QAZAQ JOURNAL OF YOUNG SCIENTIST

2026, Vol.4, No. 4 S (April)

https://qazaqjournal.kz/

УДК 004.891.3:005.963

РАЗРАБОТКА И ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА МОДЕЛИ ОЦЕНКИ ЭФФЕКТИВНОСТИ РАБОТЫ СОТРУДНИКОВ В ПРОЦЕССЕ РАЗРАБОТКИ И АВТОМАТИЗАЦИИ ИТ-ПРОЕКТОВ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Шубаев К.Н.[1], Ахметова Ж.Ж.[2]

[1] Магистрант кафедры «Информационные технологии»

[2] PhD, ассоциированный профессор, научный руководитель

Казахский университет технологии и бизнеса им. К. Кулажанова, Астана, Казахстан

В статье представлены результаты разработки и экспериментальной проверки модели оценки эффективности работы сотрудников в процессе разработки и автоматизации ИТ-проектов на основе методов машинного обучения. Сформирован набор данных из 800 наблюдений, включающий 16 проектных, процессных, инженерных и организационно-командных признаков. Проведено сравнение пяти алгоритмов классификации: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting и XGBoost. Лучший результат достигнут моделью Logistic Regression: Accuracy = 0,9500, F1-macro = 0,9511, ROC-AUC = 0,9792. Устойчивость результатов подтверждена 5-кратной стратифицированной кросс-валидацией (F1-macro = 0,9420 ± 0,0145). Анализ значимости признаков методом permutation importance показал, что наиболее значимый вклад в оценку эффективности вносят показатели выполнения задач, объема доработок, успешности тестирования и участия в code review. Предложенная модель интегрирована в программный прототип на базе FastAPI и пригодна для использования в задачах поддержки управленческих решений в ИТ-проектах.

Ключевые слова: оценка эффективности, ИТ-проекты, машинное обучение, классификация, SPACE, DORA, XGBoost, Logistic Regression, Random Forest, permutation importance.

Введение

В условиях цифровой трансформации организаций успешность реализации ИТ-проектов во многом определяется качеством работы сотрудников, соблюдением сроков, эффективностью командного взаимодействия и способностью адаптироваться к изменяющимся требованиям. Традиционные подходы к оценке персонала, основанные на периодических отчетах и экспертных суждениях, не в полной мере учитывают специфику ИТ-сферы: гибкие методологии разработки, высокую изменчивость требований, распределенную командную работу и необходимость непрерывного контроля промежуточных результатов.

Современные исследования подчеркивают, что продуктивность инженерного труда не может быть сведена к одному показателю. В модели SPACE продуктивность разработчика рассматривается как многомерная категория, включающая результативность, активность, коммуникацию и сотрудничество, поток работы, а также удовлетворенность и благополучие сотрудников. Подход DORA смещает акцент с индивидуальной активности на свойства инженерного процесса — частоту развертывания, время прохождения изменений, надежность поставки и время восстановления после сбоев.

Применение методов машинного обучения позволяет перейти от описательного анализа к предиктивному контуру управления, выявлять скрытые закономерности в цифровых следах проектной деятельности и формировать более объективные оценки эффективности. Вместе с тем применение ML в задачах HR-аналитики требует особой осторожности: алгоритмическая оценка должна оставаться интерпретируемой и не подменять собой экспертное суждение.

Целью настоящего исследования является разработка и экспериментальная проверка модели классификации уровня эффективности сотрудников в ИТ-проектах на основе методов машинного обучения, а также сравнительный анализ нескольких алгоритмов и отбор оптимальной модели для интеграции в программный прототип системы поддержки принятия управленческих решений.

Задачи исследования: сформировать экспериментальный набор данных, отражающий проектные, процессные, инженерные и организационно-командные признаки сотрудников; обучить и сравнить пять моделей машинного обучения; оценить устойчивость лучшей модели методом кросс-валидации; провести анализ значимости признаков; подготовить модель к интеграции в программный прототип.

Материалы и методы

Для проведения экспериментов сформирован набор данных, описывающий деятельность сотрудников в ИТ-проектах. Одно наблюдение соответствует работе одного сотрудника за один отчетный период (неделя, спринт или месяц).

Поскольку реальные корпоративные данные подобного типа относятся к конфиденциальной информации и недоступны для открытого научного исследования, в работе использован экспериментальный набор данных, сформированный на основе логических правил, согласованных с реальными закономерностями ИТ-проектной деятельности и положениями моделей SPACE и DORA.

Объем выборки составил 800 наблюдений, число признаков — 16. Целевая переменная efficiency_class представляет собой дискретную метку уровня эффективности, принимающую три значения: 0 — низкая эффективность, 1 — средняя, 2 — высокая. Распределение классов в генеральной совокупности составило приблизительно 30 % / 45 % / 25 %, что соответствует типичной структуре персонала в ИТ-командах. Для приближения к реальным условиям в данные внесен шум: 3 % пропущенных значений в признаках, характеризующих качество работы, и 5 % шума в метках класса, отражающего субъективность экспертных оценок.

Таблица 1 — Состав признаков экспериментального набора данных

Группа	**Признак**	**Описание**
Задачи	tasks_assigned	Количество назначенных задач
	tasks_completed	Количество выполненных задач
	on_time_ratio	Доля задач, выполненных в срок
	avg_task_duration	Среднее время выполнения задачи, ч
Качество	overdue_tasks	Число просроченных задач
	bugs_count	Количество ошибок и дефектов
	rework_count	Число доработок
	test_pass_rate	Доля успешных тестов
Инженерная активность	commits_count	Количество коммитов в репозиторий
	pull_requests_count	Количество pull request
	code_review_participation	Доля участия в code review
Процесс и KPI	workload_score	Уровень загрузки сотрудника
	kpi_completion	Выполнение KPI
	attendance_rate	Дисциплина и участие
Экспертные оценки	team_feedback_score	Оценка командного взаимодействия (1–5)
	manager_score	Экспертная оценка руководителя (1–5)

На этапе предобработки выполнена обработка пропущенных значений методом заполнения медианой по классу, удаление возможных дубликатов, а также стратифицированное разбиение выборки на обучающую и тестовую части в соотношении 80 / 20. Для моделей, чувствительных к масштабу признаков (Logistic Regression), применена стандартизация числовых признаков методом Z-нормализации с помощью StandardScaler. Для ансамблевых методов на основе решающих деревьев стандартизация не требуется в силу инвариантности таких моделей относительно монотонных преобразований признаков.

Для решения задачи многоклассовой классификации уровня эффективности сотрудников в работе сравниваются пять алгоритмов, типичных для задач HR-аналитики и прогнозирования на табличных данных:

1. Logistic Regression — базовая линейная модель, обеспечивающая высокую интерпретируемость коэффициентов и опорную точку для сравнения.
2. Decision Tree — наглядная модель правил принятия решений, склонная к переобучению при большой глубине.
3. Random Forest — устойчивый ансамблевый метод на основе bagging, эффективный на табличных данных.
4. Gradient Boosting — последовательный ансамблевый метод, итеративно обучающий слабые модели на остатках предыдущих.
5. XGBoost — современный высокопроизводительный градиентный бустинг с регуляризацией, зарекомендовавший себя как один из наиболее сильных методов для табличных данных.

Для оценки качества моделей использованы метрики Accuracy, Precision, Recall и F1-score (в усредненных вариантах macro и weighted), а также ROC-AUC в схеме One-vs-Rest. Основной акцент при интерпретации результатов сделан на метрике F1-macro, так как она более корректно отражает качество модели в условиях частично несбалансированных классов. Для проверки устойчивости моделей применена 5-кратная стратифицированная кросс-валидация. Программная реализация экспериментов выполнена на языке Python с использованием библиотек scikit-learn 1.8, XGBoost 2.1, NumPy, Pandas, Matplotlib и Seaborn.

Результаты и обсуждение

Результаты обучения и тестирования моделей на независимой тестовой выборке (160 наблюдений) приведены в таблице 2. Модели отсортированы по убыванию метрики F1-macro. Для наглядного сопоставления полученных результатов выполнена их графическая визуализация. На рисунке 1 представлен сравнительный анализ качества моделей по основным метрикам, а на рисунке 2 — ранжирование моделей по метрике F1-macro.

Таблица 2 — Результаты сравнительного анализа моделей на тестовой выборке

Модель	Accuracy	Precision (macro)	F1-macro	ROC-AUC
Logistic Regression	0,9500	0,9535	0,9511	0,9792
Random Forest	0,9437	0,9456	0,9452	0,9781
Gradient Boosting	0,9437	0,9448	0,9445	0,9769
XGBoost	0,9437	0,9441	0,9438	0,9775
Decision Tree	0,8063	0,8122	0,8099	0,8747

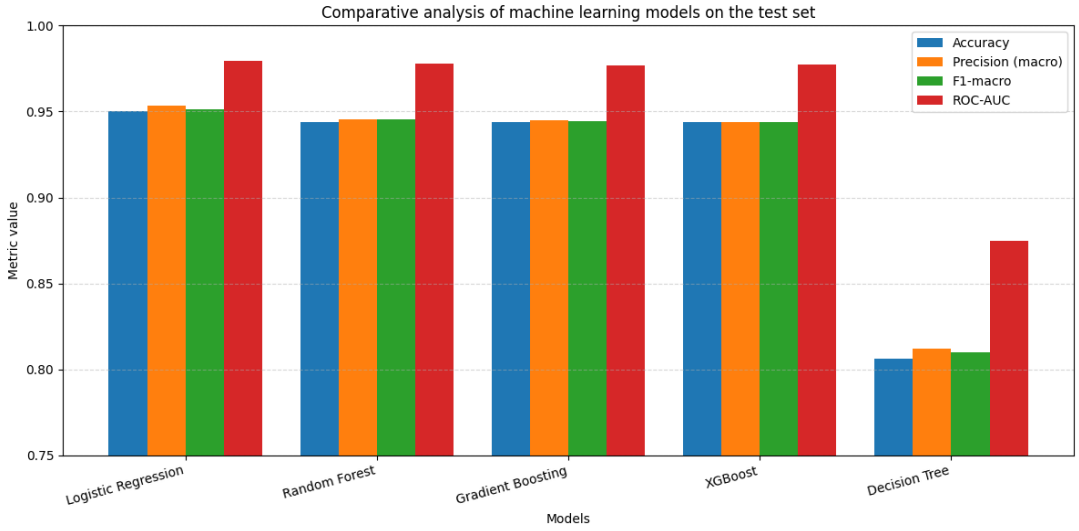

Рисунок 1 — Сравнительный анализ качества моделей машинного обучения на тестовой выборке

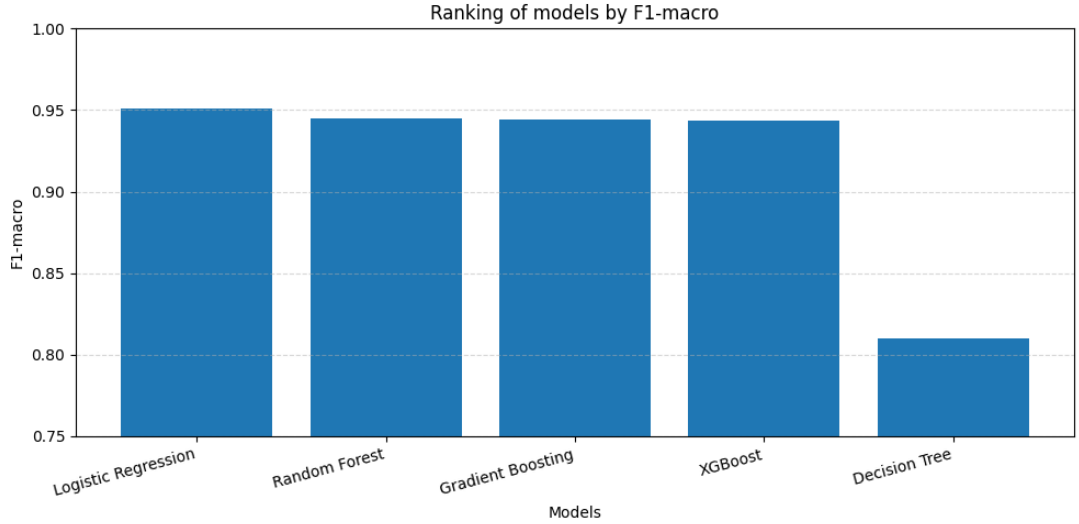

Рисунок 2 — Ранжирование моделей машинного обучения по метрике F1-macro

Наилучший результат в экспериментах продемонстрировала модель Logistic Regression со значениями Accuracy = 0,9500 и F1-macro = 0,9511. Как видно из рисунка 1, данная модель показывает наивысшие значения практически по всем рассматриваемым метрикам. Ансамблевые методы (Random Forest, Gradient Boosting, XGBoost) продемонстрировали близкие между собой показатели на уровне Accuracy = 0,9437 и F1-macro в диапазоне 0,9438–0,9452, что свидетельствует о сопоставимой эффективности этих алгоритмов на сформированном наборе данных. Это также отчетливо видно на рисунке 2, где указанные модели располагаются непосредственно за лидирующей Logistic Regression с минимальным отрывом.

Заметно ниже результаты одиночного Decision Tree (Accuracy = 0,8063), что особенно наглядно отражено на рисунках 1 и 2. Такое снижение качества согласуется с известным теоретическим положением о склонности одиночных деревьев решений к переобучению при отсутствии механизмов регуляризации.

Существенно, что модель Logistic Regression превосходит более сложные нелинейные модели при сопоставимой сбалансированности precision и recall по всем трем классам. Это объясняется структурой признакового пространства: связь между показателями проектной деятельности и классом эффективности оказывается преимущественно монотонной и близкой к линейной. В таких условиях простая модель с регуляризацией обобщает данные лучше, чем бустинги, которые на ограниченном объеме 800 наблюдений имеют тенденцию к переобучению. Полученные результаты, представленные в таблице 2 и на рисунках 1–2, поддерживают общий методологический принцип предпочтения более простых моделей при сопоставимом качестве, отмеченный в литературе по применению машинного обучения в HR-задачах.

Для проверки устойчивости полученных оценок качества применена 5-кратная стратифицированная кросс-валидация на обучающей выборке. Результаты приведены в таблице 3.

Таблица 3 — Результаты 5-кратной стратифицированной кросс-валидации

Модель	**F1-macro (среднее ± σ)**	**Разрыв train–test, Δ**
Logistic Regression	0,9420 ± 0,0145	+0,0063
Random Forest	0,9451 ± 0,0126	+0,0516
Gradient Boosting	0,9271 ± 0,0139	+0,0563
XGBoost	0,9228 ± 0,0152	+0,0563
Decision Tree	0,8326 ± 0,0381	+0,1672

Низкое стандартное отклонение оценок F1-macro по фолдам ($\sigma \leq 0{,}016$ для всех моделей, кроме Decision Tree) свидетельствует об устойчивости полученных результатов к случайным вариациям состава выборки. Разрыв между качеством модели на обучающей и тестовой выборках у Logistic Regression минимален ($\Delta = +0{,}0063$), что указывает на хорошее обобщение и отсутствие переобучения. В то же время одиночный Decision Tree демонстрирует существенный разрыв ($\Delta = +0{,}1672$), что подтверждает его чувствительность к конкретному составу обучающих данных.

Для выявления управленчески значимых факторов, влияющих на уровень эффективности сотрудников, применен метод permutation importance. Метод оценивает вклад каждого признака через падение метрики F1-macro при случайной перестановке его значений. Результаты для лучшей модели представлены в таблице 4.

Таблица 4 — Значимость признаков по результатам permutation importance

№	Признак	Снижение F1-macro
1	tasks_completed — выполненные задачи	0,0820
2	rework_count — число доработок	0,0516
3	test_pass_rate — доля успешных тестов	0,0386
4	code_review_participation — участие в code review	0,0310
5	on_time_ratio — доля задач в срок	0,0259
6	manager_score — экспертная оценка руководителя	0,0167
7	commits_count — число коммитов	0,0159

Анализ значимости признаков выявил четыре группы факторов, наиболее сильно влияющих на классификацию уровня эффективности. Первую группу образуют результативные показатели — tasks_completed и on_time_ratio, отражающие непосредственное достижение сотрудником проектных целей. Вторую группу составляют показатели качества — rework_count и test_pass_rate, фиксирующие устойчивость результата во времени и отсутствие необходимости в доработках. Третья группа — показатели вовлеченности в инженерный процесс, в первую очередь code_review_participation, характеризующие степень участия сотрудника в совместной разработке. Четвертую группу формируют экспертные оценки (manager_score).

Полученный результат согласуется с теоретическими положениями модели SPACE, в которой продуктивность разработчика складывается из нескольких измерений, включая результативность, активность и коммуникацию, а не

только из формальной производительности. Особенно показательно, что доля участия в code review вошла в первую пятерку значимых признаков: это подтверждает тезис о коллективном характере эффективности в ИТ-проектах. В то же время относительно низкая позиция показателя commits_count поддерживает известное в литературе наблюдение о том, что простое количество коммитов является слабым прокси-показателем продуктивности разработчика.

Для наилучшей модели Logistic Regression построен развернутый отчет классификации по каждому из трех классов эффективности (таблица 5). Поддержка классов (support) отражает число наблюдений соответствующего класса в тестовой выборке.

Таблица 5 — Метрики классификации по классам эффективности (Logistic Regression)

Класс	**Precision**	**Recall**	**F1-score**	**Support**
Низкая	0,9787	0,9388	0,9583	49
Средняя	0,9306	0,9571	0,9437	70
Высокая	0,9512	0,9512	0,9512	41
Macro avg	0,9535	0,9490	0,9511	160
Weighted avg	0,9506	0,9500	0,9501	160

Наибольшая точность достигнута для класса низкой эффективности (F1 = 0,9583), что управленчески важно, поскольку именно этот класс представляет наибольший интерес для руководителя с точки зрения выявления проблемных ситуаций и оперативной поддержки сотрудников. Качество классификации для средней и высокой эффективности также остается высоким (F1 = 0,9437 и 0,9512 соответственно), а близкие значения precision и recall по всем классам указывают на отсутствие систематического смещения модели в сторону того или иного класса.

На заключительном этапе лучшая модель Logistic Regression объединена в единый pipeline с этапом стандартизации признаков и сериализована средствами библиотеки joblib. Итоговый файл модели подключен к программному прототипу системы оценки эффективности, реализованному на базе фреймворка FastAPI. Прототип предоставляет REST-эндпойнт POST /predict, принимающий JSON-объект с 16 признаками сотрудника и возвращающий прогнозируемый класс эффективности, а также распределение вероятностей по трем классам.

Тестирование прототипа на контрольных примерах подтвердило корректность работы: среднее время отклика модели на одиночный запрос не

превышает 15 мс на типовой конфигурации сервера, что делает решение пригодным для интеграции в BI-панели и системы поддержки управленческих решений в режиме реального времени. Интерпретируемый вывод модели (класс + вероятности + вклад признаков) позволяет руководителю не только получить прогноз, но и понять его обоснование, что согласуется с принципом интерпретируемости, закрепленным в разделе проектирования модели.

Заключение

В статье представлены результаты разработки и экспериментальной проверки модели оценки эффективности работы сотрудников в процессе разработки и автоматизации ИТ-проектов на основе методов машинного обучения. Основные результаты исследования могут быть сформулированы следующим образом.

Во-первых, разработан экспериментальный набор данных из 800 наблюдений и 16 признаков, отражающих результативные, качественные, процессные, инженерные и организационно-командные аспекты работы сотрудников в ИТ-проектах. Состав признаков согласован с положениями моделей SPACE и DORA, что обеспечивает теоретическую обоснованность и многомерность оценки.

Во-вторых, проведен сравнительный анализ пяти моделей машинного обучения. Наилучший результат продемонстрировала модель Logistic Regression со значениями Accuracy = 0,9500, F1-macro = 0,9511 и ROC-AUC = 0,9792 при минимальном разрыве между обучающей и тестовой точностью. Устойчивость результатов подтверждена 5-кратной стратифицированной кросс-валидацией (F1-macro = 0,9420 ± 0,0145).

В-третьих, методом permutation importance выявлены наиболее значимые факторы, влияющие на уровень эффективности сотрудников: количество выполненных задач, объем доработок, доля успешных тестов, участие в code review, соблюдение сроков и экспертная оценка руководителя. Полученный результат подтверждает многомерный характер продуктивности в ИТ-сфере и согласуется с теоретическими положениями модели SPACE.

В-четвертых, лучшая модель интегрирована в программный прототип системы оценки эффективности на базе FastAPI. Прототип обеспечивает прием входных данных через REST API, возвращает прогнозируемый класс и распределение вероятностей и может быть встроен в корпоративные системы поддержки принятия решений.

Практическая значимость работы состоит в возможности применения разработанной модели в организациях, реализующих ИТ-проекты, для повышения объективности оценки персонала, раннего выявления сотрудников с риском снижения эффективности и поддержки принятия управленческих решений. Направлением дальнейших исследований является проверка модели на реальных корпоративных данных, расширение признакового пространства за

счет динамических временных характеристик и изучение применимости рекуррентных нейросетевых моделей (LSTM) для прогнозирования изменения эффективности во времени.

Список использованных источников

1. Forsgren N., Storey M.-A., Maddila C. et al. The SPACE of Developer Productivity: There's more to it than you think // ACM Queue. — 2021. — Vol. 19, No. 1. — P. 1–28. — DOI: 10.1145/3454122.3454124.

2. Rejab M.M., Omar M., Ahmad M. Agile-Compliant Performance Appraisal System: Eight Principles for Effective Performance Measurement in Agile Software Development Teams // Information and Software Technology. — 2020. — Vol. 125. — P. 106–128. — DOI: 10.1016/j.infsof.2020.106328.

3. Nayem M.A., Uddin M.A. Unbiased Employee Performance Evaluation Using Machine Learning // Journal of Open Innovation: Technology, Market, and Complexity. — 2024. — Vol. 10. — P. 100243. — DOI: 10.1016/j.joitmc.2024.100243.

4. Yanamala K.K.R. Integrating Machine Learning and Human Feedback for Employee Performance Evaluation // Journal of Advanced Computing Systems. — 2022. — Vol. 2, No. 1. — P. 1–10.

5. Huang X., Yang F., Zheng J., Feng C., Zhang L. Personalized Human Resource Management via HR Analytics and Artificial Intelligence: Theory and Implications // Asia Pacific Management Review. — 2023. — Vol. 28, No. 4. — P. 598–610. — DOI: 10.1016/j.apmrv.2023.01.001.

6. Chowdhury S., Dey P., Joel-Edgar S. et al. Unlocking the Value of Artificial Intelligence in Human Resource Management through AI Capability Framework // Human Resource Management Review. — 2022. — Vol. 32, No. 4. — Article 100899. — DOI: 10.1016/j.hrmr.2022.100899.

7. Sharma P., Bhattacharya S. HR Analytics and AI Adoption in IT Sector: Reflections from Practitioners // Journal of Work-Applied Management. — 2025. — Vol. 17. — DOI: 10.1108/JWAM-12-2024-0179.

8. Bentéjac C., Csörgő A., Martínez-Muñoz G. A Comparative Analysis of Gradient Boosting Algorithms // Artificial Intelligence Review. — 2021. — Vol. 54. — P. 1937–1967. — DOI: 10.1007/s10462-020-09896-5.

9. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. — 2016. — P. 785–794. — DOI: 10.1145/2939672.2939785.

10. Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ НЕГІЗІНДЕ АТ-ЖОБАЛАРДЫ ӘЗІРЛЕУ ЖӘНЕ АВТОМАТТАНДЫРУ ҮДЕРІСІНДЕ ҚЫЗМЕТКЕРЛЕР ЖҰМЫСЫНЫҢ ТИІМДІЛІГІН БАҒАЛАУ МОДЕЛІН ӘЗІРЛЕУ ЖӘНЕ ЭКСПЕРИМЕНТТІК ТЕКСЕРУ

Шубаев Қ.Н.[1], Ахметова Ж.Ж.[2]
[1] «Ақпараттық технологиялар» кафедрасының магистранты
[2] PhD, қауымдастырылған профессор, ғылыми жетекші
Қ. Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан

Мақалада АТ-жобаларды әзірлеу және автоматтандыру процесінде қызметкерлердің еңбек тиімділігін бағалау моделін машиналық оқыту әдістері негізінде әзірлеу және эксперименттік тексеру нәтижелері ұсынылған. 800 бақылаудан тұратын деректер жиыны қалыптастырылды, оған 16 жобалық, процестік, инженерлік және ұйымдастырушылық-командалық белгілер кіреді. Бес классификациялық алгоритм салыстырылды: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting және XGBoost. Ең жақсы нәтижеге Logistic Regression моделі қол жеткізді: Accuracy = 0,9500, F1-macro = 0,9511. Нәтижелердің тұрақтылығы 5 есе стратификацияланған кросс-валидациямен расталды.

Кілт сөздері: тиімділікті бағалау, АТ-жобалар, машиналық оқыту, классификация, SPACE, DORA.

DEVELOPMENT AND EXPERIMENTAL VALIDATION OF A MACHINE LEARNING-BASED MODEL FOR EVALUATING EMPLOYEE PERFORMANCE IN THE DEVELOPMENT AND AUTOMATION OF IT PROJECTS

Shubayev K.N., Akhmetova Zh.Zh.
[1] Master’s student of the Department of Information Technologies
[2] PhD, Associate Professor, Scientific Supervisor
K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan

The article presents the results of the development and experimental verification of a machine learning–based model for assessing the performance of employees in the development and automation of IT projects. A dataset of 800 observations including 16 project, process, engineering, and organizational–team features was constructed. Five classification algorithms were compared: Logistic Regression,

Decision Tree, Random Forest, Gradient Boosting, and XGBoost. The best result was achieved by Logistic Regression: Accuracy = 0.9500, F1-macro = 0.9511, ROC-AUC = 0.9792. The robustness of the results was confirmed by 5-fold stratified cross-validation (F1-macro = 0.9420 ± 0.0145). Permutation importance analysis showed that task completion, rework volume, test pass rate, and code review participation contribute most significantly to employee performance estimation. The proposed model was integrated into a FastAPI-based software prototype and is suitable for managerial decision support in IT projects.

Keywords: performance evaluation, IT projects, machine learning, classification, SPACE, DORA, XGBoost, Logistic Regression, Random Forest, permutation importance.