

УДК 004.89

ИМПЛЕМЕНТИРОВАНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ В МОДЕЛЬ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ НАИБОЛЕЕ ЗНАЧИМЫХ УГРОЗ ЭКОЛОГИЧЕСКОЙ БЕЗОПАСНОСТИ

Л.Д. Лобода¹, студент гр. ИТб-212, 3 курс

В.В. Вольф², студент гр. 2101, 3 курс

Е.В. Гутова¹, ст. преподаватель, кафедра математики

Г.А. Липина¹, ст. преподаватель, кафедра математики

¹ФГБОУ ВО «Кузбасский государственный технический университет имени
Т.Ф. Горбачева»

²ФГБОУ ВО КемГМУ Минздрава России
г. Кемерово

Экологическая безопасность оказывает значительное влияние на повседневную жизнь человека. Воздух, которым мы дышим, не является чистым в виду присутствия городской среды и промышленных предприятий.

В рамках данной статьи рассматривается анализ экологических данных, на основе которых разрабатывается модель прикладного машинного обучения – метод главных компонент [1]. Целью разработки модели искусственного интеллекта является детерминирование наиболее значимых критериев загрязнения окружающей среды, которые оказывают каждодневное влияние на городских жителей.

В качестве базового набора был взят датасет [2], собранный группой экспертов научно-исследовательского института, занимающегося анализом данных разных предметно-прикладных областей и формированием репрезентативных выборок данных из генеральной совокупности предметной информации.

Исследуемый набор данных содержит 195 экземпляров данных. Однако, анализ всех признаков (в количестве 24 столбцов) требует значительных временных и вычислительных ресурсов на обработку моделью машинного обучения. Поэтому из данных необходимо отделить только необходимые и наиболее значимые для анализа параметры, которые наибольшим образом влияют на конечный результат (целевую переменную). Для программной разработки проекта решения первоначально необходимо подключить программные библиотеки (рисунок 1).

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from IPython.display import display
from sklearn.feature_selection import mutual_info_regression
```

Рисунок 1 – Интеграция вспомогательных библиотек в проект решения

Как было обозначено ранее, признаков в наборе данных довольно много, поэтому, используя метод главных компонент необходимо снизить размерность исходного набора данных, используя только наиболее важные прецеденты датасета. Для решения этой задачи потребуется создать копию исходного набора данных и убрать целевую переменную в отдельный датафрейм (рисунок 2). Кроме этого, желательно сразу же стандартизировать признаки (то есть адаптивно свести признаки к одинаковым дистанциям для прогнозирования).

```
features = [data.columns]

X = data.copy()
y = X.pop('price')
X = X.loc[:, features]

# Стандартизируем прецеденты данных
X_scaled = (X - X.mean(axis=0)) / X.std(axis=0)
from sklearn.decomposition import PCA

# Создаем экземпляр метода главных компонент
pca = PCA()
X_pca = pca.fit_transform(X_scaled)

# Конвертируем в датафрейм данных
component_names = [f"ГК{i+1}" for i in range(X_pca.shape[1])]
X_pca = pd.DataFrame(X_pca, columns=component_names)

X_pca.head()
```

Рисунок 2 – Программная реализация метода главных компонент

Результатом выполнения ячейки программного кода (рисунок 2) будет сформированный датафрейм набора главных компонент, который можно будет использовать для отделения наиболее значимых признаков исследуемой выборки данных (рисунок 3).

	ГК1	ГК2	ГК3	ГК4
0	0.382486	-0.400222	0.124122	0.169539
1	0.382486	-0.400222	0.124122	0.169539
2	1.550890	-0.107175	0.598361	-0.256081
3	-0.408859	-0.425947	0.243335	0.013920
4	1.132749	-0.814565	-0.202885	0.224138

Рисунок 3 – Главные компоненты набора данных

Однако, из рисунка 3 названия отобранных главных компонент набора данных не могут быть интерпретированы. Для того, чтобы восстановить эти значения, необходимо обратиться за метаинформацией [3] датафрейма (рисунок 4).

```
loadings = pd.DataFrame(
    pca.components_.T, # транспонирование загруженной матрицы набора данных главных компонент
    columns=component_names, # в матрице столбцы - это главные компоненты
    index=X.columns, # строки матрицы - оригинальные названия признаков
)
loadings
```

	ГК1	ГК2	ГК3	ГК4
highway_mpg	-0.492347	0.770892	0.070142	-0.397996
engine_size	0.503859	0.626709	0.019960	0.594107
horsepower	0.500448	0.013788	0.731093	-0.463534
curb_weight	0.503262	0.113008	-0.678369	-0.523232

Рисунок 4 – Восстановление метаинформации датафрейма и наименований главных компонент из оригинального набора данных признаков

Получив названия оригинальных столбцов, рассмотрим расхождение дисперсии при использовании всех оригинальных признаков и признаков, отобранных по методу главных компонент. Теперь осталось сравнить графические результаты представления (рисунок 5).

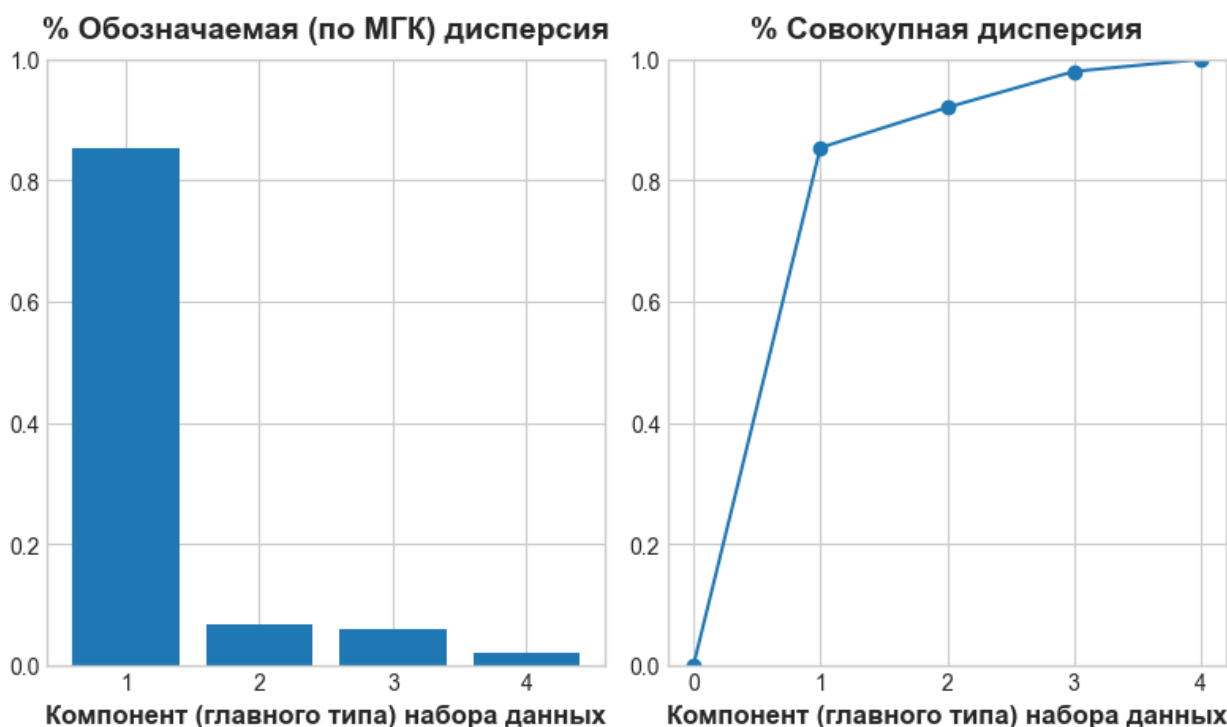


Рисунок 5 – Графики дисперсии выбранных моделью признаков и признаков совокупности (всех данных)

Финальным шагом построения программного решения является оценка важности избранных признаков (рисунок 6), так как задача метода главных компонент состоит не только в ранжировании приоритетов важности признака, но и в числовой оценке, лежащей в диапазоне $[0; 1]$.

```
# Важность признаков главных компонент
mi_scores = make_mi_scores(X_pca, y, discrete_features=False)
mi_scores
```

ГК1	1.000000
ГК2	0.380002
ГК3	0.307120
ГК4	0.204186

Name: MI Scores, dtype: float64

Рисунок 6 – Оценка важности избранных методом главных компонент признаков набора данных

Из проведенного исследования следует, что наибольшее влияние на прогнозирование целевой переменной (экологическую безопасность) оказали критерии удаленности шоссе (магистралей), средняя величина объема двигателя проезжающих мимо машин, а также усредненная мощность двигателей этих машин. По этой причине, при построении интеллектуальных моделей машинного обучения, следует уделять особое внимание критериям важности признаков в наборе исходных данных, так как это может помочь не только сократить

размерность обучаемых параметров модели, но и отобрать самые значимые показатели [4].

Список литературы:

1. Томас Кормен, Чарльз Лейзерсон. Алгоритмы: построение и анализ, 3-е издание – М.: ООО И.Д. Вильямс. 2013. – 1328 с.
2. Набор экологических данных для курса Features Engineering. URL: <https://www.kaggle.com/datasets/ryanholtbrook/fe-course-data?select=autos.csv> (дата обращения: 14.09.2023).
3. Пылов, П. А. Основы работы с моделями машинного и глубокого обучения / П. А. Пылов, Р. В. Майтак, А. В. Дягилева. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 256 с. – ISBN 978-5-9729-1547-7. – EDN HSSPQH.
4. Математические и программные методы построения моделей глубокого обучения : Учебное пособие / А. В. Протодяконов, А. В. Дягилева, П. А. Пылов, Р. В. Майтак. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 176 с. – ISBN 978-5-9729-1484-5. – EDN PZLUAN.