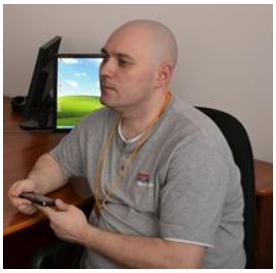
УДК 004.852

ПРОГНОЗИРОВАНИЕ СМЕРТНОСТИ У ПАЦИЕНТОВ С COVID-19 В УСЛОВИЯХ СТАЦИОНАРА С ИСПОЛЬЗОВАНИЕМ БИБЛИОТЕКИ CONFORMAL PREDICTION В KNIME



М.И. Гальченко
старший статистический аналитик HiQo Solutions Ltd; старший преподаватель ФГБОУ ВО СПбГАУ

О.Я. Порембская
кандидат медицинских наук, доцент кафедры сердечно-сосудистой хирургии СЗГМУ им. И.И.Мечникова

HiQo Solutions, Ltd, USA
Кафедра Электроэнергетики и Электротехники
ФГБОУ ВО Санкт-Петербургский Государственный Аграрный Университет, РФ
СЗГМУ им. И.И.Мечникова
E-mail: maxim.galchenko@gmail.com

М.И. Гальченко
Окончил Санкт-Петербургский Государственный Университет. Старший статистический аналитик в HiQo Solutions Ltd., старший преподаватель Санкт-Петербургского Государственного Аграрного Университета.

О.Я. Порембская
Направление деятяельности: флебология. В настоящее время работает хирургом и доцентом кафедры сердечно-сосудистой хирургии СЗГМУ им. И.И.Мечникова, научным сотрудником «Института экспериментальной медицины»

Аннотация. Прогнозирование исхода COVID-19 у стационарных больных изначально представляло серьёзную проблему. В статье анализируется набор данных, собранный на основе двух стационаров (ФГБОУ ВО «СЗГМУ им. И.И. Мечникова» Клиника имени Петра Великого и ФГБУ «Клиническая больница №1» Управления делами Президента РФ (Волынская больница) г. Москвы), всего 313 пациентов, из которых умерло 61 (19.5%). Для прогнозирования использовался KNIME Analytics Platform, библиотека Conformal Prediction. В статье показываются преимущества данного метода, рассматривается вопрос отбора базового алгоритма для конформного прогнозирования (Мондриановской классификации)

Ключевые слова: *конформное прогнозирование, смертность, COVID-19, Мондриановская классификация*

Введение.

Пандемия COVID-19 вызвала бурный рост публикаций, связанных с прогнозированием. Авторы прогнозов решали две задачи: изучение динамики процесса (пики заболеваемости, количество умерших) [1, 2], вторая — прогнозирование исхода, риска смерти в стационарных условиях [3-5]. Особый интерес представляет второй случай, так как выявление маркеров тяжёлого течения, высокого риска смерти важно с точки зрения стратегии лечения.

Авторы исследований используют достаточно большие объёмы данных (250 [5], 2782 [4] и 11 807 [3] пациентов в отобранных работах). В качестве предикторов выступаю данные

лабораторных исследований, в основном результаты анализов крови.

Алгоритмы, используемые в отобранных исследованиях, относятся к различным классам:

- Рекуррентная нейронная сеть GRU-D (AUC 0.938) [3]
- Random forest, GBM, логистическая регрессия (0.93) [4]
- SIMPLS (AUC > 0.85) [5]

Наиболее важные показатели, выделяемые авторами:

- возраст, индекс коморбидности Чарльсона, минимальная сатурация кислорода, уровни фибриногена и уровень сывороточного железа [3];
- ЛДГ, D-димер, нейтр/лимфа, нейтрофилы %, фибриноген, СРБ, рентгенограмма грудной клетки Brescia, лимфоциты %, ферритин стандарт, моноциты %[4];
- ИБС, диабет, возраст > 65 лет, AMS, деменция (топ 5) [5]

Таким образом, можно заключить, что достаточно высокие показатели качества модели могут быть достигнуты при использовании хорошо известных показателей. Однако, стоит отметить и то, что часто недоступны показатели лабораторных исследований в динамике, а некоторые из них могут быть достаточно дорогостоящими. Применяемые алгоритмы достаточно хорошо описаны, но, при этом, стоит отметить то, что различия в протоколах лечения может существенно исказить картину. Кроме того, задача классификации в её классическом виде, не выявляет пограничных состояний, то есть состояний требующих особого внимания, когда в реальности состояние пациента может приблизительно равновероятно отнести его как к классу «Выжил», так и к классу «Умер».

В такой ситуации выглядит оправданным конформное прогнозирование. «Конформное прогнозирование использует прошлый опыт для определения точных уровней достоверности прогнозов. Учитывая вероятность ошибки ε, метод, который делает предсказание y’ метки y, конформное прогнозирование создает набор меток, обычно содержащих y’, который также содержит y с вероятностью 1 − ε» [6].

Особенности и ограничения метода:

- Данные должны быть неупорядочены.
- Распределения в тестовых и обучающих данных могут различаться, но шаблоны связей между предикторами и целевыми переменными должны быть такими же.
- В качестве базового метода прогнозирования может выступать любой, дающий вероятности принадлежности к классам отдельного наблюдения (в случае классификации).
- Не требуется коррекции выборки в случае скошенности в целевой переменной.

Фактически, конформное прогнозирование — надстройка над известными методами прогнозирования, позволяющая учитывать его неопределённость. В статье исследуется применение конформного прогнозирования (а именно Мондриановской классификации) с использованием KNIME к данной задаче.

Материалы и методы.

Исходный набор данных был собран в 2021-2022 году в ФГБОУ ВО «СЗГМУ им. И.И. Мечникова» Клиника имени Петра Великого и ФГБУ «Клиническая больница №1» Управления делами Президента РФ (Волынская больница) г. Москвы Набор данных содержит информацию о 313 пациентах, заболевание которых было подтверждено ПЦР-тестом на COVID-19. Пациенты проходили стационарное лечение. Исход определялся как «выжил» либо «умер». Количество умерших составляет 19.5% (61 пациент). В качестве изучаемых предикторов были отобраны половозрастные характеристики, а также ряд показателей лабораторных исследований крови (табл. 1).

Таблица 1. Показатели, используемые в моделировании

Показатель	**Описание**
Возраст	Возраст пациента, лет
Пол	Пол пациента
ИМТ	Индекс массы тела
Лейкоциты	Группы: < $3x10^9$/л или ≥$3x10^9$/л
Лимфоциты	Группы: < $1x10^9$/л или ≥$1x10^9$/л
Тромбоциты	Группы: >$100x10^9$/л — 0 , <$100x10^9$/л — 1, <$50x10^9$/л – 2
Д-Димер, группы	Группы:Д-димер <500 мкг/л, Д-димер >500 и ≤1500; Д-димер >1500 и ≤3000; Д-димер >3000
Фибриноген, группы	Группы: 2-4 г/л; <2 г/л; >4 и ≤5 г/л; >5 и ≤9 г/л; >9 г/л
АЛТ	Группы: <40 Ед/л, ≥40 и <100 Ед/л, ≥100 и <200 Ед/л, ≥200 и <300 Ед/л, ≥300 и <400 Ед/л, ≥400 Ед/л
АСТ	Группы: <40 Ед/л, ≥40 и <100 Ед/л, ≥100-150 Ед/л, ≥150 Ед/л
Ферритин	Группы:<200 мкг/л — 0, ≥200 и <500 мкг/л – 1, ≥500 и <1000 мкг/л – 2, ≥1000 и <1500 мкг/л – 3, ≥1500 и <2000 мкг/л — 4, ≥2000 мкг/л -5

Для изучения возможности прогнозирования использовался KNIME Analytics Platform, с установленной библиотекой “Conformal Prediction” компании Redfield AB (Швеция).

Результаты.

Важность факторов была оценена с помощью Predictive Power Score (PPS) с использованием логистической регрессии (рис. 1).

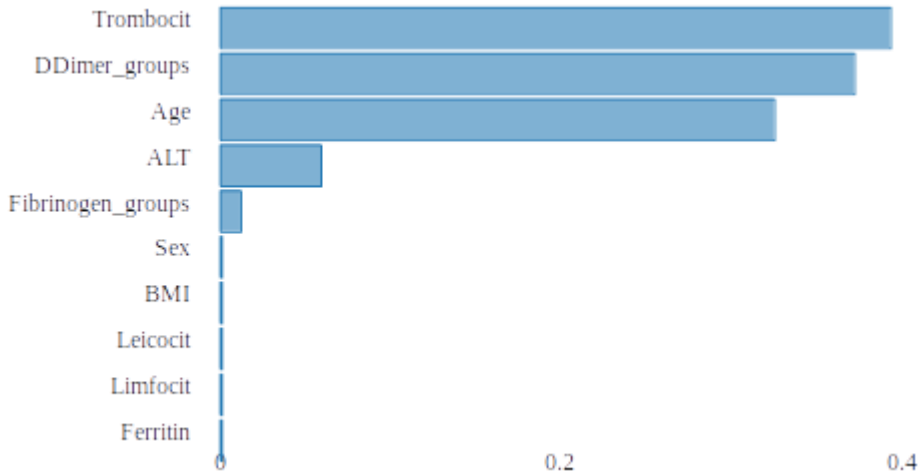


Рисунок 1. Относительная важность факторов, PPS.

Как видно, наиболее перспективными для прогнозирования являются показатели уровня тромбоцитов (0.393), уровень Д-Димера (0.372), возраст (0.325), АЛТ (0.059), уровень фибриногена (0.012). Частично показатели, используемые в данном исследовании,

отмечались и ранее другими авторами как наиболее важными в исследованиях, описанных во введении. Стоит отметить факторы Д-Димер, Фибриноген, возраст, которые входят в эти списки. Уровень тромбоцитов авторами прошлых исследований как один из наиболее важных не отмечался.

В качестве перспективных алгоритмов для моделирования были отобраны: логистическая регрессия, Gradient Boosted Trees, Random Forest и Bayesian Net. Так как наблюдается значительная скошенность в классах, предварительно обучающая выборка для Gradient Boosted Trees исправлялась с помощью алгоритма SMOTE. Тестовая выборка составляет 30% от исходной. Кросс-валидация проводилась на 10 выборках (средний размер — 50 экземпляров). Результаты моделирования (таблица 2) позволили отобрать в качестве базовых алгоритмов для конформного прогнозирования Bayesian Net и Gradient Boosted Trees Learner.

Таблица 2. Показатели качества моделей, выделены наилучшие результаты

Показатель	**Логистическая регрессия**	**Gradient Boosted Trees Learner**	**Random Forest**	**Bayesian Net**
Ошибка кросс-валидации	11.5%	**9.3%**	11.8%	**9.3%**
Точность, %	84.04%	**89.36%**	88.30%	**89.36%**
F (класс «Умер»)	66.67%	73.68%	64.52%	**77.27%**
Чувствительность (класс «Умер»), %	83.33%	77.78%	55.56%	**94.44%**
Специфичность (класс «Умер»), %	84.21%	**92.11%**	96.05%	88.16%
Каппа Коэна	0.567	0.670	0.577	**0.706**

Bayesian Net выглядит предпочтительнее за счёт высокой чувствительности по класу «Умер» (99.4%) и высокой каппы Коэна (0.706), показывающей хорошую согласованность результатов.

Конформное прогнозирование предполагает создание двух объектов для дальнейшего прогнозирования: калибровочной таблицы и модели, на основании которой она создана. Калибровочная таблица является некоторым аналогом валидационного набора данных, она выделяется из обучающей выборки и подаётся на вход для прогнозирования. Помимо прогнозируемого класса, полученного на основе модели, калибровочная таблица дополняется вероятностью принадлежности к реальному классу, полученной в результате прогнозирования. Калибровочная таблица ранжирована по классам и вероятностям принадлежности к реальному классу. Так как необходимо достигнуть устойчивости при прогнозировании процедура построения модели и калибровочной таблицы выполняется заданное количество раз (формируется цикл), деление обучающего набора на непосредственно обучающий и калибровочный осуществляется в заданном процентном соотношении. Для выполнения этих операций используются специфические узлы Conformal Calibration Loop Start и Conformal Calibration Loop End, Conformal Calibrator и узлы выбранного алгоритма прогнозирования, узел формирования таблицы моделей (рис. 2).

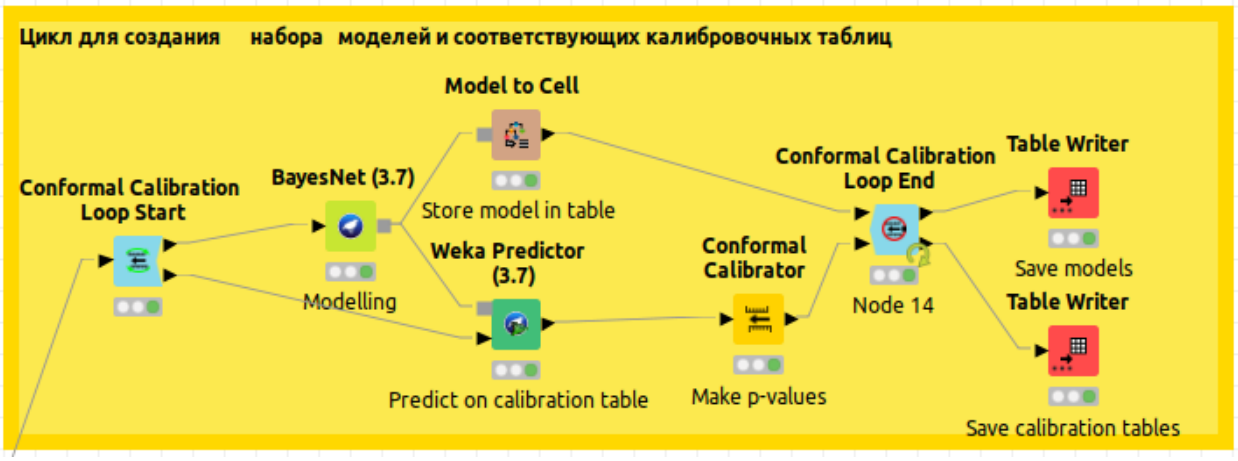


Рисунок 2. Фрагмент потока KNIME: цикл для построения объектов конформного прогнозирования (Мондриановская классификация).

Результаты могут быть сохранены в таблицах KNIME (внутренний формат, позволяющий использовать табличную структуру для сложных объектов), что даёт возможность эффективно разделять обучение и прогнозирование.

Прогнозирование осуществляется с применением Conformal Prediction Loop Start и Conformal Calibration Loop End, Conformal Predictor, Conformal Classifier и узлjd выбранного алгоритма прогнозирования (рис. 3). Особо стоит отметить необходимость задания порогового p-value, которое определяется как некоторое стандартизованное значение, либо из соображений оптимизации показателей качества. В данном исследовании пороговое значение фиксировано, равно 0.05.

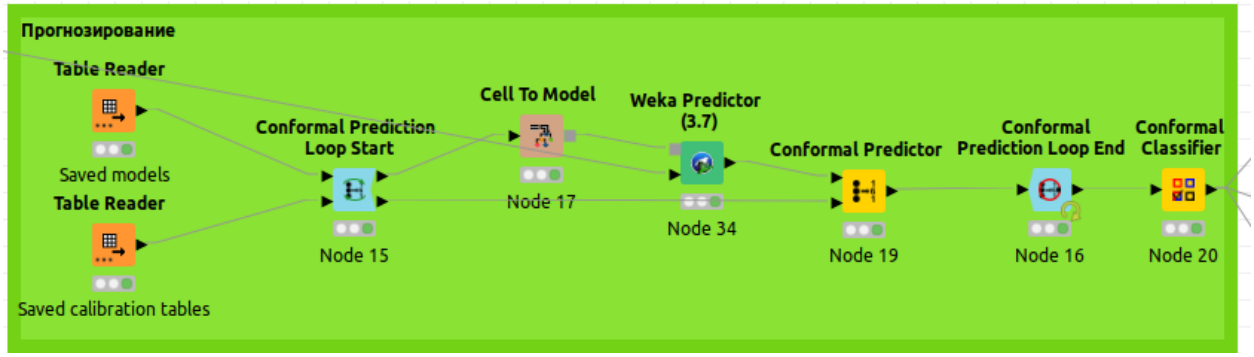


Рисунок 3. Фрагмент потока KNIME: цикл для прогнозирования на тестовой выборке/новых данных.

Для оценки качества прогноза для Мондриановской классификации определяются два показателя: Эффективность (Efficiency) - доля однозначно классифицированных записей (вне зависимости от корректности), Валидность (Validity) - доля корректно классифицированных записей. Необходимо учитывать, что если реальное значение входит в набор меток, получаемых в результате конформной классификации, прогноз считается верным. Эффективность и валидность вычисляется по классам [6]. Так как запись в тестовом наборе данных может принадлежать к нескольким классам одновременно в результате классификации классические методы оценки дают некорректные результаты.

В данном случае для класса «Выжил» эффективность 84%, валидность 98%. Для класса «Умер» эффективность 58%, валидность 92%. Если валидность расчитывать по выборке в целом и интерпретировать как точность в условиях неопределённости, можно говорить о повышении качества прогноза относительно полученной точности в 89% ранее. Неверно (однозначно) классифицированы только две записи из 63 в тестовой выборке, что соответствует 97%. Для двух пациентов с однозначной ошибочной классификацией были вычислены значения вектора Шепли (SHAP values) с использованием классического прогнозирования с применением Bayesian Net (рис. 4).

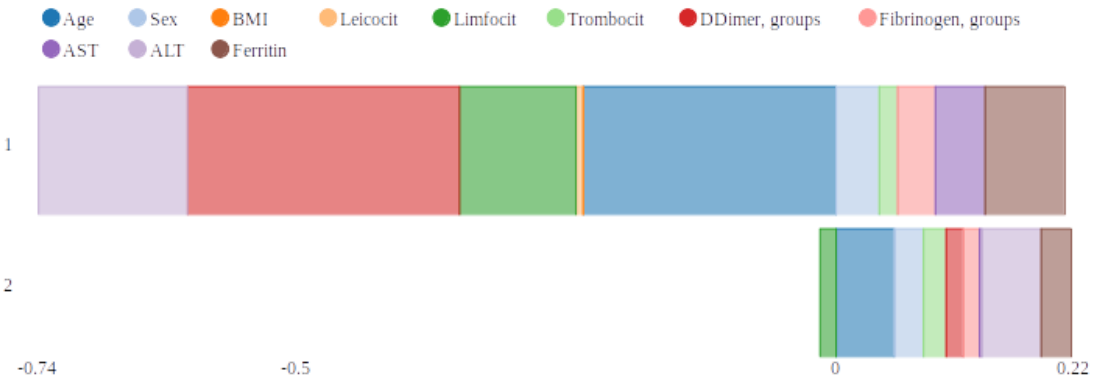


Рисунок 4. Пациент 1 выжил, хотя модель отнесла его к классу «Умер» (отрицательные значения для: Д-Димер = 5, Возраст = 78, АЛТ =0 и Лимфоциты = 1), Пациент 2 умер, значение отрицательно только для показателя Лимфоциты = 1.

Для пациента 1, ошибочно отнесённого к классу умерших высок риск летального исхода. Высокий Д-димер, фибриноген, в сочетании с ожирением. Но органной недостаточности нет, что говорит в пользу благоприятного исхода. Для пациента 2, всё же, сочетание показателей, плюс ожирение говорит в пользу неблагоприятного прогноза. Для данного пациента необходимо дополнительное исследование факторов, их соответствия прочим случаям неблагоприятного течения заболевания.

Для случаев с неопределённым исходом, когда вероятность смерти и выживания выше порогового значения, анализ показателей действительно выявляет интерпретируемые как негативные комбинации с клинической точки зрения. Так, для части пациентов комбинация отдельных факторов говорит об остром воспалительном процессе, прочих рисках. Таким образом, использование конформного прогнозирования вполне себя оправдывает.

Заключение. Конформное прогнозирование позволяет улучшить качество модели, применение связки Bayesian Net и конформного прогнозирования выглядит многообещающе в рамках данной задачи. Возникающая неопределённость в прогнозируемом исходе даёт возможность реагировать более гибким образом на возникающую возможность смертельного исхода. KNIME может быть использован как инструмент построения потока для прогнозирования исхода и сохранения результатов в привычнов виде для врачей, а именно рабочих книг формата XLSX.

Список использованных источников

[1] Barcellos, D.d.S., Fernandes, G.M.K. & de Souza, F.T. Data based model for predicting COVID-19 morbidity and mortality in metropolis. Sci Rep 11, 24491 (2021). https://doi.org/10.1038/s41598-021-04029-6 (URL: https://www.nature.com/articles/s41598-021-04029-6)

[2] Zulfany Erlisa Rasjid, Reina Setiawan, Andy Effendi, A Comparison: Prediction of Death and Infected COVID-19 Cases in Indonesia Using Time Series Smoothing and LSTM Neural Network, Procedia Computer Science, Volume 179, 2021, Pages 982-988, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.01.102. (URL: https://www.sciencedirect.com/science/article/pii/S1877050921001368)

[3] Sankaranarayanan S, Balan J, Walsh JR, Wu Y, Minnich S, Piazza A, Osborne C, Oliver GR, Lesko J, Bates KL, Khezeli K, Block DR, DiGuardo M, Kreuter J, O'Horo JC, Kalantari J, Klee EW, Salama ME, Kipp B, Morice WG, Jenkinson G COVID-19 Mortality Prediction From Deep Learning in a Large Multistate Electronic Health Record and Laboratory Information System Data Set: Algorithm Development and Validation J Med Internet Res 2021;23(9):e30157 doi: 10.2196/30157 (URL: https://www.jmir.org/2021/9/e30157)

[4] Garrafa E, Vezzoli M, Ravanelli M, et al. Early prediction of in-hospital death of COVID-19 patients: a machine-learning model based on age, blood analyses, and chest x-ray score. Elife. 2021;10:e70640. Published 2021 Oct 18. doi:10.7554/eLife.70640

[5] Banoei, M.M., Dinparastisaleh, R., Zadeh, A.V. et al. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. Crit Care 25, 328 (2021). https://doi.org/10.1186/s13054-021-03749-5

[6] Shafer G., Vovk V. A tutorial on conformal prediction //Journal of Machine Learning Research. – 2008. – Т. 9. – №. Mar. – С. 371-421.

PREDICTION OF MORTALITY FOR COVID-19 PATIENTS IN A HOSPITAL SETTING USING THE «CONFORMAL PREDICTION» LIBRARY IN KNIME

M.I. GALCHENKO
Senior Statistical Analyst at HiQo Solutions Ltd; Senior Lecturer, St. Petersburg State Agrarian University

O.Ya. POREMBSKAYA,
MD, Associate Professor of the Department of Cardiovascular Surgery, North-Western State Medical University. I.I. Mechnikova

Department of Power Engineering and Electrical Engineering
FSBEI HE St. Petersburg State Agrarian University, Russian Federation
HiQo Solutions, Ltd, USA
North-Western State Medical University. I.I. Mechnikova
E-mail: maxim.galchenko@gmail.com

Abstract. Predicting the mortality for COVID-19 patients was initially a challenge. The article analyzes a data set collected on the basis of two hospitals (FGBOU VO "North-Western State Medical University named after I.I. Mechnikov" Clinic named after Peter the Great and FGBU "Clinical Hospital No. 1" of the Administration of the President of the Russian Federation (Volynskaya Hospital) in Moscow), total 313 patients, of which 61 (19.5%) died. KNIME Analytics Platform, Conformal Prediction library was used for forecasting. The article shows the advantages of this method, considers the issue of selecting a basic algorithm for conformal forecasting (Mondrian classification)

Keywords: conformal prediction, mortality, COVID-19, Mondrian classification