РЕШЕНИЕ ЗАДАЧ ДИАГНОСТИКИ ДИАБЕТА С ПРИМЕНЕНИЕМ МАШИННОГО ОБУЧЕНИЯ

Мухамедиева Дилноз Тулкуновна
д.т.н., проф.
Национальный исследовательский университет
«Ташкентский институт инженеров ирригации
и механизации сельского хозяйства», Узбекистан
dilnoz134@rambler.ru
Турғунова Нафиса Махаммаджон кизи
Национальный исследовательский университет
«Ташкентский институт инженеров ирригации
и механизации сельского хозяйства», Узбекистан
Рауфова Мохинур Хайдар кизи
Национальный исследовательский университет
«Ташкентский институт инженеров ирригации
и механизации сельского хозяйства», Узбекистан
Нажмиддинов Ахлиддин Сирожиддин угли
Национальный исследовательский университет
«Ташкентский институт инженеров ирригации
и механизации сельского хозяйства», Узбекистан

Аннотация: В данной работе мы рассмотрим разнообразные применения машинного обучения в задачах, связанных с диагностикой диабета, а также осветим роль машинного обучения в усилении современной медицинской практики и улучшении качества жизни пациентов, страдающих диабетом.

Ключевые слова: машинное обучение, линейная регрессия, модель наивного байесовского классификатора, метод опорных векторов, среднеквадратическая ошибка, коэффициент детерминации.

Аннотация. Ушбу мақолада биз диабетни ташхислаш билан боғлиқ масалаларда машинали ўқитишни турли хил қўлланилишини кўриб чиқамиз, шунингдек, замонавий тиббий амалиётни яхшилаш ва диабет билан касалланган беморларнинг ҳаёт сифатини яхшилашда машинали ўқитишнинг ролини таъкидлаймиз.

Калит сўзлар: машинали ўқитиш, чизиқли регрессия, содда Байес классификатори модели, таянч векторлар усули, ўртача квадратик хато, детерминация коэффициенти.

Annotation: In this paper, we will look at the diverse applications of machine learning in problems related to diabetes diagnosis, and also highlight the role of machine learning in enhancing modern medical practice and improving the quality of life of patients with diabetes.

Keywords: machine learning, linear regression, Naive Bayes classifier model, support vector machine, root mean square error, coefficient of determination.

1.Введение. Диабет - это хроническое заболевание, которое влияет на миллионы людей по всему миру. Для ранней диагностики и эффективного управления диабетом медицинская община все больше обращается к методам машинного обучения и анализу данных. Актуальность машинного обучения в контексте диабета нельзя недооценивать. Это актуальное исследовательское и практическое направление, которое имеет ряд важных аспектов. Машинное обучение предоставляет инструменты для более эффективной диагностики и управления заболеванием. Ранняя диагностика диабета и его типа является ключевым фактором для успешного лечения и предотвращения осложнений. Машинное обучение позволяет разрабатывать модели для ранней диагностики. Каждый пациент уникален, и машинное обучение может помочь создавать индивидуальные планы лечения, учитывая генетические, клинические и лабораторные данные. Машинное обучение позволяет интегрировать и анализировать эти данные, создавая более полное представление о состоянии

пациентов. Осложнения диабета могут быть опасными. Модели машинного обучения могут предсказывать вероятность их развития, что позволяет предпринимать меры для их предотвращения [1-4].

Датасет Diabetes является ценным ресурсом для исследователей и практиков в области медицинского анализа данных и машинного обучения. Набор данных Diabetes включает 10 числовых признаков, которые описывают состояние пациентов. Цель применения машинного обучения в контексте диабета может варьироваться в зависимости от конкретной задачи и сценария, но общие цели включают в себя ранняя диагностика и прогнозирование риска. Целью является создание моделей, которые способны рано диагностировать диабет или предсказывать риск его развития. Это позволяет начать лечение на ранних стадиях заболевания и предотвратить его осложнения [5].

2. Методы.

Для решения задач в области диагностики, управления и исследования диабета могут использоваться различные методы машинного обучения. Вот некоторые из наиболее распространенных методов [6]:

1. Линейная регрессия может быть использована для задачи прогнозирования, например, прогнозирования уровня глюкозы в крови на основе различных параметров.

Линейная регрессия - это метод для предсказания значения зависимой переменной на основе линейной комбинации одной или нескольких независимых переменных. Вот алгоритм для обучения модели линейной регрессии:

2.Модель наивного байесовского классификатора [1].

Загружаем данные Diabetes и преобразуем задачу в бинарную классификацию на основе медианного значения.

Разделяем данные на обучающий и тестовый наборы, а также стандартизируем признаки (наивный байес не требует стандартизации, но мы это сделали для согласованности с предыдущими примерами).

Создаем и обучаем модель наивного байесовского классификатора.

Делаем прогноз на тестовом наборе данных и рассчитываем метрики.

Выводим матрицу путаницы, отчет о классификации и график AUC-ROC.

3.Метод опорных векторов (Support Vector Machines, SVM) [2] SVM может применяться как для задач классификации, так и для регрессии. Он особенно полезен в случаях, когда данные разделяются нелинейно.

Алгоритм Support Vector Machine (SVM) - это метод машинного обучения, используемый для задач классификации и регрессии. Вот общий алгоритм SVM для задачи бинарной классификации:

3.Результаты.

Результаты машинного обучения в области диабета могут быть разнообразными и зависят от конкретной задачи, используемых методов и доступных данных. Вот некоторые из типичных результатов, которые можно достичь с помощью машинного обучения в контексте диабета:

1. Для решения задачи регрессии на наборе данных Diabetes из библиотеки scikit-learn в Python, вы можете использовать следующий алгоритм:

Mean Squared Error: 2900.1732878832318

R-squared: 0.452606602161738

Среднеквадратичная ошибка (Mean Squared Error, MSE) и коэффициент детерминации (R-squared) - это две основные метрики, используемые для оценки производительности модели линейной регрессии. Ваши значения MSE и R-squared говорят о том, насколько хорошо ваша модель соответствует данным и предсказывает зависимую переменную.

Значение R-squared равно 0.4526. Это означает, что ваша модель объясняет примерно 45% вариабельности в данных. То есть модель объясняет менее половины изменений в зависимой переменной, что может быть улучшено с помощью более сложных моделей или дополнительных признаков.

2.Модель наивного байесовского классификатора

Confusion Matrix:

[[37 12]

[13 27]]

Элемент (0,0) (верхний левый угол) представляет собой количество истинно отрицательных (TN) примеров.

Элемент (0,1) (верхний правый угол) представляет собой количество ложноположительных (FP) примеров.

Элемент (1,0) (нижний левый угол) представляет собой количество ложноотрицательных (FN) примеров.

Элемент (1,1) (нижний правый угол) представляет собой количество истинноположительных (TP) примеров.

Classification Report:

	precision	recall	f1-score	support
Class 0	0.74	0.76	0.75	49
Class 1	0.69	0.68	0.68	40
accuracy			0.72	89
macro avg	0.72	0.72	0.72	89
weighted avg	0.72	0.72	0.72	89

Accuracy: 0.7191, Precision: 0.6923, Recall: 0.6750, F1-Score: 0.6835, AUC-ROC: 0.8260.

Classification Report предоставляет детальные метрики для каждого класса (Class 0 и Class 1), а также средние значения (macro avg и weighted avg). В вашем случае:

Precision (точность) измеряет, как много из объектов, которые модель предсказала как положительные, действительно являются положительными. Precision для Class 0 составляет 0.74, а для Class 1 - 0.69.

Recall (полнота) измеряет, как много из всех действительных

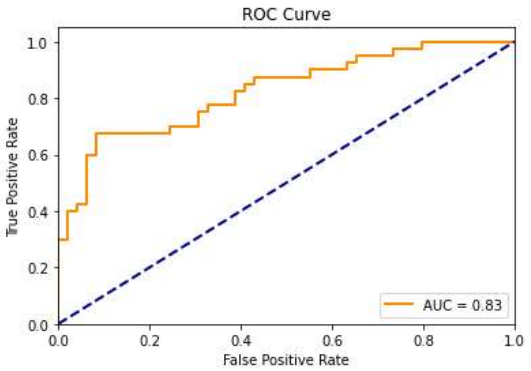
положительных объектов модель правильно классифицировала. Recall для Class 0 составляет 0.76, а для Class 1 - 0.68.

F1-Score - это гармоническое среднее между точностью и полнотой. F1-Score для Class 0 составляет 0.75, а для Class 1 - 0.68.

Accuracy (точность) - доля верно классифицированных примеров относительно общего числа примеров. Accuracy составляет 0.7191, что означает, что модель правильно классифицировала 71.91% всех примеров.

AUC-ROC (Площадь под ROC-кривой):

AUC-ROC измеряет площадь под кривой ROC (Receiver Operating Characteristic), которая отображает производительность модели при различных порогах классификации. Значение AUC-ROC близкое к 1 (ваше значение 0.8260) указывает на хорошую способность модели разделять классы.



3.SVM с линейным ядром. Загружает данные Diabetes и преобразует задачу в бинарную классификацию (на основе медианного значения). Разделяет данные на обучающий и тестовый наборы, а также стандартизирует признаки. Создает и обучает модель SVM с линейным ядром. Делает прогноз на тестовом наборе данных и рассчитывает метрики путаницы, accuracy, precision, recall, F1-score и AUC-ROC. Строит график ROC-кривой для визуализации производительности модели.

Confusion Matrix:

[[35 14]

[10 30]]

Classification Report:

	precision	recall	f1-score	support
Class 0	0.78	0.71	0.74	49
Class 1	0.68	0.75	0.71	40
accuracy			0.73	89
macro avg	0.73	0.73	0.73	89
weighted avg	0.73	0.73	0.73	89

Accuracy: 0.7303

Precision: 0.6818

Recall: 0.7500

F1-Score: 0.7143

AUC-ROC: 0.8398

Матрица путаницы показывает количество верно и ошибочно классифицированных примеров для каждого класса.

Classification Report (Отчет о классификации):

Classification Report предоставляет детальные метрики для каждого класса (Class 0 и Class 1), а также средние значения (macro avg и weighted avg). В вашем случае:

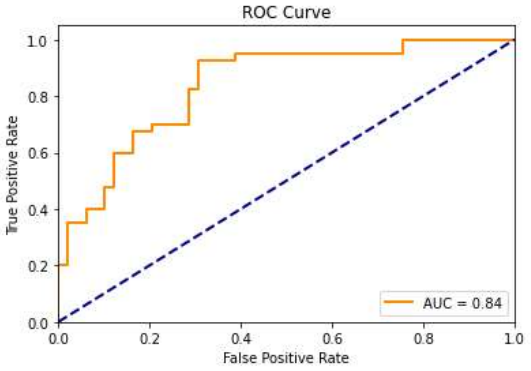
Precision (точность) измеряет, как много из объектов, которые модель предсказала как положительные, действительно являются положительными. Precision для Class 0 составляет 0.78, а для Class 1 - 0.68.

Recall (полнота) измеряет, как много из всех действительных положительных объектов модель правильно классифицировала. Recall для Class 0 составляет 0.71, а для Class 1 - 0.75.

F1-Score - это гармоническое среднее между точностью и полнотой. F1-Score для Class 0 составляет 0.74, а для Class 1 - 0.71.

Accuracy (точность) - доля верно классифицированных примеров относительно общего числа примеров. Accuracy составляет 0.7303, что означает, что модель правильно классифицировала 73.03% всех примеров.

AUC-ROC (Площадь под ROC-кривой):



Значение AUC-ROC близкое к 1 (ваше значение 0.8398) указывает на хорошую способность модели разделять классы.

4. Заключение. Машинное обучение позволяет диагностировать диабет с высокой точностью и рано выявлять риски развития заболевания. Это способствует раннему началу лечения и предотвращению осложнений.

Модели машинного обучения могут предсказывать вероятность осложнений и предостерегать врачей и пациентов. Это способствует более активному мониторингу и уходу. Машинное обучение представляет собой мощный инструмент для борьбы с диабетом, и его применение продолжит развиваться, создавая новые возможности для улучшения ухода за пациентами и научных исследований в этой области.

Литература

1.Pethunachiyar GA. Classification of diabetes patients using kernel based support vector machines. In: 2020 International Conference on Computer Communication Informatics (ICCCI). Coimbatore: IEEE (2020). p. 1–4. doi: 10.1109/ICCCI48352.2020.9104185

2. Gupta S, Verma HK, Bhardwaj D. Classification of diabetes using naïve bayes and support vector machine as a technique. In: Sachdeva A, Kumar P, Yadav OP, Garg RK, Gupta A, editors. Operations Management and Systems Engineering. Singapore: Springer (2021). p. 365–76. doi: 10.1007/978-981-15-6017-0_24

3. Рашка, С. Python и машинное обучение [Текст] / С. Рашка. – М. : ДМК Пресс, 2017. – 418 с.

4.Khattak A, Habib A, Asghar MZ, Subhan F, Razzak I, Habib A. Applying deep neural networks for user intention identification. Soft Comput. (2021) 25:2191–220. doi: 10.1007/s00500-020-05290-z

5. Chen, L, Magliano, DJ and Zimmet, PZ. (2011) The worldwide epidemiology of type 2 diabetes mellitus-present and future perspectives.// Nat Rev Endocrinol 8: 228-236

6. Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, Sherazi HHR. Machine learning based diabetes classification and prediction for healthcare applications. J Healthcare Eng. (2021) 2021:9930985. doi: 10.1155/2021/9930985