0.1. *Миронова К.Ю.* **Улучшение качества данных для систем машинного обучения с использованием методов искусственного интеллекта**

Современные модели машинного обучения сильно зависят от качества исходных данных. Однако в реальных условиях часто встречаются шумные, неполные или несбалансированные выборки, что снижает эффективность обучения. Целью данной работы стало исследование применения методов искусственного интеллекта для предобработки и улучшения качества данных в задачах классификации и регрессии.

Для решения задачи использовались два подхода. Первый — применение автоэнкодеров для восстановления пропущенных значений и удаления шума [1]. Второй — использование генеративно-состязательных сетей (GAN) для синтетического дополнения данных и балансировки классов [2]. Для оценки качества предобработки были разработаны метрики, позволяющие измерить изменения в точности моделей до и после применения указанных методов.

Результаты экспериментов представлены в таблице 1.

Таблица 1: Сравнение точности моделей машинного обучения до и после предобработки данных

Модель	**До, %**	**После, %**
Logistic Regression	72.3	80.5
Random Forest	75.8	83.1
Support Vector Machine	74.2	82.0
Простая нейронная сеть	73.5	84.7

Эксперименты показали, что предложенные методы позволили увеличить точность предсказаний моделей машинного обучения в среднем на 7–12% по сравнению с традиционными методами очистки и балансировки данных. Кроме того, удалось снизить долю ошибок классификации и повысить устойчивость моделей к выбросам и неполным данным, что согласуется с результатами, представленными в [3].

Актуальность проведённого исследования заключается в том, что во многих прикладных областях (медицина, финансы, образование) именно низкое качество данных ограничивает возможности внедрения систем искусственного интеллекта. Предложенный подход позволяет автоматизировать процесс подготовки данных и обеспечить более высокую надёжность моделей [4].

Научная новизна работы состоит в комплексном использовании автоэнкодеров и генеративных моделей для интеллектуальной предобработки данных, а также в разработке метрик оценки эффективности данного процесса.

Научный руководитель — д.т.н. Целых А. Н.

Список литературы

[1] Yao Y., Wangr X., Ma Y., Fang H., Wei J., Chen L., Anaissi A., Braytee A. Conditional Variational Autoencoder with Balanced Pre-training for Generative Adversarial Networks // CoRR (arXiv preprint). 2022.

[2] Intrator Y., Katz G., Shabtai A. MDGAN: Boosting Anomaly Detection Using Multi-Discriminator Generative Adversarial Networks // arXiv preprint (CoRR). 2018.

[3] Suryawati E., Pardede H. F., Zilvan V., Ramdan A., Krisnandi D., Heryana A., Yuwana R. S., Suryo K. R. B., Arisal A., Supianto A. A. Unsupervised feature learning-based encoder and adversarial networks // Journal of Big Data. 2021. Vol. 8. Article 118 (2021).

[4] Hwang U., Choi S., Lee H.-B., Yoon S. Adversarial Training for Disease Prediction from Electronic Health Records with Missing Data // arXiv preprint. 2017.