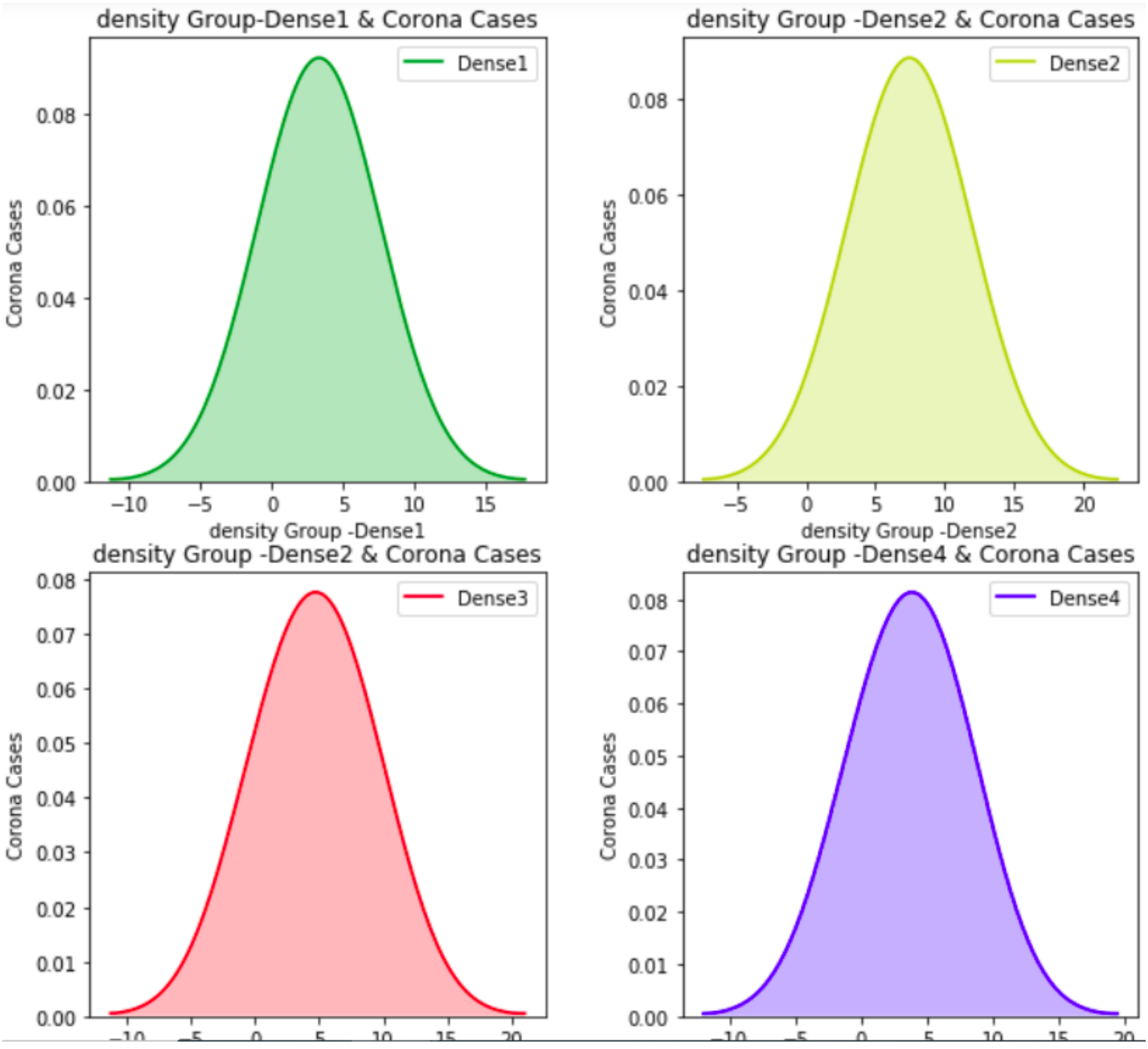
Analytics Vidhya

Understanding ANOVA in Machine Learning



What is ANOVA?

- Analysis of Variance (ANOVA) is a statistical method used to analyze differences among group means in a sample. It does this by examining the amount of variance within each group and comparing it to the amount of variance between the groups.

Why use ANOVA?

- **Feature Selection**
  ANOVA can be used in the feature selection process to determine which features are statistically significant when predicting a particular outcome.

- **Model Assumptions**
  Many machine learning algorithms assume homogeneity of variances. ANOVA can help validate this assumption.

Advantages

- **Comparison of Multiple Groups**

  Unlike the t-test which compares only two means, ANOVA can compare multiple group means.

- **Reduction of Type I Error**

  By analyzing multiple groups together instead of in pairs, ANOVA reduces the risk of a Type I error.

- **Versatility**

  ANOVA can be extended to more complex experimental designs (e.g., two-way ANOVA).

Disadvantages

- **Assumption Dependent**
  Assumes normal distribution and equal variances among the groups.

- **Sensitive to Outliers**
  Outliers can affect the sum of squares and lead to incorrect conclusions.

- **Doesn’t Identify Which Groups Differ**
  Post-hoc tests are needed to identify which groups are different if the ANOVA indicates significant differences

Implementation of ANOVA

```python
import pandas as pd
from scipy.stats import f_oneway

group1 = [84, 86, 87, 85, 88]
group2 = [91, 93, 92, 90, 94]
group3 = [78, 80, 81, 78, 79]

# Perform one-way ANOVA
f_stat, p_val = f_oneway(group1, group2, group3)

print("F-statistic:", f_stat)
print("P-value:", p_val)
```