УДК [004.774+613]:004.85

ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОКАЗАТЕЛЕЙ ЗДОРОВЬЯ С ПРИМЕНЕНИЕМ МЕХАНИЗМОВ МАШИННОГО ОБУЧЕНИЯ И ЕГО ЭРГОНОМИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

Н.А. Ванецкий
Аспирант кафедры инженерной психологии и эргономики, БГУИР

Д.А. Кислова
Студент БГУИР, кафедра инженерной психологии и эргономики
darya.student@gmail.com

А.Н. Василькова
Старший преподаватель кафедры инженерной психологии и эргономики, БГУИР
a.vasilkova@bsuir.by

Н.А. Ванецкий
Образование: Высшее; Магистратура, специальность: 7-06-1021-01 Охрана труда и эргономика (Профилизация: Управление безопасностью производственных процессов); Аспирантура, специальность: 19.00.03 - психология труда, инженерная психология, эргономика (по настоящее время);
Область профессиональных интересов / исследований: Психология труда. Инженерная психология. Эргономика. Психология управления. Юридическая психология.

Д.А. Кислова
Студентка кафедры инженерной психологии и эргономики БГУИР.
Область профессиональных интересов / исследований: языки программирования, искусственный интеллект, технологии виртуальной реальности.

А.Н. Василькова
Старший преподаватель кафедры инженерной психологии и эргономики.
Образование: 2007 - МГВРК по специальности «Программное обеспечение информационных технологий»,
2022 - магистратура БГУИР по специальности «Охрана труда и эргономика».
Область профессиональных интересов / исследований: языки программирования, искусственный интеллект, технологии виртуальной реальности.

Аннотация. В данной работе осуществлен аналитический обзор применения статистических методов и алгоритмов машинного обучения для анализа временных рядов данных о показателях здоровья. Основное внимание уделяется использованию моделей *ARIMA*, линейной регрессии, *SVR* (*Support Vector Regression*), *KNN* (*K-Nearest Neighbors*) и *Random Forest*. Подчеркивается, как применение этих методов способствует повышению точности и персонализации прогнозов здоровья, учитывая сложность и многообразие медицинских данных.

Описывается процесс оценки эффективности этих методов на основе различных метрик, включая среднюю абсолютную ошибку, среднеквадратическую ошибку и индекс рассеивания, что позволяет подчеркнуть значимость их использования в современных медицинских приложениях. Работа демонстрирует важность интеграции передовых технологий анализа данных в области здравоохранения для улучшения качества жизни пациентов.

Ключевые слова: временные ряды, показатели здоровья, *ARIMA*, машинное обучение, веб-приложения для мониторинга здоровья, медицинская диагностика.

Введение. В эпоху цифровизации медицины, методы прогнозирования показателей здоровья на основе анализа больших данных играют ведущую роль в предупреждении заболеваний и управлении лечением. В этой работе рассматривается использование различных статистических и машинно-обучающих моделей для анализа временных рядов сердечного ритма, полученных с помощью устройств актиграфии. Модель *ARIMA*, техники машинного обучения, такие как линейная регрессия, *SVR* и *KNN*, а также ансамблевые методы, включая *Random Forest*, представляют собой основные инструменты для выявления и прогнозирования ключевых показателей здоровья из множества доступных биомедицинских данных.

Основной акцент в исследовании делается на адаптации этих моделей для решения специфических задач регрессии временных рядов, актуальных в сфере здравоохранения. Важным этапом работы является корректное разделение данных на обучающую и тестовую выборки, что позволяет оценить способность моделей к обобщению и предотвращает риск переобучения. Эффективность моделей оценивается с помощью таких показателей, как средняя абсолютная ошибка, среднеквадратическая ошибка и корень из среднеквадратической ошибки, а также индекс рассеивания и коэффициент детерминации, обеспечивающие глубокий статистический анализ полученных результатов.

Процесс разработки таких приложений включает в себя не только алгоритмическое моделирование, но и тщательную работу по обеспечению качества и надежности обработки данных. В этом контексте особое внимание уделяется не только сбору и обработке информации, но и обеспечению ее конфиденциальности и безопасности. Эффективное использование статистических данных и технологий машинного обучения открывает новые горизонты в прогнозировании здоровья, что делает данную область исследований особенно актуальной и перспективной.

Важность и актуальность разработки веб-приложений для прогнозирования показателей здоровья особенно велика в контексте текущих медицинских и социальных вызовов. Такие приложения могут играть решающую роль в раннем выявлении и предотвращении ряда серьезных заболеваний, повышая качество и доступность медицинской помощи. Предсказание ключевых показателей здоровья, таких как сердечный ритм, артериальное давление, масса тела и индекс массы тела, может помочь в идентификации рисков развития сердечно-сосудистых заболеваний, диабета и других состояний, которые лучше предупредить заранее и способствовать повышению осведомленности пациентов о своем здоровье и мотивируют к более активному участию в процессе его поддержания и улучшения.

Модель *ARIMA* (*Autoregressive Integrated Moving Average*). Модель *ARIMA* представляет собой важный инструмент для анализа временных рядов, особенно актуальный в разработке веб-приложений для прогнозирования показателей здоровья. Эта модель эффективно сочетает в себе элементы авторегрессии, интегрирования и скользящего среднего, что позволяет уловить различные стандартные временные структуры в медицинских данных. Применение *ARIMA* обусловлено её гибкостью и способностью адаптироваться к различным типам временных рядов, что делает её идеальной для анализа данных о здоровье.

Основная задача модели *ARIMA* заключается в прогнозировании будущих значений на основе прошлых наблюдений, что критически важно для мониторинга здоровья. Параметры модели – p (порядок авторегрессии), d (степень интегрирования) и q (порядок скользящего среднего) – играют ключевую роль в адаптации модели к конкретным данным. Например, в контексте медицинских приложений, где важно точно

прогнозировать такие показатели, как сердечный ритм или артериальное давление, *ARIMA* может обеспечить достоверность и точность прогнозов.

Сложность настройки *ARIMA* заключается в выборе оптимальных значений этих параметров, что обычно требует итеративного подхода с методом проб и ошибок. При этом используется методология Бокса-Дженкинса для классической подгонки модели. Для определения наилучших гиперпараметров модели часто применяется автоматизированный алгоритм *GridSearch*, который позволяет оценивать модели *ARIMA* на различных комбинациях гиперпараметров, чтобы найти наиболее подходящую конфигурацию [1].

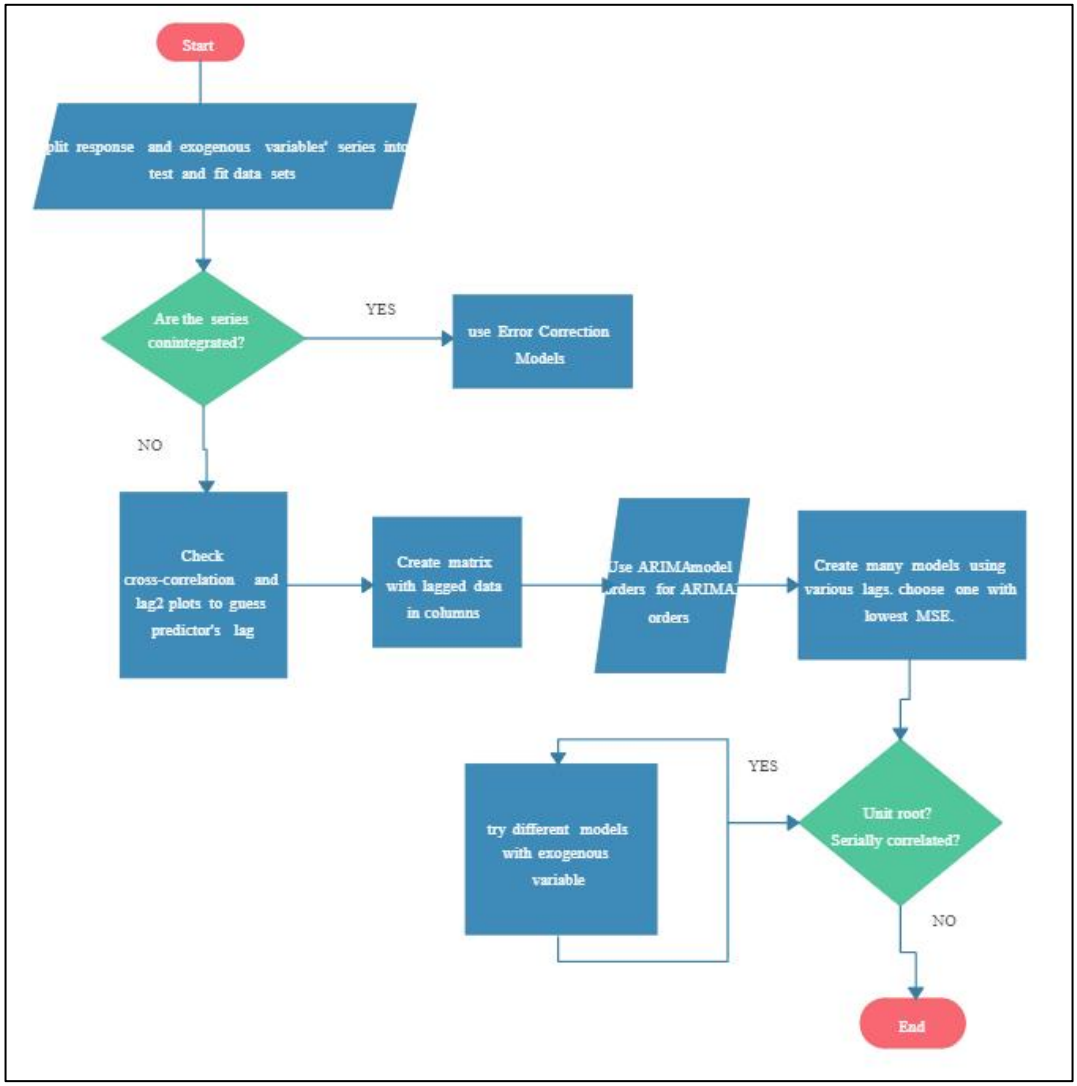

Рисунок 1. Диаграмма, иллюстрирующая алгоритм модели *ARIMA*

Техники Машинного Обучения. Техники машинного обучения играют ключевую роль в обработке и анализе временных рядов, в том числе в прогнозировании показателей здоровья. Эти модели обучаются нахождению функциональной связи между входными и выходными последовательностями (обозначаемыми как X и y соответственно), что позволяет делать прогнозы. Для работы с одномерными временными рядами, такими как данные сердечного ритма, временной ряд преобразуется в задачу обучения с учителем. Это достигается путем использования данных предыдущего временного шага ($t-1$) в качестве входных данных и данных текущего временного шага (t) в качестве выходных. Этот подход позволяет модели машинного обучения изучить и предсказать будущие значения на основе исторических данных [2].

Линейная регрессия. Линейная регрессия является одной из наиболее важных и широко используемых техник регрессии, отличающейся своей простотой. Основное преимущество этого метода заключается в легкости интерпретации результатов. Линейная регрессия стремится подогнать линейную модель с коэффициентами для минимизации суммы квадратов остатков между наблюдаемыми и предсказанными значениями. В контексте прогнозирования показателей здоровья, модель линейной регрессии обучается и адаптируется к различным размерам данных, основанным на продолжительности скользящего окна. Для каждого эксперимента модель предсказывает будущие значения показателей здоровья и рассчитывает показатели ошибок [4].

Регрессия на основе опорных векторов. *Support Vector Regression* (*SVR*) представляет собой разновидность машин опорных векторов, обеспечивающую как линейную, так и нелинейную регрессию. *SVR* используется для предсказания дискретных значений и основана на принципе поиска наилучшей линии подгонки, так чтобы ошибки не превышали определенного порога. Это особенно важно в задачах, связанных со здравоохранением, где точность прогноза критически важна.

Для достижения этой цели в статистических исследованиях в области здравоохранения параметры *SVR*, такие как ядро регрессии и параметр регуляризации *C*, могут быть тщательно настроены для улучшения точности модели. Например, настройка ядра *SVR* как линейного и параметра *C* равного 1 позволяет модели обрабатывать линейные отношения в данных. Модель тренируется на различных размерах обучающих данных, что обеспечивает гибкость в прогнозировании для различных временных интервалов, что является ключевым для динамично меняющихся медицинских данных.

Особенностью *SVR* является ее способность учитывать как шумовые, так и выбросы в данных, обеспечивая стабильность модели и надежность ее прогнозов. Это делает *SVR* мощным инструментом для прогнозирования медицинских показателей, где данные могут быть не только сложными и многомерными, но и содержать потенциальные аномалии.

Алгоритм KNN. Алгоритм *K-Nearest Neighbor* (*KNN*) важен в прогнозировании здоровья из-за его способности обрабатывать сложные шаблоны в данных о здоровье. Этот метод работает, исходя из предположения, что похожие случаи приводят к похожим результатам. Например, *KNN* может анализировать исторические данные о сердечном ритме и предсказывать будущие изменения, основываясь на сходстве с ранее наблюдаемыми случаями. Он особенно полезен, когда линейные модели оказываются недостаточными для уловления сложных взаимосвязей в данных о здоровье, и может предоставить более точные прогнозы для индивидуальных медицинских сценариев. Простая реализация регрессии *KNN* заключается в вычислении среднего значения числовой целевой переменной *K* ближайших соседей. Оптимальная эффективность алгоритма *KNN* достигается при минимизации ошибки, которая в значительной степени зависит от оптимального значения *k*. Чтобы определить это оптимальное значение *k*, применяется алгоритм *GridSearch*, который автоматически подбирает наилучшее значение *k* после определенного количества итераций на модели регрессора *KNN*.

Рисунок 2 демонстрирует данные до и после обучения. Слева показано исходное распределение с ещё не классифицированной точкой данных. Справа – результат классификации точки данных после обучения модели, где точка запроса была отнесена к определенному классу на основе ближайших соседей.

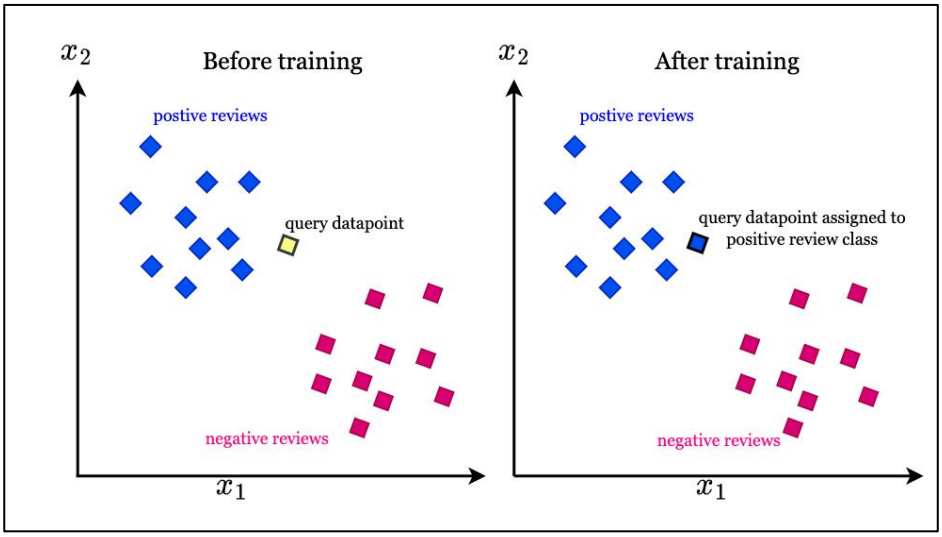

Рисунок 2. Визуализация работы алгоритма *K-Nearest Neighbors*

Дерево решений. Дерево решений представляет собой алгоритм обучения с учителем, который визуализируется как графическое представление всех возможных решений. Оно начинается с корневого узла и разветвляется, выстраивая решения на основе определённых условий. Дерево решений является эффективной моделью для решения задач как регрессии, так и классификации, используя двоичные правила для обучения связи между данными и целевой переменной [5].

В регрессии дерево решений обычно использует среднеквадратическую ошибку (*MSE*) для решения вопроса о разделении узла на два или более подузла. Важным моментом является контроль за сложностью модели, чтобы избежать её переобучения или недообучения, что может негативно сказаться на её способности к обобщению при появлении новых данных.

Оптимальный выбор глубины дерева решений может быть определён с использованием методов валидации или кросс-валидации, позволяющих оценить модель на различных уровнях сложности и выбрать наилучший баланс между способностью к обучению и обобщению.

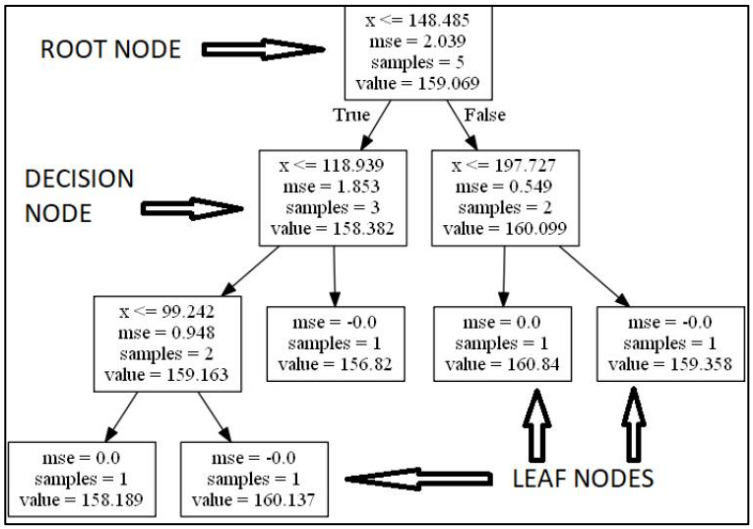

Рисунок 3. Структура дерева решений

Модель регрессии *Random Forest.* *Random Forest* используется для прогнозирования показателей здоровья, так как он обеспечивает статистическую надёжность за счёт ансамбля деревьев решений, что ведёт к уменьшению дисперсии и ошибок прогнозирования по сравнению с одним деревом решений. В контексте здравоохранения, где данные часто зашумлены и неполные, *Random Forest* эффективен в улавливании сложных нелинейных взаимосвязей между множественными предикторами и целевыми переменными.

Ключевым статистическим преимуществом *Random Forest* является его способность обрабатывать большие наборы данных с множеством входных переменных без переобучения. Это особенно ценно в медицинских приложениях, где количество возможных предикторов (например, различные биомаркеры) может быть очень велико. *Random Forest* автоматически проводит выборку переменных для каждого дерева, что помогает уменьшить корреляцию между деревьями в ансамбле и повысить точность обобщения на новые данные.

Эта модель также использует методы для уменьшения переобучения, такие как бутстрэппинг и подбор оптимального количества узлов и глубины дерева. Бутстрэппинг создаёт различные обучающие подмножества путём случайной выборки с возвращением, что позволяет каждому дереву обучаться на немного отличающихся данных. Это повышает разнообразие в ансамбле и уменьшает риск переобучения на шуме или аномалиях в обучающих данных.

Кроме того, *Random Forest* позволяет проводить оценку важности переменных, выявляя, какие предикторы наиболее значимы для прогнозирования интересующих показателей здоровья. Это информирует исследователей и клиницистов о ключевых биомаркерах и помогает в понимании биологических и патофизиологических процессов, лежащих в основе состояний здоровья и болезней.

Как показано на рисунке 4, *RF* случайным образом выбирает подмножество признаков из данных и из каждого подмножества генерирует n случайных деревьев. *RF* объединяет результаты всех деревьев решений и выдает их в итоговый результат.

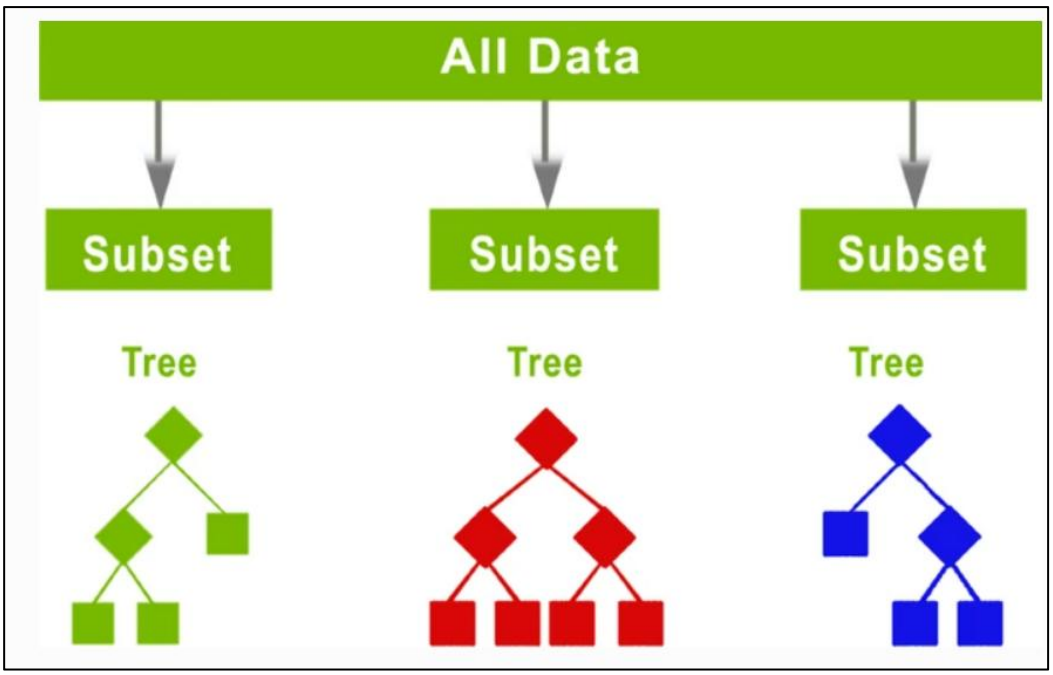

Рисунок 4. Архитектура *Random Forest*

Разделение данных. Для обучения и оценки моделей важно правильно разделить данные. В случае временных рядов сердечного ритма, полученных с помощью актиграфа,

данные были разделены таким образом, что 67% составили обучающую выборку, а оставшиеся 33% – тестовую выборку.

Все модели обучаются и оптимизируются на тренировочном наборе данных и оценены с использованием тестового набора. Это разделение данных позволяет моделям обучаться, не запоминая конкретные данные, что могло бы привести к переобучению, и в то же время обеспечивает достаточное количество данных для верификации точности модели и ее способности к обобщению на новых данных. Этап разделения данных – критически важный процесс, который напрямую влияет на надежность прогнозирующей способности модели в реальных клинических условиях [3].

Оценка модели. Для оценки моделей используются данные, из которых извлекаются временные ряды с различной продолжительностью скользящего окна: 30 секунд, 1 минута, 5 минут, 10 минут, 15 минут, 30 минут и 1 час для каждого участника. Эти данные разделяются на обучающие и тестовые сеты в соответствии с упомянутым выше соотношением. Традиционно, для оценки алгоритмов классификации временных рядов используются такие метрики, как точность и полнота, которые также применимы для расчета точности моделей классификации.

Однако, поскольку в данном исследовании акцент делается на проблемы регрессии временных рядов, эффективность моделей измеряется с использованием различных метрик, предназначенных для моделей регрессии временных рядов. Модели обучаются и используются для выполнения предсказаний для каждого скользящего окна, а затем рассчитываются средние значения средней абсолютной ошибки, среднеквадратической ошибки и корня из среднеквадратической ошибки.

Для понимания, является ли показатель хорошим или плохим, рассчитывается индекс рассеивания (*SI*), который представляет собой *RMSE*, деленный на среднее значение наблюдаемой величины. Если *SI* меньше 10%, это свидетельствует о хорошей модели, а *SI* меньше 5% указывает на очень хорошую модель. В свою очередь, модель прогнозирования должна иметь высокий коэффициент детерминации R^2 (близкий к 1), что показывает, что линия регрессии хорошо соответствует данным, и эффективность модели является высокой. Производительность каждой модели оценивается с использованием *RMSE* и *SI*. [3]

Заключение. Актуальность использования разнообразных моделей регрессии в медицине обусловлена необходимостью точного анализа сложных медицинских данных. Различные модели, такие как *ARIMA*, *Random Forest* и *SVR*, обеспечивают гибкость и адаптивность к динамичным и многомерным данным, что критически важно для точного прогнозирования медицинских показателей. Правильный выбор модели позволяет учитывать специфику данных, улучшая диагностику и разработку персонализированных лечебных подходов. Это способствует прогрессу в здравоохранении и повышает качество медицинского обслуживания, делая тему особенно значимой в современной медицине.

Список литературы

[1] How to Grid Search ARIMA Model Hyperparameters with Python [Electronic resource]. – https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/

[2] How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification [Electronic resource]. – https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/

[3] Cross Validation in Time Series [Electronic resource]. – https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4

[4] Linear Regression in Python [Electronic resource]. – https://realpython.com/linear-regression-in-python/#linear-regression

[5] sklearn.linear_model.LinearRegression [Electronic resource]. – https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Авторский вклад

Ванецкий Николай Андреевич – руководство и постановка задачи исследования BIG DATA для прогнозирования показателей здоровья с применением механизмов машинного обучения.

Василькова Анастасия Николаевна – постановка задачи исследования, описание принципа работы Big Data в улучшении эффективности прогнозирования показателей здоровья с применением механизмов машинного обучения, анализ полученных результатов, формирование структуры статьи.

Кислова Дарья Алексеевна – тестирование программного средства, описание принципов использования моделей ARIMA, Random Forest и SVR, формирование структуры статьи.

WEB APPLICATION FOR PREDICTING HEALTH INDICATORS USING MACHINE LEARNING TECHNIQUES AND ITS ERGONOMIC SOFTWARE

N.A. Vanetsky
Postgraduate student, Department of Engineering Psychology and Ergonomics, BSUIR

D.A. Kislova
BSUIR student, Department of Engineering Psychology and Ergonomics

A.N. Vasilkova
Senior Lecturer, Department of Engineering Psychology and Ergonomics, BSUIR

Abstract. This study presents an analytical review of the application of statistical methods and machine learning algorithms for analyzing time series data on health indicators. The focus is on the use of ARIMA models, linear regression, SVR (Support Vector Regression), KNN (K-Nearest Neighbors), and Random Forest. It emphasizes how the application of these methods contributes to improving the accuracy and personalization of health forecasts, considering the complexity and diversity of medical data.

The process of evaluating the effectiveness of these methods is described based on various metrics, including mean absolute error, mean squared error, and scatter index. This evaluation highlights the significance of their use in modern medical applications. The work demonstrates the importance of integrating advanced data analysis technologies in the field of healthcare to improve patient quality of life.

Keywords: time series, health indicators, ARIMA, machine learning, web applications for health monitoring, medical diagnostics.