



АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ДАННЫХ ПО УРОВНЮ ТРУДОУСТРОЙСТВА ВЫПУСКНИКОВ ОБРАЗОВАТЕЛЬНЫХ УЧРЕЖДЕНИЙ

^{1,2}Зеленина Л. И. ORCID ID 0000-0002-0155-3139,

³Соловей М. В. ORCID ID 0000-0001-5784-4077

¹Федеральное государственное автономное образовательное учреждение высшего образования
«Северный (Арктический) федеральный университет имени М. В. Ломоносова»,
Архангельск, Российская Федерация, e-mail: l.zelenina@narfu.ru;

²Федеральное государственное автономное образовательное учреждение высшего образования
«Балтийский федеральный университет имени И. Канта», Калининград, Российская Федерация;

³Федеральное государственное бюджетное образовательное учреждение высшего образования
«Калининградский государственный технический университет», Калининград, Российская Федерация

В настоящее время искусственный интеллект, являясь важным фактором развития современной экономики, широко применяется в анализе деятельности различных организаций в целях как повышения эффективности их работы, так и поддержки принятия управленческих решений. Машинное обучение, как область искусственного интеллекта, включает в себя методы и алгоритмы, предназначенные для проведения интеллектуального анализа данных из различных прикладных областей. В данной статье рассматривается применение моделей машинного обучения в исследовании данных о выпускниках средних специальных и высших учебных заведений по различным регионам Российской Федерации, различным направлениям и уровням подготовки обучающихся. Для проведения анализа использовались открытые данные о выпускниках более 5000 высших и средних специальных российских учебных заведений, сформированные из информационных систем Социального фонда и Рособнадзора. Представленное исследование отражает основные этапы проведения разведывательного анализа и предварительной обработки данных с целью решения задач предсказания и кластеризации. Группы построенных моделей, по каждой из которых определены соответствующие метрики качества, позволяют прогнозировать значения доли трудоустроенных выпускников учебных заведений и показателя уровня трудоустройства по учебным заведениям.

Ключевые слова: искусственный интеллект, машинное обучение, разведывательный анализ, классификация, регрессия, кластеризация, трудоустройство выпускников

Благодарности: Слова благодарности авторы адресуют обучающейся БФУ имени Иммануила Канта А. Д. Шумеевой, принявшей активное участие в проведенном исследовании.

MACHINE LEARNING ALGORITHMS IN DATA ANALYSIS ON THE EMPLOYMENT STATUS OF EDUCATIONAL INSTITUTION GRADUATES

^{1,2}Zelenina L. I. ORCID ID 0000-0002-0155-3139,

³Solovey M. V. ORCID ID 0000-0001-5784-4077

¹Federal State Autonomous Educational Institution of Higher Education
“Northern (Arctic) Federal University named after M. V. Lomonosov”,
Arkhangelsk, Russian Federation, e-mail: l.zelenina@narfu.ru;

²Federal State Autonomous Educational Institution of Higher Education
“Baltic Federal University named after Immanuel Kant”, Kaliningrad, Russian Federation;

³Federal State Budgetary Educational Institution of Higher Education
“Kaliningrad State Technical University”, Kaliningrad, Russian Federation

Currently, artificial intelligence, being an important factor in the development of the modern economy, is widely used in the analysis of the activities of various organizations, both in order to improve their efficiency and to support management decision-making. Machine learning, as a field of artificial intelligence, includes methods and algorithms designed to perform data mining in various applied fields. This article examines the application of machine learning models in the study of data on graduates of secondary specialized and higher educational institutions in various regions of the Russian Federation, as well as in various fields and levels of training. The analysis used data on graduates from more than 5,000 Russian higher and secondary specialized educational institutions, which were obtained from the Social Fund and Rosobnadzor information systems. The presented study reflects the main stages of conducting exploratory analysis and data preprocessing in order to solve prediction and clustering tasks. The groups of built models, for each of which the corresponding quality metrics are defined, allow to predict the values of the share of employed graduates of educational institutions and the employment rate indicator by educational institutions.

Keywords: artificial intelligence, machine learning, intelligence analysis, classification, regression, clustering, and graduate employment

Acknowledgements: The authors would like to express their gratitude to A. D. Shumeeva, a student at Immanuel Kant Baltic Federal University, who actively participated in the research.

Введение

Использование технологий искусственного интеллекта способствует развитию современной цифровой экономики, трансформации бизнес-процессов, оптимизации производственных операций, повышению эффективности управления [1–3]. Алгоритмы и методы искусственного интеллекта позволяют автоматизировать рутинные операции, сокращая временные затраты и повышая точность выполняемых действий, оптимизировать производственные процессы, осуществляя контроль качества выпускаемой продукции и способствуя оптимизации планирования производства, обрабатывать большие объемы данных, выявляя в них закономерности и определяя новые знания [4–6].

Методы и модели машинного обучения, анализируя имеющиеся данные, выявляя в них закономерности, получают новые данные, необходимые для решения актуальных задач [7–9].

Цель исследования – проведение интеллектуального анализа данных, содержащих информацию об уровне трудоустройства и заработной платы выпускников учебных заведений высшего и среднего специального образования по направлениям подготовки и уровням образования. Представленное исследование, проведенное на основе методов и алгоритмов машинного обучения, позволяет определить наиболее успешные и востребованные на рынке труда направления подготовки и учебные заведения Российской Федерации, а также определить прогнозные значения доли трудоустроенных выпускников учебных заведений и показателя уровня трудоустройства обучающихся по учебным заведениям.

Материалы и методы исследования

Алгоритмы и методы машинного обучения являются широко используемым инструментом для проведения анализа данных различных прикладных областей [10, 11]. Искусственный интеллект, включающий в себя в том числе и модели машинного обучения, предоставляет возможности по проведению разведывательного анализа и предобработки данных, эффективного решения задач интеллектуального анализа данных различного уровня сложности [12, 13].

Для знакомства с исходным датасетом и построения моделей машинного обучения использовались библиотеки языка программирования Python по управлению, анализу и визуализации данных, а также библиотеки, предоставляющие алгоритмы решения задач классификации, регрессии, кластеризации, уменьшения размерности [14, 15].

Описание используемого набора данных

Для проведения анализа использовались данные о выпускниках более 5000 высших и средних специальных российских учебных заведений, сформированные из информационных систем Социального фонда и Рособнадзора. Это выборочная совокупность о выпускниках, завершивших образование в вузе/ссузе, сгруппированная по направлениям подготовки, уровням образования, году выпуска и половой принадлежности. Используемый датасет включал в себя 1 027 682 записей и 19 параметров [16]. Данные по величине заработной платы и по трудоустройству рассматривались по состоянию на 2024 г. (дата обращения: 07.04.2026).

Описание параметров используемого датасета:

object_level: str: уровень территориальной единицы (регион)
object_name: str: название территориальной единицы (региона)
okato: int64: код ОКАТО региона
oktmo: int64: код ОКТМО региона
gender: str: пол выпускника (мужской, женский, всего)
education_level: str: уровень образования
year: int64: год выпуска
university: str: название образовательной организации
specialty_section: str: укрупненная группа направлений подготовки
specialty: str: направление подготовки
specialty_code: str: код направления подготовки
count_graduate: int64: количество выпускников
percent_employed: float64: доля трудоустроенных выпускников
average_salary_fact_avg: float64: средняя фактическая заработная плата в 2024 г.
average_salary_fact_med: float64: медианная фактическая заработная плата в 2024 г.
average_salary_norm_avg: float64: средняя зарплата, нормированная на уровень цен региона трудоустройства
average_salary_norm_med: float64: медианная зарплата, нормированная на уровень цен региона трудоустройства
average_age: float64: средний возраст выпускника
percent_combine_work_and_study: float64: доля выпускников, совмещавших учебу и работу.

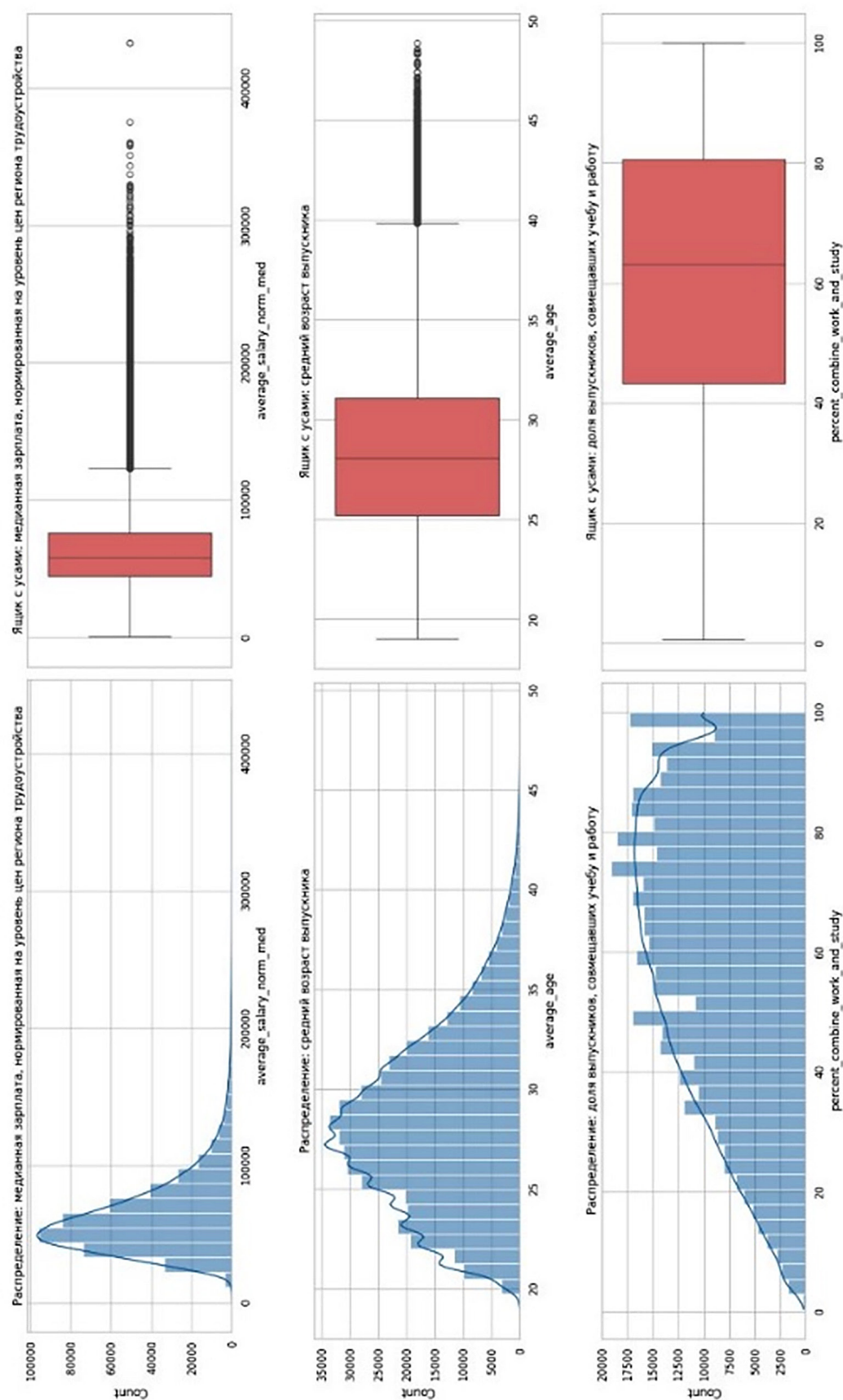


Рис. 1. Анализ распределений некоторых числовых признаков
Примечание: составлен авторами по результатам данного исследования

Результаты исследования и их обсуждение

Для знакомства со структурой используемых данных, выявления скрытых закономерностей в них и подготовки данных к последующему этапу моделирования был использован метод EDA. Итогами некоторых этапов данного процесса стал анализ распределения переменных и обработка выбросов и аномалий, представленный на рис. 1.

Интерпретация полученных результатов описательной статистики и визуализация распределений подтвердила сильную правостороннюю асимметрию (длинный хвост вправо) у таких признаков, как количество выпускников, доля трудоустроенных вы-

пускников, средняя и медианная фактическая заработная плата, средняя и медианная зарплата, нормированная на уровень цен региона трудоустройства. Распределения, близкие к симметричному, были определены для признаков Средний возраст выпускника и Доля выпускников, совмещающих учебу и работу. Для всех числовых признаков (за исключением доли выпускников, совмещающих учебу и работу) определялась доля потенциальных выбросов.

Анализ распределения категориальных признаков позволил определить топ-категории по укрупненным группам подготовки, регионам, уровням образования и полу выпускника (рис. 2).



Рис. 2. Визуализация на основе частотного распределения
Примечание: составлен авторами по результатам данного исследования

```
def winsorize_series(s, low_q=0.01, high_q=0.99): # массив, мин%, макс%
    low = s.quantile(low_q) #ниж.кван
    high = s.quantile(high_q) #верх.кван
    return s.clip(lower=low, upper=high) #Обрезка значений и возврат результата.

# логарифмирование + систематизация для сильно скошенных признаков
log_winsor_cols = [c for c in [
    "count_graduate",
    "average_salary_fact_avg",
    "average_salary_fact_med",
    "average_salary_norm_avg",
    "average_salary_norm_med"
] if c in df_clean.columns]

for col in log_winsor_cols:
    # log1p=log(1 + x) уменьшает влияние экстремально больших значений
    #Создание новой колонки с логарифмированными значениями
    df_clean[col + "_log"] = np.log1p(df_clean[col])
    # затем слегка подрезаем хвосты
    df_clean[col + "_log_w"] = winsorize_series(df_clean[col + "_log"], 0.01, 0.99)
```

Рис. 3. Программный код по обработке выбросов
Примечание: составлен авторами по результатам данного исследования

После определения аномальных значений для признаков Количество выпускников, Средняя и Медианная фактическая заработная плата, Средняя и Медианная зарплата, нормированная на уровень цен региона трудоустройства, использовалось логарифмирование и систематизация/обрезка хвостов по квантилям (статистический метод winsorization, который используется для ограничения экстремальных значений в наборе данных с целью уменьшения влияния выбросов на статистический анализ) (рис. 3).

Для остальных числовых признаков была мягко вычислена винсоризация без логарифмирования. Некоторые результаты обработки выбросов представлены на рис. 4.

Для подготовки данных к построению моделей машинного обучения, чтобы не переполнять память, использовались классы StandardScaler и MinMaxScaler для масштабирования числовых данных. Для кодирования категориальных признаков с низкой кардинальностью (≤ 30 уникальных значений) был применен метод One-hot encoding, для высококардинальных признаков – частотное кодирование (1 столбец на признак). Фрагмент обработанного датасета содержится на рис. 5.

Анализ построенной тепловой карты позволил избежать проблему мультиколлинеарности нецелевых признаков в дальнейшем. Выполненный анализ логики при-

знаков, изучение матрицы корреляции, проверка этапов предобработки, использование независимых валидационных выборок позволили решить задачу утечки данных.

По обработанным данным были решены следующие задачи предсказания данных:

- Прогнозирование доли трудоустроенных выпускников.

- Прогнозирование показателя уровня трудоустройства выпускников.

При решении задачи регрессии были построены модели LinearRegression (базовая линейная модель, которая очень быстро обучается, дает понятный ориентир «минимального уровня качества» и легко интерпретируется), DecisionTreeRegressor (простая нелинейная модель, учитывающая нелинейные зависимости и взаимодействия признаков и не требующая сложной подготовки признаков), RandomForestRegressor (усиленная версия дерева (ансамбль), обычно дающая более стабильное и точное качество, чем одно дерево, не склонная к переобучению, нежели DecisionTree). Распределение на обучающую и тестовую выборки: 80 и 20 % соответственно).

Значения метрик качества обучения данных модели продемонстрированы на рис. 7.

Лучшей по R2 оказалась модель RandomForest. Данная модель показывает и лучший баланс точности: наибольший R2 и более низкие ошибки MAE/RMSE среди рассмотренных моделей.

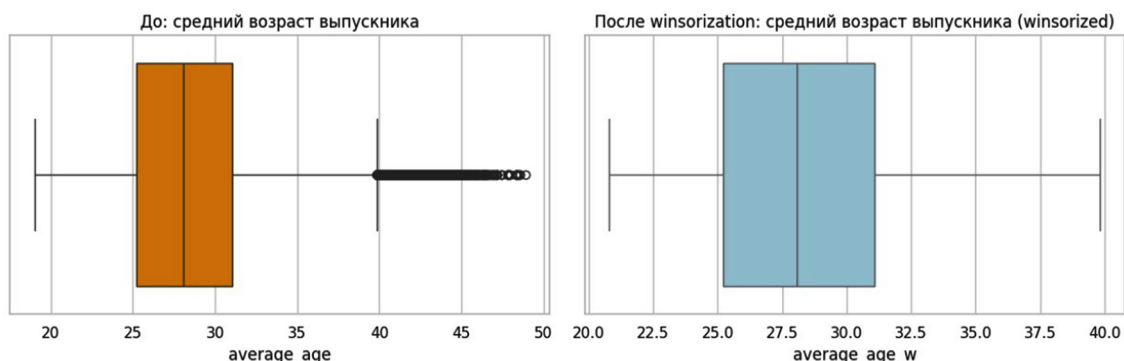


Рис. 4. Обработка выбросов

Примечание: составлен авторами по результатам данного исследования

	okato	oktmo	year	count_graduate	percent_employed	average_salary_fact_avg	average_salary_fact_med	average_salary_norm_avg	average_salary_norm_med
17	91000000	91000000	2016	-0.152017	0.832164	0.054638	-0.228368	0.493429	-0.228368
18	91000000	91000000	2016	0.086969	0.540674	-0.744166	-0.828184	-0.645669	-0.828184
19	91000000	91000000	2016	-0.263544	0.590911	-0.731389	-0.996980	-0.809680	-0.996980
20	91000000	91000000	2016	0.102902	0.670187	-0.524903	-0.867753	-0.795711	-0.867753
21	91000000	91000000	2016	-0.152017	0.563638	-0.558684	-0.658013	-0.458853	-0.658013

Рис. 5. Фрагмент обработанного датасета

Примечание: составлен авторами по результатам данного исследования

```

# Определение тестовой и обучающей выборки
X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Построение, обучение моделей, проверка качества
models = {
    "LinearRegression": LinearRegression(),
    "DecisionTree": DecisionTreeRegressor(max_depth=12, random_state=42),
    "RandomForest": RandomForestRegressor(n_estimators=60, max_depth=12, random_state=42, n_jobs=-1),
}

results = []
predictions = {}

for name, model in models.items():
    model.fit(X_train, Y_train)
    Y_pred = model.predict(X_test)

    results.append({
        "model": name,
        "MAE": mean_absolute_error(Y_test, Y_pred),
        "RMSE": mean_squared_error(Y_test, Y_pred) ** 0.5,
        "R2": r2_score(Y_test, Y_pred),
    })
    predictions[name] = Y_pred

```

Рис. 6. Фрагмент программного кода для построения моделей регрессии
Примечание: составлен авторами по результатам данного исследования

	model	MAE	RMSE	R2
0	RandomForest	6.767735	8.701873	0.495350
1	DecisionTree	7.483474	9.836038	0.355229
2	LinearRegression	8.380521	10.639238	0.245627

Рис. 7. Метрики качества моделей задачи регрессии
Примечание: составлен авторами по результатам данного исследования

	model	Accuracy	F1	ROC_AUC
0	RandomForest	0.733000	0.736989	0.819187
1	DecisionTree	0.712417	0.715194	0.782719
2	LogisticRegression	0.652083	0.647531	0.709911

Рис. 8. Анализ метрик моделей задачи классификации
Примечание: составлен авторами по результатам данного исследования

В целях решения бинарной задачи классификации представленные в датасете объекты были разбиты на два класса: 1 (если доля трудоустроенных выпускников больше медианного значения) и 0 (в остальных случаях). Качество обучения построенных моделей LogisticRegression, DecisionTree, RandomForest анализировалось по метрикам Accuracy (доля правильных ответов (0–1)), F1-мера (баланс точности и полноты (0–1)) и ROC_AUC (способность разделять классы (0–1)) (рис. 8).

Лучшей моделью выбрана RandomForest, так как она показала наибольшее значение

ROC_AUC и баланс Accuracy/F1 на тестовой выборке.

Для решения задачи кластеризации, являющейся задачей обучения без учителя, предварительно был удален признак половой принадлежности выпускника (имеющий значения: мужской, женский, всего) и построены модели KMeans и Agglomerative (с применением метода главных компонент PCA – метода снижения размерности данных с потерей наименьшего объема информации). Количество кластеров на основе elbow method было определено равным 3 (рис. 9).

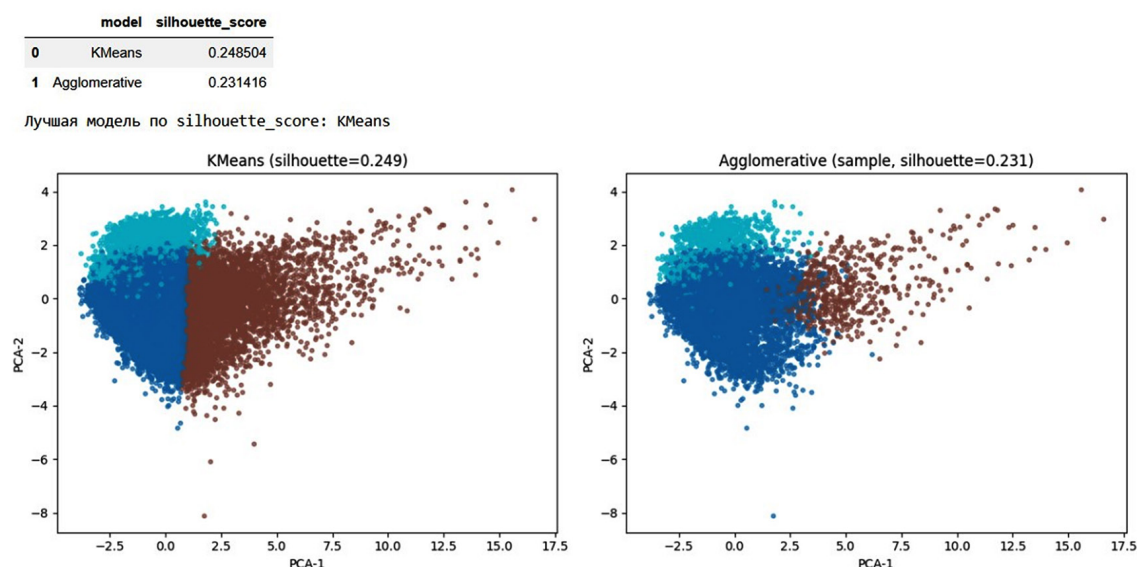


Рис. 9. Модели кластеризации
Примечание: составлен авторами по результатам данного исследования

По метрике `silhouette_score` оптимальной оказалась модель KMeans. Ее кластеры более компактны и лучше отделены друг от друга на данном наборе признаков. Полученные кластеры представляют собой группы образовательных организаций с высоким, средним и низким уровнем трудоустройства выпускников.

Заключение

Алгоритмы машинного обучения, являясь частью искусственного интеллекта, позволяют, обучаясь на данных, выявлять в них закономерности, в том числе скрытые, определять прогнозные значения исследуемых показателей на последующие временные периоды. Построенные модели машинного обучения могут использоваться как вспомогательный инструмент при условии дополнительной валидации, позволяющий оптимизировать процессы в исследуемых прикладных областях.

Представленное исследование продемонстрировало основные этапы проведения интеллектуального анализа данных об уровне трудоустройства и заработной платы выпускников высших и средних специальных учебных заведений.

Построенные модели машинного обучения могут быть применимы в решении задач предсказания исследуемых признаков и кластеризации в анализе деятельности учебных заведений.

Список литературы

1. Бозиева А. М., Цеева Ф. М., Хатухова Д. В. Применение методов машинного обучения при оценке деятельности

образовательной организации высшей школы // Известия КБНЦ РАН. 2023. № 3 (113). URL: <https://cyberleninka.ru/article/n/primenenie-metodov-mashinnogo-obucheniya-pri-otsenke-deyatelnosti-obrazovatelnoy-organizatsii-vysshey-shkoly> (дата обращения: 07.04.2026).

2. Воловик О. А., Воронина Л. В., Горохова Д. В., Губина О. В., Завалько Н. А., Зеленина Л. И., Кабанова Е. Е., Кадацкая Д. В., Ковалев В. А., Красюкова Н. Л., Кузнецова Е. К., Кузнецова С. Ю., Лаврова Ю. С., Лагунова С. В., Лысенко А. А., Москвитина Е. И., Николаев А. В., Панина О. В., Сивоброва И. А., Синицкая Н. Я., Степанова В. В., Федченко Е. А., Хаймина Л. Э., Харченко К. В., Якушева У. Е. Цифровое государство и экономика: учебник / Ред.: Прокофьев С. Е., Панина О. В., Харченко К. В. М.: КноРус, 2024. 345 с. URL: <https://book.ru/book/951781?ysclid=mrj4z12fbu781105235> (дата обращения: 13.05.2026). ISBN 978-5-406-12473-4.

3. Зеленина Л. И., Хаймина Л. Э., Деменкова Е. А., Деменков М. Е., Хаймин Е. С., Хрипунов Д. Д. Сверточные нейронные сети в задаче классификации медицинских изображений // Современные наукоемкие технологии. 2021. № 9. С. 68–73. URL: <https://top-technologies.ru/ru/article/view?id=38818> (дата обращения: 13.05.2026). DOI: 10.17513/snt.38818.

4. Кулагина Е. С., Пепельшев Д. И. Анализ возможностей искусственного интеллекта в сфере автоматизации рутинных задач (на примере образовательных учреждений) // Системный анализ в науке и образовании. 2025. № 1. С. 148–156. URL: <https://elibrary.ru/item.asp?id=80691773> (дата обращения: 07.05.2026).

5. Чернов И. В. Цифровизация как тенденция развития современного общества: специфика научного дискурса // Гуманитарий Юга России. 2021. № 1. С. 121–132. URL: https://www.isras.ru/index.php?page_id=2384&id=7833&l=&j=5&base= (дата обращения: 07.05.2026). DOI: 10.18522/2227-8656.2021.1.11.

6. Хаймина Л. Э., Зеленина Л. И., Деменкова Е. А., Деменков М. Е., Хаймин Е. С., Федькушова С. И. Современные информационные технологии в проектно-исследовательской деятельности студентов // Педагогическая информатика. 2023. № 4. С. 184–191. URL: <https://www.elibrary.ru/item.asp?id=64319136> (дата обращения: 07.05.2026).

7. Видова Т. А., Романова И. Н. Возможности применения технологий искусственного интеллекта в образова-

тельном процессе // Образовательные ресурсы и технологии. 2023. № 1 (42). С. 27–35. URL: <https://elibrary.ru/item.asp?id=52519316> (дата обращения: 07.05.2026). DOI: 10.21777/2500-2112-2023-1-27-35.

8. Баженов Р. И., Баженова Н. Г., Воротилкина И. М. Применение машинного обучения для прогнозирования у сотрудников университета степени готовности к участию в программах по здоровому образу жизни // Международный научно-исследовательский журнал. 2025. № 1 (151). URL: <https://research-journal.org/archive/1-151-2025-january/10.60797/IRJ.2025.151.12> (дата обращения: 07.04.2026). DOI: 10.60797/IRJ.2025.151.12.

9. Ромашенко А. И., Щербин С. И., Харитонов И. М., Панфилов А. Э., Жаравин Н. С. Применение статистических методов машинного обучения для прогнозирования будущей успеваемости абитуриентов в вузе // Современные наукоемкие технологии. 2025. № 9. С. 132–136. URL: <https://top-technologies.ru/article/view?id=40497> (дата обращения: 07.04.2026). DOI: 10.17513/snt.40497.

10. Щербин С. И., Харитонов И. М., Огар Т. П., Панфилов А. Э., Кравец А. Г. Применение методов кластерного анализа в задаче прогнозирования успеваемости абитуриентов в вузе // Современные наукоемкие технологии. 2024. № 5–2. С. 321–325. URL: <https://top-technologies.ru/article/view?id=40046> (дата обращения: 07.04.2026). DOI: 10.17513/snt.40046.

11. Хаймина Л. Э., Зеленина Л. И., Хаймин Е. С., Федькушова С. И. Использование интеллектуальных информационных технологий в проектной деятельности студентов // Педагогическая информатика. 2022. № 3. С. 72–81.

URL: http://pedinf.ru/content_3.22.htm (дата обращения: 07.04.2026).

12. Алпатов А. В. Применение машинного обучения для анализа образовательных результатов студентов вузов // Информационные и математические технологии в науке и управлении. 2023. № 4 (32). С. 67–78. URL: <https://cyberleninka.ru/article/n/primenenie-mashinnogo-obucheniya-dlya-analiza-obrazovatelnyh-rezultatov-studentov-vuzov> (дата обращения: 07.05.2026).

13. Сухоруков Д. С. Персонализация процесса обучения с помощью технологий искусственного интеллекта и машинного обучения: опыт образовательных платформ // Международный журнал гуманитарных и естественных наук. 2025. № 6–1 (105). С. 179–187. URL: <https://elibrary.ru/item.asp?id=82717772> (дата обращения: 07.05.2026). DOI: 10.24412/2500-1000-2025-6-1-179-187.

14. Фартушнов Н. С. Библиотеки языка Python для машинного обучения, их возможности и преимущества // Теория и практика современной науки. 2020. № 5 (59). URL: <https://cyberleninka.ru/article/n/biblioteki-yazyka-python-dlya-mashinnogo-obucheniya-ih-vozmozhnosti-i-preimuschestva> (дата обращения: 07.05.2026).

15. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем / Пер. с англ. СПб.: Альфа-книга, 2018. 688 с. ISBN 978-5-9500296-2-2.

16. Трудоустройство и зарплаты выпускников по образовательным организациям // Роструд; обработка «Если быть точным», 2024. Условия использования: Creative Commons BY 4.0. URL: https://techno.st/datasets/graduates_university (дата обращения: 07.04.2026).

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.