УДК 004.852

ГЕНЕРАЦИЯ СИНТЕТИЧЕСКОЙ ПОПУЛЯЦИИ ПАЦИЕНТОВ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ПРОГНОЗА ЦЕЛЕВЫХ ПОКАЗАТЕЛЕЙ ЛЕЧЕНИЯ COVID-19

Глазнев А.Ю. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»),
Деревицкий И.В. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»),
Научный руководитель – к. т. н., доцент Ковальчук С.В. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»),

COVID-19 — потенциально тяжёлая острая респираторная инфекция, вызываемая коронавирусом SARS-CoV-2 (2019-nCoV). Представляет собой опасное заболевание, которое может протекать как в форме острой респираторной вирусной инфекции лёгкого течения, так и в тяжёлой форме. Один из показателей течения болезни это длительность его лечения. Заблаговременное и качественное прогнозирование данного показателя снизит нагрузку на систему здравоохранения путём оптимизации траектории лечения пациентов, находящихся в стационарах или на амбулаторном лечении.

В данной работе сравнивалась точность двух регрессионных моделей для предсказания длительности лечения пациентов с COVID-19. Были разработаны три варианта обучения данных моделей: один вариант без методов предварительной обработки данных и два варианта с разными методами обработки данных и последующем использованием алгоритма генерации синтетической популяции пациентов. Первый подход заключался в обучении моделей XGBoost и Lasso на имеющихся данных с последующей оценкой этих моделей. Второй подход состоит из следующих этапов: были заполнены имеющиеся пропуски в датасете с помощью медианных значений, далее был применён алгоритм SDV для генерации синтетических данных, обучены те же самые регрессионные модели на новых данных и оценена точность обученных моделей. Третий подход подразумевал заполнение пропусков в исходном датасете используя алгоритм MISE, применение алгоритма SDV на полученных данных и обучение моделей с последующей их оценкой. В заключении работы метрики каждого метода обработки были сопоставлены.

Выводы. В результате работы было установлено, что модели, построенные на данных с заполнением пропусков с помощью алгоритма MISE и последующей генерацией синтетической популяции пациентов обладали более высокой точностью в сравнении с другими разработанными вариантами.

Глазнев А.Ю. (автор) Подпись

Ковальчук С.В. (научный руководитель) Подпись