ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
2024. Т. 6, № 3 (18). С. 95–101 http://siit.ugatu.su
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научное сообщение

УДК 004.65 DOI 10.54708/2658-5014-SIIT-2024-no3-p95 EDN VLFFLP

ПРОГНОЗИРОВАНИЕ ВЕРОЯТНОСТИ РАЗВИТИЯ ДИАБЕТИЧЕСКОЙ РЕТИНОПАТИИ У ПАЦИЕНТОВ С САХАРНЫМ ДИАБЕТОМ: АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

М. С. Зиновьев • О. С. Нургаянова

Аннотация. В статье рассматривается исследование, направленное на разработку эффективного алгоритма предсказания вероятности развития диабетической ретинопатии у пациентов с сахарным диабетом. Приводятся ряд методов машинного обучения, которые применяются для создания моделей машинного обучения, способных предсказывать вероятность развития диабетической ретинопатии в ближайшие годы. Определяется наиболее эффективная модель на основе метрик оценки эффективности моделей машинного обучения.

Ключевые слова: сахарный диабет; диабетическая ретинопатия; количественная оценка риска; машинное обучение.

Введение

Диабетическая ретинопатия (ДР) представляет собой сложное состояние, связанное с повреждением сетчатки глаза, вызванным хроническим диабетом. Это одно из наиболее серьезных осложнений сахарного диабета (СД) и является результатом диабетической микроангиопатии, которая влияет на капилляры сетчатки глаза, и, по статистике, страдает около 90% пациентов с диагностированным диабетом [1]. Это осложнение чаще всего возникает при длительном течении СД, особенно – если он плохо контролируется.

Регулярные офтальмологические обследования являются крайне важными для выявления ретинопатии на ранних стадиях [2]. Очень важно осуществлять мониторинг зрительной функции и состояния глазного дна у всех пациентов с диабетом, чтобы своевременно выявить начальные признаки ретинопатии и предотвратить ее прогрессирование. Поэтому медицинские рекомендации включают регулярные посещения офтальмолога для скрининга и лечения ретинопатии.

Нарушение зрения, вызванное ретинопатией, является одним из самых инвалидизирующих проявлений сахарного диабета [3]. Это состояние может привести к потере зрения и слепоте, и, по оценкам, риск слепоты у пациентов с диабетом выше в 25 раз по сравнению с теми, у кого нет этого заболевания [4].

Ситуацию не улучшает тот факт малого количества инструментов, которые бы позволяли эффективно предсказывать риски развития ДР у пациентов с СД [5].

В данной статье рассматривается исследование, основанное на анализе датасета, направленное на разработку эффективного алгоритма предсказания шанса развития ДР у пациентов с СД. Полученные результаты исследования могут иметь значительное практическое применение в ранней диагностике и предотвращении осложнений данного заболевания, что повышает качество жизни пациентов и снижает риски возникновения угрожающих осложнений зрительной системы.

Описание датасета

Выбранный для анализа датасет “Diabetic_Nephropathy_v1” находится в свободном доступе на ресурсе Kaggle [6]. Датасет предназначался для исследования диабетической

нефропатии (ДН), но так как среди данных указывается и ретинопатия, его можно использовать и для предсказания развития ДР.

Датасет представляет собой таблицу из 22 колонок:

1) Sex – половая принадлежность.
2) Age – возраст.
3) Diabetes duration (y) – длительность течения диабета в годах.
4) Diabetic retinopathy (DR) – наличие ДР.
5) Diabetic nephropathy (DN) – наличие ДН.
6) Smoking – употребление табачной продукции.
7) Drinking – употребление алкоголя.
8) Height (cm) – рост в сантиметрах.
9) Weight (kg) – вес в килограммах.
10) BMI (kg/m2) – индекс массы тела.
11) SBP (mmHg) – систолическое кровяное давление в мм рт. ст.
12) DBP (mmHg) – диастолическое кровяное давление в мм рт. ст.
13) HbA1c (%) – гликированный гемоглобин в крови в процентах.
14) FBG (mmol/L) – фибриноген в крови в ммоль/л.
15) TG (mmoll) – тиреоглобулин в ммоль.
16) C-peptide (ng/ml) – С-пептид в нг/мл.
17) TC (mmoll) – общий холестерин в ммоль.
18) HDLC (mmoll) – липопротеины высокой плотности в ммоль.
19) LDLC (mmoll) – липопротеины низкой плотности в ммоль.
20) Insulin – прием инсулина.
21) Metformin – прием метформин.
22) Lipid lowering drugs – прием гиполипидемических препаратов.

Датасет предоставляет разностороннюю информацию о пациентах с диабетом, как антропометрические показатели, так и анализы крови и употребляемые препараты.

Методы машинного обучения

Для обучения модели прогнозирования развития ДР необходимо выбрать метод машинного обучения, способный выдавать вероятность развития болезни в процентном виде. Поэтому для проведения анализа датасета были выбраны перечисленные ниже методы, а также методы семплирования для балансировки выборки. При проведении исследования авторы опирались на работы [6–10].

Логистическая регрессия. В статистике логистическая модель (или логит-модель) – это статистическая модель, которая моделирует логарифмический шанс события как линейную комбинацию одной или нескольких независимых переменных. Основной принцип работы линейной регрессии [11] заключается в том, чтобы аппроксимировать зависимость между переменными линейной функции. Для этого модель строит прямую линию (или плоскость в случае многомерной регрессии), которая наилучшим образом соответствует распределению данных.

Процесс обучения модели линейной регрессии состоит в подборе оптимальных коэффициентов (весов) для каждой независимой переменной таким образом, чтобы минимизировать ошибку предсказания. Это достигается с помощью метода наименьших квадратов или других оптимизационных методов. После завершения обучения модель может быть использована для предсказания значений зависимой переменной на основе новых входных данных. Ключевая

идея линейной регрессии заключается в том, чтобы найти наилучшую линейную аппроксимацию зависимости между переменными, что позволяет делать прогнозы и выявлять взаимосвязи в данных.

Метод опорных векторов. В машинном обучении метод опорных векторов (SVM, также известный как сети опорных векторов) основан на поиске оптимальной гиперплоскости, которая разделяет данные разных классов с максимальным зазором. Для нелинейных задач SVM [12] использует трюк с ядром, позволяющий перевести данные в пространство более высокой размерности, где они становятся линейно разделимыми. В процессе обучения SVM находит опорные векторы, которые играют ключевую роль в определении положения разделяющей гиперплоскости. Цель SVM – максимизация зазора между классами или минимизация ошибки в случае регрессии, что обеспечивает высокую точность и обобщение на новых данных.

Основная идея SVM заключается в поиске оптимальной разделяющей гиперплоскости, используя опорные векторы и трюк с ядром для работы с различными типами данных. Этот метод обеспечивает высокую эффективность и точность в задачах классификации и регрессии, делая его одним из наиболее широко применяемых алгоритмов в машинном обучении.

Метод случайного леса. Метод случайного леса (Random Forest) является мощным алгоритмом машинного обучения, который основан на идее ансамблирования деревьев решений. Основной принцип работы случайного леса [13] заключается в создании большого количества деревьев решений на основе случайно выбранных подмножеств признаков и обучении каждого дерева на случайной выборке данных. После обучения каждое дерево в лесу голосует за классификацию или предсказание, и результат определяется большинством голосов.

Ключевая идея случайного леса состоит в том, чтобы использовать множество слабых моделей (деревьев решений), чтобы получить сильную модель, которая способна к обобщению на новых данных. Благодаря случайному выбору признаков и данных для каждого дерева случайный лес уменьшает переобучение и повышает устойчивость модели. Этот метод широко используется в задачах классификации и регрессии благодаря своей высокой эффективности и способности работать с различными типами данных.

Градиентный бустинг. Градиентный бустинг (Gradient Boosting) – это алгоритм машинного обучения, который построен на принципе создания ансамбля слабых моделей, таких как деревья решений, и последовательном улучшении предсказательной способности путем минимизации ошибки на каждом шаге. Основной принцип работы градиентного бустинга [14] заключается в том, чтобы добавлять новые деревья к ансамблю таким образом, чтобы каждое новое дерево исправляло ошибки предыдущих деревьев.

В процессе обучения градиентного бустинга каждое новое дерево строится с учетом градиента (направления наибольшего убывания) функции потерь, что позволяет модели постепенно улучшать предсказания. Градиентный бустинг является очень эффективным методом благодаря тому, что каждое новое дерево фокусируется на ошибках предыдущих деревьев, что позволяет достичь высокой точности на обучающих данных и обобщающей способности на новых данных.

Нейронная сеть. Нейронная сеть – это алгоритм машинного обучения, который моделирует работу нейронов в человеческом мозге для выполнения различных задач, таких как классификация, регрессия и обработка изображений или текста. Принцип работы нейронной сети [15] заключается в создании сети искусственных нейронов, которые образуют слои и взаимодействуют друг с другом через веса и активационные функции.

В процессе обучения нейронная сеть адаптирует веса связей между нейронами таким образом, чтобы минимизировать ошибку предсказания на обучающих данных. Это происходит благодаря использованию таких алгоритмов оптимизации, как градиентный спуск, которые корректируют веса с учетом градиента функции потерь. Нейронные сети позволяют решать сложные задачи, включая распознавание образов, обработку естественного языка и анализ временных рядов, благодаря их способности извлекать сложные закономерности из данных.

Для анализа исследуемого датасета создана нейронная сеть с одним входным слоем, двумя скрытыми слоями и одним выходным слоем. Используется сигмоидная функция активации на выходном слое для получения вероятности.

Random Under- и -Oversampling. Random Undersampling и Random Oversampling – это техники, используемые в задачах несбалансированных данных для улучшения производительности модели машинного обучения. Принцип работы Random Undersampling [16] заключается в уменьшении количества образцов в меньшем классе данных до уровня большего класса, случайным образом отбирая данные для сохранения баланса классов. Это позволяет снизить дисбаланс и улучшить способность модели к обобщению на различные классы.

С другой стороны, Random Oversampling [16] увеличивает количество образцов в меньшем классе путем случайного дублирования существующих данных или генерации синтетических образцов, чтобы достичь баланса классов. Это также помогает улучшить обучение модели, но может привести к переобучению, если не используются дополнительные методы контроля.

Обе техники Random Undersampling и Random Oversampling широко применяются в ситуациях с несбалансированными данными, чтобы улучшить эффективность модели и добиться более точных прогнозов для всех классов.

Результаты

На исследуемом датасете были обучены 15 предсказательных моделей, то есть каждый из пяти методов был применен к исходному, сокращенному и расширенному датасетам.

Результат представлен в таблице через основные метрики.

Таблица

Результат оценки полученных моделей

Название метода	ROC-AUC	PR-AUC	Log Loss	Brier Score	Calibration Curve MSE
LogisticRegression	0.7159	0.4864	0.5663	0.1916	0.0441
LogisticRegression undersampled	0.7071	0.4616	0.6752	0.2330	0.0554
LogisticRegression oversampled	0.7372	0.5095	0.6298	0.2157	0.0551
Support Vector Machines	0.6432	0.4496	0.5920	0.2019	0.0958
Support Vector Machines undersampled	0.6158	0.3921	0.6769	0.2418	0.0488
Support Vector Machines oversampled	0.6272	0.3986	0.6750	0.2402	0.0476
RandomForest	0.6453	0.4216	0.5914	0.1993	0.0793
RandomForest undersampled	0.9299	0.8926	0.4861	0.1599	0.0703
RandomForest oversampled	0.9953	0.9906	0.2050	0.0477	0.0524
Gradient Boosting	0.6610	0.4216	0.6660	0.2180	0.0769
Gradient Boosting undersampled	0.8989	0.7316	0.5193	0.1678	0.0700
Gradient Boosting oversampled	0.8989	0.7316	0.5193	0.1678	0.0700
Neural network	0.6478	0.4281	0.6050	0.2017	0.0293
Neural network undersampled	0.5810	0.4028	1.5829	0.5150	0.1807
Neural network oversampled	0.6937	0.5220	0.7064	0.2523	0.0611

Для оценки качества модели, которая предсказывает вероятность исхода, использовались специализированные метрики, которые учитывают вероятностные предсказания модели:

- ROC-AUC (площадь под ROC-кривой) [17]: Эта метрика оценивает способность модели отделять положительные и отрицательные классы в зависимости от порога вероятности. ROC-кривая показывает изменение true positive rate (чувствительности) по отношению к false positive rate (специфичности) при изменении порога вероятности. ROC-AUC показывает общую производительность модели независимо от выбранного порога. ROC-AUC, близкая к 1, означает высокую способность модели различать классы. ROC-AUC, близкая к 0.5, указывает на случайное предсказание классов моделью.

- Precision-Recall-кривая и PR-AUC [18]: В отличие от ROC-AUC PR-AUC оценивает качество модели, сосредоточившись на precision (точности) и recall (полноте). PR-кривая показывает зависимость между precision и recall при различных порогах вероятности. PR-AUC – это площадь под PR-кривой. PR-AUC, близкая к 1, означает высокую точность и полноту модели. PR-AUC, близкая к 0, указывает на низкую точность и полноту модели.
- Log Loss (логарифмическая функция потерь) [19]: Это метрика, которая оценивает точность вероятностных предсказаний модели. Она штрафует модель за неверные предсказания, которые сильно отклоняются от истинных вероятностей. Log Loss, близкая к 0, означает высокое качество вероятностных предсказаний модели. Log Loss, близкая к бесконечности или очень высокое значение, указывает на низкое качество вероятностных предсказаний модели.
- Brier Score (оценка Бриера) [20]: Это метрика, используемая для оценки качества вероятностных предсказаний модели, особенно в задачах бинарной классификации. Эта метрика оценивает разницу между предсказанной вероятностью и истинным классом для каждого образца данных. Brier Score, близкая к 0, означает высокое качество вероятностных предсказаний модели. Brier Score, близкая к 1, указывает на низкое качество вероятностных предсказаний модели.
- Calibration Curve MSE (Среднеквадратичная ошибка калибровочной кривой) [21]. Эта метрика используется для оценки качества калибровки вероятностных предсказаний модели. Она позволяет оценить, насколько хорошо модель калибрует свои вероятностные предсказания, то есть насколько близко предсказанные вероятности соответствуют фактическим частотам событий. MSE, близкая к 0, означает маленькую разницу между предсказанными и истинными вероятностями. MSE, близкая к бесконечности или очень высокое значение, указывает на плохую калибровку модели.

Из предоставленных данных следует, что модель Random Forest (случайный лес) демонстрирует наилучшие результаты по нескольким метрикам:

- ROC-AUC и PR-AUC: Для всех трех случаев (обычной модели, модели с уменьшенной выборкой и модели с увеличенной выборкой) Random Forest показывает наивысшие значения ROC-AUC и PR-AUC (рис. 1). Это говорит о том, что модели случайного леса лучше всего разделяют классы и обладают наивысшей точностью предсказаний.
- Log Loss и Brier Score: Модели случайного леса также демонстрируют наименьшие значения Log Loss и Brier Score по сравнению с другими моделями. Это означает, что модель имеет хорошую калибровку и точность предсказаний.
- Calibration Curve MSE: Наименьшее значение Calibration Curve MSE у модели нейронной сети, что подтверждает ее калибровку и точность. Однако и у моделей случайного леса также достаточно низкие значения (рис. 2).

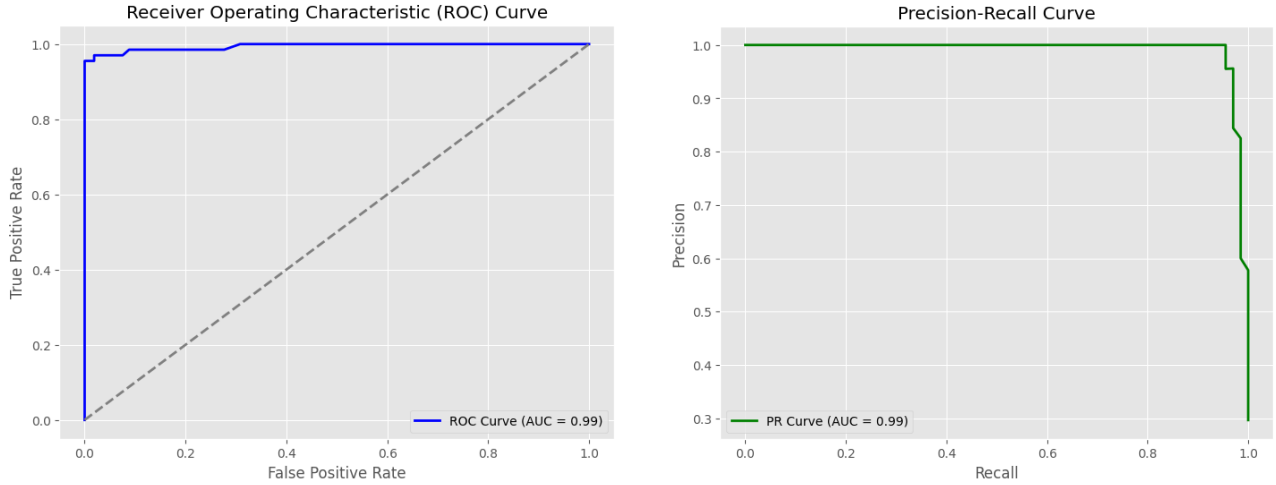

Рис. 1 ROC- и PR-кривые модели случайного леса, обученной на увеличенной выборке.

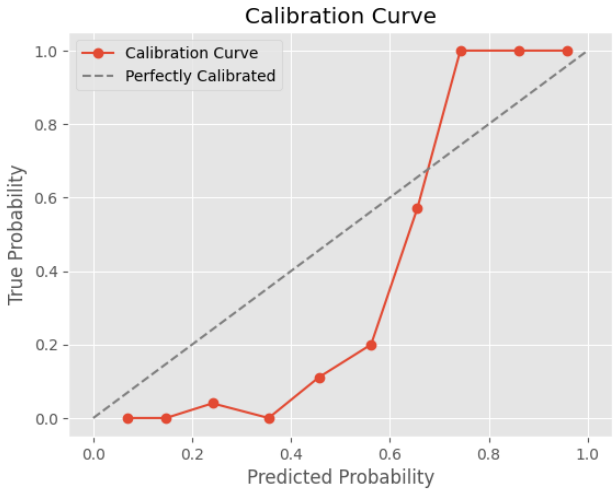

Рис. 2 Калибровочная кривая модели случайного леса, обученной на увеличенной выборке.

Анализ полученных результатов говорит о том, что в случае увеличения объема выборки модели Random Forest дают более качественный результат, что в дальнейшем может иметь большое практическое применение при работе с реальными данными.

Заключение

Таким образом, на основе полученных данных, модели Random Forest, особенно обученные на увеличенной выборке, являются наилучшими среди рассмотренных моделей машинного обучения. Они обладают высокой точностью предсказаний, хорошей калибровкой и способностью эффективно разделять классы в данных.

Благодарности и поддержка исследования

Исследование выполнено в рамках проекта при поддержке гранта РНФ 22-19-00471. Авторы выражают признательность коллегам — участникам проекта, а также другим исследователям за полезные идеи и подходы в области методов машинного обучения [18–21].

Список литературы / References

1. Касаткина Э. П. Сахарный диабет у детей. М.: Медицина, 1990. С. 206-207. [[Kasatkina E. P. Diabetes Mellitus in Children. Moscow: Medicine, 1990, pp. 206–207. (In Rssian).]]
2. Справочник педиатра-эндокринолога / Под. ред. М. А. Жуковского. М.: Медицина, 1992. С. 213-214. [[Handbook of Pediatric Endocrinologist / Ed. ed. M. A. Zhukovsky. Moscow: Medicine, 1992, pp. 213–214. (In Rssian).]]
3. Ефимов А. С., Скробонская Н. А. Клиническая диабетология. К.: Здоровья, 1998. С. 115-117. [[Efimov A. S., Skrobonskaya N. A. Clinical Diabetology. Kiev: Health, 1998, pp. 115–117. (In Rssian).]]
4. Kohner E. M. Diabetic retinopathy // Brit. Med. Bull. 1989. Vol. 5. No. 1. Pp. 148-173.
5. Зиновьев М. С., Нургаянова О. С. Оценка индивидуального риска развития сахарного диабета второго типа и возможных осложнений // СИИТ. 2023. Т. 5. № 4(13). С. 101-110. EDN HIIXFH. [[Zinoviev M. S., Nurgayanova O. S. “Assessment of individual risk of developing type 2 diabetes mellitus and possible complications” // SIIT. 2023. Vol. 5, No. 4(13), pp. 101-110. EDN HIIXFH. (In Rssian).]]
6. Ting Wang (2023). Diabetic_Nephropathy_v1. Dataset // Kaggle. URL: https://www.kaggle.com/datasets/wangting2023/diabetic-nephropathy-v1 (дата обращения: 01.03.2024).
7. Шалфеева Е. А. Методология производства жизнеспособных систем доверительного искусственного интеллекта // СИИТ. 2023. Т. 5. № 4(13). С. 28-49. EDN CJTKQH. [[Shalfeeva E. A. “Methodology for the production of viable systems of trustworthy artificial intelligence” // SIIT. 2023. Vol. 5, No. 4(13), pp. 28-49. EDN CJTKQH. (In Rssian).]]
8. Юсупова Н. И., Нургаянова О. С., Зулкарнеев Р. Х. Формализация этапов риск-анализа в СППР с учетом оценок клинических рисков при бронхолегочных заболеваниях // СИИТ. 2023. Т. 5. № 1(10). С. 11-24. EDN KHIIHT. [[Yusupova N. I., Nurgayanova O. S., Zulkarneev R. Kh. “Formalization of risk analysis stages in decision support system taking into account clinical risk assessments for bronchopulmonary diseases” // SIIT. 2023. Vol. 5, No. 1(10), pp. 11-24. EDN KHIIHT. (In Rssian).]]

9. Шахмаметова Г. Р., Христодуло А. Д., Береговая С. П. Анализ эндокринологических данных на основе моделей классификации // СИИТ. 2022. Т. 4. № 2(9). С. 30-36. EDN LBZVZL. [[Shakhmametova G. R., Khristodullo A. D., Beregovaya S. P. “Analysis of endocrinological data based on classification models” // SIIT. 2022. Vol. 4, No. 2(9), pp. 30-36. EDN LBZVZL. (In Rssian).]]
10. Насыров Р. В. Причинный подход к построению бионических вычислений на основе рекурсивных моделей анализа данных // СИИТ. 2022. Т. 4. № 1(8). С. 27-36. EDN UOMMOU. [[Nasyrov R. V. “Causal approach to the construction of bionic computations based on recursive models of data analysis” // SIIT. 2022. Vol. 4, No. 1(8), pp. 27-36. EDN UOMMOU. (In Rssian).]]
11. Tolles J., Meurer W. J. Logistic regression relating patient characteristics to outcomes // JAMA. 2016. 316 (5): 533–4. doi:10.1001/jama.2016.7653. ISSN 0098-7484. OCLC 6823603312. PMID 27483067.
12. Cortes C., Vapnik V. Support-vector networks // Machine Learning. 1995. 20 (3): 273-297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018. S2CID 206787478.
13. Ho, Tin Kam. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal. QC. 14–16 August 1995. Pp. 278-282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
14. Hastie T., Tibshirani R., Friedman J. H. 10. Boosting and Additive Trees // The Elements of Statistical Learning (2nd ed.). New York: Springer, 2009. Pp. 337-384.
15. Hardesty L. (14 April 2017). Explained: Neural networks // MIT News Office. Retrieved 2 June 2022.
16. Stuart A. Basic Ideas of Scientific Sampling. Hafner Publishing Company. New York, 1962.
17. Zweig M. H., Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine // Clinical Chemistry. 1993. Vol. 39. No. 8. Pp. 561-577. PMID 8472349.
18. Powers D. M. W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. (PDF) // Journal of Machine Learning Technologies. 2011. 2 (1): 37-63.
19. Rosasco L., De Vito E. D., Caponnetto A., Piana M., Verri A. Are loss functions all the same? // Neural Computation. 2004. 16 (5): 1063-1076. CiteSeerX 10.1.1.109.6786.
20. Brier. Verification of forecasts expressed in terms of probability // Monthly Weather Review. 1950. 78 (1): 1-3.
21. Bioanalytical Method Validation. Guidance for Industry. Center for Drug Evaluation and Research. May 2018.

Поступила в редакцию 8 мая 2024 г.

МЕТАДАННЫЕ / METADATA

Title: Assessment of the individual risk of developing type 2 diabetes mellitus and possible complications.

Abstract: The article discusses a study aimed at developing an effective algorithm for predicting the likelihood of developing diabetic retinopathy in patients with diabetes. A number of machine learning methods are presented that are used to create machine learning models that can predict the likelihood of developing diabetic retinopathy in the coming years. The most effective model is determined based on metrics for assessing the effectiveness of machine learning models.

Key words: diabetes mellitus; diabetic retinopathy; quantitative risk assessment; machine learning.

Language: Russian.

Об авторах / About the authors:

ЗИНОВЬЕВ Максим Сергеевич
ФГБОУ ВО «Уфимский университет науки и технологий», Россия.
Магистр ин-та информатики, математики и робототехники. Дипл. инж.-программист (Уфимск. гос. авиац. техн. ун-т, 2003).
E-mail: mr.zmaks@inbox.ru
ORCID: https://orcid.org/0000-0002-0723-9896

НУРГАЯНОВА Ольга Сергеевна
ФГБОУ ВО «Уфимский университет науки и технологий», Россия.
Доц. каф. вычислительной математики и кибернетики. Дипл. инж.-программист (Уфимск. гос. авиац. техн. ун-т, 2003). Канд. техн. наук по системам автоматиз. управления (там же, 2006). Иссл. в обл. новых материалов и искус. интеллекта.
E-mail: onurgayanova@yandex.ru
ORCID: http://orcid.org/0000-0003-2978-3662

ZUNOVYEV Malsim Sergeyevich
Ufa University of Science and Technologies, Russia.
Postgraduate student, Institute of Informatics, Mathematics, and Robotics. .
E-mail: mr.zmaks@inbox.ru
ORCID: https://orcid.org/0000-0002-0723-9896

NURGAYANOVA Olga Sergeyevna
Ufa University of Science and Technologies, Russia.
Docent, Dept. of Computational Mathematics and Cybernetics. Dipl. Programmer Engineer (Ufa State Aviation Tech. Uni., 2003). Cand. of Tech. Sci. (Ufa State Aviation Tech. Uni., 2006). Research: New Materials Development, Artificial Intelligence.
E-mail: onurgayanova@yandex.ru
ORCID: http://orcid.org/0000-0003-2978-3662