

How can I perform one-hot encoding in R?

Authored by
stats writer

May 12, 2024

RECOMMENDED CITATION

stats writer (2024). *How can I perform one-hot encoding in R?*. PSYCHOLOGICAL SCALES.
Retrieved from <https://scales.arabpsychology.com/?p=143901>

One-hot encoding is a data preprocessing technique used in machine learning to convert categorical variables into numerical values. In R, this can be performed using the "dummy.data.frame" function from the "dummies" package. This function creates dummy variables for each level of a categorical variable, with a value of 1 indicating the presence of that level and 0 indicating its absence. The resulting dataset can then be used for further analysis or modeling. One-hot encoding is useful for improving the performance of models that cannot handle categorical data directly, such as linear regression or logistic regression.

Perform One-Hot Encoding in R

One-hot encoding is used to convert categorical variables into a format that can be used by .

The basic idea of one-hot encoding is to create new variables that take on values 0 and 1 to represent the original categorical values.

For example, the following image shows how we would perform one-hot encoding to convert a categorical variable that contains team names into new variables that contain only 0 and 1 values:

Original Data		One-Hot Encoded Data			
Team	Points	Team_A	Team_B	Team_C	Points
A	25	1	0	0	25
A	12	1	0	0	12
B	15	0	1	0	15
B	14	0	1	0	14
B	19	0	1	0	19
B	23	0	1	0	23
C	25	0	0	1	25
C	29	0	0	1	29

The following step-by-step example shows how to perform one-hot encoding for this exact dataset in R.

Step 1: Create the Data

First, let's create the following data frame in R:

```
#create data frame
```

```
df <- data.frame(team=c('A', 'A', 'B', 'B', 'B', 'B', 'C', 'C'),
  points=c(25, 12, 15, 14, 19, 23, 25, 29))
```

```
#view data frame
```

```
df
```

```
team points
```

```
1 A 25
```

```
2 A 12
```

3 B 15

4 B 14

5 B 19

6 B 23

7 C 25

8 C 29

Step 2: Perform One-Hot Encoding

Next, let's use the `dummyVars()` function from the `caret` package to perform one-hot encoding on the 'team' variable in the data frame:

```
library(caret)
```

```
#define one-hot encoding function
```

```
dummy <- dummyVars(" ~ .", data=df)
```

```
#perform one-hot encoding on data frame
```

```
final_df <- data.frame(predict(dummy, newdata=df))
```

```
#view final data frame
```

```
final_df
```

```
teamA teamB teamC points
```

```
1 1 0 0 25
```

2 1 0 0 12

3 0 1 0 15

4 0 1 0 14

5 0 1 0 19

6 0 1 0 23

7 0 0 1 25

8 0 0 1 29

Notice that three new columns were added to the data frame since the original 'team' column contained three unique values.

Also notice that the original 'team' column was dropped from the data frame since it's no longer needed.

The one-hot encoding is complete and we can now feed this dataset into any machine learning algorithm that we'd like.

Note: You can find the complete online documentation for the `dummyVars()` function .

The following tutorials offer additional information about working with categorical variables: