РЕЗЮМЕ

Сердечно-сосудистые заболевания занимают большую долю в заболеваемости людей и находятся в статистике по уровню смертности в первой тройке мест. Для выявления этих заболеваний применяют специальные тесты, однако начало заболевания не всегда возможно точно определить ввиду неспецифичности симптомов и их слабой выраженности.

Язык Python – универсальный современные язык программирования, который все больше используется в мире для решения медицинских задач. В частности, в Великобритании на текущий момент 32% медицинских приложений выпускаются с использованием языка Python. Задачей современной медицины на стыке с информационными технологиями является получение информации из разных источников, объединение этой информации и использование ее для построения моделей машинного обучения с целью раннего выявления сердечно-сосудистых заболеваний для снижения смертности людей от них.

Гипотезой исследования является широкая применимость языка Python для выявления болезней сердца. Альтернативной гипотезой является отсутствие пользы Python в диагностике болезней сердца.

В литературе описаны работы, в которых успешно используется Python для выявления болезней сердца на основании ряда параметров. Точность некоторых моделей достигает 87% (метод K-NN), 83% (вспомогательный вектор), 84% (случайный лес), 79% (дерево решений). Язык Python позволяет шифровать данные пациентов, используя специальные пакеты из области кибербезопасности. В литературе не описаны точные методики создания моделей для предсказания заболевания сердца.

В данной публикации детально описаны этапы построения модели обучения, базирующейся на ансамблевой методике «Случайный лес», в которой в качестве ветвей обучения представлено определенное количество деревьев решений, которые объединяются в формировании классификации данных. Методика идеально подходит для задачи классификации, которая и поставлена (детекция болезни сердца).

Анализ данных разбит на следующие этапы:

1. Импорт библиотек
2. Импорт данных
3. Эксплоративный анализ данных (исследование данных)
4. Оценка данных
5. Очистка данных
6. Оценка взаимоотношений между исследуемыми признаками
7. Преобразование и подготовка данных для построения модели
8. Построение модели машинного обучения
9. Оценка и подбор гиперпараметров модели
10. Деплой модели

Авторы разбирают математические механизмы основных методов машинного обучения, используемых для задачи классификации: метод ближайших соседей, деревья решений, метод поддерживающих векторов, случайный лес, логистическая регрессия. Авторы подчеркивают наибольшую универсальность и высокую точность метода «случайный лес», которая достигает 83%.

Выводы:

Язык Python – универсальный язык, включающий в себя объектно-ориентированный подход к программированию, высокоуровневый язык, имеющий множество пакетов для анализа мед. данных и построения моделей машинного обучения.

Предсказание болезни сердца на основании определенных параметров позволяет быстро и эффективно выявить наличие болезни на ранних стадиях, что улучшает прогноз для пациентов и снижает смертность.

Применение для задачи классификации в медицине алгоритма машинного обучения «Случайный лес» позволяет достичь удовлетворительной точности предсказания заболевания сердца (83%).

Применение языка Python в анализе медицинских данных и построение моделей машинного обучения для предсказания заболевания сердца с использованием достоверных и точных данных потенциально может снизить применение инвазивных методик определения заболевания – таких, как ангиография. Соответственно снизится количество вероятных осложнений, связанных с проведением инвазивных процедур.