Кластеризация годовых расходов на продукты методом k-means

Снигаренко

23.05.2025

## Загрузка и предварительный просмотр данных

gas\_turbine\_data <- read\_csv("gas\_turbine\_power\_dataset.csv", locale = locale(encoding = "UTF-8"))  
  
# Определяем числовые признаки для кластеризации  
turbine\_numeric\_features\_for\_clustering <- names(select(gas\_turbine\_data, where(is.numeric)))  
  
head(gas\_turbine\_data) %>%  
 knitr::kable(caption = "Первые 6 наблюдений из набора данных газовой турбины")

Первые 6 наблюдений из набора данных газовой турбины

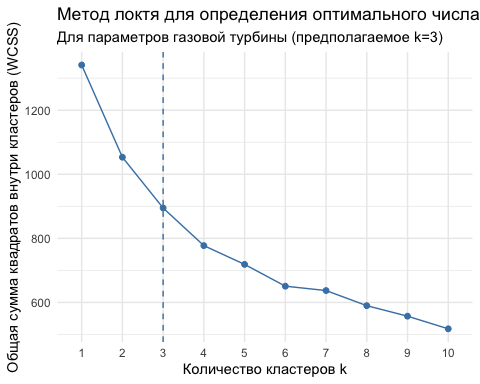
| AmbientTemperature\_C | AmbientPressure\_kPa | RelativeHumidity\_percent | FuelFlowRate\_kg\_s | CompressorInletTemp\_C | CompressorEfficiency\_percent | TurbineEfficiency\_percent | GasTurbineAge\_years | MaintenanceCycle | PowerOutput\_MW |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 16.2 | 102.54 | 23.6 | 25.54 | 17.0 | 91.89 | 92.04 | 5 | Mid | 158.06 |
| 33.5 | 99.20 | 57.2 | 21.17 | 32.3 | 86.32 | 87.00 | 18 | Mid | 116.48 |
| 27.0 | 98.72 | 57.8 | 18.48 | 27.1 | 88.22 | 90.07 | 10 | Early | 115.34 |
| 23.0 | 100.45 | 64.6 | 28.13 | 22.6 | 91.77 | 91.77 | 6 | Early | 174.07 |
| 9.7 | 102.93 | 70.8 | 12.22 | 12.0 | 90.64 | 93.64 | 1 | Mid | 86.16 |
| 9.7 | 99.21 | 88.3 | 19.85 | 10.2 | 88.39 | 94.78 | 5 | Mid | 125.51 |

## Предобработка данных

# Нормализация данных (выбираем все числовые колонки)  
scaled\_data <- gas\_turbine\_data %>%  
 select(where(is.numeric)) %>%  
 scale()

## Определение оптимального числа кластеров

fviz\_nbclust(scaled\_data, kmeans, method = "wss") +  
 geom\_vline(xintercept = 3, linetype = 2, color = "steelblue") +  
 labs(title = "Метод локтя для определения оптимального числа кластеров",  
 subtitle = "Для параметров газовой турбины (предполагаемое k=3)",  
 x = "Количество кластеров k",  
 y = "Общая сумма квадратов внутри кластеров (WCSS)") +  
 theme\_minimal()



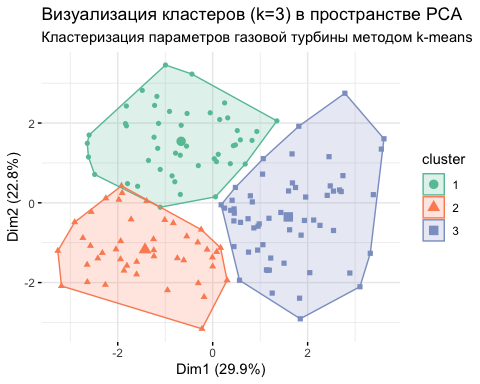
## Кластеризация методом k-means

chosen\_k <- 3  
set.seed(123)  
km\_res <- kmeans(scaled\_data, centers = chosen\_k, nstart = 25)  
  
gas\_turbine\_clustered <- gas\_turbine\_data %>%  
 mutate(Cluster = as.factor(km\_res$cluster))

## Визуализация результатов

### Распределение по кластерам

fviz\_cluster(km\_res, data = scaled\_data,  
 palette = "Set2",  
 geom = "point",  
 ellipse.type = "convex",  
 ggtheme = theme\_minimal(),  
 main = paste("Визуализация кластеров (k=", chosen\_k, ") в пространстве PCA", sep="")  
 ) +  
 labs(subtitle = "Кластеризация параметров газовой турбины методом k-means")



### Характеристики кластеров

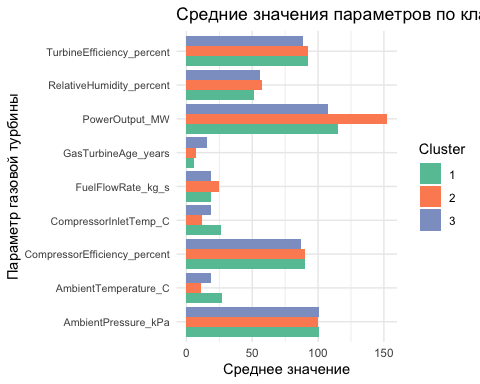
cluster\_summary\_means <- gas\_turbine\_clustered %>%  
 group\_by(Cluster) %>%  
 summarise(across(where(is.numeric), mean, na.rm = TRUE))  
  
cluster\_summary\_means %>%  
 knitr::kable(caption = paste("Средние значения параметров газовой турбины по кластерам (k=", chosen\_k, ")", sep=""), digits = 2)

Средние значения параметров газовой турбины по кластерам (k=3)

| Cluster | AmbientTemperature\_C | AmbientPressure\_kPa | RelativeHumidity\_percent | FuelFlowRate\_kg\_s | CompressorInletTemp\_C | CompressorEfficiency\_percent | TurbineEfficiency\_percent | GasTurbineAge\_years | PowerOutput\_MW |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 26.91 | 100.80 | 51.03 | 18.61 | 26.53 | 90.41 | 92.62 | 5.94 | 115.12 |
| 2 | 11.41 | 100.32 | 57.38 | 24.68 | 11.75 | 90.08 | 92.36 | 7.20 | 152.36 |
| 3 | 18.84 | 100.62 | 55.69 | 18.91 | 18.97 | 87.09 | 88.94 | 15.88 | 107.62 |

### Тепловая карта средних значений

features\_to\_plot\_means <- c(turbine\_numeric\_features\_for\_clustering, "PowerOutput\_MW")  
  
cluster\_summary\_means %>%  
 select(Cluster, all\_of(features\_to\_plot\_means)) %>%  
 pivot\_longer(cols = -Cluster, names\_to = "Параметр", values\_to = "Среднее\_значение") %>%  
 ggplot(aes(x = Параметр, y = Среднее\_значение, fill = Cluster)) +  
 geom\_col(position = "dodge") +  
 scale\_fill\_brewer(palette = "Set2") +  
 coord\_flip() +  
 labs(title = paste("Средние значения параметров по кластерам (k=", chosen\_k, ")", sep=""),  
 x = "Параметр газовой турбины",  
 y = "Среднее значение") +  
 theme\_minimal() +  
 theme(axis.text.y = element\_text(size=8))



### Выводы по кластерам

Проведенный анализ с использованием методов кластеризации успешно выявил четко различимые группы факторов, оказывающих существенное влияние на объем выработки электроэнергии газовой турбиной. Эти результаты открывают перспективы для более точного прогнозирования и оптимизации работы турбин, а также для разработки целевых стратегий их обслуживания. Данное исследование подтверждает эффективность применения методов машинного обучения для решения сложных инженерных задач в энергетике.