ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ
INFORMATION TECHNOLOGY AND TELECOMMUNICATIONS

УДК004.056.5:621.398.96
CC BY 4.0
DOI: 10.21822/2073-6185-2024-51-4-87-98 Оригинальная статья /Original article

Системный анализ и обработка информации для задачи выявления поломки информационного накопителя компьютера

Н.М. Кодацкий, Е.А. Ревякина, А.Р. Газизов
Донской государственный технический университет,
344002, г. Ростов-на-Дону, пл. Гагарина, 1, Россия

Резюме. Цель. Одним из важных аспектов поддержания эффективной работы информационных систем является состояние оборудования компьютеров. Цель исследования заключается в описании метода эффективного системного анализа состояния информационного накопителя компьютера. **Метод.** Исследование основано на использовании алгоритмов машинного обучения для анализа и интерпретации данных, полученных из SMART-тестов. Включает в себя комплексный анализ и экспериментальное исследование. Представляет собой проведение экспериментов с набором данных и облачной средой Google Colab, создание и анализ модели машинного обучения, оценку эффективности и качества обучения. **Результат.** Разработан инструмент оценки состояния компьютерного оборудования на основе алгоритма Случайного леса, используя исторические данные SMART-тестов. **Вывод.** Результат данной работы не только позволяет внедрять рабочий инструмент анализа данных в сферу обслуживания компьютерного оборудования, но и предлагает практический пример использования для повышения надежности и эффективности работы информационных систем. Результаты исследования могут быть полезны как для ИТ-специалистов, ответственных за техническую поддержку, так и для организаций, стремящихся оптимизировать процессы обслуживания оборудования и улучшить свою конкурентоспособность на рынке.

Ключевые слова: информационные технологии, надежность, производительность, инновационные методы, мониторинг, прогнозирование, оборудование, SMART-тесты, модель оценки, случайный лес, риск, отказ, анализ данных, повышение эффективности

Для цитирования: Н.М. Кодацкий, Е.А. Ревякина, А.Р. Газизов. Системный анализ и обработка информации для задачи выявления поломки информационного накопителя компьютера. Вестник Дагестанского государственного технического университета. Технические науки. 2024; 51(4):87-98. DOI:10.21822/2073-6185-2024-51-4-87-98

System analysis and information processing to solve the problem of detecting breakdowns of computer information storage

N.M. Kodatsky, E.A. Revyakina, A.R. Gazizov
Don State Technical University,
1 Gagarina Square, Rostov-on-Don, 344002, Russia

Abstract. Objective. One of the important aspects of maintaining the efficient operation of information systems is the condition of computer equipment. The purpose of the study is to describe a method for effective system analysis of the state of a computer's information storage device. **Method.** The study is based on the use of machine learning algorithms to analyze and interpret data obtained from SMART tests. Includes a comprehensive analysis and experimental study. It involves conducting experiments with a data set and the Google Colab cloud environment, creating and analyzing a machine learning model, and evaluating the effectiveness and quality of training. **Result.** A tool for assessing the state of computer equipment based on the Random Forest algorithm has been developed using historical data from SMART tests. **Conclusion.** The results not only allow the implementation of a working data analysis tool

in the field of computer equipment maintenance, but also contain a practical example of increasing the reliability and efficiency of information systems. The results are useful for IT specialists and for organizations optimizing equipment maintenance processes and increasing competitiveness.

Keywords: information technology, reliability, performance, innovative methods, monitoring, forecasting, equipment, SMART tests, evaluation model, random forest, risk, failure, data analysis, efficiency improvement

For citation: N.M. Kodatsky, E.A. Revyakina, A.R. Gazizov. System analysis and information processing to solve the problem of detecting breakdowns of computer information storage. Herald of Daghestan State Technical University. Technical Sciences. 2024;51(4):87-98. DOI:10.21822/2073-6185-2024-51-4-87-98

Введение. Причинами потери данных могут быть не только «человеческие ошибки, кибератаки», но и «поломки оборудования» [1,3], включая компьютеры и другие устройства хранения информации, которые гораздо чаще происходят даже в штатном режиме [2,12]. Одним из основных факторов потери данных является поломка оборудования компьютеров [3]. Одним из главных устройств компьютера, относительно хранения информации, является его накопитель, без которого просто невозможно оперировать информацией. Принято считать, что компьютерное оборудование – это стандартизированные устройства, которые «должны работать с удовлетворимым коэффициентом амортизации элементов» и минимальным, но «достаточным временем наработки на отказ» [2].

Постановка задачи. Существуют различные факторы, которые могут привести к неисправности средств вычислительной техники. В этой связи, важно иметь «систему контроля состояния оборудования» [1], чтобы предотвратить потерю данных и обеспечить непрерывность работы компьютерной системы.

Методы исследования. Одним из эффективных методов решения данной проблемы является использование интеллектуальных методов прогнозирования выхода из строя различных устройств. Современные «интеллектуальные методы оценки состояния оборудования компьютеров», использующие «машинное обучение и искусственный интеллект» [1], могут помочь определить вероятность возникновения поломки оборудования до того, как она произойдет. Так можно выполнить профилактические меры и заменить неисправное оборудование, прежде чем оно приведет к потере данных, что делает исследование актуальным.

В качестве решения поставленной цели будут использоваться современные инструменты и подходы, такие как облачные вычисления Google Coola, включая Python и библиотеки машинного обучения, такие как Scikit-learn, а также другие библиотеки для работы с данными и их оперированием, такие как Numpy и Pandas. В работе также будут использованы данные и результаты экспериментов, полученные на реальных компьютерных системах. Результаты данной работы представляют собой значимый вклад в область информационной безопасности и технологий. Применимость работы распространяется на различные сферы общества, включая бизнес, науку, медицину и другие области, где широко используется компьютерная техника [12, 16]. Полученные результаты позволяют разработать и реализовать собственный отечественный метод мониторинга состояния информационных накопителей для предприятий по результативности не хуже зарубежных аналогов. Такой подход к проблеме обеспечения безопасности данных и контроля за хранением информации может заменить использование зарубежных аналогов с закрытым исходным кодом. Исследование способствует укреплению информационной безопасности организаций [11, 12] и дает возможность иметь более надежные инструменты для контроля за информационными ресурсами в контексте сложных политических обстоятельств и геополитических взаимоотношений с сфере продуктов и разработок.

Ручной анализ данных. Жесткие диски (ЖД) являются одним из наиболее распространенных устройств хранения данных в компьютерах и серверах. Они играют ключевую роль в сохранении, доступе и обработке информации. В связи с этим, понимание состоя-

ния и производительности жестких дисков имеет важное значение для эффективного функционирования компьютерных систем. Для оценки состояния и работы жестких дисков используется различная статистика, включая показатели SMART (Self-Monitoring, Analysis, and Reporting Technology). SMART представляет собой технологию, которая «позволяет мониторить и анализировать работу дискового устройства, предупреждать о возможных проблемах и предсказывать отказы» [7]. Изучение статистики SMART является важным аспектом для эффективного управления и поддержки жестких дисков. Понимание того, что эти показатели говорят о состоянии диска, «позволяет принять своевременные меры для предотвращения потери данных и снижения риска отказа системы» [7]. Анализ статистики SMART имеет также существенное значение в области информационной безопасности. Определение состояния жесткого диска с помощью показателей SMART позволяет выявить потенциальные проблемы, связанные с целостностью данных или возможными нарушениями безопасности. Использование статистики SMART не только способствует эффективному управлению жесткими дисками, но также играет важную роль в обеспечении информационной безопасности и защите данных от потенциальных угроз. Каждый диск включает в себя технологию самоконтроля, анализа и отчетности (SMART), которая предоставляет внутреннюю информацию о диске. SMART представляет систему мониторинга, встроенную в жесткие диски, которая сообщает о различных атрибутах состояния данного диска. Доступно более 70 статистических данных SMART, но для задачи прогнозирования поломки ЖД будет достаточно знать только пять. Описание атрибутов представлено в табл. 1 [7, 5]. Когда значение одного из этих атрибутов больше нуля, возникает повод обратить на это внимание.

Таблица 1. Атрибуты SMART-теста и его описание
Table 1. SMART test attributes and its description

Атрибут/ Attribute	Описание/ Description
CMAPT 5	Количество перераспределенных секторов/ Reallocated Sectors
CMAPT 187	Сообщенные о неисправимых ошибках/ Reported Uncorrectable Errors
CMAPT 188	Прерывание команды/ Command Abort
CMAPT 197	Текущее количество ожидающих секторов/ Current Pending Sectors
CMAPT 198	Неисправимое количество секторов/ Uncorrectable Sectors

Также «необходимо отслеживать состояние RAID-массива для выявления потенциальных проблем с диском» [6]. Эти инструменты обычно сообщают только об исключениях, поэтому в любой момент количество расследований можно контролировать, даже несмотря на значительное количество ЖД. Согласно статистике, предоставляемой компанией Backblaze Vault [8], по анализу состояния ЖД (табл. 2), можно сказать, что процент рабочих ЖД, у которых один или несколько из пяти показателей SMART больше нуля составил 4,2 %, когда неисправные диски по тем же показателям больше нуля составили 76,7 %.

Таблица 2. Процент неисправных и работающих дисков по выделенным атрибутам [4]
Table 2. Percentage of faulty and working hard drives by selected attributes [4]

Состояние ЖД HDD status	SMART 5	SMART 187	SMART 188	SMART 197	SMART 198
Исправны/ Good	1,1%	0,5 %	4,8 %	0,7 %	0,3 %
Неисправны/ Faulty	42,2 %	43,5%	44,8%	43,1 %	33 %

Наличие данного показателя со значением, превышающим ноль, в данный момент может ничего не значить. Например, диск может иметь необработанное значение SMART 5, равное двум, что означает, что два сектора диска были переназначены. Сама по себе такая ценность мало что значит, пока она не будет объединена с другими факторами. В процессе оценки «может потребоваться изрядное количество интеллекта» (как человеческого, так и искусственного), чтобы прийти к выводу о том, «что операцион-

ный двигатель выйдет из строя» [9, 7]. Единственный способ, который помогает – наблюдаются несколько ошибок SMART.

Рассмотрим корреляцию между наблюдаемыми атрибутами в табл. 3 [9]. Табл. 3 отражает взаимоотношение между рассматриваемыми атрибутами, что, помогает в определении, какие конкретные параметры SMART в большей степени связаны друг с другом, а также выявляет противоречия или зависимости между различными атрибутами.

Таблица 3. Корреляция между статистическими данными атрибутов SMART
Table 3. Correlation between SMART attribute statistics

	SMART 5	SMART 187	SMART 188	SMART 197	SMART 198
SMART 5	1	0,034	0,026	0,064	0,043
SMART 187	0,034	1	0,007	0,025	0,033
SMART 188	0,026	0,007	1	0	0,006
SMART 197	0,064	0,025	0	1	0,808
SMART 198	0,043	0,033	0,006	0,808	1

В большинстве случаев статистика мало коррелирует и может считаться независимой. Только SMART 197 и 198 имеют хорошую корреляцию [7], то есть их можно рассматривать как один индикатор вместо двух. Но целесообразно продолжать собирать эти статистики вместе по двум причинам: «корреляция не идеальна», поэтому есть место для ошибок и «не все производители накопителей сообщают об обоих атрибутах» [9].

Понимание корреляции или ее отсутствия может помочь в принятие решения как поступать с ЖД. Например, накопитель сообщил о необработанном значении SMART 5, равном 10, и необработанном значении SMART 197, равном 20. Из этого исходит, что ЖД изнашивается и его следует его заменить. Принимая во внимание, что, если тот же диск имел необработанное значение SMART 197, равное 5, и необработанное значение SMART 198, равное 20, и никаких других ошибок, можно было бы отложить замену диска, ожидая получения дополнительных данных, таких как частота возникновения ошибок. Статистика SMART, приведенная выше, за исключением SMART 197, носит накопительный характер, то есть необходимо учитывать период времени, в течение которого были зарегистрированы ошибки, а не просто количество показателей, отличных от нуля.

Правильное использование статистики SMART требует систематического мониторинга и анализа показателей. Регулярный контроль состояния дисков и применение рекомендаций по использованию статистики SMART позволяют повысить надежность системы и обеспечить безопасность хранения данных. Понимание значения и применения статистики SMART является неотъемлемой частью управления и поддержки жестких дисков. Анализ показателей SMART помогает предотвращать потерю данных, предупреждать о возможных проблемах и улучшать безопасность системы. Использование данной информации позволяет принять своевременные меры и обеспечить надежную работу компьютерных систем в условиях быстро меняющейся информационной среды.

Данные для обучения. Одним из эффективных методов решения исследуемой проблемы является использование интеллектуальных методов прогнозирования выхода из строя различных устройств. Цель данного раздела заключается в обоснование эффективности интеллектуального метода оценки состояния оборудования компьютера с использованием алгоритма Случайного Леса (Random Forest) и подходов **B**agging и **B**oosting для предотвращения потери данных на информационных накопителях компьютера.

Источником данных выступает операционная система устройства (Windows или Linux), а именно следующие источники системы: технологии мониторинга SMART-атрибутов (smartmontools или CrystalDiskInfo), журнал событий операционной системы и утилиты командной строки. Наибольший интерес для определения выхода из строя жесткого диска представляют 5 атрибутов данных SMART-статистики, а именно [9]:

- SMAPT 5 – количество перераспределенных секторов;
- SMAPT 187 – сообщения об неисправностях;

- SMAPT 188 – требуемое время на ожидание команд;
- SMAPT 197 – текущее количество секторов в режиме ожидания;
- SMAPT 198 – количество некорректируемых секторов.

Для более корректного обучения модели было принято решение воспользоваться собранными необработанными тестовыми данными жестких дисков за 2022 – 2023 года от компании Backblaze, использующих эти данные в своих центрах обработки 67 814 жестких дисков. [9]. Для сбора данных SMART-статистики компания использует Smartmontools [5]. Сбор производится один раз в день для каждого жесткого диска.

Таким образом, «добавляются несколько элементов», таких как модель диска, серийный номер и т. д., и «создается строка в ежедневном журнале для каждого диска» [10]. Диски, которые вышли из строя, помечаются как таковые, и их данные больше не регистрируются. Иногда «диски удаляются из эксплуатации», даже если он не вышел из строя, например, когда компания «обновляет Storage Pod», заменяя диски емкостью 1 ТБ на диски емкостью 4 ТБ [9]. В этом случае диск объемом 1ТБ не помечается как неисправный, но данные SMART больше не регистрируются. Каждый день центр обработки данных Backblaze делает снимок каждого работающего жесткого диска своего исследовательского центра. Этот «набор данных включает основную информацию о диске», а также его статистику SMART [9, 10]. Ежедневный снимок одного диска - это «одна запись или строка данных». Все снимки дисков за определенный день собираются в файл, состоящий из строк для каждого активного жесткого диска. Этот файл имеет формат «*.csv» (значения, разделенные запятыми). Так, каждый день формируется файл с названием в формате ГГГГ-ММ-ДД.csv, например, 2023-07-01.csv и имеет следующую структуру [8]:

1. первая строка каждого файла содержит имена столбцов, остальные строки - это исторические данные состояния накопителя.
2. столбцы содержат следующую информацию:

- дата - дата файла в формате гггг-мм-дд;
- серийный номер - серийный номер диска, присвоенный производителем;
- модель - номер модели привода, присвоенный производителем;
- емкость - емкость диска в байтах;
- сбой - содержит «0», если диск в порядке. Содержит «1», если это последний день работы диска перед сбоем.

Выбор интеллектуального метода. Для задачи прогнозирования поломки жестких дисков на основе размеченных данных SMART можно использовать различные алгоритмы машинного обучения. Каждая модель имеет свои преимущества и недостатки, и выбор конкретной модели зависит от специфики задачи, доступности данных, требований к точности и интерпретируемости. Мною были выделены некоторые из подходящих:

1. Случайный лес (Random Forest). Способен достичь высокой точности и устойчивости к переобучению благодаря ансамблю деревьев решений. Может обрабатывать большие объемы данных с множеством признаков. Возможна оценка важности признаков, что полезно для понимания, какие атрибуты SMART наиболее влияют на выход из строя. Но такие модели на реальных и глобальных задачах могут быть достаточно сложными и требовать значительных вычислительных ресурсов для обучения и предсказаний. А ее интерпретируемость может быть ниже, чем у некоторых других подходов.
2. Градиентный бустинг (Gradient Boosting). Одна из наиболее мощных и широко используемых техник, которая часто показывает высокую точность в задачах классификации и регрессии. Метод строит модель последовательно, что позволяет ему эффективно корректировать ошибки предыдущих моделей. К том же, требует тщательной настройки гиперпараметров для достижения оптимальной производитель-

ности. Может быть более подвержен переобучению по сравнению с случайным лесом при недостаточном количестве данных.

3. Логистическая регрессия. Проста в интерпретации результатов. Обладает быстрой скоростью обучения и предсказания по сравнению с более сложными моделями. Хорошо подходит для задач бинарной классификации. Однако, предполагает линейную зависимость между признаками и целевой переменной, что может не всегда быть верным для данных SMART.
4. Метод опорных векторов (SVM). Эффективны в высокоразмерных пространствах и при наличии четкого разделения классов. Обладают версатильностью (способность обрабатывать широкий спектр данных) благодаря различным ядрам для обработки нелинейных зависимостей, но требуют интенсивных вычислений при больших объемах данных, а выбор и настройка ядра могут значительно повлиять на производительность модели.

Нами была выбрана модель случайного леса (Random Forest) для предсказания поломки жесткого диска на основе размеченных данных SMART тестов. На наш взгляд, это мощный алгоритм, который объединяет преимущества множества деревьев решений для повышения точности и устойчивости модели. Случайный лес хорошо справляется с большими наборами данных и может автоматически учитывать важность признаков, что делает его особенно подходящим для анализа данных SMART. Случайный лес является ансамблевым методом, который строит множество деревьев решений при обучении и выдает средний прогноз для классификации или регрессии. Это позволяет достигнуть высокой точности предсказаний, снижая при этом риск переобучения благодаря механизмам случайности при выборе признаков и образцов для построения деревьев. Данные SMART тестов часто включают множество различных атрибутов, отражающих состояние жестких дисков. Случайный лес может эффективно обрабатывать такие наборы данных, автоматически определяя наиболее значимые признаки для предсказания отказов. Одним из основных принципов работы такого алгоритма является использование подвыборок признаков для каждого дерева, что позволяет снизить влияние нерелевантных или слабо влияющих на целевую переменную признаков и увеличить общую точность модели. Хотя сама по себе модель случайного леса может казаться менее интерпретируемой по сравнению с одиночным деревом решений, она предоставляет полезную информацию о важности признаков. Понимание того, какие атрибуты SMART наиболее влияют на прогнозы, может быть важно для дальнейшего анализа и понимания причин отказов жестких дисков.

Эффективное использование алгоритма Random Forest с подходами Bagging и Boosting предоставляет «значительные возможности для прогнозирования» [13 - 15] выхода из строя оборудования компьютеров и предотвращения потенциальной потери данных. Этот комбинированный подход объединяет преимущества методов Bagging (усреднение результатов множества моделей) и Boosting (адаптивное взвешивание ошибок моделей), что способствует созданию надежных и точных моделей. Такой подход позволяет построить сильный классификатор, который «способен адаптироваться к различным данным и условиям» [13], повышая точность предсказаний. Подходы Bagging и Boosting улучшают обобщающую способность модели, позволяя ей эффективно обрабатывать сложные данные и быстро реагировать на изменения в состоянии оборудования.

К преимуществам алгоритма отнесем [13]:

- имеет высокую точность прогнозирования (для большинства задач работает лучше линейных алгоритмов, а точность сравнима с точностью boosting);
- практически не имеет чувствительности к выбросам в данных из-за случайного сэмлирования выборок методом bootstrap;
- не чувствителен к масштабированию (любым монотонным преобразованиям) значений признаков, связано с выбором случайных подпространств;
- не требует тщательной настройки параметров, хорошо работает «из коробки»;

- способен эффективно обрабатывать данные с большим числом признаков и классов;
- редко переобучается, на практике добавление деревьев почти всегда только улучшает композицию (до определенного предельного уровня);
- хорошо работает с пропущенными данными, сохраняет хорошую точность, если большая часть данных пропущена;
- могут быть расширены до неразмеченных данных, что приводит к возможности делать кластеризацию и визуализацию данных, обнаруживать выбросы;
- можно легко распараллелить и масштабировать (увеличить число деревьев и их глубину).

К недостаткам алгоритма отнесем [13]:

- в отличии от одного дерева, результаты случайного леса сложно интерпретировать;
- алгоритм работает хуже многих линейных методов, когда в выборке очень много разреженных признаков (тексты, bag of words);
- не умеет экстраполировать данные, в отличии от линейной регрессии;
- склонен к переобучению на некоторых задачах, где данные не сильно зашумлены;
- для данных, включающих категориальные переменные с различным количеством уровней, случайный лес предвзят в пользу признаков с большим количеством уровней: когда у признака много уровней, дерево будет сильнее подстраиваться именно под эти признаки, так как на них можно получить более высокое значение оптимизирующего функционала (информационный выигрыш);
- большой размер получающихся моделей, что требует $O(N \cdot K)$ памяти для хранения моделей, где K – число деревьев.

Выбор случайного леса для предсказания поломки жесткого диска на размеченном наборе данных SMART тестов обусловлен его способностью к обработке больших и сложных наборов данных, высокой точностью и устойчивостью к переобучению, а также возможностью интерпретации важности признаков. Эти качества делают случайный лес подходящим для задач, где требуется надежное и точное прогнозирование на основе большого количества признаков, как и в случае с данными SMART тестов жестких дисков. Комбинация таких методов не только обеспечивает высокую точность прогнозирования поломок оборудования, но и позволяет оперативно реагировать на любые изменения или неполадки, минимизируя вероятность серьезных сбоев. Такой подход обеспечивает стабильность работы компьютерных систем и обеспечивает высокий уровень безопасности данных, что является критически важным аспектом в современной информационной технологии и бизнес-среде.

Перейдем к реализации самой модели интеллектуального метода в задаче прогнозирования выхода из строя информационного накопителя компьютера.

Использование алгоритма Случайного леса в задаче прогнозирования выхода из строя информационного накопителя компьютера. Отобразим матрицу диаграмм рассеивания (рис. 1), чтобы установить зависимости между 5-ю параметрами SMART используемого набора данных.

Каждая диагональ матрицы показывает распределение одного параметра, а вне диагонали - диаграммы рассеяния для каждой пары параметров. Заметим, что как было описано ранее, данные 5 и атрибутов тесно связаны друг с другом и коррелируют. Важно отметить, что для обучения из набора данных [7] останутся только значения, обладающие высокой корреляцией на отказ и размеченные для них метки классов: SMART 5 – количество перераспределенных секторов; SMART 187 – сообщенные о неисправимых ошибках; SMART 188 – тайм-аут команды; SMART 197 – количество текущих ожидающих секторов; SMART 198 – количество неисправных секторов. Метка класса состояния диска (0 – вышедший из строя, 1 – функционирующий исправно).

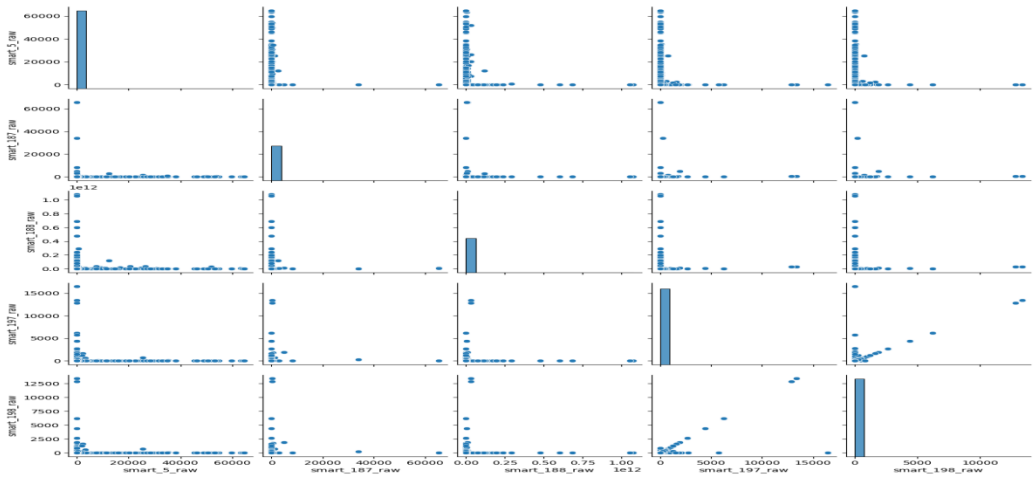

Рис. 1. Матрица рассеивания 5 SMART атрибутов тренировочной выборки
Fig. 1. Scatter matrix of 5 SMART attributes of the training sample

По результатам обучения модели важно оценить ее качественные показатели. Рассмотрим 5 метрик, позволяющих оценить корректность принятых решений моделью и ее предвзятость к данным. Каждая из этих метрик предоставляет информацию о различных аспектах производительности модели, а именно:

- точность (Accuracy) – доля правильно предсказанных классов относительно всех примеров;
- точность (Precision) – доля истинно положительных примеров среди всех положительных предсказаний;
- полнота (Recall) – доля истинно положительных примеров, которые были предсказаны правильно;
- F1-мера (F1-Score) – гармоническое среднее точности и полноты;
- коэффициент корреляции Мэтьюса (Matthews correlation coefficient) – мера качества бинарной классификации, учитывающая ложные положительные и ложные отрицательные прогнозы.

Также отобразим матрицу ошибок (confusion matrix) для оценки производительности модели классификации (рис. 2).

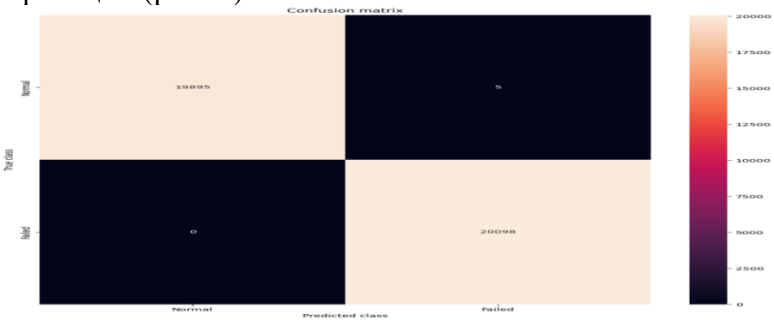

Рис. 2. Матрица ошибок модели
Fig. 2. Model error matrix

Полученные результаты метрик в ходе исследования вышли следующими:

[Model] >>> Random Forest classifier
[The accuracy] >>> 0.9998749937496875
[The precision] >>> 0.9997512809033477
[The recall] >>> 1.0
[The F1-Score] >>> 0.9998756249844531
[The Matthews correlation coefficient] >>> 0.9997500123101655

Обсуждение результатов. Далее произведем формализацию полученных результатов и сделаем вывод о проделанной работе. По полученным результатам можно сделать следующие выводы о производительности обученной модели Random Forest:

1. Точность. С высокой точностью около 99.99% модель правильно классифицирует примеры из тестового набора данных. Это означает, что модель правильно предсказывает класс примера в почти всех случаях.
2. Точность. Precision также очень высока и составляет около 99.98%. Это указывает на то, что из всех примеров, которые модель предсказала как положительные, почти все действительно являются положительными.
3. Полнота. Recall равен 1.0, что означает, что модель правильно классифицирует все истинно положительные примеры из тестового набора данных. Это хороший показатель и говорит о том, что модель эффективно обнаруживает все положительные случаи.
4. F1-мера. F1-мера, которая является гармоническим средним между точностью и полнотой, также очень высока и составляет около 99.99%. Это указывает на сбалансированность между точностью и полнотой модели.
5. Коэффициент корреляции Мэтьюса. MCC также близок к единице, что указывает на очень сильную корреляцию между предсказаниями модели и фактическими значениями классов.

В целом, эти результаты свидетельствуют о том, что модель Random Forest хорошо обучена и демонстрирует высокую производительность на тестовом наборе данных. Рассматривая матрицу ошибок, где строки представляют фактические классы, а столбцы - предсказанные классы. Приведем интерпретацию результатов:

1. True Positives (TP), верхняя левая клетка (19895) соответствует ситуации, когда модель правильно классифицировала нормальные (негативные) примеры как нормальные.
2. False Positives (FP), верхняя правая клетка (5) соответствует ситуации, когда модель неправильно классифицировала нормальные примеры как отказавшие (ложно положительные).
3. False Negatives (FN), нижняя левая клетка (0) соответствует ситуации, когда модель неправильно классифицировала отказавшие примеры как нормальные (ложно отрицательные).
4. True Negatives (TN), нижняя правая клетка (20098) соответствует ситуации, когда модель правильно классифицировала отказавшие примеры как отказавшие.

Исходя из матрицы на рис. 2, общее количество правильно классифицированных примеров (истинно нормальных и истинно отказавших) составляет 39893, а количество ошибочно классифицированных примеров (ложно нормальных и ложно отказавших) составляет 5. Также, интерес представляет визуализации элементов ансамблей деревьев. Продемонстрирую полученные 1, 2, 3 и… 50-ые построенные деревья решений на рис. 3 – 6. Метод случайного леса обладает сложной формализацией, что является его главным минусом. Но оценивая рис.3-6 в общем, можно заметить, что деревья претерпевают изменения и стараются выявить наиболее коррелирующие атрибуты для прогнозирования.

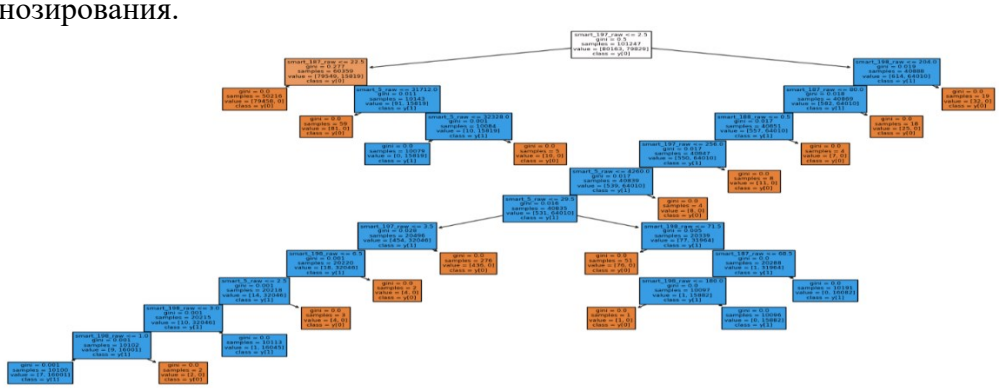

Рис. 3. Первое решающее дерево
Fig. 3. The first decision tree

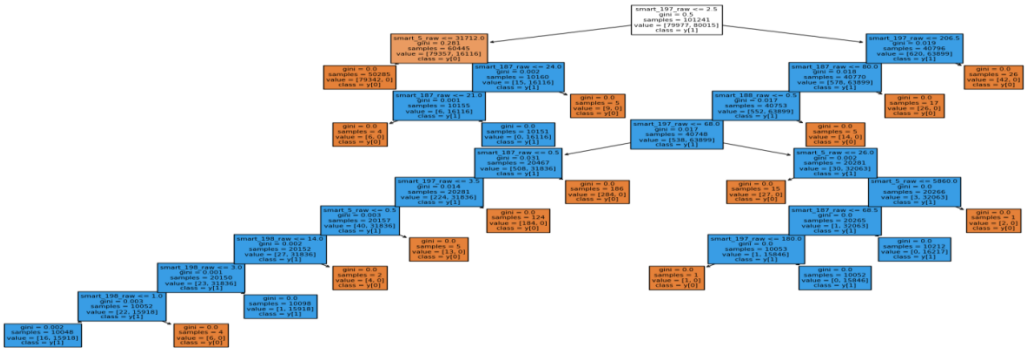

Рис. 4. Второе решающее дерево
Fig. 4. The second decision tree

Рис. 5. Третье решающее дерево
Fig. 5. The third decision tree

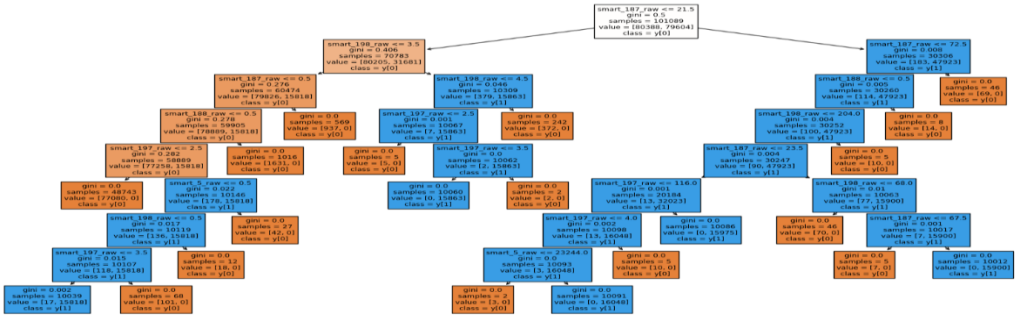

Рис. 6. Пятидесятое решающее дерево
Рис. 6. Пятидесятое решающее дерево

А в рамках данной задачи точное объяснение выдвинутого моделью решения не является ключевым аспектом, куда важнее заблаговременно узнать об ошибках и постараться выявить проблему диска, с чем модель отлично справляется.

Вывод. Таким образом, в данной работе был разработан инструмент оценки состояния компьютерного оборудования на основе алгоритма Случайного леса, используя исторические данные SMART-тестов.

Результаты работы показали, что разработанная модель обладает высокой точностью и способностью правильно классифицировать как нормальные, так и отказавшие компоненты. Анализ confusion matrix подтвердил, что модель демонстрирует хорошую способность к правильному выявлению обоих классов. Модель имеет высокую способность правильно идентифицировать нормальные состояния (True Negatives), что является хорошим показателем. Также у модели есть некоторое количество ложно положительных результатов (False Positives), что может привести к излишней тревоге или ненужным проверкам оборудования. Модель всегда удается точно предсказать отказы (False Negatives), что исключает пропуску реальных проблем. Важно, отметить, что обучение происходило в условиях сильной ограниченности вычислительной мощности. Набор данных был взят лишь за два дня наблюдений SMART-тестов 2023 года, поскольку количество оперативной памяти не позволяло оперировать большим набором, а входе эксперимента был взят максимум из предоставляемых виртуальных ресурсов вычислительной машины (рис. 7).

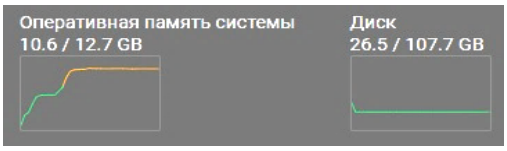

Рис. 7. Используемые вычислительные ресурсы в ходе обучения
Fig. 7. Computing resources used during training

В целом, результаты работы указывают на перспективность использования алгоритма Случайного леса для создания системы мониторинга состояния компьютерного оборудования. Дальнейшие исследования и улучшения модели могут помочь еще более точно и надежно определять отказавшие компоненты и повысить эффективность обслуживания оборудования, предоставляя результаты не хуже аналоговых систем, но используя отечественную собственную разработку, которую можно интегрировать и использовать под нужды каждого отдельного предприятия.

Библиографический список:

1. Национальный стандарт Российской Федерации ГОСТ Р ИСО 13381-1-2011 Контроль состояния и диагностика машин. Прогнозирование технического состояния. Часть 1. Общее руководство. Дата введения 2012-12-01. Подготовлен Автономной некоммерческой организацией «Научно-исследовательский центр контроля и диагностики технических систем» (АНО «НИЦ КД). Внесен Техническим комитетом по стандартизации ТК 183 «Вибрация, удары и контроль технического состояния». Утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 16 ноября 2011 г. № 553-ст.
2. Шкляр, В.Н. Надежность систем управления: учебное пособие – Томск: Томский политехнический университет, 2009. – 126 с.
3. Вострецова, Е.В. Основы информационной безопасности: учебное пособие – Екатеринбург: Издательство Уральского университета, 2019. – 208.
4. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 15.
5. Официальный сайт Smartmontools – URL http://www.smartmontools.org/ (дата обращения: 20.09.2023) – Текст электронный.
6. Официальный сайт компании Backblaze. Статья «Backblaze Vaults: Zettabyte-Scale Cloud Storage Architecture». – URL https://www.backblaze.com/blog/vault-cloud-storage-architecture/ (дата обращения: 20.09.2023) – Текст электронный.
7. Официальный сайт компании Backblaze. Статья «Hard Drive SMART Stats». – URL https://www.backblaze.com/blog/hard-drive-smart-stats/(дата обращения:22.09.2023)Текст электронный.
8. Официальный сайт компании Backblaze. Статья «Hard Drive Data and Stats». Набор данных для обучения – URL https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data#downloading-the-raw-hard-drive-test-data (дата обращения: 19.09.2023) – Текст электронный.
9. Официальный сайт компании Backblaze. Статья «What SMART Stats Tell Us About Hard Drives». – URL https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/ (дата обращения: 21.09.2023) – Текст электронный.
10. Официальный сайт компании Backblaze. Статья «Hard Drive Data and Stats». – URL https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data/ (дата обращения: 19.09.2023) – Текст электронный.
11. Российская Федерация. Законы. О персональный данных: Федеральный закон Российской Федерации № 152-ФЗ. [принят Государственной думой 8 июля 2006 года: одобрен Советом Федерации 14 июля 2006 года]. – Москва: Кремль: Кодекс, 2021. – 24 с. // КонсультантПлюс.
12. Российская Федерация. Законы. Об информации, информационных технологиях и о защите информации: Федеральный закон Российской Федерации № 149-ФЗ: текст с изменениями и дополнениями на 20 марта 2021 года: [принят Государственной думой 8 июля 2006 года: одобрен Советом Федерации 14 июля 2006 года]. – Москва: Кремль: Кодекс, 2021. – 24 с. // КонсультантПлюс.
13. Breiman, L. Random forests. Machine Learning. Нью-Йорк: Springer, 2001, 32 с. (дата обращения: 18.03.2024) – Текст электронный.
14. Liaw, A., Wiener, M. Classification and regression by random forest. Лондон: R Foundation, 2002, 22 с. (дата обращения: 18.03.2024) – Текст электронный.
15. Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J. Random forests for classification in ecology. Нью-Йорк: Springer, 2007, 582 с. (дата обращения: 18.03.2024) – Текст электронный.

16. Короченцев Д.А. и др. Импортозамещающие технологии обеспечения информационной безопасности и защиты данных. – 2021.

References:

1. National Standard of the Russian Federation GOST R ISO 13381-1-2011 Condition monitoring and diagnostics of machines. Prediction of technical condition. Part 1. General guidance. Date of introduction 2012-12-01. Prepared by Autonomous Non-Commercial Organization "Research Center for Control and Diagnostics of Technical Systems" (ANO SIC CD). Introduced by the Technical Committee for Standardization TC 183 "Vibration, Shocks and Technical Condition Control". Approved and put into effect by the Order of the Federal Agency for Technical Regulation and Metrology dated November 16, 2011; 553 (In Russ).
2. Shklyar V.N. Reliability of control systems: textbook . Tomsk Polytechnic University, 2009;126 (In Russ).
3. Vostretsova, E.V. Fundamentals of information security: textbook - Yekaterinburg: Ural University Press, 2019; 208. (In Russ).
4. Chandola, V., Banerjee, A., & Kumar, V. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 2009;41(3):15.
5. 5Smartmontools official website . URL http://www.smartmontools.org/(acces. 20.09.2023) Text electronic.
6. Backblaze official website. Article "Backblaze Vaults: Zettabyte-Scale Cloud Storage Architecture." URL https://www.backblaze.com/blog/vault-cloud-storage-architecture (date of reference: 20.09.2023)Text electronic.
7. Official website of Backblaze. Article "Hard Drive SMART Stats." - URL https://www.backblaze.com/blog/hard-drive-smart-stats/ (date of reference: 22.09.2023) - Text electronic.
8. Official website of Backblaze. Article "Hard Drive Data and Stats. Training dataset - URL https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data#downloading-the-raw-hard-drive-test-data (access date: 19.09.2023) - Text electronic.
9. Backblaze official website. Article "What SMART Stats Tell Us About Hard Drives." - URL https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/ (accessed 21.09.2023) - Text electronic.
10. Backblaze official website. Article "Hard Drive Data and Stats." - URL https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data/ (accessed 19.09.2023) - Text electronic.
11. Russian Federation. Laws. On personal data: Federal Law of the Russian Federation No. 152-FZ. [adopted by the State Duma on July 8, 2006: approved by the Federation Council on July 14, 2006]. - Moscow: Kremlin: Codex, 2021; 24. ConsultantPlus. (In Russ).
12. Russian Federation. Laws. About information, information technologies and about protection of information: the Federal law of the Russian Federation No. 149-FZ: the text with amendments and additions for March 20, 2021: [adopted by the State Duma on July 8, 2006: approved by the Federation Council on July 14, 2006]. - Moscow: Kremlin: Codex, 2021; 24 ConsultantPlus. (In Russ).
13. Breiman, L. Random forests. Machine Learning. New York: Springer, 200;32. (date of reference: 18.03.2024) - Text electronic.
14. Liaw, A., Wiener, M. Classification and regression by random forests. London: R Foundation, 2002; 22. (date of reference: 18.03.2024) - Text electronic.
15. Cutler D.R., Edwards Jr, T.C., Beard K.H., Cutler A., Hess, K. T., Gibson, J., Lawler, J. J. Random forests for classification in ecology. New York: Springer, 2007;582(date of reference: 18.03.2024) Text electronic.
16. Korochentsev D. A. et al. Import-substituting technologies for information security and data protection. - 2021. (In Russ).

Сведения об авторах:

Кодацкий Никита Максимович, студент; nickitadatsky@gamil.com; ORCID: 0009-0001-3726-0178

Ревякина Елена Александровна, кандидат технических наук, доцент кафедры «Кибербезопасность информационных систем», revyelena@yandex.ru

Газизов Андрей Равильевич, кандидат педагогических наук, заведующий кафедрой «Вычислительные системы и информационная безопасность»; gazandre@yandex.ru

Information about authors:

Nikita M. Kodatsky, Student; nickitadatsky@gamil.com; ORCID: 0009-0001-3726-0178

Elena A. Revyakina,. Cand. Sci. (Eng.), Assoc.Prof., Department "Cyber Security of Information Systems"; revyelena@yandex.ru

Andrey R. Gazizov, Cand. Sci. (Pedag.), Head of the Department "Computing Systems and Information Security"; gazandre@yandex.ru

Конфликт интересов/Conflict of interest.

Авторы заявляют об отсутствии конфликта интересов/The authors declare no conflict of interest.

Поступила в редакцию/ Received 27.05.2024.

Одобрена после рецензирования/ Reviced 20.06.2024.

Принята в печать/ Accepted for publication 15.10.2024.