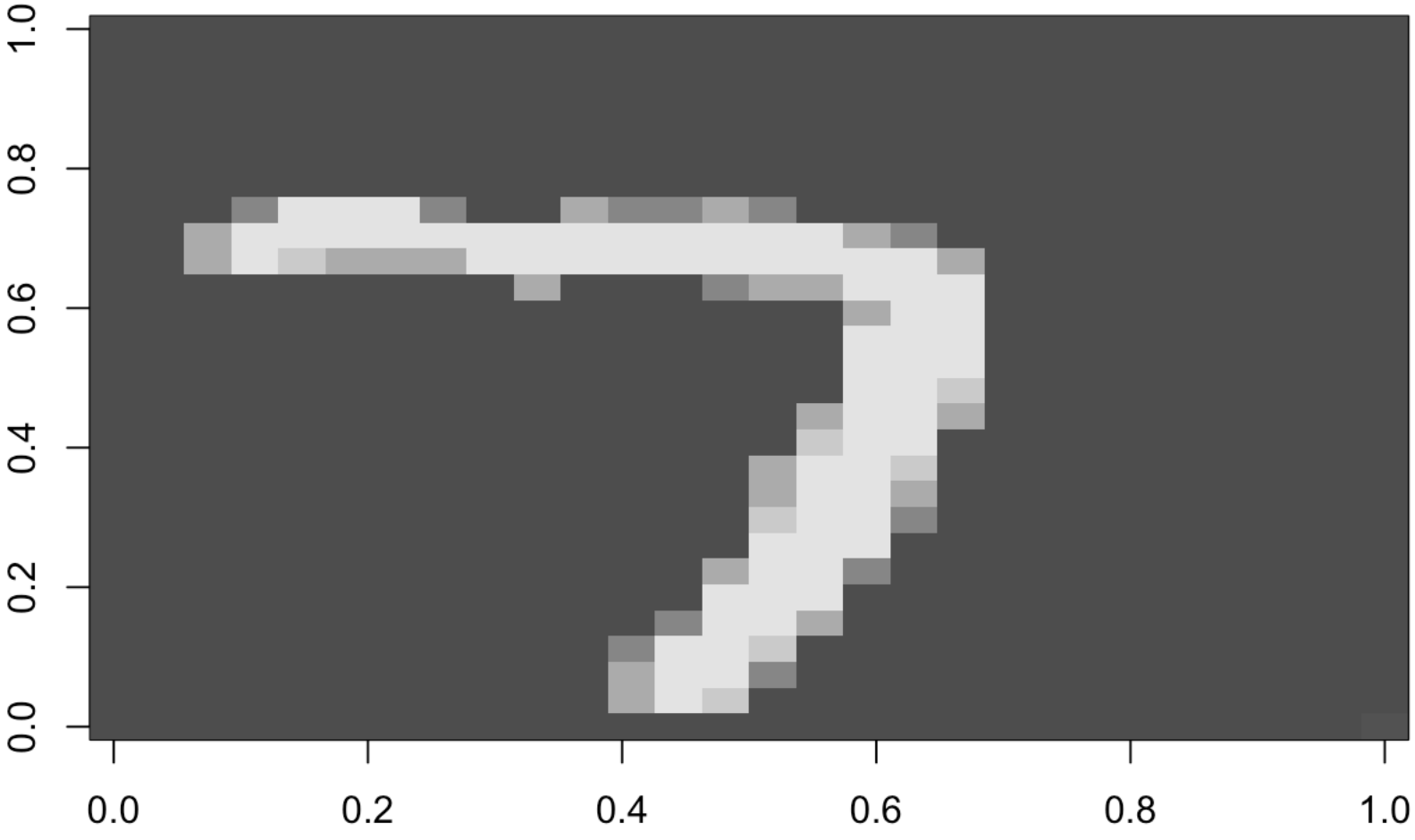
```
# Image 27 in test dataset
m27 = rotate(matrix(unlist(test.df[27, -1]), nrow = 28, byrow = T))
image(m27, col = grey.colors(255))
```




```
# Prediction for image 27 using the random forest classifier
print(pred.test[27])
```

```
## 27
##  2
## Levels: 0 1 2 3 4 5 6 7 8 9
```

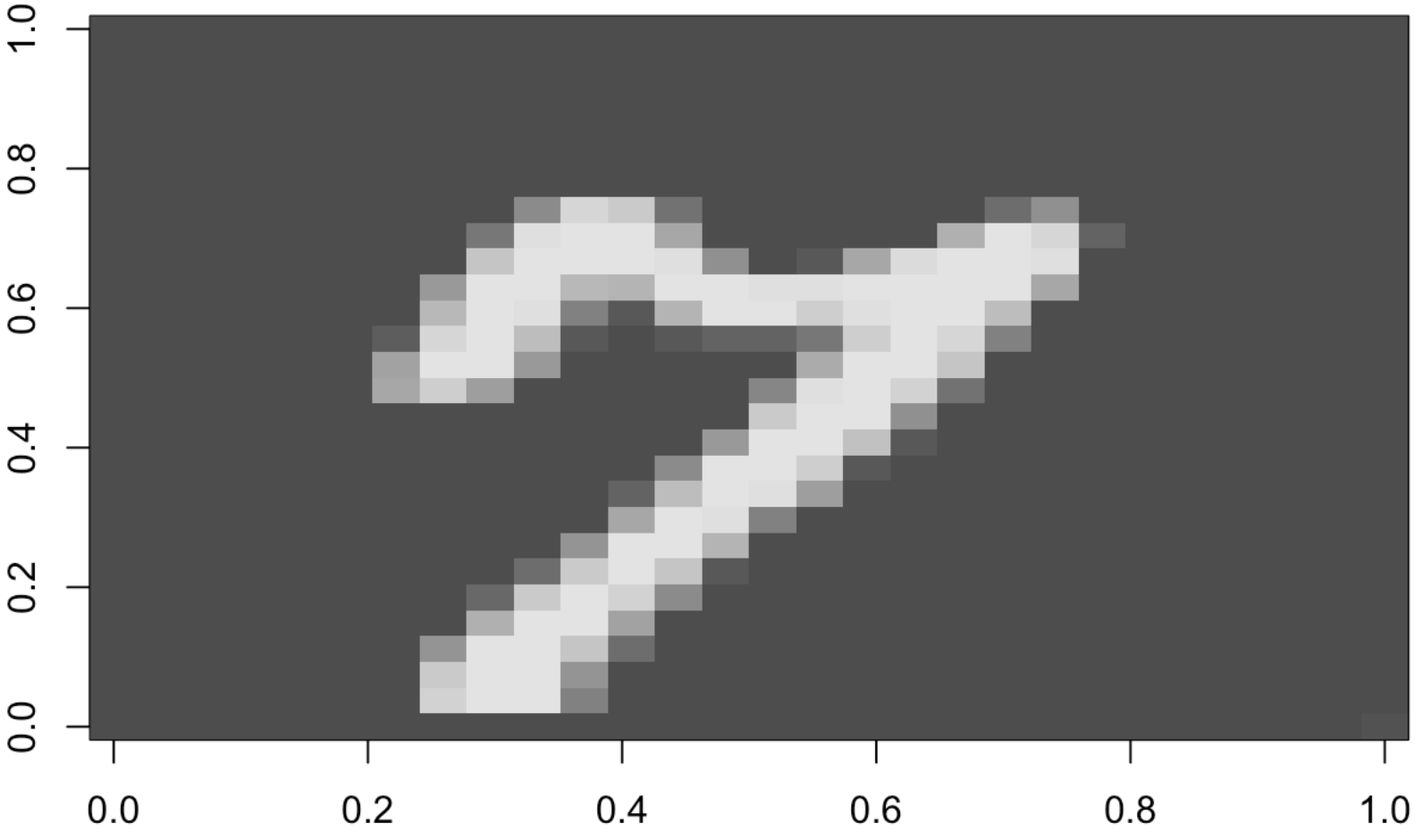
```
# Image 42 in test dataset
m42 = rotate(matrix(unlist(test.df[42, -1]), nrow = 28, byrow = T))
image(m42, col = grey.colors(255))
```




```
# Prediction for image 42 using the random forest classifier
print(pred.test[42])
```

```
## 42
##  7
## Levels: 0 1 2 3 4 5 6 7 8 9
```

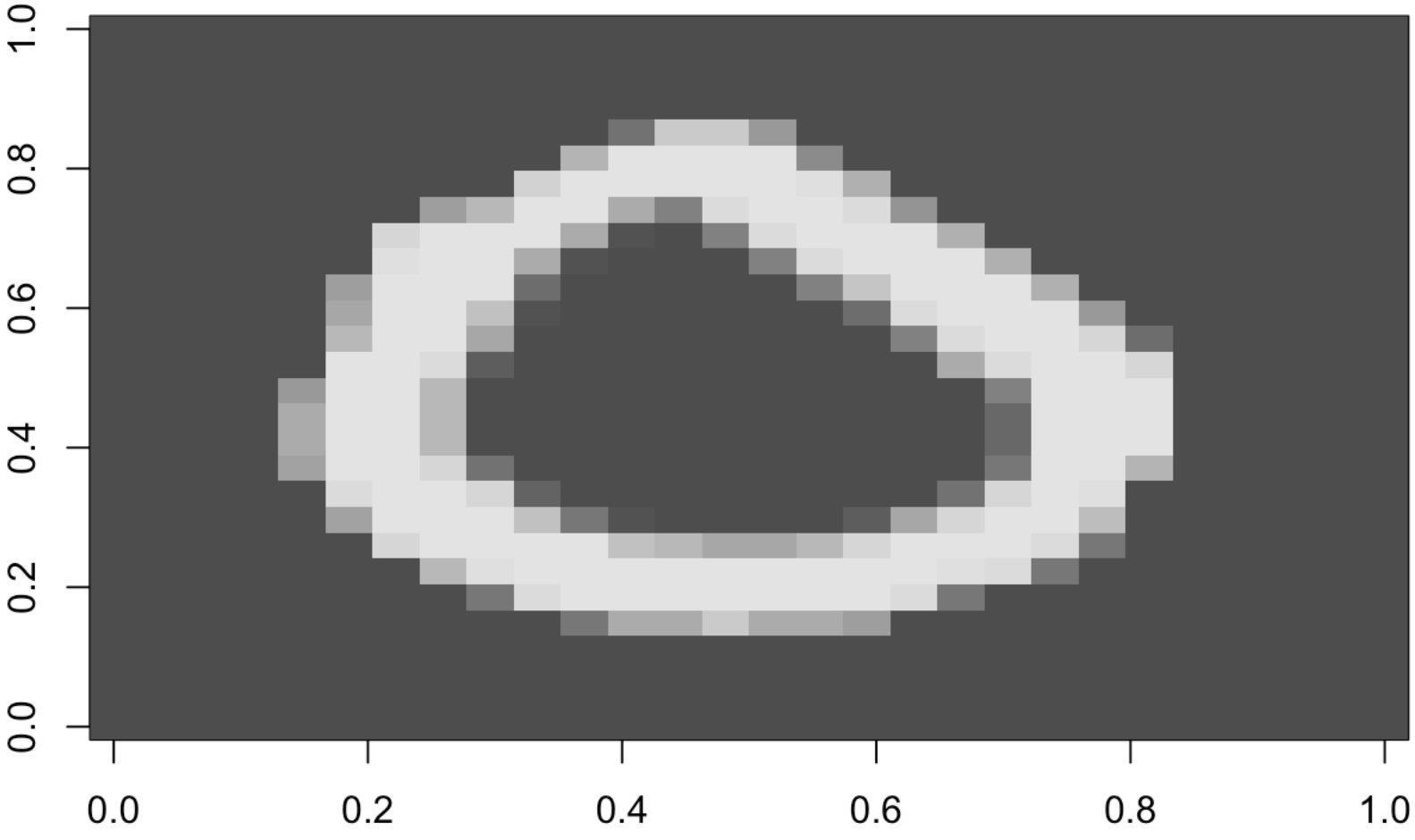
```
# Image 105 in test dataset
m105 = rotate(matrix(unlist(test.df[105, -1]), nrow = 28, byrow = T))
image(m105, col = grey.colors(255))
```




```
# Prediction for image 105 using the random forest classifier
print(pred.test[105])
```

```
## 105
##   7
## Levels: 0 1 2 3 4 5 6 7 8 9
```

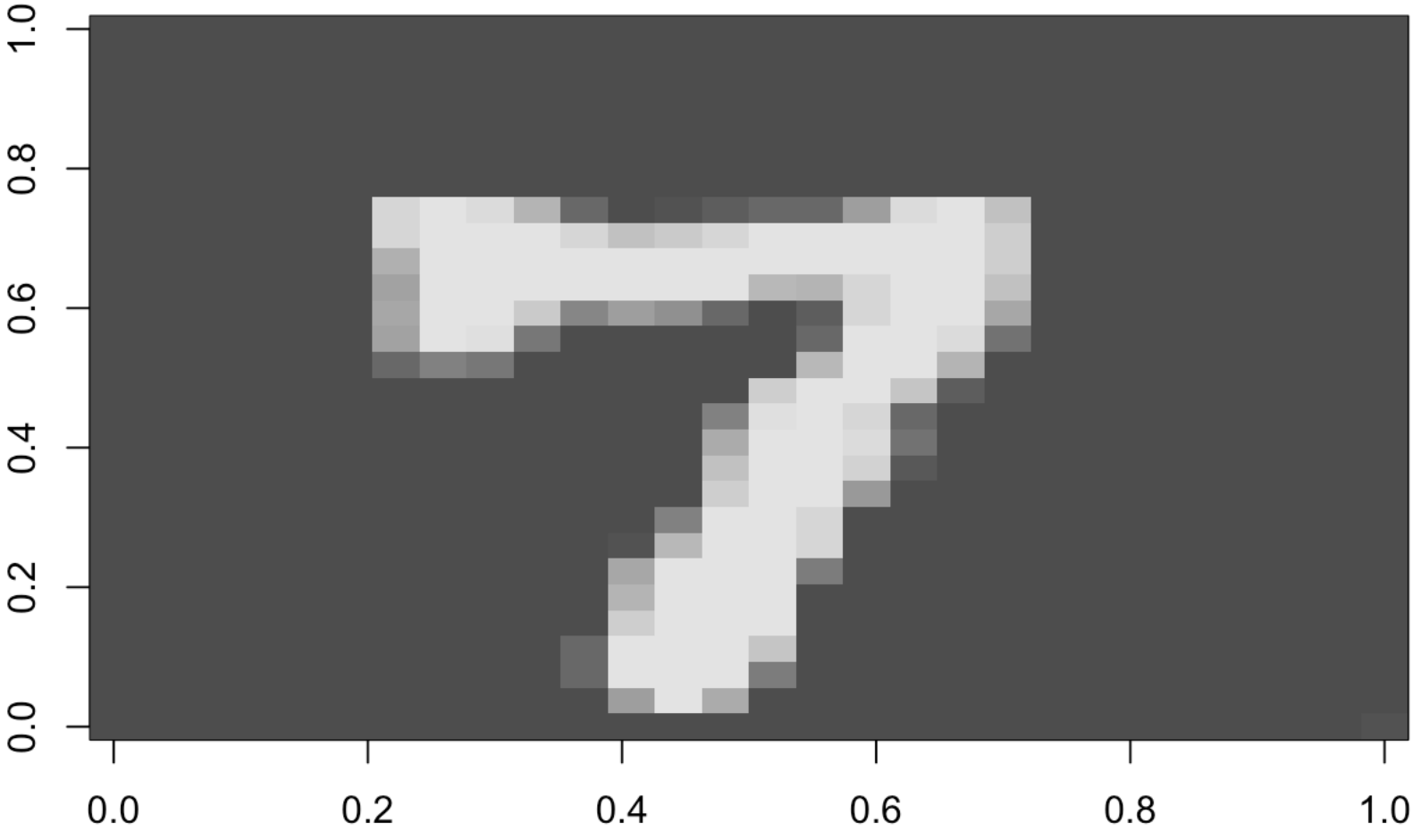
```
# Image 107 in test dataset
m107 = rotate(matrix(unlist(test.df[107, -1]), nrow = 28, byrow = T))
image(m107, col = grey.colors(255))
```




```
# Prediction for image 107 using the random forest classifier
print(pred.test[107])
```

```
## 107
##   0
## Levels: 0 1 2 3 4 5 6 7 8 9
```

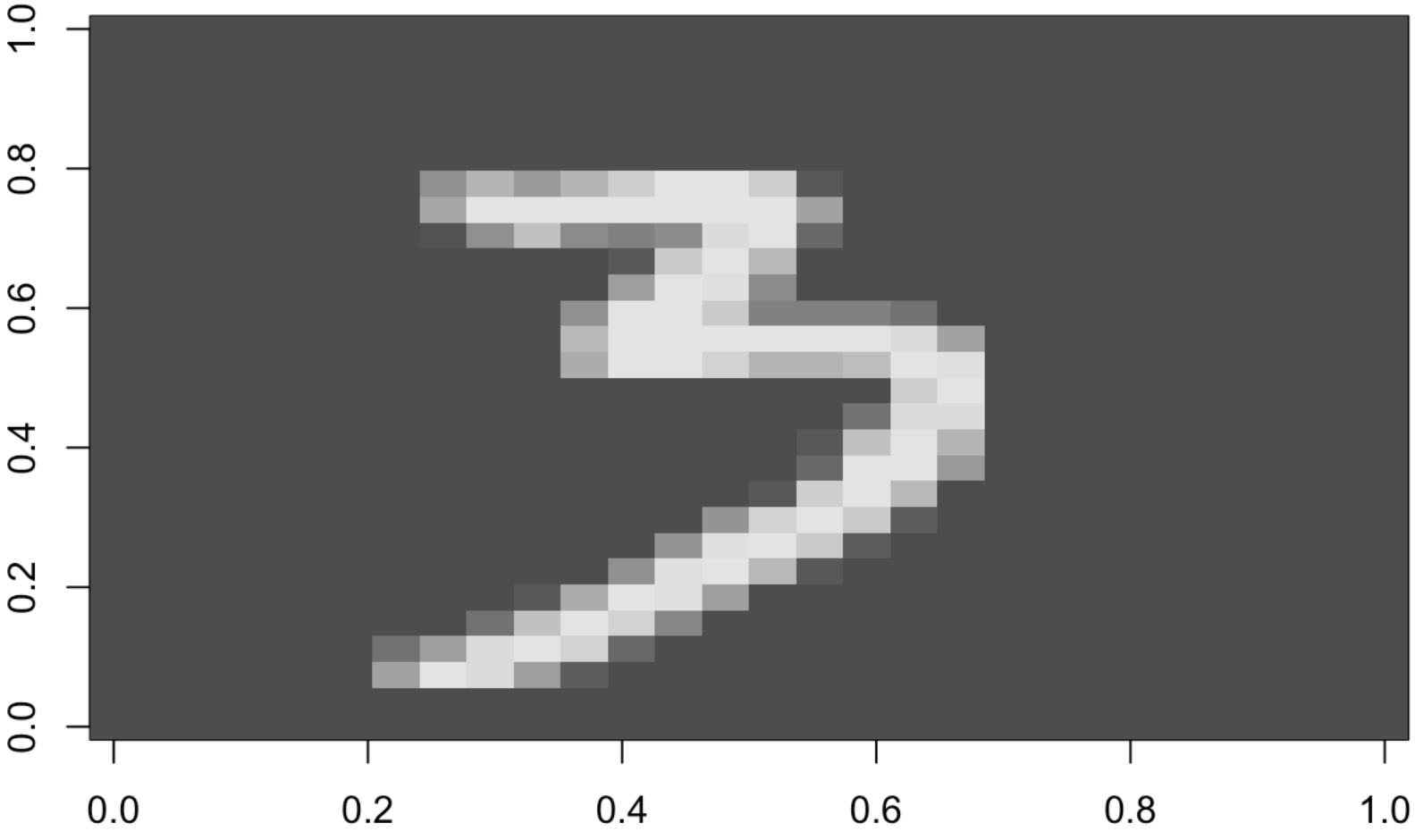
```
# Image 198 in test dataset
m198 = rotate(matrix(unlist(test.df[198, -1]), nrow = 28, byrow = T))
image(m198, col = grey.colors(255))
```




```
# Prediction for image 198 using the random forest classifier
print(pred.test[198])
```

```
## 198
##   7
## Levels: 0 1 2 3 4 5 6 7 8 9
```

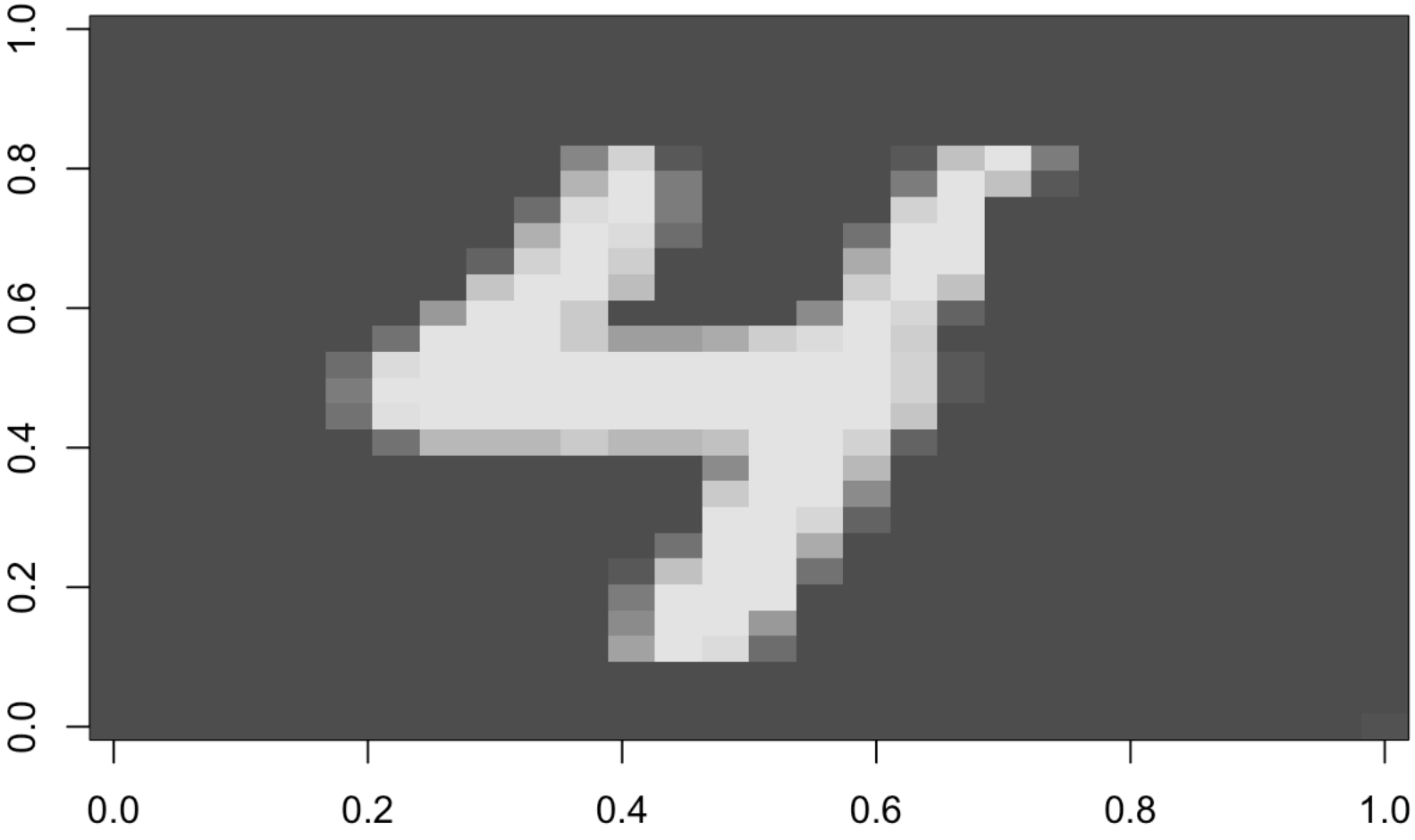
```
# Image 254 in test dataset
m254 = rotate(matrix(unlist(test.df[254, -1]), nrow = 28, byrow = T))
image(m254, col = grey.colors(255))
```




```
# Prediction for image 254 using the random forest classifier
print(pred.test[254])
```

```
## 254
##   3
## Levels: 0 1 2 3 4 5 6 7 8 9
```

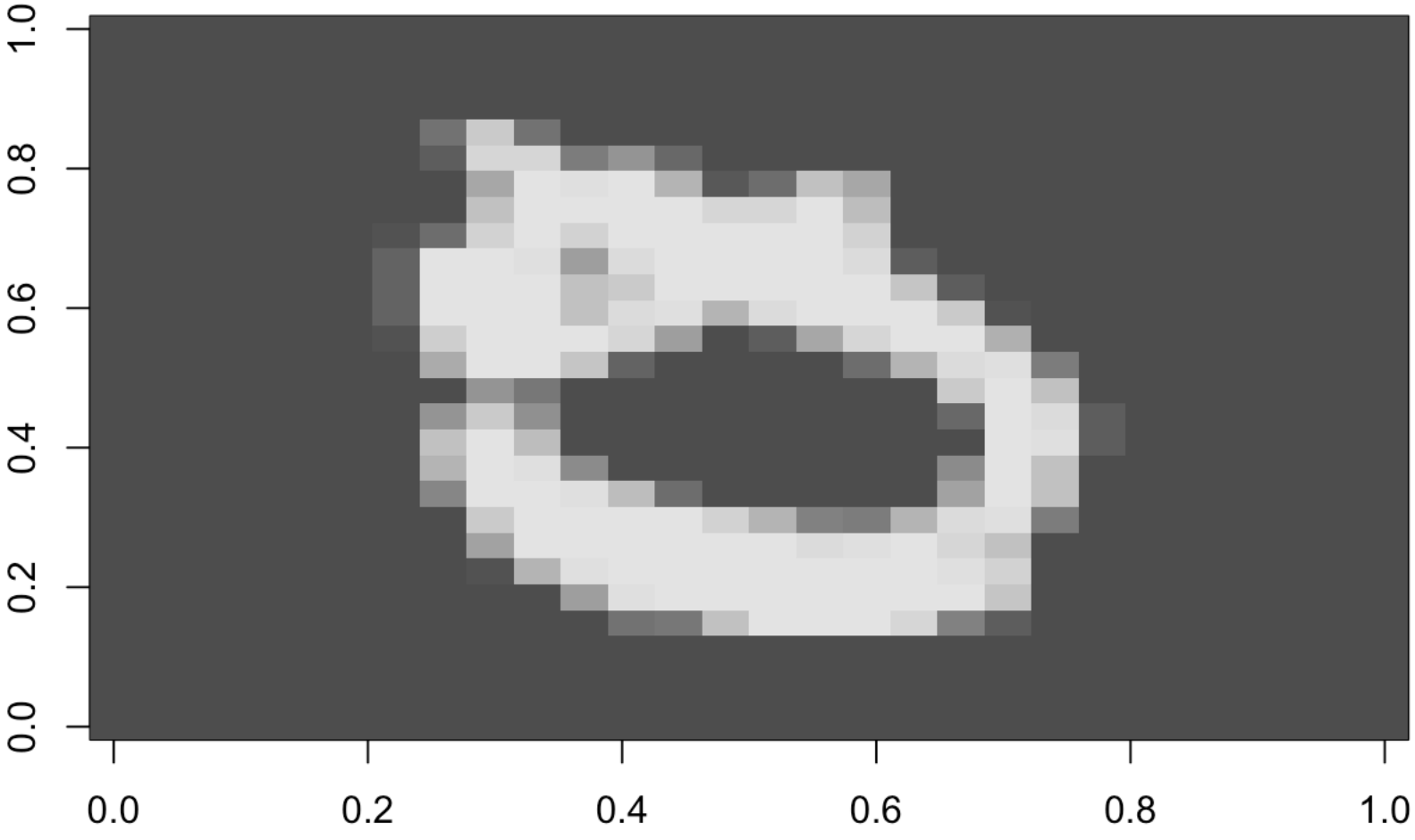
```
# Image 444 in test dataset
m444 = rotate(matrix(unlist(test.df[444, -1]), nrow = 28, byrow = T))
image(m444, col = grey.colors(255))
```




```
# Prediction for image 444 using the random forest classifier
print(pred.test[444])
```

```
## 444
##   4
## Levels: 0 1 2 3 4 5 6 7 8 9
```

```
# Image 555 in test dataset
m555 = rotate(matrix(unlist(test.df[555, -1]), nrow = 28, byrow = T))
image(m555, col = grey.colors(255))
```




```
# Prediction for image 555 using the random forest classifier
print(pred.test[555])
```

```
## 555
##   0
## Levels: 0 1 2 3 4 5 6 7 8 9
```

As we can see here, the first 9 predictions are correct. And the tenth one is difficult to determine. I feel its 9 (written in a very bad way :)), but the prediction is 0 (which maybe correct.). But this result reaffirms the accuracy of the model at around 90-95%.