МИНИСТЕРСТВО СПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Российский университет спорта «ГЦОЛИФК»

ТЕХНОЛОГИИ КОНВЕРГИРОВАНИЯ СРЕДСТВ И МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ПРОЦЕСС СПОРТИВНОЙ ПОДГОТОВКИ

Монография

Москва, 2024

УДК

Технологии конвергирования средств и методов искусственного интеллекта в процесс спортивной подготовки/ А.В Ермаков – Москва,2024 – 154 с.

Рекомендовано к изданию:

Экспертно-методическим советом

Протокол № от

Оглавление

[1. Конвергенция в социальных, экономических и биологических исследованиях 4](#_Toc160218086)

[2. Государственное регулирование развития технологий в области искусственного интеллекта и практика использования предлагаемых технологий в сфере физической культуры и спорта 14](#_Toc160218087)

[3. Анализ больших объёмов данных в сфере физической культуры и спорта 35](#_Toc160218088)

[3.1. Интернет вещей и носимые технологии источник данных в спорте 46](#_Toc160218089)

[3.2. Исследование, подготовка и статистический анализ данных 52](#_Toc160218090)

[3.3. Описательная статистика и особенности распределений наблюдений в исследованиях спортивной деятельности 58](#_Toc160218091)

[3.4. Распределение веса спортсменов 67](#_Toc160218092)

[3.5. Распределение значений роста спортсменов 82](#_Toc160218093)

[3.5. Статистическая оценка для данных большого размера в сфере физической культуры и спорта 111](#_Toc160218094)

[4. Регрессионный и классификационный анализ 121](#_Toc160218095)

[4.1. Линейная регрессия 123](#_Toc160218096)

[4.2. Логистическая регрессия 144](#_Toc160218097)

[4.3. Ансамблевые методы улучшения производительности математических и компьютерных моделей в сфере спорта 159](#_Toc160218098)

[Заключение 173](#_Toc160218099)

# 1. Конвергенция в социальных, экономических и биологических исследованиях

Конвергенция в качестве подхода, направленного на сближение различных систем, технологических решений из малосвязанных в прошлом областей знания, объединения и интеграции прочно заняла свое место в современном научном исследовании. Зачастую она понимается узко специально в отдельных областях знания, как например в экономике часто рассматривают конвергенцию именно как частное явление в качестве процесса сходимости показателей экономического роста бедных и богатых стран за счёт ускоренного роста последних[[1]](#footnote-1). Стоит отдельно упомянуть, что даже в подобных случаях узкого понимания конвергенции, авторами используется конвергенция их научных подходов с информационными технологиями, а именно используются такие популярные алгоритмы машинного обучения и интеллектуального анализа данных, как регрессионные модели для получения научного результата, а значит происходит именно конвергенция экономики и технологий искусственного интеллекта[[2]](#footnote-2). Теория конвергенции также широко используется в исследованиях коммуникации, культуры и бизнеса для изучения слияния и интеграции различных средств и методов объектной области, например медиа платформ и информационных технологий[[3]](#footnote-3). Конвергенция в этом случае может означать технологическое объединение сосуществующих медиа систем, реализуемых посредством информационных технологий и различных платформ на которых происходит их реализация. Утверждается, что конвергенция информационных потоков уже изменила процесс потребления контента пользователями и взаимодействие множества медиаканалов и медиа продуктов[[4]](#footnote-4). В качестве сложной многогранной концепции конвергенцию в медиа пространстве можно проиллюстрировать синергией средств массовой информации (например, предоставление аудитории доступа к новостям, музыке и видео через одну и ту же медиа-платформу), интеграцией технологий (например, функции отслеживания сердечного ритма и фитнес-трекеры, интегрирования их в смартфон, и обеспечения связи с носимыми устройствами), слияние медиа культуры и продуктов (например, сюжеты и персонажи видеоигр, воспроизводятся в фильмах, печатной продукции и спортивных событиях и пр.), слияние рыночных и потребительских баз (например, коллаборация спортивных организаций и производителей компьютерных игр) и т. д. Культура конвергенции представляет собой и потребительский сдвиг, побуждающий потребителя непрерывно искать новую информацию и для этого устанавливать связи с рассредоточенным медиа-контентом[[5]](#footnote-5). Культура конвергенции по мнению исследователей понятия состоит из различных процессов: конвергенции медиа пространств, культуры участия и так называемого коллективного разума, рождающегося в информационном пространстве. Конвергенция в медиа пространстве представляет собой поток контента между медиа платформами, при этом культура участия позволяет осуществлять превращение потребителей в активную составляющую процесса, а коллективный разум является результатом того, что потребление осуществляется совместно целыми сообществами. Описываемая исследователями культура конвергенции также приводит нас к тому, что спортивные команды должны учитывать, что медиа пространства состоят как из технологий, так и из культурных составляющих, а также осознавать, изменение, распространение и контроль спортивного контента под воздействием информационных технологий). В целом экосистема спортивных медиа изменилась в результате появления новых технологий, снижающих затраты на производство и потребление контента, а главное в результате возникла потребность улучшения взаимодействия с потребителем и создания для него дополнительных ценностей за счёт анализа и воспроизведения контента уже внутри самого коммьюнити. Иными словами исследователи предлагают расширение точек соприкосновения с клиентами вместо попытки максимально широкого, но поверхностного охвата[[6]](#footnote-6).

В спортивной же науке конвергенция так же может принимать самое разнообразное значение, в том числе и сходное с проявляющимся в экономике. Можно например, обратиться к области спортивного менеджмента и также определить конвергенцию как сближение спортивной отрасли и индустрии развлечений[[7]](#footnote-7). Возможно как отдельные группы исследователей напрямую перенести упоминаемое выше понятие конвергенции в экономике на развитие национального спортивного движения рассматривая государственное управление спортивной отраслью[[8]](#footnote-8). В этом случае изучается сближение различных спортивных систем в процессе их перспективного развития, как итог - глобализация. Обобщая, можно заметить, что здесь нами упоминается конвергенция, как некоторый перенос имеющихся тенденций из одной области знаний на другую по признаку сходства объектной области исследования.

Кроме вышеупомянутого почти прямого переноса понятия конвергенции из экономики, в современной спортивной науке мы можем наблюдать и некоторое разнообразие в его использовании. Исследователи например, отмечают конвергенцию между традиционно практикуемыми видами спорта и компьютерными играми[[9]](#footnote-9). Именно она создала, по их мнению, компьютерный спорт как отдельное уникальное явление в современной спортивной науке и практике.

В другой, сходной со спортивной наукой объектной области исследования, биологии, конвергенция может означать сходство различных организмов в результате их продолжительного обитания в сходных условиях[[10]](#footnote-10)[[11]](#footnote-11). Несмотря на устоявшийся термин с узким значением, сейчас, когда говорят о конвергенции в биологии, то часто речь идёт не об описываемом выше явлении, а именно об интеграции традиционных подходов исследования с современными технологическими решениями. Инновационные технологические решения зачастую рассматриваются в качестве того связующего звена, которое обеспечивает конвергенцию между различными направлениями исследования. В качестве иллюстрации можно рассмотреть нанотехнологии, которые по мнению исследователей обеспечивают процесс конвергенции между биологией и медициной и сами развиваются на основе передовых информационных технологий[[12]](#footnote-12). В свою очередь сами нанотехнологии находятся по мнению других исследователей в процессе сближения с биологией[[13]](#footnote-13). В результате возникает нанобиотехнология, в которой применяются нанотехнологии для изучения и трансформации биосистем, и которая использует биологические принципы и материалы для создания новых технических устройств и систем, интегрированные на нано уровне. Ожидается, что в ближайшем темпы развития информационных технологий и когнитивной науки дающей сейчас направления для исследования в этой области только ускорятся[[14]](#footnote-14). Соответственно успешность или не успешность научно –исследовательской работы всё более зависит от сближения избранного направления научного исследования и актуальных информационных технологий.

В этой ситуации более значимым направлением конвергенции информационных технологий в биологии является так называемое «машинное обучение», направление развития искусственного интеллекта создающее, применяющее и оценивающее алгоритмы, предназначенные для прогнозирования, классификации, распознавания и решения других задач с использованием математических моделей, полученных на основе наблюдаемых данных. При этом важным фактором является автоматизация интеллектуального труда, что значительно повышает качество и масштабность исследовательской работы[[15]](#footnote-15).

Существует и обратная зависимость, в основе некоторых подходов машинного обучения лежит биологическая составляющая. Так в основе модели глубокого обучения лежит объект, названный персептроном, который представлял собой попытку смоделировать реальное поведение нейронов, и в результате этой попытки возникла область проектирования искусственных нейронных сетей[[16]](#footnote-16). Ещё в 1982 году алгоритм персептрона был использован для изучения бактерий Escherichia coli[[17]](#footnote-17). В свою очередь архитектуры искусственных нейронных сетей были созданы с обращением к организации зрительной нервной системы[[18]](#footnote-18)[[19]](#footnote-19). Считается, что данный процесс будет только развиваться и технологии станут настолько «органическими», что различие между технологическим и биологическим будет провести не так просто[[20]](#footnote-20). За прошедшие годы адаптивность методов машинного обучения росла совместно с ростом средств и методов математического обоснования как отдельных алгоритмов, так и данного подхода в целом, и вполне естественно надеяться, что методы машинного обучения в будущем значительно повысят эффективность процесса решения научных проблем и позволят достичь более глубокого понимания непрерывно растущих по своему объему и сложности биологических данных.

Безусловно спорт в качестве отдельно рассматриваемого явления сочетает в себе и социальную, и биологическую деятельность, что позволяет использовать все вышерассмотренные подходы к конвергенции. И действительно существует большое количество работ, которые относят конвергенцию в спорте к вопросам прежде всего спортивного менеджмента[[21]](#footnote-21) и зачастую стратегического менеджмента[[22]](#footnote-22)[[23]](#footnote-23)[[24]](#footnote-24)[[25]](#footnote-25). Другая часть исследователей рассматривает конвергенцию в спорте с позиций использования информационных технологий с упором на медиа технологии[[26]](#footnote-26)[[27]](#footnote-27) или же информационные технологии взятые во всём их многообразии где ставится вопрос о полном изменении понятия спорт («Sport 2.0») в результате этой конвергенции[[28]](#footnote-28). Изучается даже «эмоциональная конвергенция» описывающая эмоциональное сближение в зависимости от типа и роли людей в спортивной деятельности[[29]](#footnote-29). Рассматривается гендерная конвергенция [[30]](#footnote-30). Также популярной темой для дискуссий является сближение цифровой реальности не только в виде компьютерного спорта, но и всего спектра цифровых технологий с традиционными видами спорта[[31]](#footnote-31)[[32]](#footnote-32). Особенно это касается спорта высших достижений о котором прямо заявляют, что он уже «встроен в технологические структуры» и о том, что надо ставить вопрос о «конвергентных технологических приложениях» в спорте[[33]](#footnote-33)[[34]](#footnote-34). В качестве закономерного итога зарубежные исследователи приходят к выводу что само существование спорта как явления теперь полностью зависит от развития информационных технологий, которые за счёт высоких скоростей своей работы позволяют полностью преобразить спортивную деятельность[[35]](#footnote-35)[[36]](#footnote-36)[[37]](#footnote-37). Как актуальная проблема рассматривается также конвергенция любительского и профессионального спорта (в первую очередь по схожести развития медико-биологического обеспечения)[[38]](#footnote-38), а также спорта и искусства в максимально широком его понимании[[39]](#footnote-39).

В целом несмотря на разнообразие приложений понятия «конвергенция» общим в этом процессе остаётся сближение современных информационных технологий, компьютерного моделирования и всего спектра технологий искусственного интеллекта включая машинное и глубокое обучение (искусственные нейросети). Конвергенция информационных технологий в процесс спортивной подготовки была обоснована ещё советскими учёными во второй половине прошлого века. Обсуждение конвергенции информационных технологий в сферу физической культуры и спорта проходило в ходе научных конференций и научно-методических семинаров под названием "Кибернетика и спорт"[[40]](#footnote-40)[[41]](#footnote-41). Тесное взаимопроникновение биологического и машинного начал в кибернетике была отмечена ещё Н. Виннером в самом начале существования этой науки[[42]](#footnote-42)[[43]](#footnote-43). Возникали и более частные решения, как например фокусировка внимания на проблеме управления движениями в спортивной деятельности[[44]](#footnote-44). Это безусловно было обосновано тем, что частное понимание кибернетики на тот момент могло быть заужено именно до науки об управлении на основе информации, вместо более широкого понимания как науки об информации во всём её многообразии. Значительные усилия были приложены чтобы обеспечить эффективную конвергенцию трёх составляющих: биологии, информационных технологий (на тот момент под названием кибернетики) и спортивной науки[[45]](#footnote-45)[[46]](#footnote-46). При этом уже тогда ставился вопрос о конвергенции не только средств и методов отдельных наук, а именно технологий и в первую очередь конечно информационных технологий[[47]](#footnote-47).

Например Чхаидзе Л.В., ссылаясь работы Н. А. Бернштейна[[48]](#footnote-48)[[49]](#footnote-49)[[50]](#footnote-50), на одиннадцатой странице своей работы прямо заявляет, что «нельзя отрешиться от общих закономерностей построения управляемых систем и процессов управления, установленных бурно развивающейся кибернетикой»[[51]](#footnote-51). Тем самым он практически декларирует конвергенцию информационных технологий, математического моделирования и спортивной науки.

В результате появилось достаточно большое количество работ углублённо исследующих наиболее общие подходы к проблеме конвергенции кибернетики и спорта[[52]](#footnote-52)[[53]](#footnote-53). Кроме затрагивания наиболее общих вопросов в опубликованных диссертационных исследованиях затрагивались и более конкретные области конвергенции этих наук[[54]](#footnote-54). Отдельной подобной проблемой рассматривалась необходимость разработки математического аппарата для конвергенции информационных технологий и спортивной науки[[55]](#footnote-55)[[56]](#footnote-56)[[57]](#footnote-57).

Значительный вклад в эту работу внёс Зациорский В.М (1969) указав на необходимость обеспечения конвергенции информационных технологий, математического моделирования и спортивной науки[[58]](#footnote-58).

В качестве одной из важных целей конвергенции информационных технологий была сформулирована автоматизация в том числе и научно-исследовательской деятельности для повышения её эффективности[[59]](#footnote-59)[[60]](#footnote-60). Автоматизация с применением информационных технологий коснулась в том числе управления психофизическим состоянием спортсмена [[61]](#footnote-61). В целом автоматизировались различные аспекты физкультурно-спортивной деятельности[[62]](#footnote-62).

Спортивные учёные затрагивали темы моделирования с использованием языковых моделей, которые стали так популярны сегодня после удачного дебюта Chat GPT[[63]](#footnote-63). Другим популярным ныне направлением к которому обращались исследователи было вероятностное моделирование[[64]](#footnote-64)[[65]](#footnote-65). Отдельно стоит отметить практическую реализацию конвергенции спортивной науки и информационных технологий с самого начала этого процесса, так уже в 1977 году Ботвинник М. М. реализовал древовидные модели в программном решении анализа шахматных задач противопоставляя их алгоритму полного перебора вариантов и пытаясь приблизить алгоритм к человеческому мышлению и принятию решений[[66]](#footnote-66). При этом основатель отечественного научного направления изучения искусственного интеллекта Поспелов Г.С. используя аналогию именно с шахматами утверждал, что какой бы развитой не была искусственного интеллекта, всё равно [[67]](#footnote-67).

В дальнейшем, обобщая многие отдельные подходы, Шестаковым М.П.(1997) была обоснована конвергенция машинного (Machine Learning — ML) и глубокого обучения (искусственные нейросети, Deep Learning — DL) и процесса спортивной подготовки на примере лёгкой атлетики[[68]](#footnote-68). Развитие идей конвергенции современных информационных технологий и процесса спортивной подготовки уже на примере тяжёлой атлетики получило продолжение в работах Фураева А.Н. (2022)[[69]](#footnote-69).

Отдельно строит отметить работы посвящённые конвергенции информационных технологий и не только непосредственно физкультурно-спортивного образования[[70]](#footnote-70)[[71]](#footnote-71)[[72]](#footnote-72), а также преподаванию предметов связанных с физкультурой и спортом[[73]](#footnote-73)[[74]](#footnote-74). Проводились попытки систематизировать возможности использования информационных технологий в высшем физкультурно-спортивном образовании, к данным технологиям относили: «программированное обучение, имитационное моделирование, автоматический анализ функционального состояния спортсменов и применение экспертных систем» [[75]](#footnote-75).

В итоге рассмотрения вышеизложенного вопроса можно прийти к следующим выводам:

Явление конвергенции широко распространено как в социальных так и в биологических науках, не исключением является и сфера физической культуры и спорта.

Актуальный вид конвергенции в большинстве сфер человеческой деятельности включая спортивную это конвергенция информационных технологий в общем и искусственного интеллекта (как классического машинного обучения, так и глубокого обучения) в частности с традиционными для этих сфер средств и методов.

В настоящее время информационные технологии и непосредственно технологии искусственного интеллекта используются как в процессе спортивной деятельности, так и при получении физкультурно-спортивного образования

# 2. Государственное регулирование развития технологий в области искусственного интеллекта и практика использования предлагаемых технологий в сфере физической культуры и спорта

Развитие искусственного интеллекта в качестве явления значимого для отечественной экономики и признание его таковым государством можно считать оформленным в первой редакции национальной программе "Цифровая экономика Российской Федерации"[[76]](#footnote-76) которая в свою очередь результатом реализации «Стратегии развития информационного общества в Российской Федерации на 2017 - 2030 годы»[[77]](#footnote-77). Если в Стратегии говорилось о построении «Общества знаний» и «Цифровой экономики», то уже Программа непосредственно манипулировала понятием искусственный интеллект, затронув вопросы прав собственности на системы интеллектуального анализа данных и машинного обучения, которые являются реализациями систем искусственного интеллекта в федеральном проекте "Информационная инфраструктура" и обеспечение безопасности при разработке и использовании технологий искусственного интеллекта в федеральном проекте "Информационная безопасность". В ходе дискуссий о развитии программы были озвучены планы о создании целых «умных» городов и использования технологий блокчейна, определялось тотальное изменение профессиональных компетенций. В итоге основными сквозными цифровыми технологиями Программы в 2017 году помимо прочих стали большие данные, искусственный интеллект, технологии виртуальной и дополненной реальностей. Была утверждена дорожная карта направления «Кадры и образование» разработанная Агентством стратегических инициатив включающая базовую модель и перечень ключевых компетенций цифровой экономики, персональный профиль компетенций и траектории развития обеспечение запросов компаний цифровой экономики, например, WorldSkills Russia. Так же было озвучено улучшение цифровой грамотности населения и поддержка образовательных проектов и отдельных образовательных решений, связанных с информационными технологиями. Для WorldSkills Russia аспирантами ВНИИФК был разработан модуль «С» компетенции "Физическая культура, спорт и фитнес". История развития компетенции «Физическая культура, спорт и фитнес» по стандартам Worldskills Russia до внедрения в сезоне 2019-2020 методики использования носимых гаджетов и программного обеспечения для консультирования представлена в таблице (таб.1)[[78]](#footnote-78)

Таблица. 1 - История развития компетенции «Физическая культура, спорт и фитнес» по стандартам Worldskills Russia.

|  |  |  |  |
| --- | --- | --- | --- |
| **Дата события** | **Событие** | **Количество представленных регионов** | **Вид репрезентации компетенции** |
| Апрель 2016 год | I Открытый региональный чемпионат WorldSkills Russia | 14 | Презентационно |
| Июль 2017 год | Проектная сессия экспертов по компетенции | 20 | Проектная сессия |
| 2016-2017 год | 14 региональных чемпионатов по РФ | 14 | Презентационно |
| Май 2017 год | Финал V Национального чемпионата «Молодые профессионалы» (WorldSkills Russia) | 7 | Презентационно |
| Июнь 2017 год | Пилотные проекты по сдаче ГИА в формате демонстрационного экзамена | 2 | Пилотный проект |
| Август 2017 год | Проектная сессия экспертов по компетенции | более 30 | Проектная сессия |
| Октябрь 2017 год | Компетенция «Физическая культура и спорт» переведена из разряда презентационных в основной состав |  | Официальный |
| 2017-2018 год | Соревнования по компетенции «Физическая культура и спорт» по стандартам WorldSkills Russia | 32 | Официальный |
| 2-9 мая 2018 года | Отборочные соревнования на право участия в Финале VI Национального чемпионата «Молодые профессионалы» (WorldSkills Russia) | 31 | Официальный |
| Июнь 2018 | Пилотные проекты по сдаче ГИА в формате демонстрационного экзамена | 12 | Пилотный проект |
| 2018-2019 год | Соревнования по компетенции «Физическая культура и спорт» по стандартам WorldSkills Russia | 45 | Официальный |
| Апрель 2019 год | Отборочные соревнования VII Национальный чемпионат «Молодые профессионалы» (WorldSkills Russia) | 42 (основная категория), 16 (юниоры 14-16 лет) | Официальный |
| Май-июнь 2019 год | ГИА в формате демонстрационного экзамена | 25 | Официальный |
| Август 2019 год | Проектная сессия экспертов по компетенции | более 40 | Проектная сессия |

Как можно заметить несмотря на краткое время своего существования данная компетенция развивалась достаточно быстро стала внедрять актуальные информационные технологии и производить анализ с помощью информационных технологий данных. Так в данном модуле методика использования носимых гаджетов для оценки интенсивности физической нагрузки в демонстрационном экзамене использовалась для разработки и проведения экспресс-консультаций с заявленной в модуле целью по привлечению различных возрастных групп населения к здоровому образу жизни с использованием современных цифровых технологий (рис. 1) [[79]](#footnote-79).

В дальнейшем, продолжая сотрудничество с производителем носимых гаджетов ONETRAK, нами было проведено уже более масштабной исследование в течение 2018-2019 года с сентября по сентябрь было произведено фиксирование двигательной активности 40145 пользователей носимых гаджетов ONETRAK в 951 населённом пункте России, что составляет 85,14% от всех городов России. При этом выяснилось, что при разбиении испытуемых по группам, мы получаем неравномерное распределение по возрастам и различия в средних значениях количества шагов для каждой группы.

Безусловно использование носимых гаджетов порождает большое количество вопросов к методикам их использования (так как проконтролировать каждого конкретного пользователя на этот предмет маловероятно), их точности (так как это не лабораторное оборудование и специального отношения к точности замера здесь не присутствует) и непосредственно вопросов к интерфейсу и первичной обработки данных (так как производитель разрабатывает эти шаблоны и методы самостоятельно и они не всегда соответствуют принятым в спортивной науке).

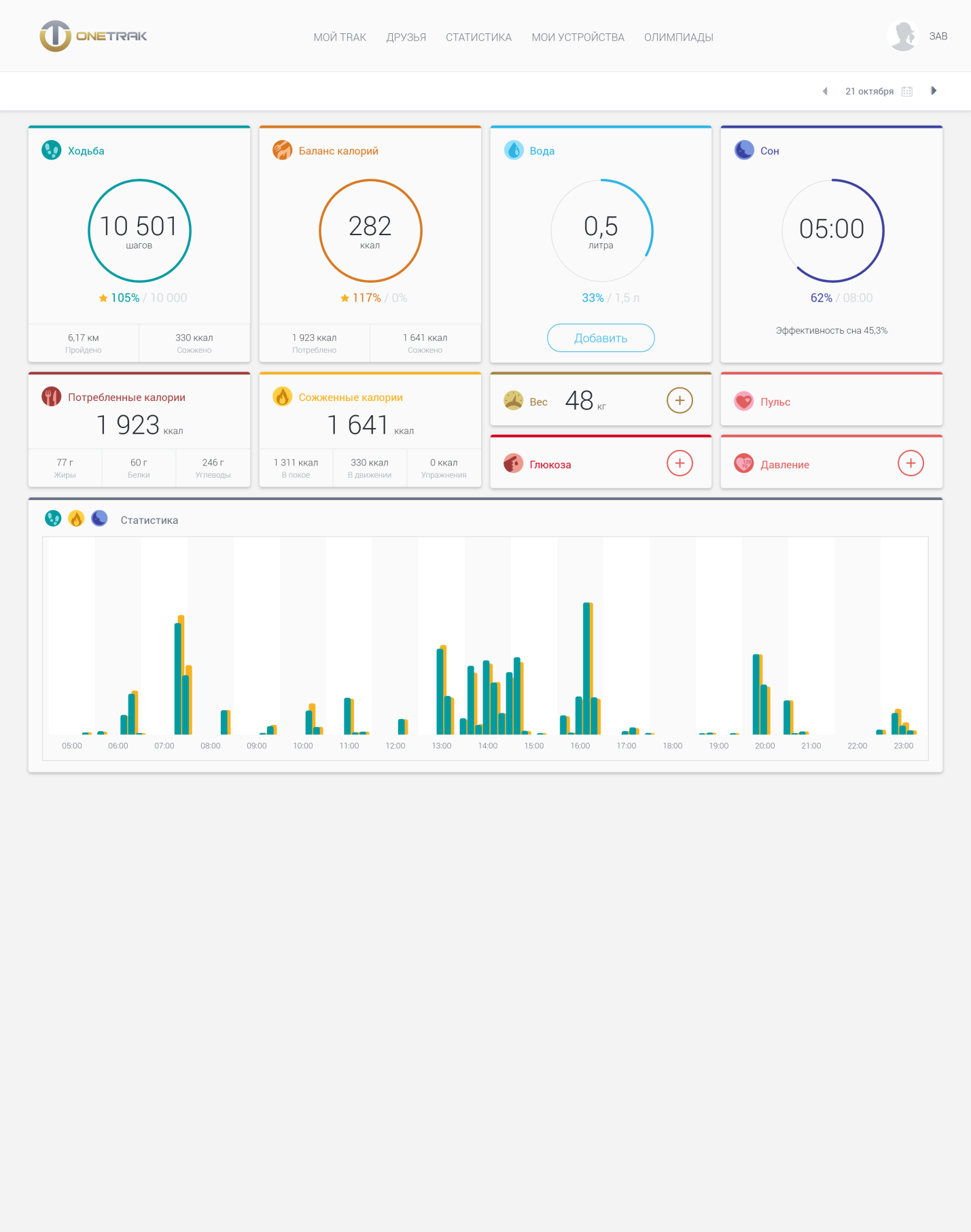


Рис.1 - Пример отображения параметров в мобильном приложении ONETRAK в задании соревнующихся WorldSkills Russia

С полученными в ходе исследования распределениями по гендерно-возрастным признакам имеющимся сезонным закономерностям и прочим аспектам использования носимых гаджетов при фиксации двигательной активности можно ознакомиться в таблицах и диаграммах расположенных ниже[[80]](#footnote-80). В таблице (таб.2) представлено распределение дневной двигательной активности по возрастному признаку.

Более наглядно результаты представлены на рисунке (рис.2) где можно увидеть, что различные возрастные группы различаются не только различным количеством среднего количества шагов в день, но и популярность использования гаджетов различна у разных возрастных групп.

Таблица 2. Распределение дневной двигательной активности по возрастному признаку (Ермаков А.В., 2020)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Возраст (лет)** | **До 16 лет** | **16-27** | **28-45** | **Старше 45** |
| **Количество (человек)** | 3361 (8%) | 7877 (20%) | 16394 (41%) | 12513 (31%) |
| **Среднее кол-во шагов в день** | 3362,6 | 3851,3 | 4378,2 | 4475,3 |

Анализируя таблицу и диаграмму можно обратить внимание, что если широта использования (количество гаджетов) присуща группе 28-45 лет, то активность использования больше у старшей возрастной группы. Однако в дальнейших главах используя принципы интеллектуального анализа данных (Data mining) мы показали, что на самом деле ситуация гораздо сложнее и зависимости дают более интересную пищу для размышлений.

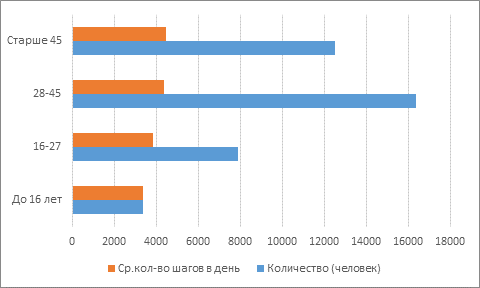


Рис.2 - Распределение испытуемых по возрастным группам (кол-во испытуемых, среднее кол-во шагов в день) (Ермаков А.В., 2020)

Вопрос с сезонностью решался в два этапа и, если на первом этапе мы видим общую картину увеличения среднего количества шагов в «тёплые» месяцы с небольшим провалом в сентябре связанном вероятно с началом учебного года и окончанием отпускного периода. Это отображено на рисунке (рис.3). Однако в дальнейшем будет показано, что ситуация гораздо более сложна и приводит к неоднозначным, но весьма продуктивным умозаключениям.

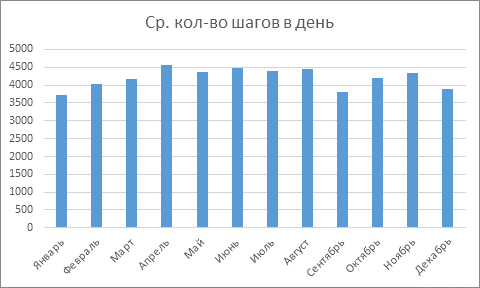


Рис.3 - Распределение ежедневной двигательной активности населения в течение года (сентябрь 2018-сентябрь 2019) (Ермаков А.В., 2020)

Попытка найти зависимость от локализации или количества населения в том или ином населённом пункте не привела нас к какому-нибудь результату (табл. 3) и здесь уже требуются средства и методы, позволяющие находить более глубокие закономерности и которые сейчас относят к технологиям искусственного интеллекта.

Таблица.3 Распределение ежедневной двигательной активности населения в отдельных городах (данные о численности населения получены с помощью ресурса «Статистика населения»[[81]](#footnote-81)).

|  |  |  |  |
| --- | --- | --- | --- |
| **Город** | **Численность населения (человек)** | **Кол-во испытуемых (человек)** | **Среднее количество шагов в день сделанных одним испытуемым** |
| **Москва** | 11 514 330 | 1882 | 4149,8 |
| **Санкт-Петербург** | 4 848 742 | 1397 | 4539,2 |
| **Барнаул** | 612 091 | 371 | 4216,7 |
| **Владивосток** | 592 069 | 275 | 3669,6 |
| **Кемерово** | 532 884 | 347 | 3742,3 |
| **Астрахань** | 520 662 | 322 | 3899,4 |
| **Архангельск** | 348 716 | 451 | 4623,7 |
| **Владимир** | 348 256 | 297 | 4927,1 |
| **Волжский** | 314 436 | 17 | 4758,1 |
| **Владикавказ** | 311 635 | 93 | 4167,3 |
| **Балашиха** | 215 353 | 118 | 3910,0 |
| **Благовещенск** | 214 397 | 95 | 3358,0 |
| **Арзамас** | 106 367 | 22 | 5655,8 |
| **Биробиджан** | 73 623 | 15 | 4723,5 |
| **Лабытнанги** | 26211 | 27 | 5141,6 |

Дальнейшее развитие идеи использования искусственного интеллекта для отраслевых и государственных нужд было упомянуто в паспорте национального проекта Национальная программа "Цифровая экономика Российской Федерации"[[82]](#footnote-82). В качестве продолжения развития данного тренда была опубликована национальная стратегия развития искусственного интеллекта на период до 2030 года и указ президента России Владимира Путина об утверждении этой стратегии[[83]](#footnote-83). Были определены технологии искусственного интеллекта, среди которых определялись традиционные, к которым относятся компьютерное зрение, обработка естественного языка (Natural Language Processing, NLP), интеллектуальная поддержка принятия решений и другие. Также были выделены «перспективные» технологии сильного искусственного интеллекта среди которых кроме можно было увидеть: «автономное решение различных задач, автоматический дизайн физических объектов, автоматическое машинное обучение, алгоритмы решения задач на основе данных с частичной разметкой и (или) незначительных объемов данных, обработка информации на основе новых типов вычислительных систем, интерпретируемая обработка данных и другие методы». Были введены в действие дорожные карты, как например «Дорожная карта развития «сквозной» цифровой технологии «Нейротехнологии и искусственный интеллект»», «Дорожная карта развития «сквозной» цифровой технологии «Технологии виртуальной и дополненной реальности»»[[84]](#footnote-84) и другие (таб. 4).

Законодательство в этой области непрерывно совершенствуется и сейчас уже принят Указ Президента Российской Федерации от 15.02.2024 № 124 "О внесении изменений в Указ Президента Российской Федерации от 10 октября 2019 г. № 490 "О развитии искусственного интеллекта в Российской Федерации" и в Национальную стратегию, утвержденную этим Указом" [[85]](#footnote-85)

Таблица 4. Ключевые технические характеристики технологий искусственного интеллекта в дорожной карте развития «сквозной» цифровой технологии «Нейротехнологии и искусственный интеллект»[[86]](#footnote-86)

|  |  |
| --- | --- |
| Направление развития | Ключевые технические характеристики |
| 1.Компьютерное зрение | − Скорость обработки и передачи информации  − Требования к качеству фото и видео данных  − Объем данных для обучения  − Точность анализа (вероятность ошибки, по сравнению с человеком и др. устройствами)  − Оптические возможности (определение цветов, расстояний и размеров, поиск по шаблонам и др.)  − Требования к аппаратному обеспечению (увеличение разрешения видеосенсоров, динамический диапазон и объем вычислительной мощности для обработки) |
| 2.Обработка естественного языка | − Скорость обработки и передачи информации  − Необходимый объем текстовых библиотек для обучения системы  − Точность анализа (вероятность ошибки, по сравнению с человеком и др. устройствами)  − Требования к аппаратному обеспечению (качество и количество устройств ввода, требования к памяти, CPU)  − Пословная ошибка |
| 3.Рекомендательные системы и системы поддержки принятия решений | − Скорость обработки и передачи информации  − Необходимое качество данных для обучения  − Объем данных для обучения  − Точность (качество вывода)  − Интерпретируемость ответа (объяснимость)  − Способность адаптироваться к изменениям входных данных (самообучаемость)  − Требования к аппаратному обеспечению  − Потенциал масштабирования |
| 4.Распознавание и синтез речи | − Объем данных для обучения  − Требования к качеству аудио данных  − Скорость обработки данных  − Пословная ошибка  − Точность анализа (вероятность ошибки, по сравнению с человеком и др. устройствами)  − Акустические возможности (определение частоты, тембра, силы., исключение шумов)  − Требования к аппаратному обеспечению (качество и количество устройств ввода, требования к памяти, CPU) |
| 5.Перспективные методы и технологии в ИИ | − Энергопотребление  − Решение задач экспоненциальной сложности  − Количество задач, решаемых с помощью ИИ  − Количество успешных решений с применением перспективного метода  − Требования к аппаратному обеспечению |

Предварительное исследование показало возможности использования информационных технологий в сфере спорта, ознакомиться с результатами можно в таблице 5.[[87]](#footnote-87) Безусловно это было только предварительное распределение, но уже оно позволило нам сформировать подход к подготовке кадров для цифровой экономики в сфере физической культуры и спорта.

Таблица 5. Области применения цифровых технологий в спорте.

|  |  |  |
| --- | --- | --- |
| №пп | Цифровая технология | Область применения |
| 1 | Big data (большие данны) | Управление отраслью (объектами и субъектами,); спортивная подготовка (создание модельных характеристик, как спортсмена, так и команды в целом; автоматизированный анализ соревновательной деятельности); разработка новых проектов и процессов (на основе вовлеченности коммьюнити); иное. |
| 2 | AI (искусственный интеллект) | Анализ и предложение решений на различных уровнях (от отдельного спортсмена, до массовых физкультурно-спортивных мероприятий). Благодаря технологии трансформируется деятельность ключевых профессий отрасли «физическая культура и спорт» - от тренера до спортивного судьи. |
| 3 | Cloud Storage (облачные хранилища данных и облачные технологии в целом) | Открытость и доступность отраслевых услуг. Снятие барьеров цифрового неравенства за счет аутсорсинга отдельных (часто дорогостоящих) информационных и технических процессов. С помощью технологии обеспечивается доступ к передовым технологическим ращениям физкультурно-спортивной сферы. Рекомендуется в первую очередь для удаленных и сельских территорий, а также локальных спортивных учреждений. |
| 4 | GPS trackers, носимые гаджеты и сенсорные визуализаторы | Занятия массовой физической культурой и спортом. Технологии способствуют динамической диагностике различных показателей пользователя (физической активности), с одной стороны, а с другой – создают коммуникативные площадки пользователей, вовлекая их в различные активности (в том числе соревновательные). Являются инструментом самодиагностики. |
| 5 | VR и AR (технологии виртуальной и дополненной реальности) | Технологии используются как для моделирования условий, необходимых как для тренировочной, так и соревновательной деятельности. Применение технологий позволяет проводить дистанционные соревнования с использованием физической активности. |
| 6 | iTable, i-wall и т.п. (интерактивные устройства) | Технологии обеспечивают дополнительный функционал спортивным сооружениям и оборудованию, чем способствуют психоэмоциональному включению занимающихся физической культурой и спортом. Эффективно использовать как с начинающими, так и со спортсменами высокого класса. |
| 8 | Компьютерные игры | На основе этих технологий сформирован отдельный вид спорта, а также дисциплины других видов спорта. Например, «интерактивный футбол», «интерактивный бокс», «интерактивный баскетбол» и т.д. Применяется в качестве спортивного инвентаря. Также некоторые компьютерные игры могут использоваться в качестве тренажерных технологий для развития конкретных качеств спортсменов различного уровня. |

Исходя из полученных результатов исследования и поставленных в описываемых нормативных документах задач, как например «Обеспечение цифровой экономики компетентными кадрами» указанное в Паспорте федерального проекта "Кадры для цифровой экономики"[[88]](#footnote-88) в Российском университете спорта «ГЦОЛИФК» была разработана и внедрена в практику образовательная программа магистратуры направления подготовки 49.04.03 «Спорт» была реализована направленность «Цифровая трансформация спорта» в которую было включено значительное количество учебных дисциплин, связанных с как это следует из названия с цифровизацией и информатизацией спортивной сферы и конкретно средствами и методами интеллектуального анализа данных и искусственного интеллекта[[89]](#footnote-89). Распределение зачётных единиц и академических часов по части дисциплин можно увидеть в таблице 6.

Таблица 6. Количественное распределение зачётных единиц и академических часов по части дисциплин образовательной программы магистратуры программы «Цифровая трансформация спорта».

|  |  |  |
| --- | --- | --- |
| **Наименование дисциплины** | **Зач. единиц** | **Акад. часов** |
| **Итого по всем дисциплинам (модулям)** | 78 | 2808 |
| Информационные технологии в науке и спортивной практике | 4 | 144 |
| Цифровые технологии в системе подготовки спортсменов | 10 | 360 |
| Управление цифровыми проектами в сфере спорта | 5 | 180 |
| Цифровая трансформация международного спортивного и олимпийского движения | 4 | 144 |
| Технологии работы с большими данными в спорте | 5 | 180 |
| Защита информации | 3 | 108 |
| Искусственный интеллект и машинное обучение в спорте | 2 | 72 |
| Технологии видеосъемки в спорте | 3 | 108 |
| Цифровые технологии в компьютерном спорте | 3 | 108 |
| Цифровая трансформация государственного управления спортом | 4 | 144 |
| Цифровые коммуникационные технологии в спорте | 2 | 72 |
| Математические методы обработки информации | 3 | 108 |

Для определения возможностей дальнейшей исследовательской работы мы произвели анализ часто встречающихся задач в сфере спорта, соотнеся их с технологиями искусственного интеллекта, которые применялись исследователями в различных видах спорта. В ходе рассмотрения полученных результатов (таблица 7), можно увидеть, что практически все найденные задачи имеют универсальный характер и представлены во всех видах спорта, такой же универсальный характер имеют и технологии искусственного интеллекта, и практически все задачи могут решаться всеми представленными методами.

Таблица 7. Применение технологий искусственного интеллекта для исследований в области спорта

|  |  |  |
| --- | --- | --- |
| **Задача** | **Технология** | **Группы видов спорта** |
| Прогнозирование развития спортивной карьеры | Линейная (полиномиальная) регрессия | Зимние виды спорта, игровые виды, лёгкая атлетика[[90]](#footnote-90)[[91]](#footnote-91)[[92]](#footnote-92)[[93]](#footnote-93) |
| Прогнозирование спортивного результата по характеристикам спортсмена | Логистическая регрессия, древовидные модели, градиентный бустинг | Компьютерный спорт, игровые виды, лёгкая атлетика [[94]](#footnote-94)[[95]](#footnote-95)[[96]](#footnote-96) |
| Выявление неочевидных особенностей спортивной деятельности | Древовидные модели, кластерный анализ | Единоборства, игровые виды[[97]](#footnote-97)[[98]](#footnote-98)[[99]](#footnote-99)[[100]](#footnote-100) |
| Изучение психологических аспектов спортивной деятельности | Кластерный анализ, линейная регрессия | Игровые виды, без дифференциации[[101]](#footnote-101)[[102]](#footnote-102)[[103]](#footnote-103) |
| Вклад отдельных признаков соревновательной деятельности или действий в спортивный результат | Логистическая регрессия, линейная регрессия, древовидные модели | Единоборства, игровые виды[[104]](#footnote-104)[[105]](#footnote-105) |
| Определение взаимовлияния спортивных упражнений (действий) | Линейная (полиномиальная) регрессия, метод опорных векторов | Игровые виды, зимние виды[[106]](#footnote-106)[[107]](#footnote-107) |
| Определение баланса в правилах и регламентах состязаний | Метод опорных векторов, байесовская классификация | Единоборства, компьютерный спорт [[108]](#footnote-108)[[109]](#footnote-109) |
| Профилактика травматизма | Кластерный анализ, метод опорных векторов, глубокое обучение (нейросети) | Зимние виды спорта, игровые виды, [[110]](#footnote-110)[[111]](#footnote-111)[[112]](#footnote-112) |
| Анализ спортивной техники | Глубокое обучение (компьютерное зрение) | Единоборства, гимнастика, игровые виды [[113]](#footnote-113)[[114]](#footnote-114)[[115]](#footnote-115) |

В дальнейшем для получения достаточного количества уникальных данных для исследовательской работы нами была разработана и запатентована программа по визуальному определению амплитуды движений в основных суставах человека при выполнении двигательных спортивных и прикладных упражнений с использованием искусственного интеллекта [[116]](#footnote-116). Для создания программы нами были выбраны продукты, реализованные его на языке Python для упрощения дальнейшей работы с программой и данными получаемыми в результате и имеющие открытый исходный код, а также лицензию предусматривающую свободное распространение продукта. Поэтому для обработки данных внутри программы и вывода результата нами использовалась библиотека Pandas[[117]](#footnote-117) которая предназначена для анализа данных. Для захвата изображения (ввода данных в видеоформате) и дальнейшее преобразование видео осуществлялась с помощью библиотеки «компьютерного зрения» OpenCV[[118]](#footnote-118), которая обеспечивает видеоанализ с алгоритмами оценки движения, возможность обнаруживать людей и отдельных их частей, например, глаз, изменение и вывод изменённых изображений. Для более точной детекции основных точек на теле человека использовался фреймворк MediaPipe[[119]](#footnote-119). Он работает с кроссплатформенными настраиваемыми решениями машинного обучения для уже подготовленных мультимедиа файлов, что оптимально соответствует исследовательским целям, так как позволяет выделить точно требуемый участок видео для работы, однако существует возможность такой же обработки потокового видео.

Фреймворк MediaPipe позволяет осуществлять обнаружение координат до 33 точек на теле человека в двух и трёхмерном пространстве (рис. 4 )

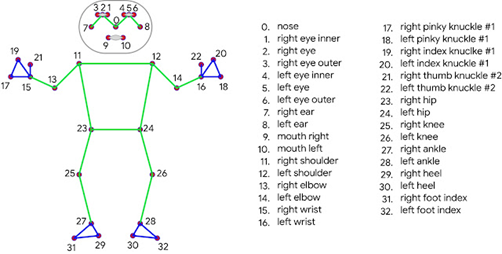


Рис.4 – Координаты точек на теле человека определяемые алгоритмами MediaPipe[[120]](#footnote-120)

После первоначального определения этих точек их координаты начинают рассчитываться автоматически и возможно использование их координат в трёхмерном пространстве для вычисления требуемых метрик движения. И если две координаты естественным образом получаются при определении местоположения точки на плоской матрице, то третьи координаты получаются с помощью модели BlazePose GHUM[[121]](#footnote-121). В результате на каждый фрейм получается массив 33x5 значений. Модель была проверена на 1400 выборках из 14 различных регионов и получили среднюю точность совпадения с локализациями суставов выполненных экспертами было от 89,7 до 96,2% в зависимости от модели (их использовалось три варианта - heavy, full и lite). Оценка производилась на основе набора данных фотографий с задней камеры смартфона и была получена средняя производительность для тяжелой модели 94,2% +/- 1,3% с диапазоном [91,4%, 96,2%] по регионам, средняя производительность для полной модели 91,8% +/- 1,4% с диапазоном [89,2%, 94,0%] по регионам и для облегченной модели средней производительностью 87,0% +/-2,0% с диапазоном [83,2%, 89,7%] по регионам. Оценка средней точности определения точки сустава при сравнении с определением экспертов по набору данных фотографий с задней камеры смартфона для тяжелой модели дает среднюю производительность 93,6% при диапазоне [89,3%, 95,0%], для полной модели среднюю производительность 91,1% при диапазоне [85,9%, 92,9%], для облегченной модели со средней производительностью 86,4% при диапазоне [80,5%, 87,8%]. Максимальное расхождение между категориями худших и наилучших характеристик составляет 5,7% для тяжелой модели, 7,0% для полной модели и 7,3% для облегченной модели. Оценка по признаку пола дает среднюю производительность 94,8% при диапазоне [94,2%, 95,3%] для тяжелой модели, среднюю производительность 92,3% при диапазоне [91,2%, 93,4%] для полной модели и в среднем 83,7% с диапазоном [86,0%, 89,1%] для облегченной модели. Максимальное расхождение составляет 1,1% для тяжелой модели, 2,2% для полной модели и 3,1% для облегченной модели[[122]](#footnote-122)[[123]](#footnote-123).

Сравнение продуктивности MediaPipe (трёх описываемых выше моделей BlazePose GHUM) с другими моделями было произведено авторами фреймворка по 17 ключевым точкам топологии COCO[[124]](#footnote-124) на наборах данных содержащих движения танца, выполнения упражнений йоги и высокоинтенсивного интервального тренинга (HIIT).

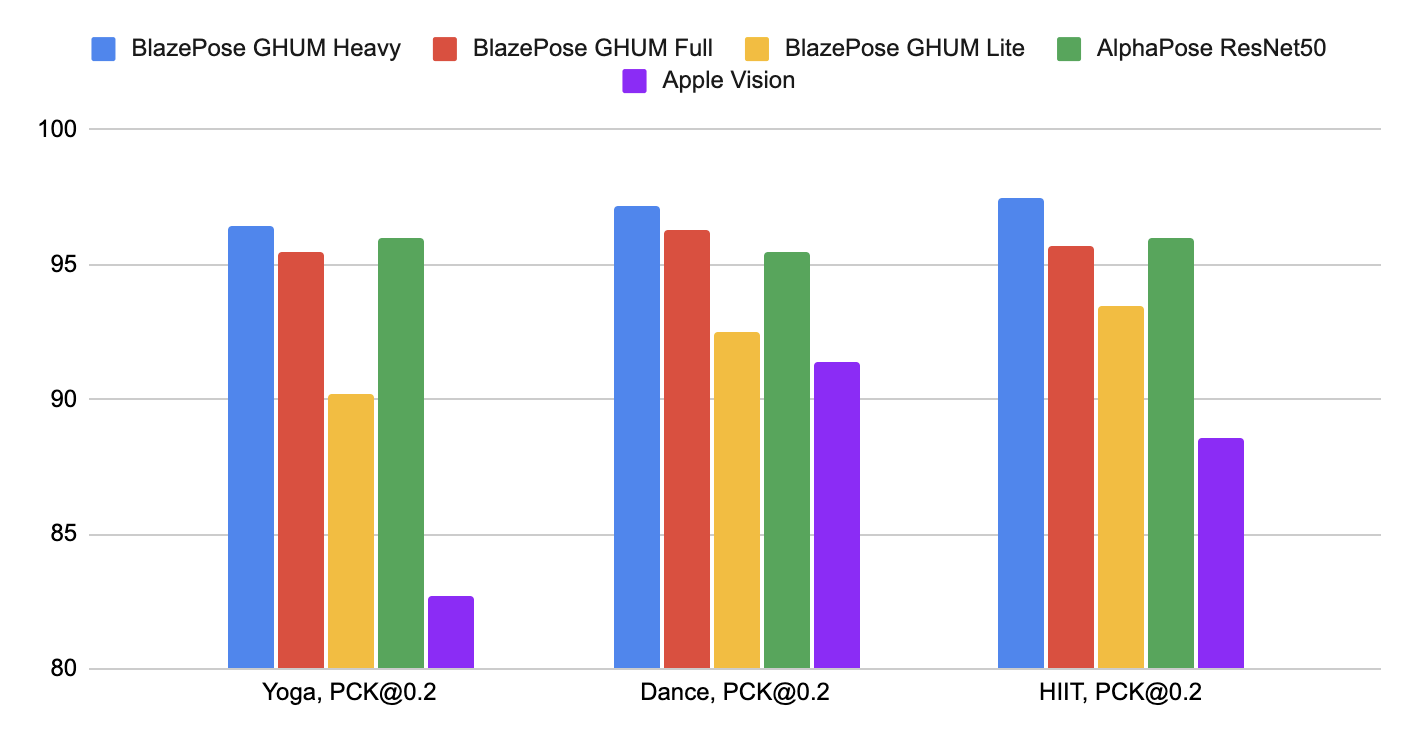


Рис.5 - Результаты сравнения MediaPipe (трёх моделей BlazePose GHUM) с другими моделями

Результаты данного сравнения можно увидеть на рис. 5[[125]](#footnote-125)

Другое исследование так же показало преимущество алгоритмов MediaPipe перед другими средствами оценки человеческой позы. в результате проведённых исследований было определено, что MediaPipe показывает наименьшую среднеквадратичную ошибку при определении углов в суставах человека (рис.6)[[126]](#footnote-126).

Безусловно это позволяет достаточно оперативно получать важную информацию о выполнении спортивного упражнения, при этом постоянно повышающаяся продуктивность в работе технологий искусственного интеллекта позволяет предположить, что в обозримом будущем будет возможным не только проводить исследование быстрее и точнее, но и сделает их неотъемлемой частью всех аспектов спортивной подготовки и физкультурно-спортивного образования.

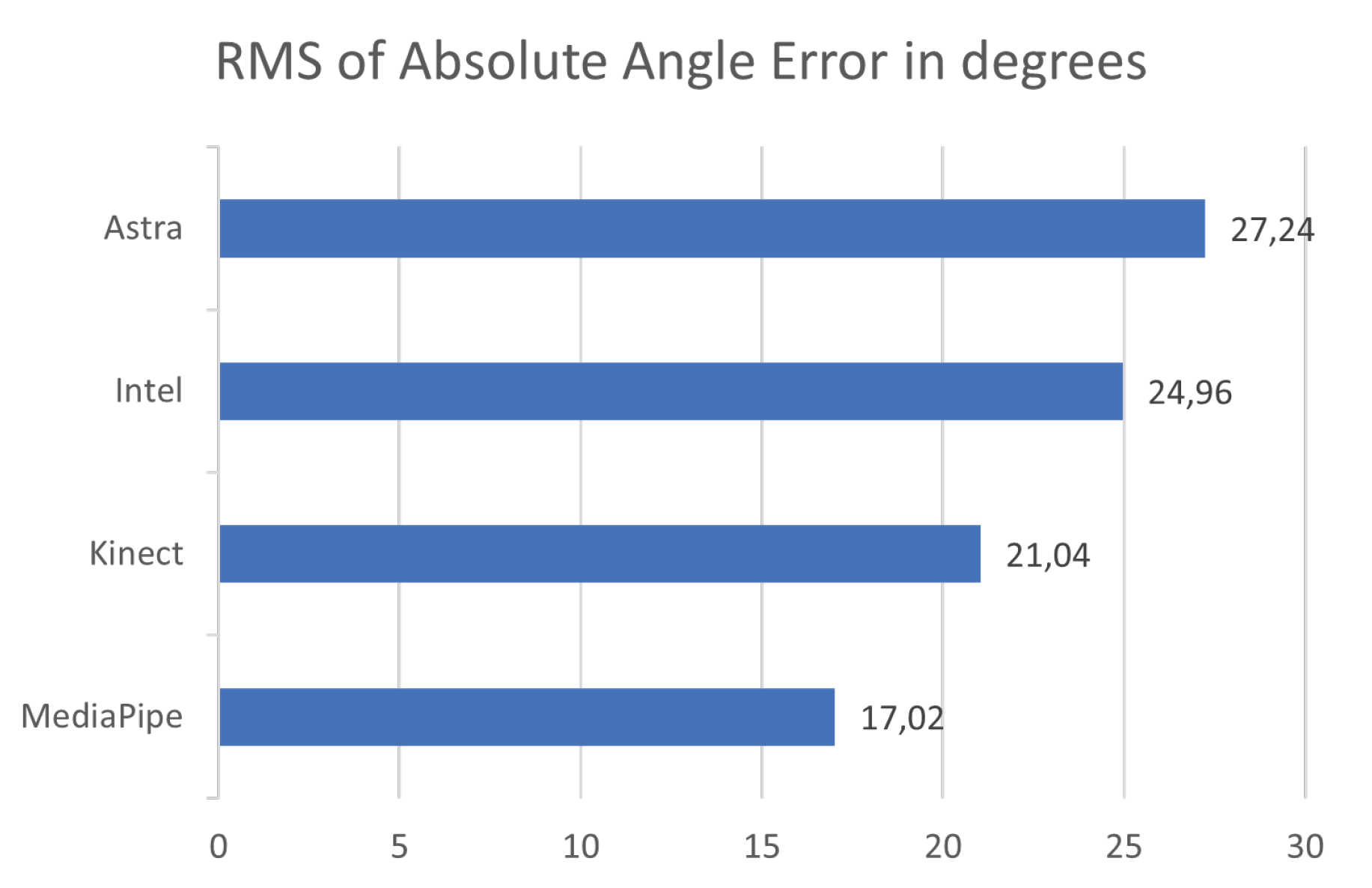


Рис.6 – Среднеквадратичная ошибка при определении углов в суставах человека при помощи MediaPipe и других моделей (Lafayette, T.B.G., Kunst, V.H.L., Melo, P.V.S, 2022)

Получив таким образом координаты определённых точек на теле человека мы используя средства и методы аналитической геометрии соединив точки между собой сообразно проекционным линиям соединяющим основные суставы мы определили вектора и используя скалярное произведение с помощью библиотеки NumPy определили углы между этими векторами[[127]](#footnote-127).

В дальнейшем определив углы в лучезапястном, локтевом, плечевом, тазобедренном, коленном и голеностопном суставе в двух и трёхмерной проекциях, а также угла ротации туловища. Всего от 25 параметров движения спортсмена определяются в среднем 30 раз в секунду (модель рассчитывалась до 40 фреймов).

Поэтому даже 10 секунд видео дают вывод в таблицу 300х25, что снабжает исследователя достаточным количеством данных для анализа и построения моделей.



Рис.7 – Пример использования искусственного интеллекта в ходе оценки проведённого учебно-тренировочного занятия

Кроме прочего на имеющееся изображение наносится дополнительно скелетная структура с численным обозначением величин углов в локтевых, плечевых, тазобедренных и коленных суставов, номером фрейма и значением частоты кадров для данного фрейма (рис. 7). При анализе получаемой информации возможно сделать предварительные выводы об отдельных характеристиках движения. В результате любой тренер работающий со спортсменами любого уровня подготовки может генерировать данные достаточно большого объёма, анализ которых в свою очередь может быть основой для принятия обоснованного решения как об организации спортивной подготовки спортсмена, так и оптимизации его соревновательной деятельности.

Безусловно описанные случаи не ограничивают технологии искусственного интеллекта. Согласно ГОСТ Р 59277-2020 «Системы искусственного интеллекта. Классификация систем искусственного интеллекта»[[128]](#footnote-128) их разннобразие велико (рис.8)

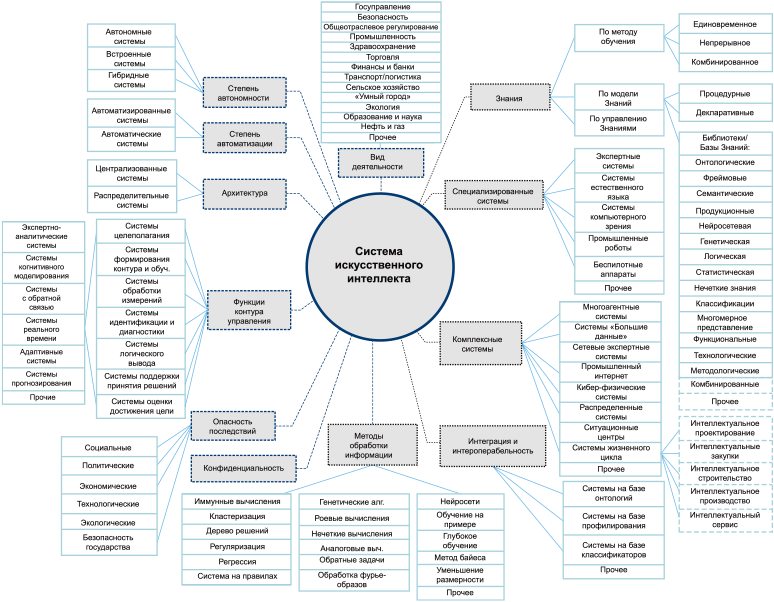


Рис.8 – Технологии составляющую систему искусственного интеллекта (по ГОСТ Р 59277-2020 Системы искусственного интеллекта. Классификация систем искусственного интеллекта)

Наиболее же распространённым и изученным направлением в технологиях искусственного интеллекта, которое по мнению исследователей является основой для применения в сфере спорта является машинное обучение (Machine Learning)[[129]](#footnote-129). С упрощенной его структурой можно ознакомиться на схеме (рис.9)

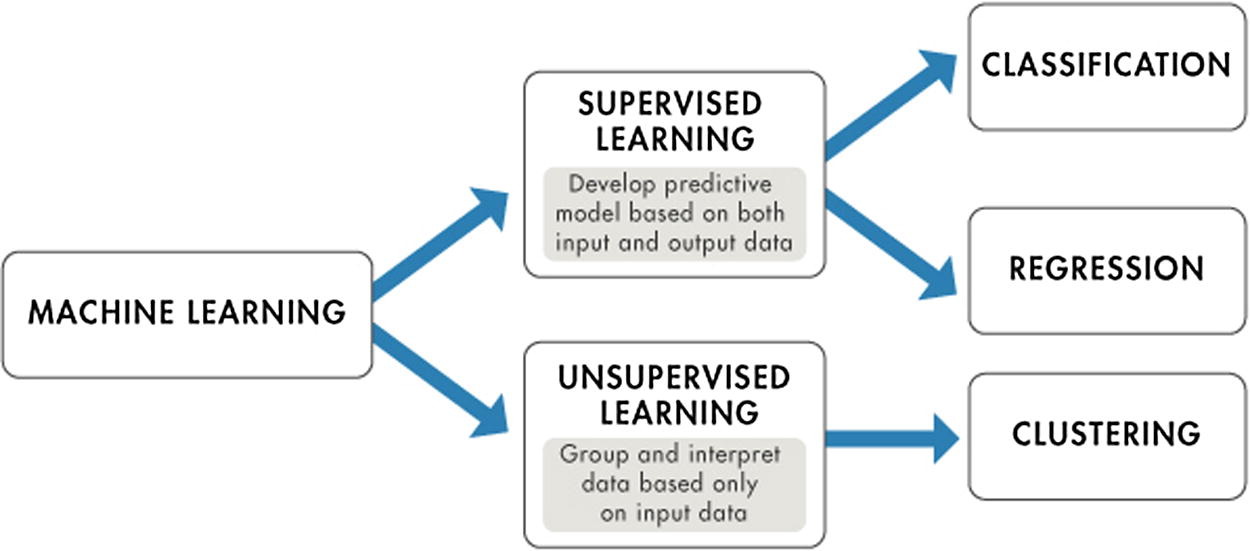


Рис.9 – Структура системы машинного обучения для использования в сфере спорта (по Bunker R. P., Thabtah F., 2019)

Согласно приведённой структуры в спортивных исследованиях решаются следующие основные задачи – регрессии, классификации и кластеризации. Первые две из них относятся к так называемому обучению с учителем (Supervised learning - SL). Этот набор методов машинного обучения характеризуется тем, что разрабатываемая модель обучается на размеченных данных (истинный результат для зависимой переменной известен модели), а применяется уже на данных результат для которых (зависимая переменная) ещё неизвестна[[130]](#footnote-130). В свою очередь третья задача, которая обозначена как кластеризация, то есть объединение данных в группы по выбранным критериям сходства относится к так называемому обучению без учителя (Unsupervised learning).

Задача кластеризации очень требовательна к квалификации исследователя и это накладывает ограничения на использование её в повседневной практике спортивной подготовки. Поэтому несмотря на всю ценность, которую она имеет при формировании гипотез при анализе неочевидных закономерностей в данных данная задача в этой работе рассматриваться не будет и данная проблема будет изучаться в дальнейших работах, позволяя сконцентрироваться на более универсальных в своей применимости для работы в спортивной сфере задачах регрессии и классификации.

В результате вышеизложенного в данном разделе можно сделать следующие выводы:

Развитие технологий искусственного интеллекта сейчас является одним из стратегических направлений развития Российской Федерации. Нормативно правовые акты, определяющие место информационных технологий и искусственного интеллекта в частности постоянно совершенствуется и влияет на все социально экономические сферы нашей страны.

Конвергенция информационных технологий и искусственного интеллекта в частности с физкультурно-спортивной сферой может позволить последней совершить качественный скачок в совершенствовании и достичь имеющихся целей быстрее и надёжнее.

Искусственный интеллект может быть использован для совершенствования спортивной подготовки, соревновательной деятельности и физкультурно-спортивного образования.

Перспективным можно считать использование технологий искусственного интеллекта для автоматизации исследований на основе анализа данных, в том числе и больших объёмов для использования в ежедневной спортивной деятельности нга всех её уровнях.

# 3. Анализ больших объёмов данных в сфере физической культуры и спорта

Большие данные (big data)в качестве термина уже давно используется как в научном так и нормативно-правовом поле. Однако точное однозначное определение всё ещё составляет проблему. «Большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа[[131]](#footnote-131)». Поэтому не представляется возможным однозначно отнести к большим датам конкретный набор данных определённого размера. Отсутствие подобного однозначного толкования данного термина порождает ситуацию невозможности обеспечивать корректную коммуникацию между специалистами, исследователями, различными ветвями власти и прочими заинтересованными сторонами в сфере физической культуры и спорта.

Под большими данными часто понимаются «массивы данных (dataset), которые не могут быть восприняты, получены, управляемы и обработаны традиционными ИТ и программно-аппаратными средствами в течение допустимого времени»[[132]](#footnote-132). Подобную сложность определения термина «большие данные» отмечали многие авторы вне зависимости от сферы в которой большие данные определялись[[133]](#footnote-133)[[134]](#footnote-134).

Попытки определить границы больших данных в зависимости от их объёма не увенчались успехом. В разное время их определяли как данные размером более 100 Гб[[135]](#footnote-135), или как «те для обработки которых требуется минимум два сервера и, соответственно это данные более 3 терабайтов»[[136]](#footnote-136).

Таким данным определение больших данных через их объём изменялось от килобайт, мегабайт, гигабайт до многих экзабайт[[137]](#footnote-137).

Отдельные исследователи пытаются количественно определить большие данные в зависимости от общего количества данных генерируемых человечеством в совокупности на момент определения. Однако, из 4,4 ЗБ данных, собранных человечеством на 2013 год, только 1,5% из этого объёма обладала информативной ценностью. Более подробно эти данные представлены в таблице (таб.1).

Таблица 1. Объем данных, накопленных человечеством[[138]](#footnote-138).

|  |  |
| --- | --- |
| Год | Объём данных генерируемый человечеством |
| 2003 г. | 5 экзабайт (1 ЭБ = 1 миллиард гигабайт). |
| 2008 г. | 0,18 зеттабайта (1 ZB = 1024 экзабайта). |
| 2015 г. | более 6,5 зеттабайт |
| 2020 г. | 40–44 зеттабайта |
| 2025 г. | 440 зеттабайта объем вырастет еще в 10 раз(прогноз). |

Отмечается так же, что большая часть данных будет генерироваться не потребителями, а разнообразными гаджетами в автоматическом режиме (интернетом вещей).

Рост данных не прекращается, например, к 2020 году в перерасчёте на каждого человека генерируется около 1,7 мегабайт новой информации в секунду и не менее трети всех этих данных обрабатывается с помощью облачных технологий[[139]](#footnote-139).

Исходя из вышеизложенного можно выделить два основных подхода к определению. Первый из них определяет большие данные согласно ограничениям технологий, предназначенных для получения, хранения и обработки данных. Второй подход использует для определения больших данных их собственный объём, общий объём данных генерируемых человечеством и полезную часть генерированных данных. Однако присутствует мнение, что необходимо учитывать и собственные ограничения, человека обеспечивать необходимое качество восприятия только ограниченного объёма информации. Проблема ограничений человека в возможностях потреблять и понимать большие данные была поднята задолго до их экспоненциального роста и несмотря на то, что развитие технологий позволяет хранить и обрабатывать все большие объемы данных, но наши сенсорные и когнитивные способности обрабатывать этот поток данных ограничены возможностью обрабатывать около семи фрагментов информации в своей кратковременной памяти[[140]](#footnote-140).

В результате отсутствия продуктивности определения больших данных исключительно через их объём исследователи предприняли попытку использовать именно характеристики ограничений, накладываемых размером данных на эффективное взаимодействие с ними.

Модель чаще всего имеет название "3V" модель (volume, velocity, variety), как очевидно из названий – это объём, скорость и вариативность, изначально авторы использовали «3D-модель». В дальнейшем количество параметров получило тенденцию к увеличению и появились 4V, 5V, 6V, 7V, 9V[[141]](#footnote-141). Более подробно с V-моделями различной размерности можно ознакомиться в таблице (табл.1)

Таблица1. Содержание V-моделей различной размерности

|  |  |  |
| --- | --- | --- |
| 3Vs | Volume, Velocity, Variety[[142]](#footnote-142) | Объем, скорость, разнообразие |
| 4Vs | Volume, Velocity, Variety, Value[[143]](#footnote-143)[[144]](#footnote-144) | Объем, скорость, разнообразие, ценность |
| 5Vs | Volume, Velocity, Variety, Value, Veracity[[145]](#footnote-145)[[146]](#footnote-146) | Объем, скорость, разнообразие, ценность, достоверность |
| 6Vs | Volume, Velocity, Variety, Veracity, Validity, Volatility[[147]](#footnote-147) [[148]](#footnote-148) | Объем, Скорость, Разнообразие, Достоверность, Валидность, Изменчивость |
| 7Vs | Volume, Velocity, Variety, Veracity, Value, Validity, Volatility[[149]](#footnote-149) | Объем, Скорость, Разнообразие, Достоверность, Ценность, Валидность, Изменчивость |
| 9Vs | Volume, Velocity, Variety, Veracity, Value, Variability, Validity, Volatility, Visualization[[150]](#footnote-150) | Объем, Скорость, Разнообразие, Достоверность, Ценность, Изменчивость, Валидность, Волатильность, Визуализация |

Рассмотреть всю вариативность подходов к определению больших данных можно обратиться к анализу приведённому в исследовании De Mauro[[151]](#footnote-151) (таблица 2).

Таблица 2. Вариативность определений больших данных в зарубежной научной литературе по De Mauro, 2015.

|  |  |
| --- | --- |
| Источник | Определение |
| M. A. Beyer and D. Laney, The Importance of “Big Data”: A Definition, 2012[[152]](#footnote-152) | Большой объем, скорость и разнообразие информационных активов, которые требуют экономически эффективных, инновационных форм обработки информации для повышения качества извлекаемых знаний и принятия решений |
| Dijcks J. P. Oracle: Big data for the enterprise, 2012[[153]](#footnote-153) | Четыре характеристики, определяющие большие данные, — это объем, скорость, разнообразие и ценность. |
| Peer Research, Big Data Analytics, 2012[[154]](#footnote-154) | Сложные, неструктурированные или большие объемы данных |
| Suthaharan S. Big data classification: Problems and challenges in network intrusion prediction with machine learning, 2014[[155]](#footnote-155) | Мощность (объём), непрерывность (невозможность разделить на части) и сложность |
| Schroeck M. et al. Analytics: The real-world use of big data //IBM Global Business Services, 2012.[[156]](#footnote-156) | Большие данные — это сочетание объема, разнообразия, скорости и достоверности, которое дает организациям возможность получить конкурентное преимущество на современном цифровом рынке |
| Framework D.N.B.D.I. NIST Big Data Interoperability Framework, 2014[[157]](#footnote-157) | Большие массивы данных, характеризуются в первую очередь объемом, скоростью и/или разнообразием, которые требуют масштабируемой архитектуры для эффективного хранения, обработки и анализа |
| Ward J. S., Barker A. Undefined by data: a survey of big data definitions, 2013[[158]](#footnote-158) | Хранение и анализ больших и/или сложных массивов данных происходит с использованием ряда методов, включая, помимо прочего: NoSQL, MapReduce и машинное обучение |
| Customers Rapidly Adopting Big Data Solutions — Driven By Marketing, Sales and More — Reports New Microsoft Research, 2013[[159]](#footnote-159) | Процесс применения большой вычислительной мощности, последних достижений в области машинного обучения и искусственного интеллекта, для обработки больших и часто очень сложных наборов информации |
| Dumbill E. Making sense of big data //Big Data, 2013[[160]](#footnote-160) | Данные, которые превышают возможности обработки обычных систем баз данных. |
| Fisher D. et al. Interactions with big data analytics //interactions, 2012[[161]](#footnote-161) | Данные, которые не могут быть обработаны и обработаны простым способом. |
| Shneiderman B., in Proc. 2008 ACM SIGMOD Int. Conf. Manag. Data, 2008[[162]](#footnote-162) | Массив данных слишком велик для размещения на одном экране |
| Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C.,Big data: The next frontier for innovation, competition, 2011[[163]](#footnote-163) | Массивы данных, размер которых превышает возможности стандартных инструментов программного обеспечения баз данных для сбора, хранения, управления и анализа |
| Chen, H., Hsinchun, R., Chiang, R.H.L. and Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact, 2012[[164]](#footnote-164) | Массивы данных и методы их анализа настолько велики и сложны, что требуют передовых и уникальных технологий хранения, управления, анализа и визуализации данных. |
| Boyd D., Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, 2012[[165]](#footnote-165) | Культурное, технологическое и научное явление, основанное на взаимодействии технологии, анализа и мифологии. |
| Mayer-Schönberger V., Cukier K. Big data: A revolution that will transform how we live, work, and think, 2013[[166]](#footnote-166). | Феномен, который приводит к трем ключевым изменениям в способах анализа информации, меняющим наше понимание и организацию общества. |

Рассматривая сферу физической культуры и спорта можно заметить, что наиболее распространённым средством работы с данными является приложение, предназначенное для работы с электронными таблицами Microsoft Excel в которой существует ограничение на 1 048 576 строк и 16 384 столбцов[[167]](#footnote-167). Можно взять эти ограничения за определение больших данных, но в сфере физической культуры и спорта пока редко выходят за значение тысячи строк, как пример можно привести количественный обзор данных приведённых в диссертациях по направлениям физической культуры и спорта[[168]](#footnote-168)

Таблица 3. Количественный анализ данных представленных в табличном виде в диссертационных исследованиях, защищённых в марте 2021 года (Ермаков А.В., 2022)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № исследования | Количество испытуемых | Количество измерений/переменных | | | | | | | |
| **измерения** | | | | **переменные** | | | |
| **максимум** | **медиана** | **среднее** | дисперсия, | **максимум** | **медиана** | **среднее,** | дисперсия, |
| 1 | 20, 26 | **12** | **5** | **6.2** | 11.74 | **6** | **4** | **4.47** | 1.84 |
| 2 | 214, 63 | **42** | **4** | **7** | 73 | **15** | **5** | **5.69** | 9.65 |
| 3 | 15 | **11** | **3** | **5.21** | 10.03 | **14** | **8** | **7.16** | 13.72 |
| 4 | 21,174 | **30** | **6** | **9.25** | 90.33 | **6** | **3** | **3.19** | 0.83 |
| 5 | 38 | **38** | **7** | **9.46** | 82.26 | **9** | **6** | **5.46** | 4.36 |
| 6 | 12,26 | **24** | **7** | **8.37** | 25.69 | **20** | **7** | **7.8** | 18.37 |
| 7 | 10,10 | **30** | **8** | **9.06** | 36.46 | **8** | **6** | **5.62** | 3.18 |
| 8 | 919 | **105** | **13** | **17.55** | 354.29 | **19** | **8** | **10.94** | 26.71 |

Из приведённых данных можно сделать вывод, что данные в таблице из тысячи строк и десяти столбцов сейчас достаточно большие для исследований в сфере физической культуры и спорта, однако такая ситуация лишь показывает отставание и потребность его компенсировать.

Существует много инструментов для создания достаточно больших наборов данных для сферы физической культуры и спорта. К очевидным относятся обработка видео, данные с носимых гаджетов или инструменты социальных сетей. Так, например, в условиях пандемии Covid-19 было затруднено проведение исследований в связи с накладываемыми ограничениями на контакты между людьми, однако средства социальных сетей предоставили возможность провести исследование двигательной активности населения. В итоге в опросе только в одной из групп социальной сети «В контакте» приняло участие 1754 человека из нескольких стран (рис.1)[[169]](#footnote-169).



Рис.1 - Страница опроса, проведённого в социальной сети «В контакте» (Lyakh Y., Ermakov A, Sarafanova E., 2021)

Таким образом использование информационных технологий значительно расширяет возможность получать данные в том числе и большого размера для дальнейшего их анализа. При этом анализ больших данных естественно будет иметь определённые особенности, затрудняющие его, которые будут рассмотрены позже. При этом в некоторых случаях величина данных может значительно облегчить анализ. Например, если доступна визуализация данных, то даже из простых, но больших по размеру данных можно извлечь достаточно много неочевидных закономерностей. Рассматриваемое ранее исследование, проведённое на данных производителя носимых гаджетов, где фиксировалось количество пройдённых шагов в день, пол, возраст и место жительства испытуемых при визуализации с разделением по гендерному признаку позволяет сделать интересное наблюдение (рис.1)

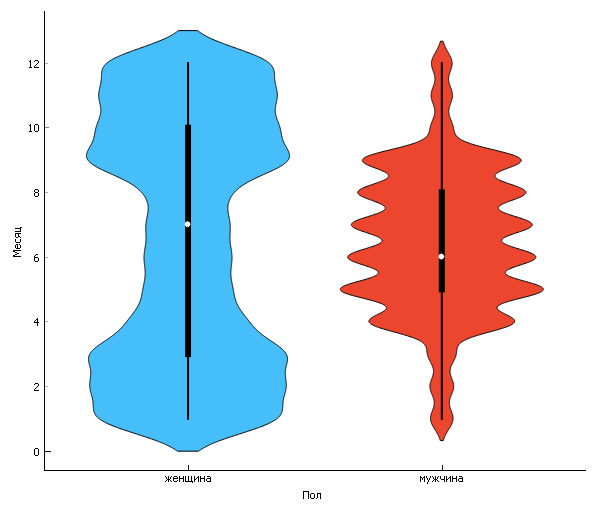


Рис.1 – Анализ двигательной активности населения с разделением по гендерному признаку (Ермаков, 2020)

Наблюдение состоит в том, что мужчины и женщины предпочитают различные сезоны для двигательной активности, и если мужчины чаще всего предпочитают «теплые» месяцы, то женщины напротив активизируются в более «холодные» месяцы. Таким образом возникает знание об особенности, которую можно использовать при планировании деятельности отдельно для мужчин и женщин увеличивая её эффективность (например, планирование отдельных мероприятий, предложения услуг, пропаганды здорового образа жизни и многого другого).

Если же сделать визуализацию более сложной и включить в неё разделение не только по гендерному, но и по возрастному признаку, то можно выделить информацию дающие ещё более глубокие знания этого явления. (рис.1)

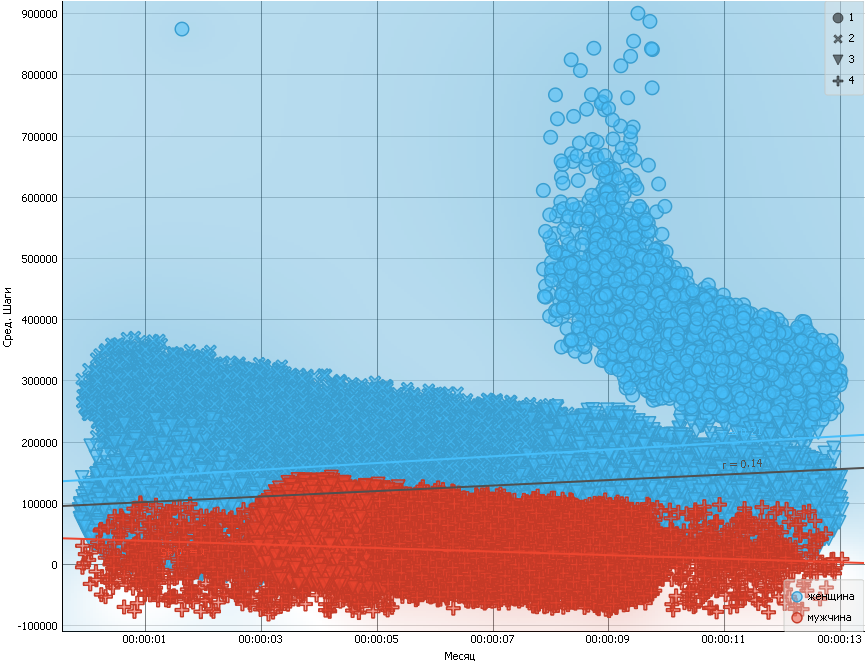


Рис.1 - Анализ двигательной активности населения с разделением по гендерному и возрастному признаку (Ермаков, 2020)

В результате визуального анализа можно однозначно зафиксировать всплеск интереса к увеличению двигательной активности у девушек до 18 лет начиная с второй половины августа с завершением в ноябре-начале декабря, что безусловно может быть использовано с пользой в большом количестве случаев.

Таким образом можно сделать следующие выводы

Большие данные в сфере спорта несут в себе как сложности в их обработке, так и новые возможности для получения знаний. Эти сложности и возможности относятся как к получению данных, так и к их анализу.

На текущий момент времени объём данных в сфере физической культуры и спорта по отношению к другим сферам человеческой деятельности относительно невелико и таблица размером более 1000 строк может уже считаться большой.

Действительно большие данные можно получать при помощи технических средств: обработки потокового видео, информации из социальных сетей, данные носимых гаджетов.

Автоматизация получения данных связана с автоматизацией их обработки и позволяет создавать решения, делающие повседневную спортивную активность предметом научного исследования.

# 3.1. Интернет вещей и носимые технологии источник данных в спорте

Большие объёмы данных должны порождаться также технологично так, как только автоматизация этого процесса может создать действительно большие потоки данных обработка которых сможет принести новое качество как спортивной науке, так и спортивной практике. Одним из таких решений является описываемая выше обработка видеоматериала, а другим носимые технологии, которые способны непрерывно генерировать данные о состоянии спортсмена или группы спортсменов. Исследователи считают, что эти интернет вещей, «умные» и носимые технологии третьей волной ИТ, способной трансформировать многие отрасли включая спортивную[[170]](#footnote-170). Когда-то используемые нами устройства состояли из механических, а потом и из электрических частей, но теперь они превратились в сложные системы, сочетающие в себе аппаратную составляющую, датчики, хранилища данных, микропроцессоры, программное обеспечение и средства связи множеством способов. Таким образом выделяются три основных компонента интернета вещей:

* физический включающий механические и электрические элементы;
* «умный» включающий датчики, микропроцессоры, хранилища данных, элементы управления, программное обеспечение;
* обеспечивающий связь проводным и/или беспроводным способом включающий порты, антенны, протоколы связи по схемам один с одним, один с многими и многие с многими

Понятие «интернет вещей» возникло, чтобы отразить растущее число «умных» продуктов с выходом в компьютерную сеть и прежде всего глобальную сеть интернет и подчеркнуть новые возможности, которые они могут представлять. Интернет, независимо от того, касается ли он людей или вещей, — это просто средство для передачи информации. Фундаментальные различия кроются в природе тех самых «вещей», которые должны быть «умными». Возможности «умных» технологий можно сгруппировать по четырем направлениям: мониторинг, контроль, оптимизация и автономность. Каждый основывается на предыдущем; Например, чтобы иметь возможность управления, устройство должно иметь возможность мониторинга. Носимые устройства как правило позволяют осуществлять комплексный мониторинг состояния спортсмена, выполнение им отдельных упражнений и состояния внешней среды с помощью датчиков или внешних источников данных. Сейчас используется большое количество устройств, как правило носимых, основной целью которых является мониторинг. Например, GPS-трекеры[[171]](#footnote-171)[[172]](#footnote-172) или мониторы сердечного ритма, акселерометры, в том числе использующие для обработки полученной информации и такие сложные методы как нейронные сети[[173]](#footnote-173). Обрабатывая получаемые данные, устройство может предупреждать пользователей и других лиц об изменениях состояния внешней или внутренней среды, тем самым предотвратить вероятный вред здоровью или скорректировать тренировочный план. «Умными» устройствами можно управлять с помощью удаленных команд или алгоритмов, встроенных в устройство или находящихся в облаке продукта. Наиболее часто это может встречаться при управлении нагрузкой на спортивном тренажёре. Поток данных мониторинга от «умных» устройств в сочетании с возможностью контроля работы устройства позволяет оптимизировать. Например, порекомендовать новый тренировочный план или автоматически подобрать характеристики работы устройства в зависимости от индивидуальных особенностей пользователя. Возможности мониторинга, контроля и оптимизации в совокупности позволяют интеллектуальным подключенным продуктам достичь высокого уровня автономности. Достаточно сложные устройства способны не просто работать автономно выполняя предусмотренный заранее алгоритм, но и изучать внешнюю и внутреннюю среду, самостоятельно диагностировать свои потребности в услугах и адаптироваться к предпочтениям пользователей. Автономные устройства способны действовать в координации с другими продуктами и системами. Ценность этих возможностей может расти в геометрической прогрессии по мере того, как все больше и больше устройств подключаются к сети. В итоге возникает возможность создания полноценного искусственного тренера управляющего процессом спортивной подготовки через «умные» устройства, носимые технологии и другие подобные решения высокой автономности.

На изображении (рис.1) представлен пример принципиальной схемы «умного» устройства применяемого для получения данных в спортивных соревнованиях [[174]](#footnote-174). Приведённая схема состоит из нескольких уровней, включая аппаратное обеспечение, встроенное программное обеспечение, возможности подключения, облако продуктов, состоящее из программного обеспечения, работающего на удаленных серверах, набор инструментов безопасности, шлюзы для внешних источников информации и интеграцию с другими системами.

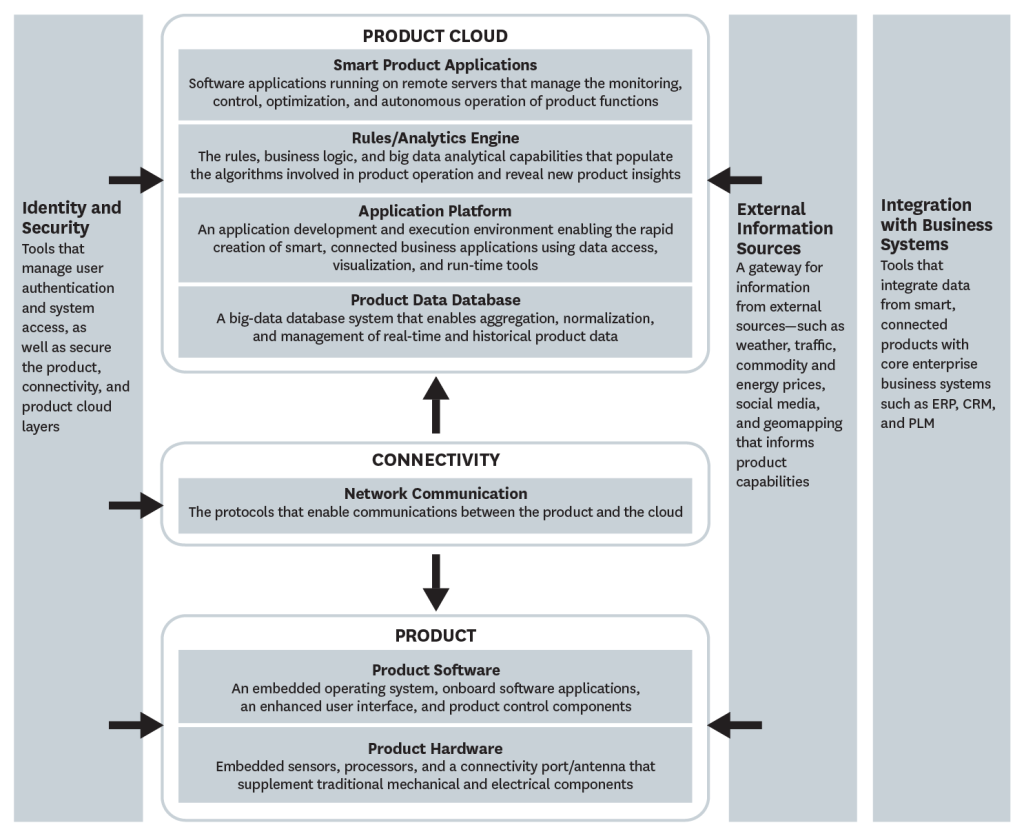


Рис. Принципиальная схема работы «умного» устройства в спорте (Porter M. E, 2014)

Роль обратной связи в спорте имеет большое значение, и как тренеры, так и спортсмены могут извлечь из нее пользу как средство улучшения спортивных результатов или минимизации риска травм. Тренерский процесс может быть весьма субъективным и получение объективных данных может значительно его улучшить[[175]](#footnote-175). Носимые технологии в этом процессе, представляют собой альтернативный подход, который может преодолеть ограничения, накладываемые традиционно используемым лабораторным оборудованием и сделать получение данных не отдельным меропрриятием, а включить его в повседневную спортивную практикую носимые датчики спроектированы так, чтобы быть небольшими, легкими, беспроводными не ограничивающими движений. Это и дает возможность наблюдать за спортсменами за пределами лабораторных условий и в их естественной тренировочной деятельности. Датчики включающие в себя акселерометры используются в таких видах спорта, как лыжные гонки, сноуборд и плавание, которые проводятся в особых средовых условиях условиях и обладают дополнительными характеристиками, такими как водонепроницаемость или способность выдерживать низкие температуры при записи данных[[176]](#footnote-176)[[177]](#footnote-177)[[178]](#footnote-178)[[179]](#footnote-179)[[180]](#footnote-180)[[181]](#footnote-181). . Однако они не лишены ограничений: наличие сильных магнитных полей может исказить измерения, полученные с помощью инерциальных систем[[182]](#footnote-182), а также могут быть ошибки при попытке экстраполяции координат[[183]](#footnote-183)[[184]](#footnote-184). Кроме того, использование беспроводного метода передачи данных может привести к потере сигнала во время записи или помехам со стороны мобильных телефонов или других устройств, которые могут находиться на той же частоте передачи[[185]](#footnote-185).

Несмотря на описываемые недостатки мониторинг рабочей нагрузки игроков устройств глобальной системы позиционирования (GPS) позволяет некоторым исследователям анализировать нагрузку, которую ранее выполняли с использованием мониторов сердечного ритма, так как они считают, что использование только частоты пульса не является точным методом определения нагрузки, поскольку оно не учитывает скорость и расстояние, пройденное во время игры[[186]](#footnote-186).

Носимые датчики использовались в американском футболе для мониторинга сотрясений мозга путем измерения линейного и углового ускорения головы при ударе[[187]](#footnote-187). Датчики были интегрированы в подкладки и каппы шлемов, что подчеркивает их незаметность, что является ключевым фактором при их использовании на тренировках и соревнованиях.

Носимые устройства уже давно способны осуществлять достаточно сложные биомеханические измерения спортивных движений в том числе и в игровых видов спорта для оценки игры игрока на поле при контактном взаимодействии с соперниками[[188]](#footnote-188). Существует широкий спектр датчиков, которые в настоящее время используются психофизиологами, которые информируют о когнитивных процессах, эмоциональных состояниях и поведенческих моделях неинвазивными способами. Некоторые из них появились несколько десятилетий назад, а многие новые технологии в последние годы позволили создать различные конструкции датчиков: электрические, магнитные, электрохимические, механические, тепловые и оптические. Сейчас доступны портативные, безопасные и недорогие сенсорные системы[[189]](#footnote-189). Уже давно для увеличения эффективности и точности регистрации и обработки полученных данных используются такие сложные цифровые инструменты глубокое обучение (нейронные сети), вероятностные модели и другие технологии искусственного интеллекта. В качестве примера можно привести систему для регистрации движений глаз[[190]](#footnote-190) или классификацию движений в крикете[[191]](#footnote-191), особенностей техники бегунов[[192]](#footnote-192), и использования сложных систем из носимых устройств и видеокамер для оценки биомеханических характеристик пловцов[[193]](#footnote-193). Носимые технологии для контроля используют и при мониторинге соблюдений правил соревнований[[194]](#footnote-194)[[195]](#footnote-195)[[196]](#footnote-196).

Подводя итог можно сделать следующие выводы:

Носимые технологии позволяют собирать данные о большом спектре двигательных активностей спортсмена.

Носимые технологии могут активно использоваться в повседневной учебно-тренировочной деятельности, не нарушая её и тем самым служить источником больших данных для исследований в сфере физической культуры и спорта.

Качество данных получаемых от носимых технологий постоянно повышается и усложняется их структура.

Технологии искусственного интеллекта и облачные технологии оптимально подходят для объединения в единую систему с технологиями искусственного интеллекта для конвергенции с сферой физической культуры и спорта.

# 3.2. Исследование, подготовка и статистический анализ данных

Интеллектуальный анализ данных (DM, Data Mining)[[197]](#footnote-197) рассматривается как одну из фаз процесса обнаружения знаний в базах данных (KDD, Knowledge discovery data-bases)[[198]](#footnote-198). При этом важнейшей задачей считается автоматизация исследовательской деятельности за счёт автоматизации обработки «сырых» данных, которая в свою очередь как сокращает количество информации до размеров удобных для восприятия и последующего анализа, так и преобразует её в удобный для анализа вид[[199]](#footnote-199). Этап же интеллектуального анализа данных касается, главным образом, извлечения из данных, а точнее создание новых знаний, которые представляют из себя более сложные закономерности, чем те, которые обычно получают, используя только методы описательной статистики и сравнения распределений. В литературе не всегда однозначно различаются эти подходы[[200]](#footnote-200). Однако при появлении всё больших объёмов данных потребность в более совершенных инструментах анализа возрастала и основной целью стало недопущение возможной потери знаний при обработке всё больших объёмов информации[[201]](#footnote-201)[[202]](#footnote-202). С конца прошлого века наблюдается рост количества и попыток консолидации средств и методов, предназначенных для интеллектуального анализа данных[[203]](#footnote-203). Предпринимаются некоторые усилия, направленные на установление стандартов в этой области в попытке объединить работу ученых и разработчиков бизнес структур[[204]](#footnote-204). Безусловно в первую очередь было использовано бизнес-моделирование, так как математические модели в области экономики давно уже являются популярными инструментами[[205]](#footnote-205). Непосредственно для интеллектуального анализа данных данная консолидация породила несколько подходов, которые следуя англоязычной традиции часто именуют методологиями. SEMMA(Sample, Explore, Modify, Model, and Assess.) — это аббревиатура, означающая «Выборка», «Исследование», «Изменение», «Моделирование» и «Оценка». Это список последовательных шагов, разработанный Институтом SAS.

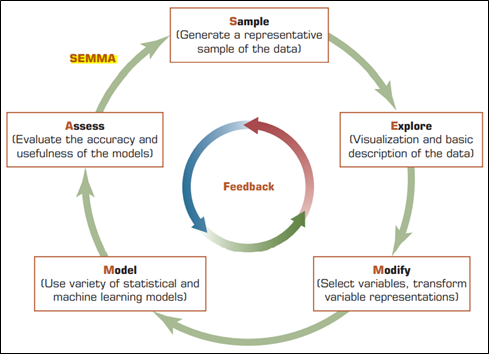
Выборка (Sample). Шаг включает в себя извлечение из большого набора данных надежной и статистически репрезентативной выборкой экземпляров вместо анализа всего объема данных (генеральной совокупности или возможного получения подобной выборки из просто большого набора данных, который не позволяет исследовать его целиком из-за технических или алгоритмических ограничений) с тем чтобы оптимизировать затраты на исследование и его производительность.

Исследование (Explore). Данные исследуются путем поиска ошибок, пропущенных значений, закономерностей (в том числе и не очевидных), и необычных паттернов, выбросов т т.д. Это может улучшить понимание содержания и структуры данных. Для лучшего понимания используется визуализация и могут использоваться исследования с помощью статистических методов, кластерного, факторного анализа и т. д.

Изменение (Modify). Данные изменяются путем извлечения признаков, преобразования и создания новых признаков (часто в виде неких интегральных коэффициентов) или их значений. Например, для масштабирования данных, их стандартизации и нормализации требуется их изменить, признак при этом останется тем же самым, но значения его будут уже изменены.

Моделирование (Model). Моделирование, как это следует из названия посвящается созданию одной, а чаще нескольких моделей в ходе поиска оптимальной. Модель призвана выявить и использовать закономерности, имеющиеся в данных автоматизируя процесс их использования. Модели решают задачи регрессии, классификации, кластеризации и т.д.

Оценка (Assess). Необходимо определить, являются ли выводы из данных достаточно ценными (полезными) и надежными. На этом этапе можно также оценить данные о том, насколько хорошо он работает. Если модель данных действительна, она должна нормально работать как с зарезервированной выборкой, так и с построенной выборкой.



Sharda R., Delen D., Turban E. Business intelligence, analytics, and data science: a managerial perspective. – Pearson Education. New Jersey. 2018., ISBN: 9780134633282

Данный метод применяется в научных исследованиях, так например можно увидеть использование SEMMA в диссертационном исследовании посвященного построению прогностических моделей в профессиональном хоккее (NHL)[[206]](#footnote-206) или спортивную деятельность в футболе[[207]](#footnote-207).

CRISP-DM [[208]](#footnote-208)

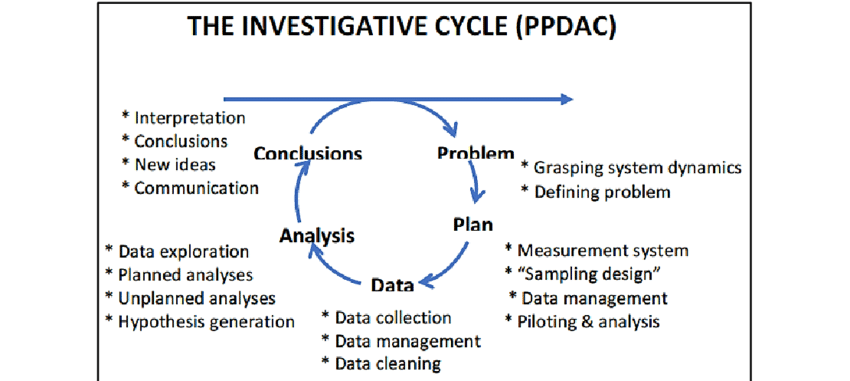
Увеличение эффективности исследовательской деятельности с помощью применения языков программирования для первичной обработки данных и последующего статистического анализа является широко распространённой идеей и можно наблюдать как издаются учебники посвящённые непосредственно этой теме[[209]](#footnote-209). Однако исследователями отмечается, что качество статистического анализа в спортивных исследованиях до сих пор достаточно низкое и требуется даже привлечение профильных специалистов даже для проведения диссертационных исследований и публикации их результатов в изданиях с высоким импакт-фактором (исследование охватывало десятилетие 2003-2013 гг.). При этом утверждается, что сложность применяемых методов начинает возрастать со временем и в большинстве работ использовались по нескольку методов одновременно, хотя наиболее популярными по-прежнему остаются дисперсионный анализ, критерий Стьюдента и регрессионные модели[[210]](#footnote-210).

Понимание того, что подразумевается под статистикой и статистическим анализом, за последние несколько десятилетий претерпело значительные изменилось. Например, определение, которое дал Морис Кендалл «Статистика — это вид научного метода, который имеет дело с данными, полученными путем подсчета или измерения свойств групп природных объектов одного вида (в оригинале: «the properties of populations of natural phenomena»)[[211]](#footnote-211). Более современное определение дал Дэвид Хэнд: «Статистика – это: удовольствие от поиска закономерностей в данных; удовольствие от открытия; постановка важных философских вопросов; способность объяснить значимые решения и способность управлять этими решениями... в бизнесе, науке, правительство, медицине, промышленности...»[[212]](#footnote-212).

Таким образом с течением времени статистический анализ проник во все сферы жизнедеятельности и начал играть роль обоснования для принятия решений. Особенностью этих изменений прослужило большое количество информации, доступной всем отраслям экономики и каждому члену общества. Процесс понимания получаемой информации (превращение её в знания) и принятие обоснованных решений на основе такого понимания является основной функцией современных статистических методов[[213]](#footnote-213). Если просмотреть функциональные возможности специалистов, размещённых на своём сайте Американской статистической ассоциацией (American Statistical Association, ASA)[[214]](#footnote-214), то можно увидеть там дизайн экспериментов, анализ временных рядов, многомерный анализ, алгоритмы машинного обучения, глубокое обучение (нейронные сети), моделирование и многое другое.

Естественно такое расширение статистического анализа нельзя считать оптимальным и уже построение регрессионной модели мы в нашей работе будем относить к задаче машинного обучения, а не статистического анализа.

Наиболее общий алгоритм для проведения современного статистического анализа и связи его с научным методом исследования (PPDAC -The Data Problem Solving Cycle) был представлен R. J. MacKay, R. W. Oldford в 2000 году (см таблицу1)[[215]](#footnote-215)



Безусловно первым шагом для понимания собранного набора данных является соотношение выборки, которая составляет исследуемый нами набор данных и генеральной совокупности, которую нам сложно изучить во всей её полноте. В этом случае сам выбор данных для исследования может составлять проблему.

Поскольку статистика включает в себя анализ данных, а процесс получения данных включает в себя какой-то процесс измерения, важно хорошее понимание измерений. В следующих подразделах мы обсуждаем вопрос об измерениях и шкалах измерений, а также о том, как измеренные данные можно сгруппировать в простые классы для получения распределений данных. Наконец, мы представляем две проблемы, которые служат для маскировки или изменения результатов измерений несколько неожиданным образом. Первым из них является так называемый эффект статистической группировки, при котором сгруппированные данные дают результаты, которые неочевидным образом отличаются от несгруппированных данных. Вторым из них является пространственный эффект, при котором выбор конкретного расположения пространственных группировок (например, переписных округов) может радикально изменить получаемые результаты.

Возможно, одной из наиболее горячо обсуждаемых тем в последние годы стал рост так называемых «больших данных». В статье «Большие данные: совершаем ли мы большую ошибку?» в Financial Times, март 2014 г., Тим Харфорд рассматривает эти и многие другие проблемы, подчеркивая некоторые из менее очевидных проблем, связанных с большими данными. Возможно, главной среди них является предвзятость, которая обнаруживается во многих таких наборах данных. Такие предубеждения могут быть незаметными, их трудно выявить и с ними невозможно справиться. Например, почти все большие данные, связанные с Интернетом, по своей сути предвзяты в пользу тех, кто имеет доступ к Интернету и использует его больше всего, с присущей им демографической и географической предвзятостью. То же самое относится и к конкретным сервисам, таким как Google, Twitter, Facebook, сети мобильной связи, добровольные онлайн-опросы, электронные письма с добровольной подпиской — примеров много, и они разнообразны, но проблемы во многом такие же, как те, которые знакомы статистикам для более века. Большие данные не подразумевают качественных или объективных данных, а большие данные создают и другие проблемы — слишком легко сосредоточиться на исследовании данных и обнаружении закономерностей, выявлении корреляций, которые вполне могут быть ложными — в результате огромного объема данных и количество измеряемых событий и переменных. При наличии достаточного количества данных и сравнений статистически значимые результаты неизбежны, но они не обязательно обеспечивают реальное понимание, понимание или выявление причинно-следственных связей. Конечно, существует множество важных и интересных наборов данных, где сбор и хранение являются гораздо более систематическими, менее подверженными предвзятости, с прямой записью переменных, с «полными» и «чистыми» записями. Такие данные хорошо хранятся и управляются, и, как правило, их собирают агентства, которые дополняют данные метаданными (данными о данных) и информацией об обеспечении качества.

# 3.3. Описательная статистика и особенности распределений наблюдений в исследованиях спортивной деятельности

Приступая непосредственно к работе, мы проводим статистический анализ данных для более глубокого ознакомления с нашими данными и определение возможностей для дальнейшего построения математических и компьютерных (в том числе и цифровых двойников) моделей на основе имеющегося набора данных. В ходе нашей деятельности мы обращались как к наборам данных собранными самостоятельно, так и с собранными другими исследователями. Если наши данные как правило собирались в результате непосредственного анализа спортивной и тренировочной деятельности студентами, магистрантами и аспирантами, то заимствованные данные были собраны в результате парсинга специализированных интернет ресурсов агрегирующих спортивную статистику. Приведём несколько примеров исследования спортивной деятельности с определением распределений случайной величины (наблюдений) и описательных статистик. Возьмём набор данных собранных методом парсинга ресурсов Olympedia[[216]](#footnote-216), Olympics[[217]](#footnote-217) и Sports Reference[[218]](#footnote-218). Основа выбранного набора данных была собрана под названием «120 years of Olympic history: athletes and results» и распространяется по свободной лицензии CC0: Public Domain[[219]](#footnote-219). В результате мы получаем для анализа таблицу со структурой как на рис.

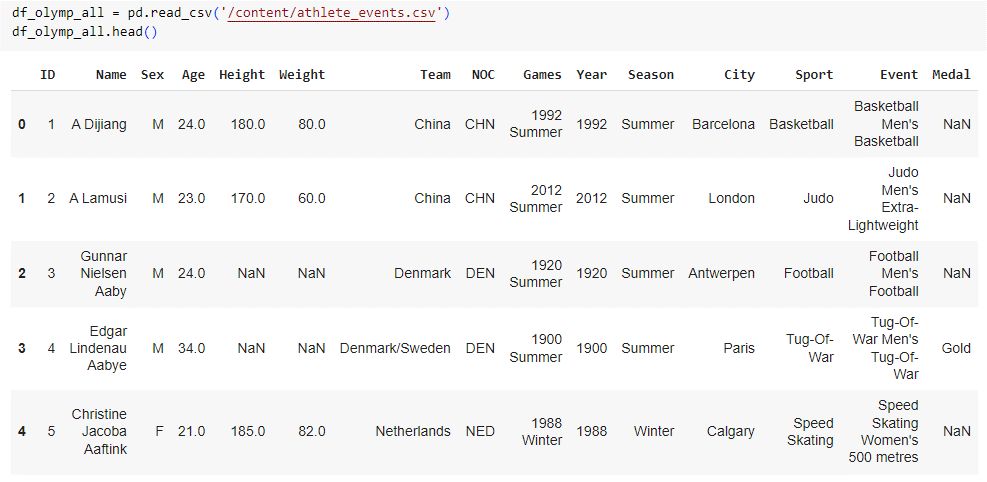


Рис. – Структура набора данных «120 years of Olympic history: athletes and results»

Таблица содержит 271116 (двести семьдесят одну тысячу сто шестнадцать) записей (строк) и 15 параметров (столбцов). Среди параметров:

ID - уникальный номер для каждого спортсмена

Name - имя спортсмена

Sex - пол

Age - возраст

Height – рост в сантиметрах

Weight – вес в килограммах

Team – страна/страны за чью команду выступал спортсмен

NOC - трехбуквенный код национального олимпийского комитета

Games – год и сезон проведения (лето, зима)

Year – год проведения игр

Season - сезон проведения (лето, зима)

City – город в котором проводились игры

Sport – вид спорта

Event – спортивная дисциплина

Medal - золотая, серебряная, бронзовая или NA

Продолжая процедуру ознакомления с данными, мы проводим исследовательский анализ данных и в данном случае проверим полноту и типы данных в нашем наборе. Результаты исследования приведены на рис.

Можно заметить, что данные разделены у нас на две традиционные группы. Присутствуют как численные данные представленные натуральными числами – это идентификационный номер спортсмена и год проведения события (ID, Year), десятичными дробями или числами с плавающей точкой возраст, рост в сантиметрах, вес в килограммах (Age, Height, Weight), так и категориальные данные к которым относятся - имя спортсмена, пол, страна/страны за чью команду выступал спортсмен, трехбуквенный код национального олимпийского комитета, год и сезон проведения, город в котором проводились игры (лето, зима) вид спорта, спортивная дисциплина и завоёванная награда (Name, Sex, Team, NOC, Games, Year, Season, City, Sport, Event, Medal).

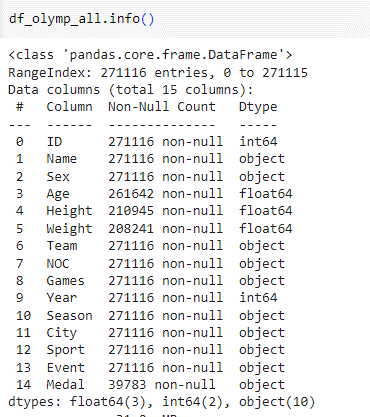


Рис. – полнота и тип данных согласно имеющимся 15 параметрам.

Рассмотрев полноту мы видим, что в основном данные полные. Исключение составляют завоёванные награды (Medal), что вполне объяснимо, так как в наборе присутствуют все участники олимпийских игр, а не только призёры. Другим важным недостатком являются неполные данные по возрасту, росту и весу спортсменов (Age, Height, Weight). В данном исследовании мы сосредоточимся на закономерностях распределения величин рост и веса спортсменов, поэтому возраст не будет для нас значимым. Традиционно считается, что такие характеристики человека как рост (в первую очередь) и вес при измерении их на большом контингенте испытуемых подчиняются закономерностям нормального распределения. Проверим полноту данных по интересующим нас параметрам. По росту полнота составляет 77,8%, а по весу 76,8%, что безусловно значительный недостаток, но общее количество записей составляет по росту 210945 (двести десять тысяч девятьсот сорок пять) записей и по весу 208241 (двести восемь тысяч двести сорок одна) записей и это более чем достаточно для проведения качественного анализа. Для принятия окончательного решения обратимся к визуальному анализу данных. На рис. Отображена полнота записей по росту и весу в процентах.

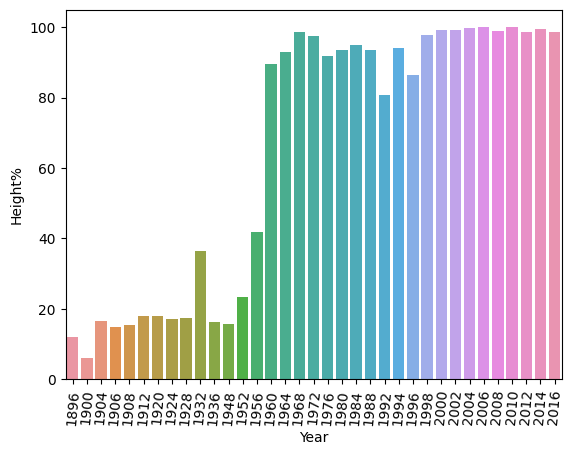




Рис.1 –Полнота набора данных по весу и росту спортсменов, где Weight% - полнота заполнения данных по росту спортсменов в процентах и Height% - полнота заполнения по росту

Как можно заметить заполняемость по обоим параметрам изменялась похожим образом и непрерывно росла вплоть до 1968 года, потом произошла стабилизация. Более подробно с результатами можно ознакомиться в таблице 1, для периода с 1896 по 1964 гг. и в таблице 1 для периода с 1968 по 2016 гг.

Таблица 1. Полнота данных периода с 1896 по 1964 гг.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **ID** | **Age** | **Height** | **Weight** | **Medal** | **Height%** | **Weight%** | **Medal%** |
| **1896** | 380 | 217 | 46 | 49 | 143 | **12.1** | **12.9** | 37.6 |
| **1900** | 1936 | 1146 | 116 | 79 | 604 | **6.0** | **4.1** | 31.2 |
| **1904** | 1301 | 1027 | 213 | 147 | 486 | **16.4** | **11.3** | 37.4 |
| **1906** | 1733 | 990 | 257 | 205 | 458 | **14.8** | **11.8** | 26.4 |
| **1908** | 3101 | 2452 | 475 | 483 | 831 | **15.3** | **15.6** | 26.8 |
| **1912** | 4040 | 3884 | 721 | 596 | 941 | **17.8** | **14.8** | 23.3 |
| **1920** | 4292 | 3447 | 767 | 471 | 1308 | **17.9** | **11.0** | 30.5 |
| **1924** | 5693 | 4551 | 974 | 690 | 962 | **17.1** | **12.1** | 16.9 |
| **1928** | 5574 | 4611 | 975 | 718 | 823 | **17.5** | **12.9** | 14.8 |
| **1932** | 3321 | 2991 | 1213 | 550 | 739 | **36.5** | **16.6** | 22.3 |
| **1936** | 7401 | 7188 | 1192 | 987 | 1025 | **16.1** | **13.3** | 13.8 |
| **1948** | 7480 | 6304 | 1169 | 1151 | 987 | **15.6** | **15.4** | 13.2 |
| **1952** | 9358 | 9081 | 2188 | 2187 | 1033 | **23.4** | **23.4** | 11.0 |
| **1956** | 6434 | 5796 | 2686 | 2680 | 1043 | **41.7** | **41.7** | 16.2 |
| **1960** | 9235 | 9014 | 8274 | 8187 | 1058 | **89.6** | **88.7** | 11.5 |
| **1964** | 9480 | 9424 | 8799 | 8772 | 1215 | **92.8** | **92.5** | 12.8 |

Таблица 1. Полнота данных периода с 1968 по 2016 гг.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **ID** | **Age** | **Height** | **Weight** | **Medal** | **Height%** | **Weight%** | **Medal%** |
| **1968** | 10479 | 10361 | 10326 | 10310 | 1256 | **98.5** | **98.4** | 12.0 |
| **1972** | 11959 | 11863 | 11658 | 11570 | 1414 | **97.5** | **96.7** | 11.8 |
| **1976** | 10502 | 10450 | 9626 | 9582 | 1531 | **91.7** | **91.2** | 14.6 |
| **1980** | 8937 | 8750 | 8349 | 8341 | 1602 | **93.4** | **93.3** | 17.9 |
| **1984** | 11588 | 11372 | 10990 | 10985 | 1698 | **94.8** | **94.8** | 14.7 |
| **1988** | 14676 | 14566 | 13743 | 13748 | 1845 | **93.6** | **93.7** | 12.6 |
| **1992** | 16413 | 16369 | 13238 | 13256 | 2030 | **80.7** | **80.8** | 12.4 |
| **1994** | 3160 | 3158 | 2973 | 2971 | 331 | **94.1** | **94.0** | 10.5 |
| **1996** | 13780 | 13772 | 11909 | 11959 | 1842 | **86.4** | **86.8** | 13.4 |
| **1998** | 3605 | 3603 | 3521 | 3519 | 440 | **97.7** | **97.6** | 12.2 |
| **2000** | 13821 | 13820 | 13698 | 13695 | 2004 | **99.1** | **99.1** | 14.5 |
| **2002** | 4109 | 4109 | 4080 | 4062 | 478 | **99.3** | **98.9** | 11.6 |
| **2004** | 13443 | 13443 | 13407 | 13406 | 2001 | **99.7** | **99.7** | 14.9 |
| **2006** | 4382 | 4382 | 4376 | 4366 | 526 | **99.9** | **99.6** | 12.0 |
| **2008** | 13602 | 13600 | 13451 | 13443 | 2048 | **98.9** | **98.8** | 15.1 |
| **2010** | 4402 | 4402 | 4400 | 4378 | 520 | **100.0** | **99.5** | 11.8 |
| **2012** | 12920 | 12920 | 12752 | 12560 | 1941 | **98.7** | **97.2** | 15.0 |
| **2014** | 4891 | 4891 | 4871 | 4673 | 597 | **99.6** | **95.5** | 12.2 |
| **2016** | 13688 | 13688 | 13512 | 13465 | 2023 | **98.7** | **98.4** | 14.8 |

Рассматривая сгруппированные подобным образом данные можно заметить, что, действительно начиная с 1968 года полнота данных стабильно превышает 90% и при анализе мы можем по этому набору данных правомочно делать выводы, касающиеся всей генеральной совокупности спортсменов, принимавших участие в олимпийских играх. Также в качестве основной гипотезы можно предположить, что распределение величин веса и роста спортсменов совпадают с нормальным распределением.

Дополнительно рассмотрим описательные статистики для категориальных данных (табл. 1).

Таблица 1. Описательные статистики категориальных данных.

| Для п=271116 | **Name** | **Sex** | **Team** | **NOC** | **Games** | **Season** | **Sport** | **Event** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Уникальных значений** | 134732 | 2 | 1184 | 230 | 51 | 2 | 66 | 765 |
| **Мода** | Robert Tait McKenzie | M | United States | USA | 2000 Summer | Summer | Athletics | Football Men's Football |
| **Частота моды** | 58 | 196594 | 17847 | 18853 | 13821 | 222552 | 38624 | 5733 |
| **%** | 0.02 | 72.51 | 6.58 | 6.95 | 5.10 | 82.09 | 14.25 | 2.11 |

При более подробном изучении можно заметить, что уникальных значений имён меньше чем количество записей и можно сразу заметить этому объяснение. Наиболее часто, 58 раз в данном наборе данных упомянут Robert Tait McKenzie. При осуществлении поиска можно получить следующую выборку (пример на рис.1). Сначала необходимо определить уникальный идентификационный номер для этого спортсмена и далее произвести фильтрацию по данному номеру. В результате можно убедиться, что действительно соревнования по различным видам искусства проводились на Олимпийских играх 1912, 1920, 1924, 1928, 1932, 1936, 1948 и 1952 годов. Победители соревнований также, как и спортсмены других видов спорта награждались золотыми, серебряными и бронзовыми медалями. Мероприятия были вдохновлены Пьером де Кубертеном, который пожелал совместить соревнования с двигательной активностью с соревнованиями по искусству. Среди участников и медалистов был и Robert Tait McKenzie[[220]](#footnote-220)[[221]](#footnote-221).

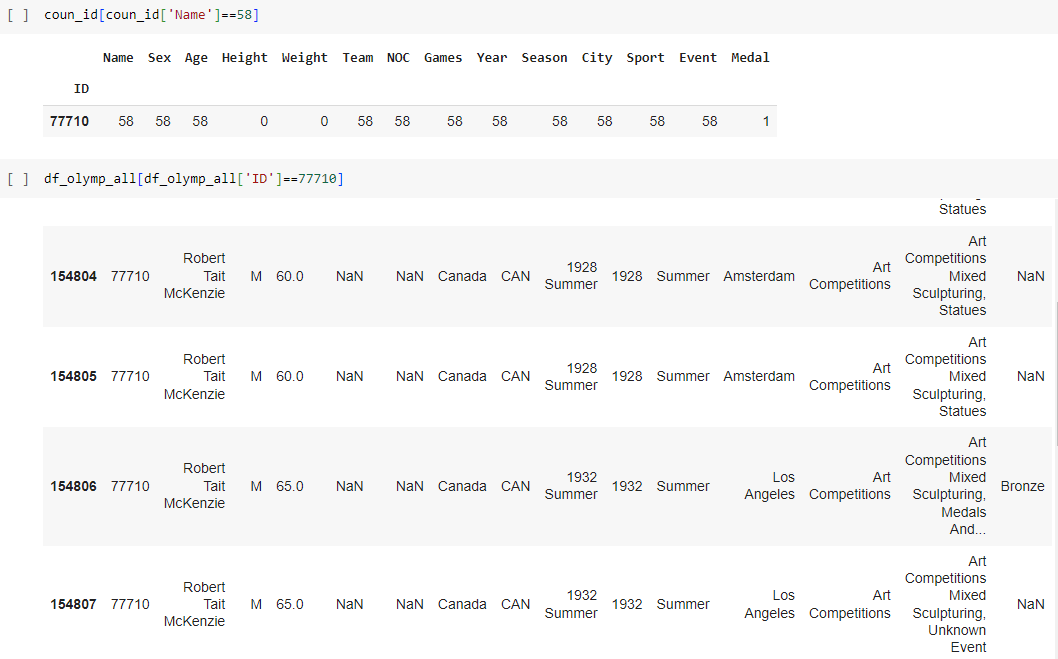


Рис.1 - Результаты поиска спортсмена, упоминаемого 58 раз в наборе данных

Соревнования по различным видам искусства были исключены из олимпийской программы, поскольку большинство художников были профессионалами, против чего выступил международный олимпийский комитет (МОК). Конкурсы проводились по архитектуре, литературе, музыке, живописи и скульптуре.

Два золотых медалиста в области искусства также завоевывали медали в других спортивных дисциплинах кроме соревнований по искусству. Alfred Hajós из Венгрии выиграл две золотые медали по плаванию в 1896 году и Walter Winans выиграл медали по стрельбе по бегущему оленю 100 метров в 1908 и 1912 годах[[222]](#footnote-222).

В результате можно заметить, что данные 58 записей составляют только 0,02% от общего количества записей и в ней отсутствуют интересующие нас величины роста и веса, что характерно для данного временного периода, остальные же подобные множественные записи меньше и они также не приносят значительных проблем для анализа. Например, близкий по количеству записей (30) имеет известный американский пловец Michael Fred Phelps, результаты поиска его записей на рис.1

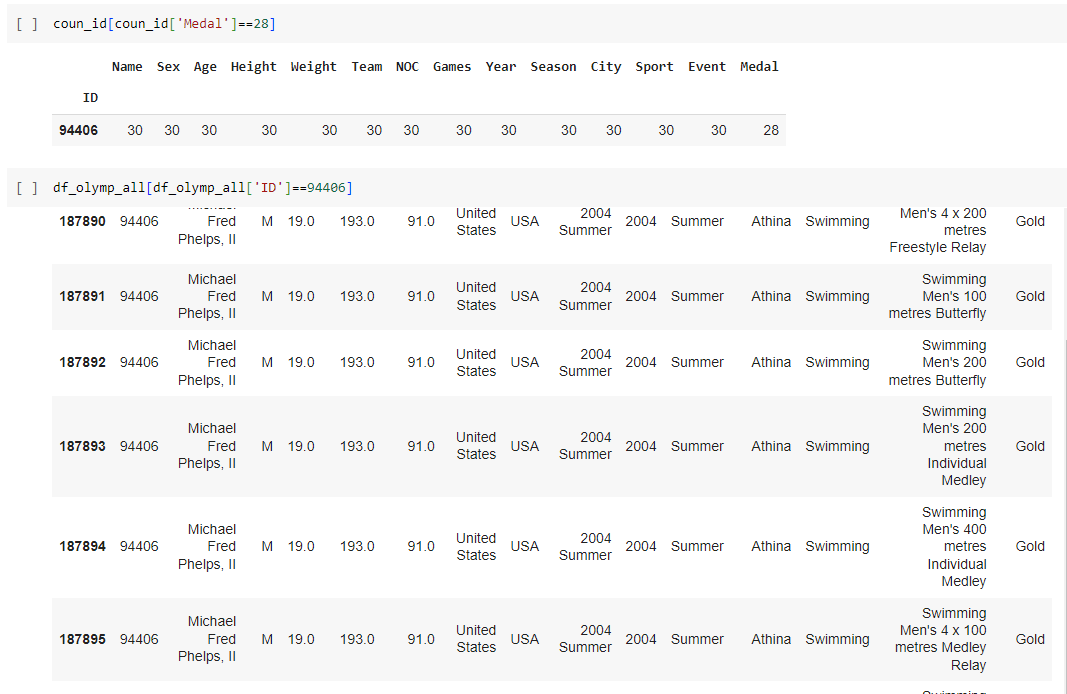


Рис.1 - Результаты поиска спортсмена Michael Fred Phelps, упоминаемого 30 раз в наборе данных

Кроме выявленной особенности имеется значительная неравномерность в других параметрах. Так в гендерном распределении и большая часть 72,51% записей об участниках мужского пола, 82,09% записей относится к летним олимпийским играм, 14,25% записей относится к лёгкой атлетике и почти 7% к спортсменам, выступающим за США. Эти обстоятельства вносят своё влияние на имеющийся набор данных, но оснований для введений поправочных коэффициентов не представляется необходимым, так как эти неравномерности отражают реальное развитие олимпийских игр, то есть именно генеральную совокупность.

Следующим шагом мы перейдём к анализу количественных данных, которые представлены в таблице 1.

Таблица 1. Описательная статистика для исследуемых количественных данных.

|  | **Возраст (Age)** | **Рост (Height)** | **Вес (Weight)** | **Год (Year)** |
| --- | --- | --- | --- | --- |
| **Количество наблюдений (count)** | 261642 | 210945 | 208241 | 271116 |
| **Средн.арифметическое (mean)** | 25.56 | 175.34 | 70.70 | 1978.39 |
| **Средн.квадратичное откл (std)** | 6.39 | 10.52 | 14.35 | 29.88 |
| **Минимальное значение**  **(min)** | 10 | 127.00 | 25.00 | 1896 |
| **25% процентиль** | 21 | 168.00 | 60.00 | 1960 |
| **50% процентиль**  **(медиана)** | 24 | 175.00 | 70.00 | 1988 |
| **75% процентиль** | 28 | 183.00 | 79.00 | 2002 |
| **Максимальное значение (max)** | 97 | 226.00 | 214.00 | 2016 |

Обращает внимание, то, что по росту и весу выборки среднее арифметическое и медиана отличаются на десятые и это прямо показывает, что данные симметричны и равномерно распределены вокруг среднего значения. Это свидетельствует в пользу наличия нормального распределения, но не позволяет пока делать однозначные выводы.

# 3.4. Распределение веса спортсменов

Далее отдельно рассмотрены распределения по весу и росту. Вес в процессе спортивной тренировки может меняться более быстро и разнонаправленно, в отличии от роста. Поэтому первым было рассмотрено распределение веса спортсменов и результаты подверглись сравнению с распределением роста. Гистограмму распределения веса можно посмотреть на рис 1. Сразу можно обратить внимание на то, что если принять значение математического ожидания за центр, то в этом случае имеется очень длинный ряд небольших по частоте, но принимающих значение на достаточно большом интервале от 79 до 214 килограмм, как это и видно в таблице, что создаёт так называемые «выбросы», которые могут повлиять на распределение величины веса спортсменов.

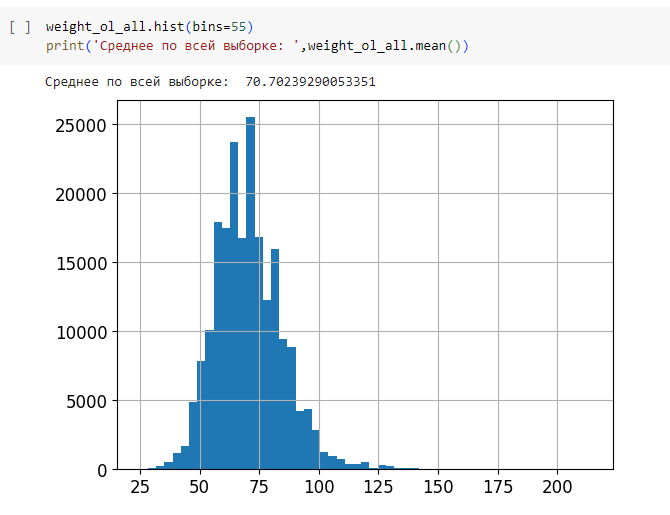


Рис. 1 – Гистограмма распределения значений веса спортсменов

Для большей наглядности можно рассмотреть другой тип графика – Boxplot (называют ещё ящиком с усами), который как раз является графической демонстрацией разброса и асимметрии числовых данных (как правило относительно значения медианы) через их квартили и соответственно межквартильный размах (вычисляется по формуле IQR = Q3 -Q1), а также наличие и характер выбросов, которые значительно отличаются от остального набора данных, могут быть отображены в виде отдельных точек. Диаграммы Boxplot непараметричны, они отображают дисперсии в выборках статистической совокупности без каких-либо предположений об основном статистическом распределении. Выбор границ «усов» основан на значении 1,5\*IQR. Сверху верхнего квартиля (Q3) измеряется расстояние, в 1,5 раза превышающее IQR, и проводится линия до наибольшей наблюдаемой точки данных из набора данных, попадающей в пределы этого расстояния. Аналогичным образом, расстояние, в 1,5 раза превышающее IQR, измеряется ниже нижнего квартиля (Q1), и к самой нижней наблюдаемой точке данных из набора данных, попадающей в пределах этого расстояния, рисуется «ус».

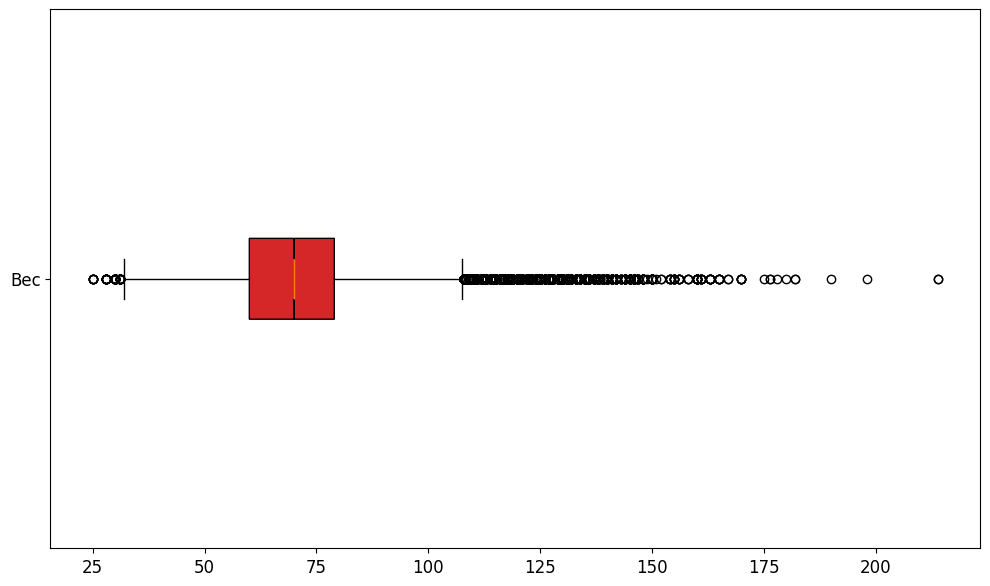


Рис.1 – Демонстрация веса распределения веса спортсменов, асимметрии, выбросов и межквартильного размаха с помощью диаграммы Boxplot

Поскольку усы должны заканчиваться в наблюдаемой точке данных, длина усов может выглядеть неодинаковой, хотя 1,5 IQR одинаково для обеих сторон. Все остальные наблюдаемые точки данных за пределами границ усов нанесены на график как выбросы. Как можно заметить, количество выбросов действительно может повлиять на определение распределения веса спортсменов. Для более наглядного представления и сравнения распределений вероятностей используется Violinplot. Он добавляет на Boxplot оценку плотности ядра или иногда пишут ядерную плотность (KDE, kernel density estimation). Это непараметрический метод оценки функции плотности вероятности случайной величины, используется для сглаживания данных, когда выводы о совокупности делаются на основе конечной выборки данных. В некоторых областях, таких как обработка сигналов и эконометрика, его также называют методом окна Парцена-Розенблатта (Parzen–Rosenblatt window) в честь Эмануэля Парцена (Emanuel Parzen) и Мюррея Розенблатта (Murray Rosenblatt), которым обычно приписывают независимое создание его в его современной форме. Распределение плотности венроятности для веса спортсменов можно видеть на рис.1

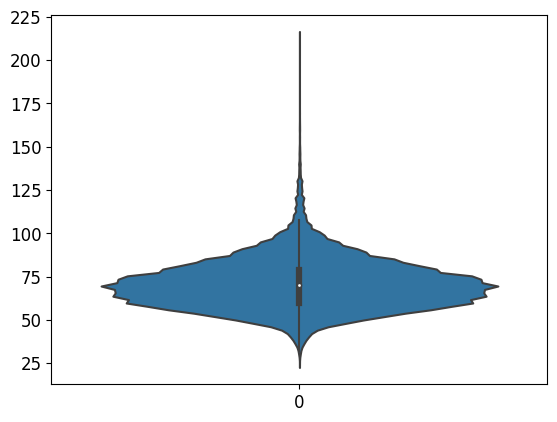
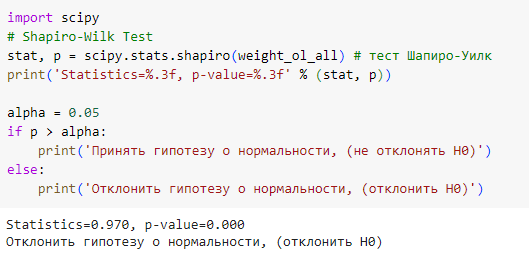


Рис.1 – Демонстрация распределения значений веса спортсменов совместно с плотностью распределения вероятности с помощью диаграммы Violinplot

Violinplot показывает полное распределение данных и эту даграмму можно использовать в мультимодальных данных (более одного пика). В этом случае скрипичный график показывает наличие различных пиков, их положение и относительную амплитуду. На приведённой диаграмме отчётливо видно, что действительно присутствует смещение в распределении веса спортсмена и вероятно, что данное распределение не соответствует нормальному распределению несмотря на наличие признаков, позволяющих предположить такую похожесть. Для более объективной оценки необходимо произвести оценку с использованием статистических критериев.

Для более точной оценки использовалось несколько популярных критериев используемых для нормальности распределения из различных библиотек научных функций языка Python. Все три выбранных метода – критерий Шапиро-Уилка, критерий согласия Пирсона, критерий согласия Колмогорова-Смирнова при применении показали, что распределение веса спортсмена не является нормальным с высокой степенью достоверности p-value=0.000 (рис.1).



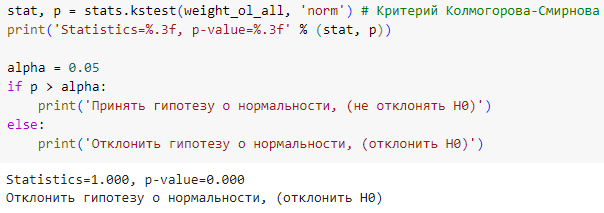
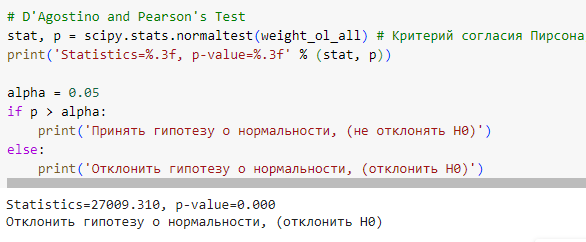


Рис.1 – Исследование нормальности распределения веса спортсменов в полном наборе данных

Если критерии не позволяют назвать распределения веса спортсмена соответствующее нормальному, то возникает закономерный вопрос какому из распределений оно максимально близко. Произведя проверку на соответствие были получены следующие пять наиболее близких распределения (рис.1)

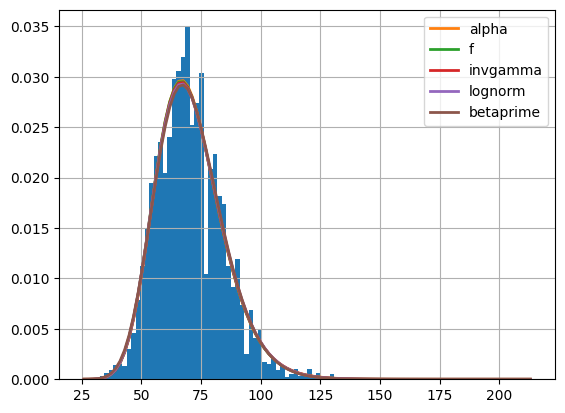


Рис.1 – Соответствие распределения веса спортсменов полного набора данных пяти распределениям

Более подробно с результатами сравнения можно ознакомиться в таблице 1.

Таблица 1. Результаты сравнения распределения веса спортсменов полного набора данных пяти распределениям

| **Вид** | **Средн. квадрат. ошибка (Sumsquare error)** | **Информационные критерии** | | **Критерий согласия Колмогорова** | |
| --- | --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **ks\_statistic** | **ks\_pvalue** |
| **alpha** | 0.000352 | 1912.166018 | 1942.905372 | 0.027012 | 1.959805e-132 |
| **f** | 0.000353 | 1997.038039 | 2038.023844 | 0.028140 | 1.091873e-143 |
| **invgamma** | 0.000353 | 1979.011455 | 2009.750809 | 0.026090 | 1.419494e-123 |
| **lognorm** | 0.000355 | 2029.380406 | 2060.119760 | 0.025149 | 7.523571e-115 |
| **betaprime** | 0.000356 | 2055.056953 | 2096.042758 | 0.024988 | 2.206514e-113 |

Рассмотрев ближе те распределения, которые максимально близки исследуемому можно увидеть, что самые близкие имеют некоторую связь с нормальным распределением.

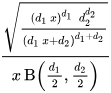
Alpha - распределение — это непрерывное распределение вероятностей, которое часто используется для моделирования неотрицательных данных, таких как срок службы, интенсивность отказов и других подобных наборов данных. Неотрицательные данные относятся к значениям данных, которые являются либо положительными (больше нуля), либо равными нулю, то есть они не содержат отрицательных значений. Такой тип данных обычно встречается в ситуациях, когда отрицательные значения невозможны или не имеют смысла, например, при измерении времени, подсчете объектов или количественном определении таких величин, как расстояние, вес или размер. Это распределение определяется двумя параметрами: параметром формы альфа и параметром масштаба бета. Параметр альфа определяет форму распределения вероятностей, а параметр бета определяет его разброс или размер распределения по горизонтальной оси.

(Плотность вероятности - )

Где Φ — стандартное нормальное распределение. Существует сходство между этой плотностью вероятности и функцией обратной нормальной плотности. Это связано с тем, что плотность вероятности — это функция плотности X = 1/Y, когда Y имеет нормально распределенную случайную величину, усеченную слева от нуля с α = ξ, σ и β = 1/ σ.

F-распределение Фишера-Снедекора, представляет собой непрерывное распределение вероятностей где независимые случайные величины имеют распределение хи-квадрат, которое часто возникает как нулевое распределение, в дисперсионном анализе (ANOVA)

(Функция распределения - 

плотность вероятности - )

Связано с экспоненциальным и гамма распределениями

Invgamma – обратное гамма-распределение представляет собой двухпараметрическое семейство непрерывных распределений вероятностей является обобщением экспоненциального распределения.

(Функция распределения

Плотность вероятности - )

Применяют обычно для определения времени достижения события. Основываясь на центральной предельной теореме (ЦПТ), при больших значениях параметра k гамма-распределение может приближаться к нормальному распределению.

Lognorm (логнормальное) распределение — это непрерывное распределение вероятностей случайной величины, логарифм которой имеет нормальное распределение. Таким образом, если случайная величина X имеет логнормальное распределение, то Y = ln(X) имеет нормальное распределение.

(Функция распределения - 

Плотность вероятности - )

Аналогично, если Y имеет нормальное распределение, то показательная функция Y, X = exp(Y) , имеет логарифмически нормальное распределение. Случайная величина, имеющая логарифмически нормальное распределение, принимает только положительные действительные значения. Это удобная и полезная модель для измерений в точных и технических науках, а также в медицине, экономике и других областях (например, энергии, концентрации, длины, цены финансовых инструментов и других показателей).

Betaprime (также известное как инвертированное бета-распределение или бета-распределение второго рода) — это абсолютно непрерывное распределение вероятностей, определяется двумя параметрами α и β

(Функция распределения - 

Плотность вероятности - )

В то время как соответствующее бета-распределение представляет собой сопряженное априорное распределение параметра распределения Бернулли, выраженное как вероятность, Betaprime представляет собой сопряженное априорное распределение параметра распределения Бернулли, выраженное в шансах.

Таким образом можно увидеть, что все наиболее подходящие распределения хотя и имеют связь с нормальным, но считать их нормальным будет несколько опрометчиво.

Вслед за этим был проведён эксперимент по случайному извлечению выборки в 30 наблюдений из полного набора данных. Результаты представлены на рис.1.

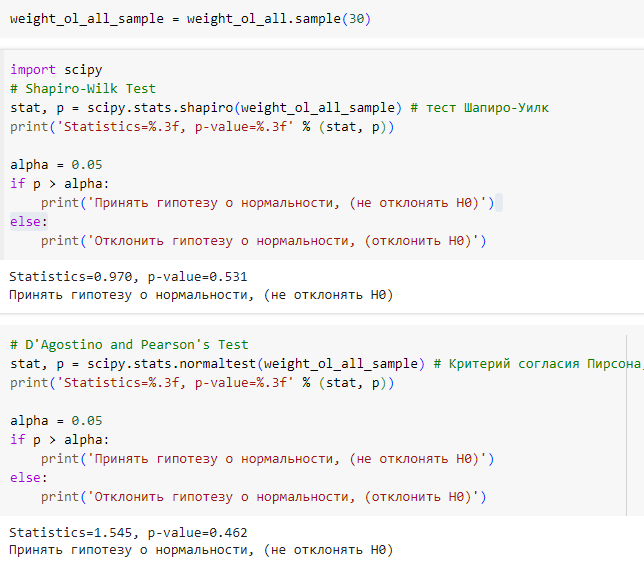


Рис.1 – Проверка нормальности распределения случайной выборки (n=30) из полного набора данных

Обращает своё внимание тот момент, что в случае случайного извлечения выборки тесты определяют распределение уже как нормальное. В результате отбора выбросы имеют гораздо меньшую вероятность попасть выборку вследствие их малочисленности и поэтому критерии определяют выборочное распределение соответствующее нормальному.

Проведя подбор пяти оптимально соответствующих распределений можно убедиться в причинах произошедшего (рис.1)

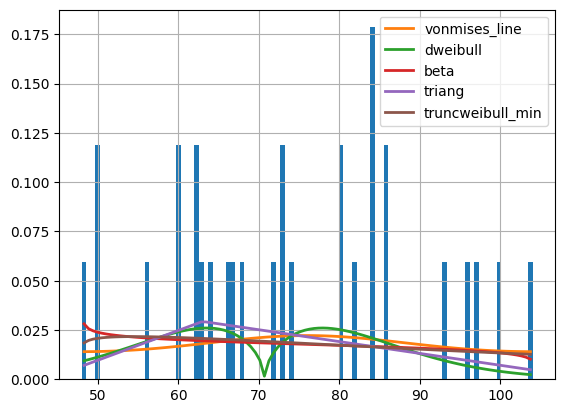


Рис.1 - Соответствие распределения веса спортсменов (n=30) из полного набора данных пяти распределениям

Как можно увидеть несмотря на результаты применения критериев визуально кривая Гаусса не является самой подходящей для описания выборочного набора данных и более того, столь небольшая выборка даёт значительный разброс в использовании известных распределений для её описания. Для получения более точной информации необходимо обратиться к таблице 1, где приведены численные данные сравнения.

Таблица 1. Результаты сравнения распределения веса спортсменов выборочного набора данных пяти распределениям

| **Вид** | **Средн. квадрат. ошибка (Sumsquare error)** | **Информационные критерии** | | **Критерий Согласия Колмогорова - Смирнова** | |
| --- | --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **ks\_statistic** | **ks\_pvalue** |
| **dweibull** | 0.137747 | 862.070135 | 866.273727 | 0.133920 | 0.607640 |
| **beta** | 0.137760 | 816.586055 | 822.190844 | 0.133238 | 0.613950 |
| **triang** | 0.137806 | 837.452088 | 841.655681 | 0.129523 | 0.648433 |
| **truncweibull\_min** | 0.137943 | 818.157755 | 825.163742 | 0.122371 | 0.714627 |

Необходимо более подробно рассмотреть пять наиболее соответствующих распределений:

Dweibull распределение представляет собой непрерывное распределение вероятностей. Он моделирует широкий диапазон случайных величин, в основном таких как время до отказа или время между событиями. Примерами являются максимальное количество осадков за один день и время, которое пользователь проводит на веб-странице. Впервые применено для описания распределения частиц по размерам[[223]](#footnote-223).

(Функция распределения - 

Плотность вероятности - )

Применяется в прогнозировании различных изменений. Связано с гамма и экспоненциальным распределением, в некоторых случаях может представлять из себя прямую, но в исследуемом случае распределение имеет два явных пика, что может отражать влияние гендерных особенностей.

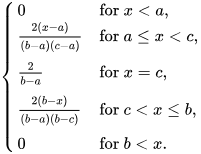
Beta представляет собой семейство непрерывных распределений вероятностей, определенных на интервале [0, 1] или (0, 1) с помощью двух положительных параметров, обозначаемых альфа (α) и бета (β), которые появляются как показатели степени переменной и ее дополняют до 1 соответственно и управляют формой распределения. Бета-распределение используется для моделирования поведения случайных величин, ограниченных интервалами конечной длины. Бета-распределение является подходящей моделью случайного поведения процентов и пропорций.

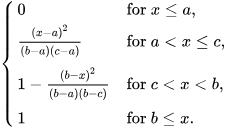
(Функция распределения - 

Плотность вероятности - )

В байесовской статистике бета-распределение представляет собой сопряженное априорное распределение вероятностей для распределения Бернулли, биномиального, отрицательного биномиального и геометрического распределений.

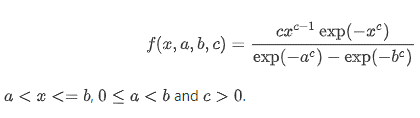
Triang - треугольное распределение представляет собой непрерывное распределение вероятностей с нижним пределом a, верхним пределом b и пиковым значением (которое является модой) c, где a < b и a ≤ c ≤ b.

(Функция распределения - 

Плотность вероятности -  )

Треугольное распределение обычно используется как субъективное описание совокупности, для которой имеются лишь ограниченные выборочные данные, и особенно в случаях, когда взаимосвязь между переменными известна, но данных мало (возможно, из-за невозможности сбора полного набора). Оно основано на знании минимума и максимума и «предположении основанном на вдохновении»[[224]](#footnote-224) относительно модального значения.

Truncweibull min усеченное модифицированное распределение, как и следует из названия относится к семейству распределения Weibull

(Функция распределения - )

Вычисляется методом максимального правдоподобия из двух параметров распределения Weibull с учетом меньшего усечения.

В результате можно увидеть, что с одной стороны критерии демонстрируют наличие нормального распределения, а с другой стороны визуально данное распределение отличается от нормального и наиболее подходящие по критерию Колмогорова-Смирнова и информационным критериям распределения не являются нормальными. Для углубленного понимания было проведено моделирование, в ходе которого 1000 раз была произведена выборка (n=30) из полного набора данных (N=208241) оценённая одновременно двумя критериями Шапиро-Уилка и Пирсона (рис.1).

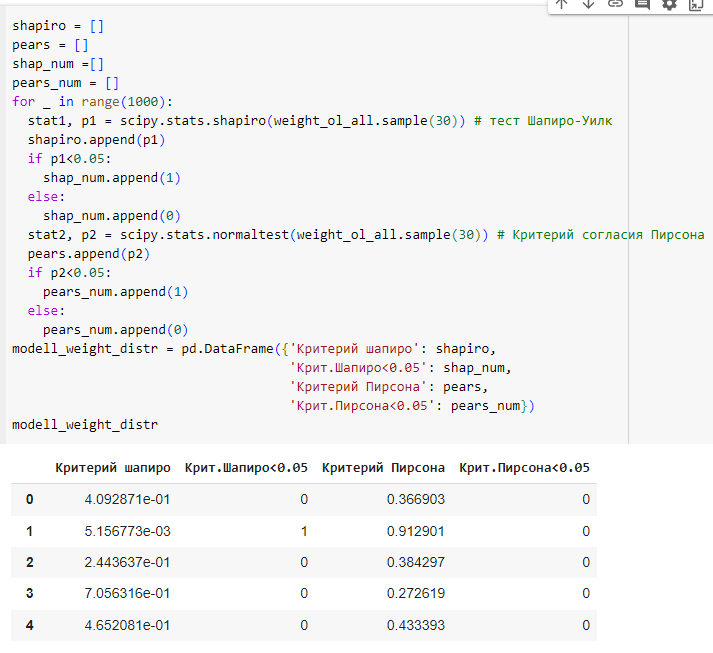


Рис.1 – Моделирование оценки 1000 выборок (n=30) из полного набора данных (N=208241) двумя критериями Шапиро-Уилка и Пирсона

С численными данными можно ознакомиться в таблице 1.

Таблица 1 - Оценка соответствия нормальному распределению моделирования случайных выборок из полного набора данных

| **р<0.05** | **Критерий Шапиро-Уилка** | **Критерий согласия Пирсона** |
| --- | --- | --- |
| **нет** | 764 | 736 |
| **да** | 236 | 264 |

При оценке полученных данных можно увидеть, что действительно в большей части производимых замеров согласно выбранным двум критериям выборочные данные распределяются нормально с вероятностью ≈0,75. Вероятность же того, что в выборку попадут данные из множества «выбросов», составляет только 3237/208241 ≈0.0156 (отношение выбросов к общему количеству значимых значений). Это позволяет использовать параметрические критерии для оценки и использовать нормальное распределение для моделирования процессов связанных с весом спортсменов. Для более точных моделей желательно в больших выборках распределения с длинным правым хвостом.

# 3.5. Распределение значений роста спортсменов

Вслед за весом разумно рассмотреть рост спортсменов, значения которого в большей степени должны соответствовать нормальному распределению[[225]](#footnote-225). При этом исследования не вполне подтверждают подобные утверждения[[226]](#footnote-226). Проведя аналогичные исследованию распределения значений веса был произведён визуальный анализ распределения значений роста спортсменов (рис.1)

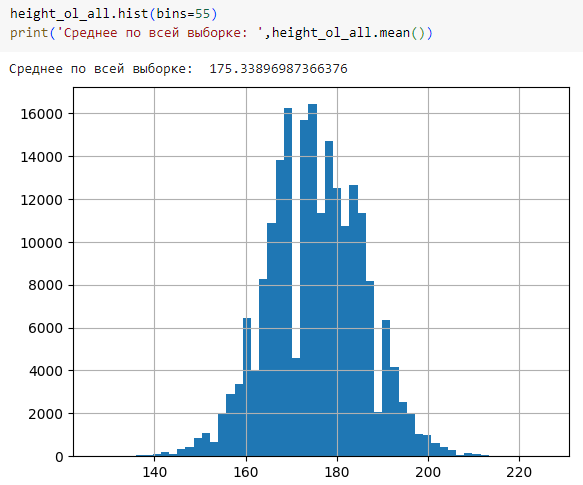


Рис.1 – Гистограмма значения роста спортсменов

Общая форма гистограммы может быть описана кривой Гаусса более точно, чем гистограмма значений веса и можно сформулировать популярную гипотезу, что значения роста спортсменов подчиняются закономерностям нормального распределения. Для более наглядного представления симметричности, выбросов и межквартильного размаха этого набора данных была применена диаграмма Boxplot (рис.1)

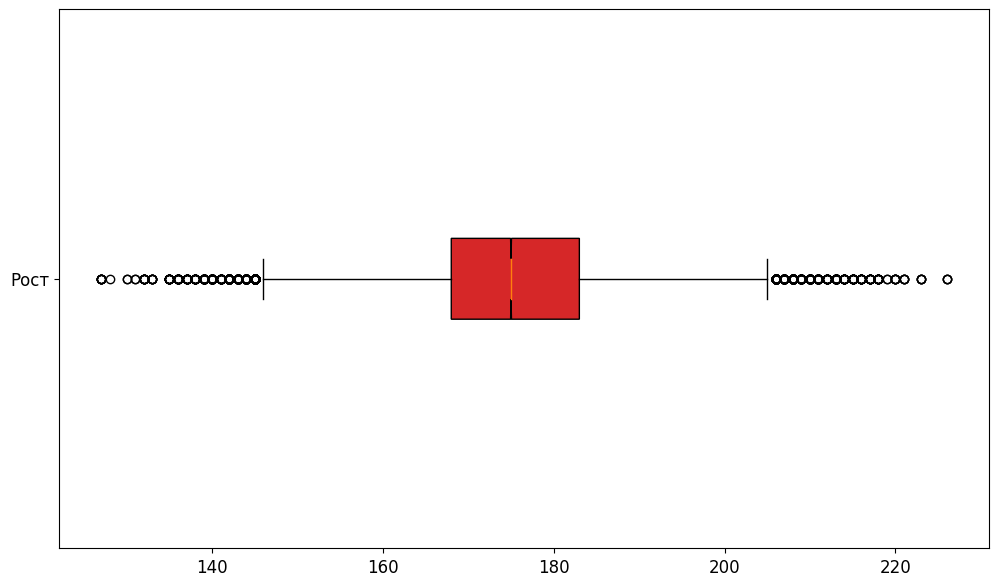


Рис.1 - Демонстрация распределения роста спортсменов, асимметрии, выбросов и межквартильного размаха с помощью диаграммы Boxplot

На данной диаграмме видно, что распределение роста спортсменов более симметрично, чем распределение веса, средний рост спортсмена не значительно отличается от роста среднего человека в большую сторону, что обусловлено меньшей зависимости изменений роста чем веса человека под воздействием спортивной тренировки. Среднее арифметическое 175.3 отличается от медианного значения 175.0 только на три десятых. Межквартильный размах (25-75%) представляет собой диапазон от 168 до 183 сантиметров. Выбросы составляют только 1338 значения из 210945, что составляет только 0,6%. Всё это свидетельствует в пользу первоначально выдвинутой гипотезы о соответствия распределения значения роста спортсменов закономерностям нормального распределения.

Для более наглядного рассмотрения распределения значений роста спортсменов совместно с плотностью распределения вероятности была использована, как и в случае оценки распределения значений веса спортсменов, диаграмма Violinplot (рис.1)

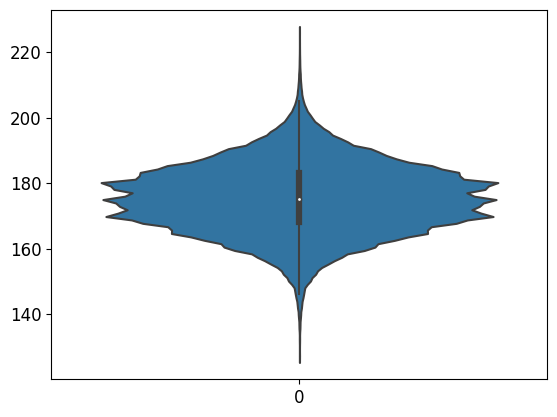


Рис.1 – Демонстрация распределения значений роста спортсменов совместно с плотностью распределения вероятности с помощью диаграммы Violinplot

На Violinplot можно заметить, как подтверждение изначальной гипотезы о нормальном распределении, так и возможность дальнейшего углубленного исследования распределения значений роста спортсменов при помощи кластерного анализа, так как на диаграмме отчётливо выделяются несколько наиболее вероятных значений.

Дальнейший шаг был проверкой на соответствие нормальному распределению уже с помощью статистических критериев Шапиро-Уилка, критерия согласия Пирсона и критерия Колмогорова-Смиронова. Результаты проверки можно увидеть на рисунке 1.

Можно увидеть, что применение всех трёх критериев дало один согласованный ответ, что вопреки первоначально выдвинутой гипотезе распределение значений роста спортсменов не подчиняется закономерностям нормального распределения случайной величины.



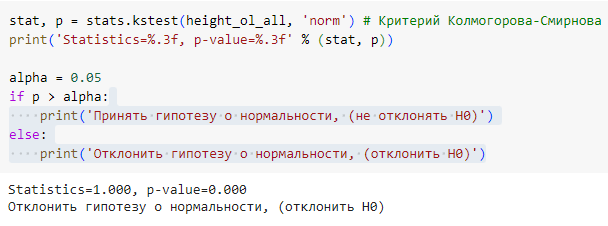


Рис.1 – Исследование нормальности распределения роста спортсменов в полном наборе данных

Согласно уже применённому подходу по изучению распределения значений веса спортсменов было проведено сопоставление полного набора данных с 80 наиболее часто встречающихся распределений случайной величины выбрав пять наиболее соответствующих (рис.1).

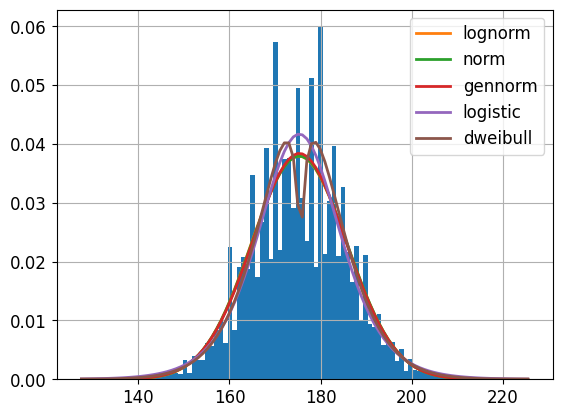


Рис.1 – Соответствие распределения роста спортсменов полного набора данных пяти распределениям

Более подробно с результатами сравнения можно ознакомиться в таблице 1.

Таблица 1. Результаты сравнения распределения роста спортсменов полного набора данных пяти распределениям

| **Вид** | **Средн. квадрат. ошибка (Sumsquare error)** | **Информационные критерии** | | **Критерий согласия Колмогорова** | |
| --- | --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **ks\_statistic** | **ks\_pvalue** |
| **lognorm** | 0.003415 | 1398.862722 | 1429.640780 | 0.034494 | 1.680390e-218 |
| **norm** | 0.003415 | 1397.779716 | 1418.298421 | 0.034353 | 1.017618e-216 |
| **gennorm** | 0.003424 | 1385.325533 | 1416.103591 | 0.035912 | 8.503765e-237 |
| **logistic** | 0.003634 | 1227.705647 | 1248.224352 | 0.048541 | 0.000000e+00 |
| **dweibull** | 0.003645 | 1284.200814 | 1314.978873 | 0.044829 | 0.000000e+00 |

Более подробное рассмотрение пяти распределений случайно величины наиболее соответствующих распределению роста спортсменов даёт следующие результаты:

Lognorm (или логнормальное) распределение — это непрерывное распределение вероятностей случайной величины, логарифм которой имеет нормальное распределение. Таким образом, если случайная величина X имеет логнормальное распределение, то Y = ln(X) имеет нормальное распределение. Эквивалентно, если Y имеет нормальное распределение, то показательная функция Y, X = exp(Y), имеет логарифмически нормальное распределение[[227]](#footnote-227).

(Функция распределения - 

Плотность вероятности -  )

Случайная величина, имеющая логарифмически нормальное распределение, принимает только положительные действительные значения. Это удобная и полезная модель для неотрицательных распределений точных науках, а также в медицине, экономике и других областях (например, энергии, концентрации, длины, цены финансовых инструментов и других показателей).

Norm — это тип непрерывного распределения вероятностей для действительной случайной величины[[228]](#footnote-228).

(Функция распределения - 

Плотность вероятности -  )

Нормальные распределения важны в статистике и часто используются в естественных и социальных науках для представления действительных случайных величин, распределение которых неизвестно[[229]](#footnote-229). Их важность частично обусловлена центральной предельной теоремой. Теорема утверждает, что при некоторых условиях среднее значение многих выборок (наблюдений) случайной величины с конечным средним значением и дисперсией само по себе является случайной величиной, распределение которой сходится к нормальному распределению по мере увеличения количества выборок. Следовательно, физические величины, которые, как ожидается, будут суммой многих независимых процессов, таких как ошибки измерения, часто имеют распределение, близкое к нормальному[[230]](#footnote-230).

Более того, гауссовы распределения обладают некоторыми уникальными свойствами, которые ценны в аналитических исследованиях. Например, любая линейная комбинация фиксированного набора независимых нормальных отклонений является нормальным отклонением. Многие результаты и методы, такие как распространение неопределенности и подбор параметров методом наименьших квадратов, могут быть получены аналитически в явной форме, когда соответствующие переменные имеют нормальное распределение[[231]](#footnote-231).

Нормальное распределение иногда неофициально называют колоколообразной кривой. Однако многие другие распределения как было видно из исследования распределений значения веса спортсменов имеют колоколообразную форму (например, распределение Коши, распределение Стьюдента и логистическое распределение).

Gennorm - обобщенное нормальное распределение или обобщенное распределение Гаусса (GGD) представляет собой одно из двух параметрических непрерывных распределений вероятностей. Чтобы различать эти два семейства они называются «симметричными» и «асимметричными. Симметричное обобщенное нормальное распределение, также известное как экспоненциальное распределение степени или обобщенное распределение ошибок, представляет собой параметрическое семейство симметричных распределений. Оно включает в себя все нормальные распределения и распределения Лапласа, а в качестве предельных случаев — все непрерывные равномерные распределения на ограниченных интервалах вещественной прямой. Оно включает нормальное распределение.

(Функция распределения - 

Плотность вероятности -  )

Асимметричное обобщенное нормальное распределение представляет собой семейство непрерывных распределений вероятностей, в которых параметр формы распределения (β) может использоваться для введения асимметрии[[232]](#footnote-232).Когда параметр формы распределения равен нулю, получается нормальное распределение. Положительные значения параметра формы распределения дают смещенные влево распределения, ограниченные вправо, а отрицательные значения параметра формы распределения дают смещенные вправо распределения, ограниченные влево. Только когда параметр формы распределения равен нулю, функция плотности для этого распределения положительна по всей действительной оси: в этом случае распределение является нормальным распределением, в противном случае распределения являются смещенными и, возможно, обратными логнормальными распределениями.

(Функция распределения - 

Плотность вероятности -  )

Logistic (логистическое распределение) представляет собой непрерывное распределение вероятностей. Его функция распределения — это логистическая функция, которая активно используется для классификации в логистической регрессии и как функцию активации в нейронных сетях прямого распространения. По форме оно напоминает нормальное распределение, но имеет более тяжелые хвосты (более высокий эксцесс). Логистическое распределение является частным случаем лямбда-распределения Тьюки.

(Функция распределения - 

Плотность вероятности -  )

В исследуемом случае похожесть на логистическое распределение отражает наличие только положительных значений наблюдений

Dweibull распределение уже было рассмотрено при изучении распределения значений веса спортсменов и в этом случае у него те же самые основания и особенности применения[[233]](#footnote-233).

Из оценки наиболее соответствующих распределений случайной величины можно увидеть, что нормальное распределение является одним из лидеров на наибольшее соответствие с распределением значений роста спортсменов полного набора данных. При этом согласно проведённым тестам с использованием критерия Шапиро-Уилка, критерия согласия Пирсона и критерия Колмогорова-Смирнова исследуемые данные не соответствуют закономерностям нормального распределения. Это происходит несмотря на визуальное соответствие и высокую схожесть с нормальным распределением по отношению к другим распределениям по информационным критериям и лучшие результаты по критерию Колмогорова-Смирнова. Такая ситуация принуждает отвергнуть расхожее мнение, о том, что рост человека — это величина нормально распределённая или принять некоторое допущение, что несмотря на несоответствие критериям рост спортсмена всё же можно оценивать, как распределённый близко к нормальному. Существуют исследования, которые утверждают, что до шести лет рост детей подчиняется нормальному распределения[[234]](#footnote-234)

Согласно высказыванию приписываемому Пуанкаре и часто приводимому для иллюстрации сложившейся ситуации, нормальное распределение является не вполне определённым для реальных исследований: «Физики полагают, что закон Гаусса доказан в математике, а математики полагают, что он экспериментально установлен в физике»[[235]](#footnote-235). Следуя шагам исследования распределения значений веса спортсменов произведём оценку случайной выборки (n=30) из общего набора данных (рис.1)



Рис.1 - Проверка нормальности распределения случайной выборки значений роста спортсменов (n=30) из полного набора данных

Как и в случае с распределением значений веса спортсменов случайная выборка 30 наблюдений из общего набора данных демонстрирует соответствие закономерностям нормального распределения согласно критериям Шапиро-Уилка и Пирсона в отличии от ситуации с проверкой полного набора данных.

Повторяя устоявшуюся схему было произведено сравнение значений полученной выборки на соответствие восьмидесяти наиболее часто встречающихся распределений случайной величины с последующей выборкой пяти наиболее соответствующих распределений. Результаты визуального отображения прошедшего сравнения представлено на рис.1.

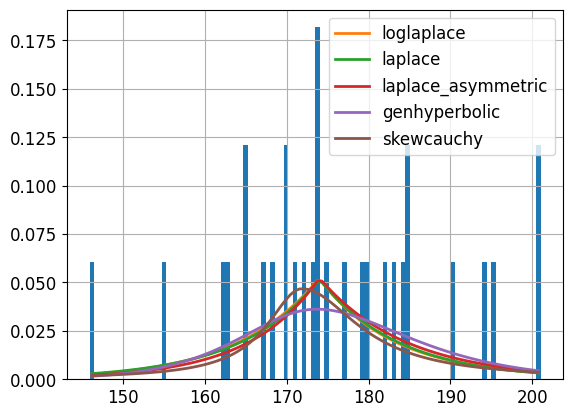


Рис.1 – Соответствие распределения роста спортсменов выборки (n=30) из полного набора данных пяти распределениям

Более подробно с результатами сравнения можно ознакомиться в таблице 1.

Таблица 1. Результаты сравнения распределения роста спортсменов выборки (n=30) из полного набора данных пяти распределениям

| **Вид** | **Средн. квадрат. ошибка (Sumsquare error)** | **Информационные критерии** | | **Критерий согласия Колмогорова** | |
| --- | --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **ks\_statistic** | **ks\_pvalue** |
| **loglaplace** | 0.117393 | 884.452194 | 888.655787 | 0.112577 | 8.013240e-01 |
| **laplace** | 0.117494 | 880.942006 | 883.744401 | 0.114782 | 7.824885e-01 |
| **laplace\_asymmetric** | 0.117565 | 888.166219 | 892.369811 | 0.085627 | 9.669481e-01 |
| **genhyperbolic** | 0.118119 | 879.699570 | 886.705557 | 0.566667 | 1.255522e-09 |
| **skewcauchy** | 0.118484 | 926.448751 | 930.652343 | 0.110181 | 8.211846e-01 |

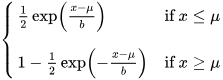
Loglaplace — это распределение вероятностей случайной величины, логарифм которой имеет распределение Лапласа. Если X имеет распределение Лапласа с параметрами µ и b, то Y = eX имеет логарифмическое распределение Лапласа. Свойства распределения можно получить из распределения Лапласа.

(Функция распределения - 

Плотность вероятности - )

Также существуют версии логарифмического распределения Лапласа, основанные на асимметричном распределении Лапласа. В зависимости от параметров, включая асимметрию, логарифм Лапласа может иметь или не иметь конечное среднее и конечную дисперсию[[236]](#footnote-236).

Laplace (распределение Лапласа) — это непрерывное распределение вероятностей. Его также иногда называют двойным экспоненциальным распределением, потому что его можно рассматривать как два экспоненциальных распределения (с дополнительным параметром местоположения), соединенных вместе по оси абсцисс, хотя этот термин также иногда используется для обозначения распределения Гамбеля. Разница между двумя независимыми одинаково распределенными экспоненциальными случайными величинами определяется распределением Лапласа, как и броуновское движение, оцениваемое в экспоненциально распределенном случайном времени.

(Функция распределения - 

Плотность вероятности -  )

Плотность вероятности распределения Лапласа выражается через абсолютную разницу со средним значением. Следовательно, распределение Лапласа имеет более толстые хвосты, чем нормальное распределение. Это частный случай обобщенного нормального распределения и гиперболического распределения. Непрерывные симметричные распределения, которые имеют экспоненциальные хвосты, такие как распределение Лапласа, включают логистическое распределение, и распределение Champernowne.

Laplace\_asymmetric (ALD) — это непрерывное распределение вероятностей, которое является обобщением распределения Лапласа. Точно так же, как распределение Лапласа состоит из двух экспоненциальных распределений одинакового масштаба, расположенных рядом друг с другом относительно x=m, асимметричное распределение Лапласа состоит из двух экспоненциальных распределений неравного масштаба, расположенных рядом друг с другом около x=m, скорректированных для обеспечения непрерывности и нормализации. Разница двух переменных, экспоненциально распределенных с разными средними значениями и параметрами, будет распределяться в соответствии с ALD[[237]](#footnote-237).

(Функция распределения - 

Плотность вероятности - )

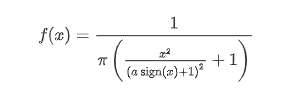
Здесь m — параметр местоположения, λ > 0 — параметр масштаба, а κ — параметр асимметрии. Когда κ = 1, (x-m)s κs упрощается до |x-m| и распределение упрощается до распределения Лапласа.

Genhyperbolic - обобщенное гиперболическое распределение (GH) непрерывное распределение вероятностей, определяемое как нормальная смесь дисперсии и среднего значения, где распределение смешивания представляет собой обобщенное обратное распределение Гаусса (GIG). Его функция плотности вероятности задана в терминах модифицированной функции Бесселя второго рода.

(Плотность вероятности -  )

Как следует из названия, Genhyperbolic имеет очень общую форму, являясь суперклассом, среди прочего, t-распределения Стьюдента, распределения Лапласа, гиперболического распределения, нормального обратного гауссова распределения и дисперсионно-гамма-распределения.

Skewcauchy – скошенное обобщенное t-распределение представляет собой семейство непрерывных распределений вероятностей. Распределение используется в большом количестве различных приложениях[[238]](#footnote-238)[[239]](#footnote-239). Существуют различные параметризации для искаженного обобщенного t-распределения и одно из них скошенное распределение Коши (Skewed Cauchy)

(Плотность вероятности -  )

Среднее значение, дисперсия, асимметрия и эксцесс асимметричного распределения Коши не определены.

При рассмотрении этих распределений случайной величины можно обратить внимание на то, что достоверность их соответствия выборочному набору данных действительно высока, в выборочном наборе данных присутствует асимметрия в распределении и присутствует связь с нормальным распределением, но само нормальное распределение возможно только при определённых ситуациях.

Как и в случае с распределением значений веса спортсменов было проведено моделирование в ходе которого было проведено тысячу случайных выборок из полного набора данных и оценка на соответствие закономерностям нормального распределения с использованием критерия Шапиро-Уилка и критерия согласия Пирсона. Ход этого эксперимента можно увидеть на рис.1.

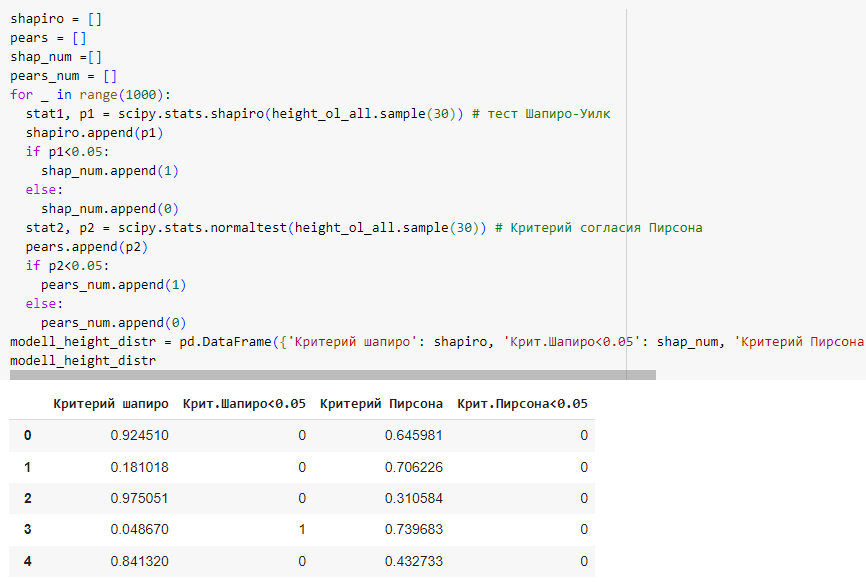


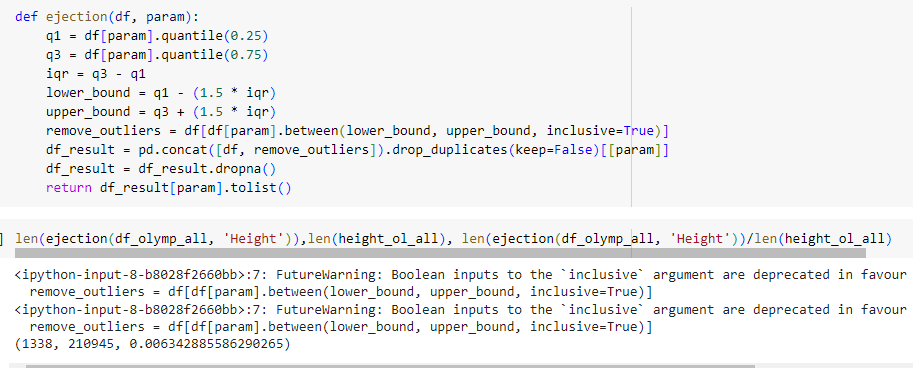
Рис.1 – Моделирование оценки 1000 выборок (n=30) из полного набора данных (N=210945) двумя критериями Шапиро-Уилка и Пирсона

С результатами проведённого моделирования в численном виде данными можно ознакомиться в таблице 1.

Таблица 1 - Оценка соответствия нормальному распределению моделирования случайных выборок из полного набора данных

| **р<0.05** | **Критерий Шапиро-Уилка** | **Критерий согласия Пирсона** |
| --- | --- | --- |
| **нет** | 925 | 921 |
| **да** | 75 | 79 |

Обращает на себя внимание то, что в данном случае ≈ 92% случаев распределение выборочных данных оценивалось как нормальное, что значительно превосходит результаты для оценки распределения значений веса спортсменов. В исследовании распределений значения веса спортсменов было выдвинуто предположение, что существовала связь с вероятностью появления выбросов в случайной выборке. Подобная же операция была проведена и для значений роста. Оценка количества выбросов распределения значений роста спортсменов представлена на рисунке 1.



Рнис.1 – Определение вероятности появления выброса при случайном выборе значения из полного набора данных

Было проведено определение вероятности попадания выбросов в выборку из 30 случайно взятых наблюдений рис.1

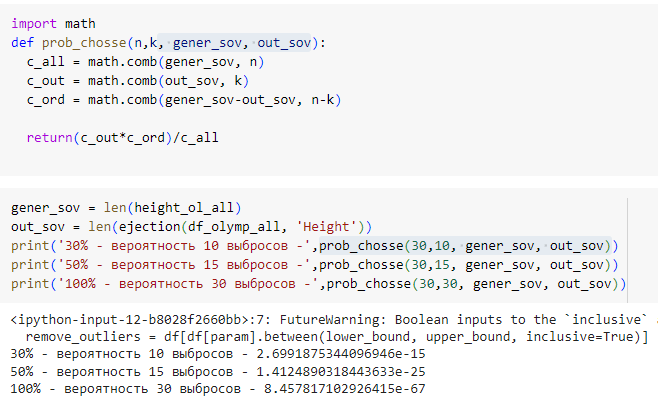


Рис.1 – Определение вероятности попадания выбросов в выборку (n=30) из общего набора данных

Даже простое вычисление 1338/210945 ≈ 0.0063 показывает, что вероятность выбрать значение из полного набора данных и оно принадлежало бы множеству выбросов достаточно мала, в свою очередь выбор 30%, 50% или даже 100% слабо отличается от нуля и это действительно маловероятное явление.

Следующим шагом были проверены уже не случайные выборки, а отдельные виды спорта. Так как нормальное распределение соответствовало небольшим по количеству наблюдений выборкам равным 30, то следующим были проверены записи, относящиеся к виду спорта регби, так как количество записей было рвано 62, что несколько больше 30, но гораздо меньше записей по другим видам спорта. Распределение по виду спорта регби можно увидеть на рис.1

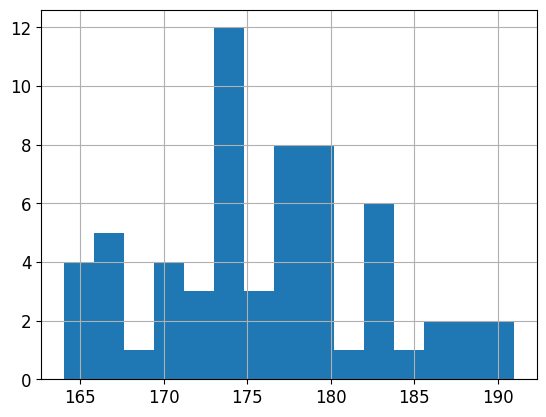


Рис.1 – Гистограмма распределения значений роста для спортсменов регби

По внешнему виду распределение похоже на нормальное, среднее арифметическое 176,1 приблизительно равно медиане 175,0 что является обнадёживающим признаком. Размах выборки с минимальным 164,0 и максимальным значением равным 191,0 говорит об отсутствии выбросов и значит они не смогут повлиять на стремление распределения к нормальному.

Проверка на нормальность также была успешной и оба используемых критерия показали соответствие выборочного распределения закономерностям нормального (рис.1)

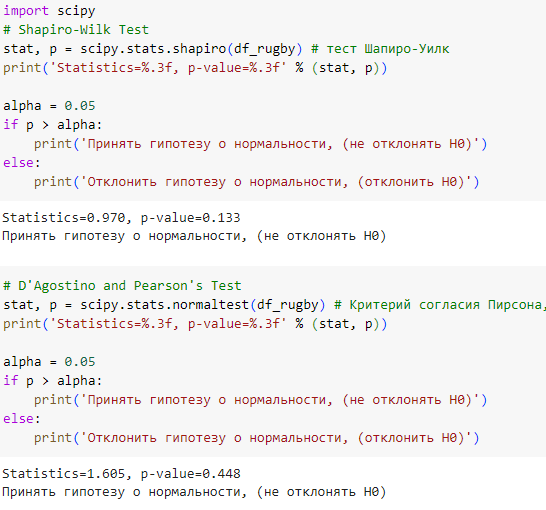


Рис.1 - Проверка нормальности распределения выборки значений роста спортсменов регби из полного набора данных

Следует обратить внимание, что случайная выборка 30 наблюдений демонстрирует лучшее соответствие закономерностям нормального распределения согласно критериям Шапиро-Уилка и Пирсона в отличии от данной выборке со спортсменами одного вида спорта и не включающую в себя выбросов.

Далее было произведено сравнение значений выборки спортсменов регби на соответствие восьмидесяти наиболее часто встречающихся распределений случайной величины с последующей выборкой пяти наиболее соответствующих распределений. Результаты визуального отображения прошедшего сравнения представлено на рис.1.

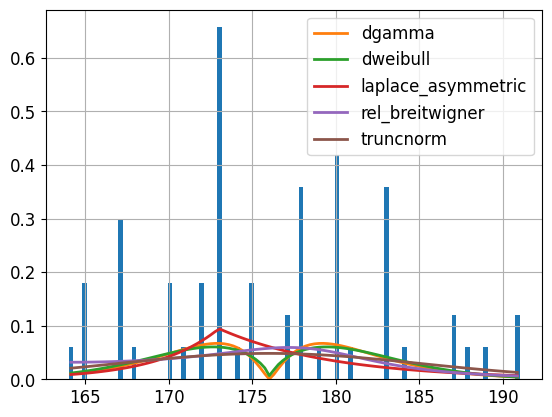


Рис.1 – Соответствие распределения выборки роста спортсменов регби

Можно обратить внимание на асимметричность распределения. Более подробно с результатами сравнения можно ознакомиться в таблице 1.

Таблица 1. Результаты сравнения выборки роста спортсменов регби

| **Вид** | **Средн. квадрат. ошибка (Sumsquare error)** | **Информационные критерии** | | **Критерий согласия Колмогорова** | |
| --- | --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **ks\_statistic** | **ks\_pvalue** |
| **dgamma** | 0.987202 | 729.254733 | 735.636136 | 0.090121 | 0.661636 |
| **dweibull** | 0.991645 | 713.142058 | 719.523461 | 0.089954 | 0.663865 |
| **laplace\_asymmetric** | 0.997851 | 733.451189 | 739.832592 | 0.140517 | 0.156967 |
| **rel\_breitwigner** | 1.004499 | 705.877207 | 712.258610 | 0.119985 | 0.308816 |
| **truncnorm** | 1.004649 | 689.850036 | 698.358574 | 0.145873 | 0.129195 |

Рассматривая полученные данные можно обратить внимание, что полученные распределения близки к нормальному и повторяют в большей или меньшей степени ситуацию со случайной выборкой.

Следующей выборкой были значения роста для спортсменов дзюдо. Распределение представлено на рис.1

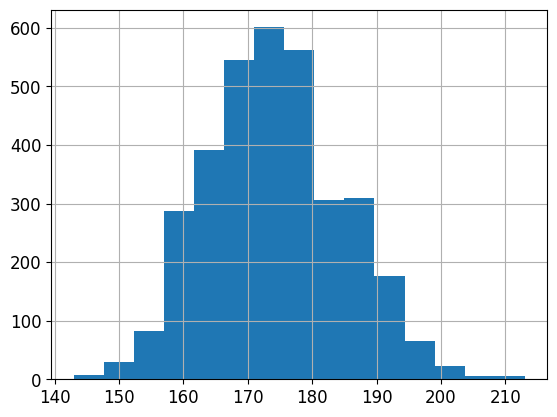


Рис.1 – Гистограмма распределения значений роста для спортсменов дзюдо

Дзюдо появилось на олимпийских играх 1964 года в Токио и это предполагает с одной стороны полноту данных согласно нашим предыдущим наблюдениям, а с другой стороны более узкий временной диапазон позволяет говорить о меньшей изменчивости таких антропометрических характеристик как рост. Среднее арифметическое 174,2 ещё точнее соответствует медиане 174,0, размах выборки. Размах выборки 143-213 см говорит о некотором количестве выбросов в выборке. Общее количество наблюдений значительно больше чем для регби – 3400.

Проверка на нормальность распределения приведена на рис.1

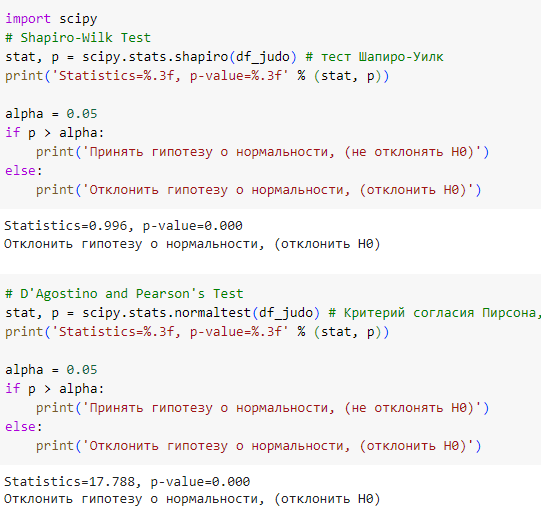


Рис.1 - Проверка нормальности распределения выборки значений роста спортсменов дзюдо из полного набора данных

Было произведено сравнение значений выборки спортсменов дзюдо из полного набора данных на соответствие восьмидесяти наиболее часто встречающихся распределений случайной величины с последующей выборкой пяти наиболее соответствующих распределений. Результаты визуального отображения прошедшего сравнения представлено на рис.1.

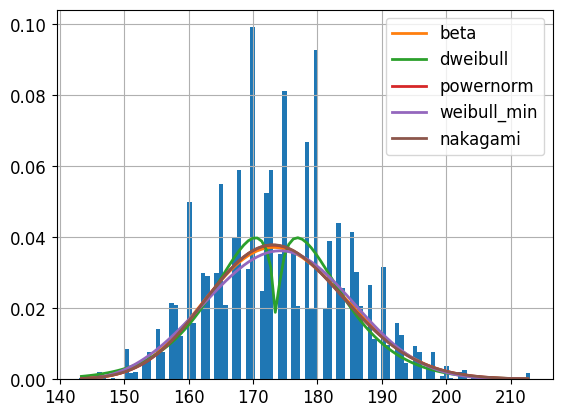


Рис.1 – Соответствие распределения выборки роста спортсменов дзюдо

Можно обратить внимание на то, что с увеличением количества наблюдений распределение значений всё более соответствует нормальному распределению, при этом согласно проведённым оценкам соответствия самому нормальному распределению не наблюдается. Для значений роста, как и для значений веса спортсменов характерна некоторая асимметричность, которая вероятно и мешает соответствию с нормальным распределением. Более подробно с результатами сравнения можно ознакомиться в таблице 1.

Все наиболее соответствующие выборочному распределению относятся к семействам ранее рассмотренных и являются вариациями уже упомянутых выше и все они имеют связь с нормальным распределением.

Таблица 1. Результаты сравнения выборки роста спортсменов дзюдо

| **Вид** | **Средн. квадрат. ошибка (Sumsquare error)** | **Информационные критерии** | | **Критерий согласия Колмогорова** | |
| --- | --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **ks\_statistic** | **ks\_pvalue** |
| **beta** | 0.028375 | 1050.014541 | 1074.540664 | 0.040268 | 3.157122e-05 |
| **dweibull** | 0.028376 | 1028.865858 | 1047.260450 | 0.056819 | 5.547241e-10 |
| **powernorm** | 0.028406 | 1036.933653 | 1055.328246 | 0.044136 | 3.415363e-06 |
| **weibull\_min** | 0.028408 | 1062.226738 | 1080.621330 | 0.041466 | 1.620514e-05 |
| **nakagami** | 0.028411 | 1035.919093 | 1054.313685 | 0.044601 | 2.578073e-06 |

Наиболее соответствующие выборочному распределения в той или иной степени близки к нормальному, при этом критерий Колмогорова-Смирнова не подтверждает совпадения, однако это лучшее из имеющихся часто встречающихся распределений. Дополнительно средне квадратичная ошибка и информационные критерии также показывают достаточно хорошее соответствие.

Следующим этапом была исследована выборка, содержащая спортсменов спортивной борьбы. Эта выборка уже гораздо больше предыдущих и насчитывает 5346 записей, что уже предполагает, что значения выборки вероятно распределяются нормально. Значение среднего арифметического - 172,4 сантиметра приблизительно соответствует значению медианы – 172 сантиметра. Размах выборки достаточно широк и включает в себя некоторое количество выбросов. Минимальное значение составляет – 137 сантиметров, а максимальное – 214 сантиметров. 25% и 75% перцентили составляют – 165 и 180 сантиметров соответственно, что дополнительно свидетельствует о достаточно высокой сбалансированности значений в выборке.

Визуально распределение значений роста спортсменов спортивной борьбы можно оценить на рис.1.

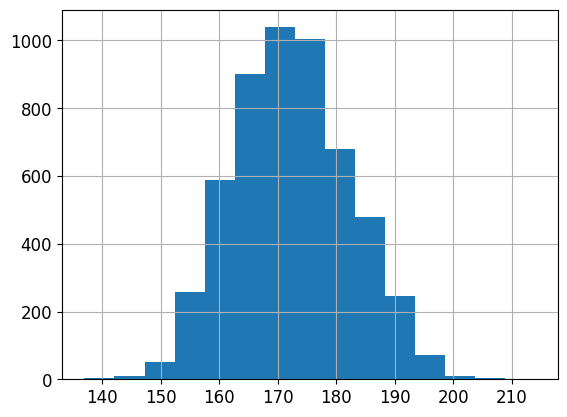


Рис.1 – Гистограмма распределения значений роста для спортсменов спортивной борьбы

Следующим шагом была проверка выборочного распределения на соответствие закономерностям нормального распределения критерием Шапиро-Уилка и критерием согласия Пирсона (рис.1). В ходе проверки оба критерия позволили утверждать, что выборочное распределение не соответствует закономерностям нормального. Кроме прочего можно обратить внимание на то, что критерий Шапиро-Уилка, как это указано в документации не самым лучшим образом справляется с данными размерностью больше чем 5000 наблюдений. При этом критерий согласия Пирсона демонстрирует устойчивую тенденцию увеличения с ростом количества наблюдений: для 210945 наблюдений - 246.893, для 5346 наблюдений - 71.688, для 3400 наблюдений - 17.788, для 62 наблюдений - 1.605, то у критерия Шапиро-Уилка это выглядит по-другому: для 210945 наблюдений - 0.998, для 5346 наблюдений - 0.992, для 3400 наблюдений - 0.996, для 62 наблюдений - 0.970

Стоит обратить внимание на то, что несмотря на такую особенность критерии дали согласованный ответ и с одной стороны это может означать, что несмотря на заявленные ограничения критерий Шапиро-Уилка пригоден и для оценки массивов данных более 5000 записей, а с другой, что количество записей оказывает влияние на оба критерия.

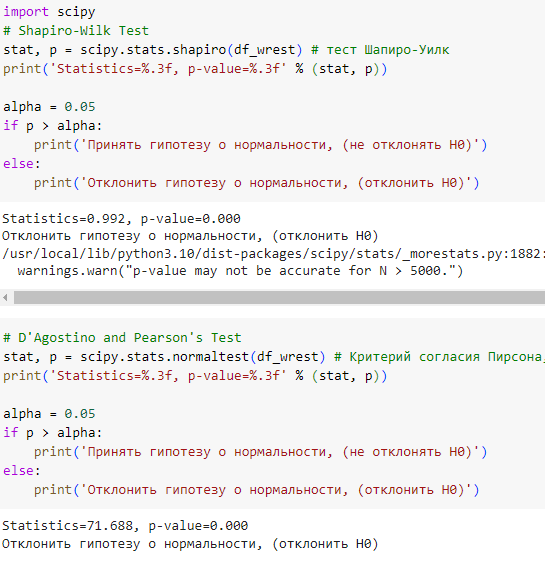


Рис.1 - Проверка нормальности распределения выборки значений роста спортсменов спортивной борьбы из полного набора данных

Было произведено сравнение значений выборки спортсменов дзюдо из полного набора данных на соответствие восьмидесяти наиболее часто встречающихся распределений случайной величины с последующей выборкой пяти наиболее соответствующих распределений. Результаты визуального отображения прошедшего сравнения представлено на рис.1.

В этот раз распределение визуально более всего напоминает нормальное и в этом можно убедиться на рис.1

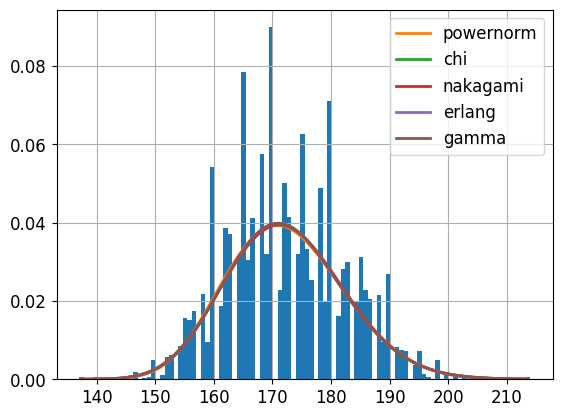


Рис.1 – Соответствие распределения выборки роста спортсменов спортивной борьбы

Можно заметить, что графики распределения практически совпадают, но сами распределения случайной величины различны. С численными оценками можно ознакомиться в таблице 1.

Таблица 1. Результаты сравнения выборки роста спортсменов спортивной борьбы

| **Вид** | **Средн. квадрат. ошибка (Sumsquare error)** | **Информационные критерии** | | **Критерий согласия Колмогорова** | |
| --- | --- | --- | --- | --- | --- |
| **AIC** | **BIC** | **ks\_statistic** | **ks\_pvalue** |
| **powernorm** | 0.019564 | 1163.821048 | 1183.573360 | 0.049946 | 4.977146e-12 |
| **chi** | 0.019634 | 1160.730227 | 1180.482539 | 0.054450 | 3.235065e-14 |
| **nakagami** | 0.019634 | 1160.730110 | 1180.482422 | 0.054450 | 3.234650e-14 |
| **erlang** | 0.019649 | 1145.749360 | 1165.501672 | 0.052938 | 1.841459e-13 |
| **gamma** | 0.019649 | 1145.749260 | 1165.501571 | 0.052938 | 1.841504e-13 |

Powernorm – это обобщение нормального распределения, является скошенным и часто используется в эмпирических исследованиях[[240]](#footnote-240), например для оценки высоты деревьев[[241]](#footnote-241), рост человека чаще оценивается логнормальным распределением, но и Powernorm находит применение в оценке роста[[242]](#footnote-242).

(Плотность вероятности - )

При с=1 будет получено нормальное распределение

Chi - это непрерывное распределение вероятностей по неотрицательной действительной прямой. Это распределение положительного квадратного корня из суммы квадратов независимых гауссовых случайных величин. Эквивалентно, это распределение евклидова расстояния между многомерной гауссовой случайной величиной и началом координат. Таким образом, он связан с распределением хи-квадрат, описывая распределение положительных квадратных корней из переменной, подчиняющихся распределению хи-квадрат.

(Функция распределения - 

Плотность вероятности -  )

Оно связано с нормальным распределением  в определённых условиях.

Nakagami - это двухпараметрическое распределение вероятностей, связанное с гамма-распределением[[243]](#footnote-243). Он используется для моделирования физических явлений с затуханием колебаний, таких как те, которые обнаруживаются при медицинском ультразвуковом исследовании, связи, метеорологии, гидрологии и сейсмологии.

(Функция распределения - 

Плотность вероятности -  )

При 2m = k распределение Nakagami дает распределение Chi. При m = 1/2 распределение Nakagami дает полунормальное распределение.

Erlang - представляет собой двухпараметрическое семейство непрерывных распределений вероятностей. Распределения Эрланга и Пуассона дополняют друг друга: в то время как распределение Пуассона подсчитывает количество событий, происходящих за фиксированный промежуток времени, распределение Эрланга подсчитывает количество времени до возникновения фиксированного количества событий.

(Функция распределения - 

Плотность вероятности - )

Возрастное распределение заболеваемости раком часто следует распределению Эрланга, также используется как приближение распределения времени клеточного цикла[[244]](#footnote-244)[[245]](#footnote-245).

Gamma - представляет собой двухпараметрическое семейство непрерывных распределений вероятностей. Экспоненциальное распределение, распределение Эрланга и распределение хи-квадрат являются частными случаями гамма-распределения[[246]](#footnote-246).

(Функция распределения - 

Плотность вероятности -  )

Гамма-распределение часто используется для моделирования времени ожидания. Например, при тестировании жизни время ожидания до смерти является случайной величиной, которую часто моделируют с помощью гамма-распределения

Завершая рассмотрение распределения значений роста спортсменов можно прийти к следующим выводам:

Бытующее предположение, что значения роста при оценке больших групп соответствует закономерностям нормального распределения не соответствует действительности, более точным будет утверждение, что оно стремится к приблизится нормальному распределению.

Действующие подходы к оценке показывают, что при уменьшении наблюдений случайной выборки из совокупности чьё распределение стремится к нормальному вероятность получить собственно нормальное распределение гораздо выше, чем при увеличении количества наблюдений. Можно сделать закономерное предположение, что в общем случае можно обращаться с данными, как если бы их значения распределялись, подчиняясь закономерностям нормального распределения если случайные выборки небольшого размера из общего набора распределены согласно закономерностям нормального распределения.

Один и тот же параметр (в исследуемом случае рост) при оценке спортсменов отдельных видов спорта или групп видов спорта может подчиняться разным законам распределения случайной величины, даже при сходстве видов спорта или нахождение их в одной группе. Так в проведённом исследовании в одной и той же группе видов спорта – единоборства, значения роста для спортсменов различных видов – дзюдо и спортивная борьба значения роста спортсменов стремились к различным видам распределений случайной величины. Это безусловно имеет значение для качества математического и компьютерного моделирования в сфере физической культуры и спорта.

Проведённое исследование было возможно только с применением информационных технологий и технологий искусственного интеллекта, так как без автоматизации проведения вычислений и поддержки принятия решений обработка такого количества данных заняла бы недопустимо большое время.

# 3.5. Статистическая оценка для данных большого размера в сфере физической культуры и спорта

Получив оценки распределения значений для роста спортсменов возник вопрос о возможности использования как параметрических, так и непараметрических статистических критериев для оценки данных размер которых становится всё более значительным.

Традиционно в сфере физической культуры и спорта наибольшее распространение получил критерий Стьюдента (Student's t-test) который позволяет определить статистическую значимость различий двух значений среднего арифметического для одной или двух зависимых, или независимых выборок.

Исходя из этой структуры первым был использован одновыборочный критерий Стьюдента. Для определения признака сравнения было проведено обращение к глобальному ресурсу «Сотрудничество по факторам риска неинфекционных заболеваний» (NCD-RisC) который представляет глобальную сеть ученых-медиков, которая располагает данными более чем 3300 исследований населения из 196 стран, начиная с 1957 года, в которых приняли участие почти 200 миллионов участников где в качестве параметров используется рост и вес исследуемого. Данные по росту и весу сосредоточены на подрастающем поколении (от пяти до девятнадцати летнего возраста) и были использованы для исследований, опубликованных в Lancet.

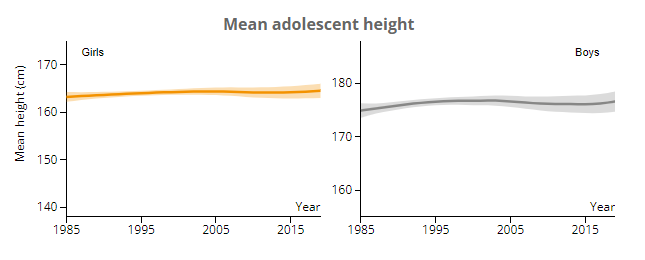


Рис.1 – Изменение среднего возраста подрастающего поколения в России (по данным NCD Risk Factor Collaboration)

Для исследования была проведена выгрузка данных для России[[247]](#footnote-247) (ри.1). Исходя из полученных данных можно заметить, что наибольшее соответствие среднему арифметическому значений роста спортсменов из полного набора данных (175.34 сантиметра) соответствует рост мальчиков и юношей из набора данных NCD-RisC по России с 1985 по 2019 гг. Поэтому были сформированы три выборки:

1. Значения роста по всем данным
2. Значения роста для 19 летнего возраста
3. Значения роста для юношей 19 летнего возраста

Таблица 1. – Описательные статистики для значений роста юношей из набора данных NCD-RisC по России

|  | **Выборка** | | |
| --- | --- | --- | --- |
| **№1** | **№2** | **№3** |
| **Количество наблюдений (count)** | 1050.0 | 70.0 | 35.0 |
| **Средн.арифметическое (mean)** | 146.108553 | 170.184581 | 176.293300 |
| **Средн.квадратичное откл (std)** | 20.421521 | 6.166802 | 0.482826 |
| **Минимальное значение (min)** | 109.098216 | 163.235882 | 174.966888 |
| **25% процентиль** | 127.087641 | 164.181730 | 176.166216 |
| **50% процентиль (медиана)** | 150.078566 | 169.742130 | 176.343921 |
| **75% процентиль** | 163.563122 | 176.327943 | 176.672164 |
| **Максимальное значение (max)** | 176.830582 | 176.830582 | 176.830582 |

Первая выборка бракуется сразу так как разность средних арифметических слишком велика и для исследования берутся вторая и третья выборки.

Результаты сравнения одновыборочным критерием Стьюдента среднего роста спортсменов со средним ростом 19 -летних юношей и девушек России на рис.1



Рис.1 - Результаты сравнения одновыборочным критерием Стьюдента среднего роста спортсменов со средним ростом 19летних жителей России

Сравнение с результатами второй выборки можно видеть на рис.1



Рис.1 - Результаты сравнения одновыборочным критерием Стьюдента среднего роста спортсменов со средним возрастом юношей России

Так как средний рост спортсменов в полном наборе данных равен 175.34 сантиметра, то различие со средним арифметическим второй выборкой составляет ≈ 5,34 сантиметра, а со средним арифметическим третьей выборки ≈ 0,96 сантиметра. Можно обратить внимание, что различие даже менее единицы демонстрирует, что присутствует достоверное различие.

Далее была попытка уточнить полученные результаты и из полного набора данных были выбраны данные начиная с 1968 года средне арифметическое которых равно 175.54916 сантиметра. (Рубеж 1968 года был определён ранее при произведении исследовательского анализа данных) Результаты представлены на рис.1

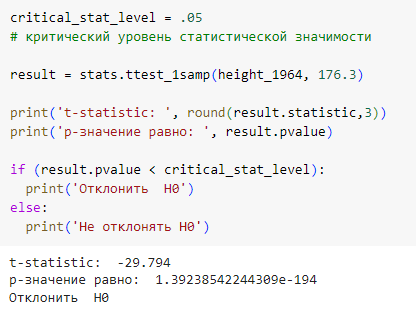


Рис.1 - Результаты сравнения одновыборочным критерием Стьюдента среднего роста спортсменов в играх с 1968 года со средним возрастом юношей России

Даже в этом случае, когда различие ещё меньше ≈ 0,75 можно увидеть достоверное различие и возникает вопрос возможно действительно данные достаточно большого размера не имеет смысла исследовать с помощью параметрических критериев. Однако существует популярная сейчас альтернатива - Z-тест. Z-тест — это статистический метод, для которого распределение тестовой статистики при нулевой гипотезе может быть аппроксимировано нормальным распределением. Z-тест проверяет среднее значение распределения. Для каждого уровня значимости в доверительном интервале Z-критерий имеет одно критическое значение, что делает его более удобным, чем t-критерий Стьюдента, критические значения которого определяются размером выборки (через соответствующие степени свободы). Кроме того он более устойчив для малых (n≤10) или для больших (n≤10) наборов данных[[248]](#footnote-248)[[249]](#footnote-249). Однако Z-тест требует знания дисперсии или среднеквадратичного отклонения поэтому реже используется на практике. В исследуемом случае среднеквадратичное отклонение известно и поэтому был использован Z-тест и результаты можно увидеть на рис.1

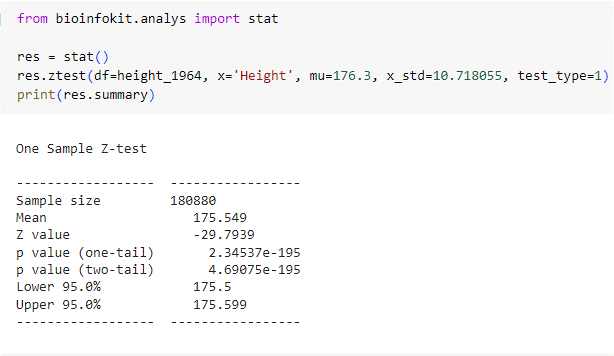


Рис.1 - Результаты сравнения с использованием Z-теста среднего роста спортсменов в играх с 1968 года со средним возрастом юношей России

Далее был использован t-критерий Стьюдента для зависимых выборок и в этом случае сравнивался рост спортсменов раннего и современного периода, который мы определили границей между 1964 и 1968 годом. Поэтому первая группа включала в себя записи до 1968 года и начиная с этого года до 2016 года. Так как записей, включающих в себя год не больший 1964 было только 30065 более поздний период, пришлось ограничить случайной выборкой такого же числа наблюдений. Результаты исследования можно увидеть на рис.1

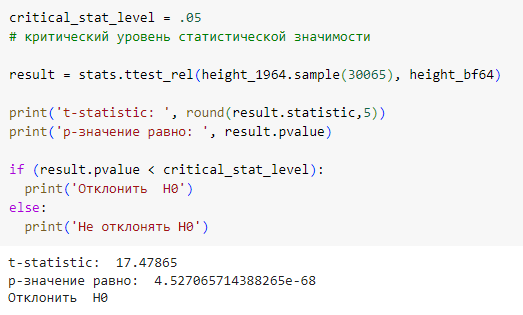


Рис.1 - Результаты сравнения роста спортсменов раннего и современного периодов с использованием t-критерия Стьюдента

Согласно результатам применения критерия можно утверждать, что средний рост спортсменов изменился. Далее был использован t-критерий Стьюдента для независимых выборок (рис.1)

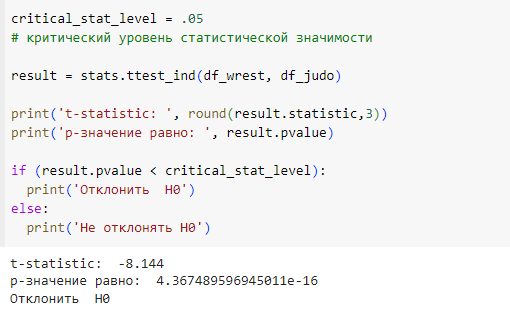


Рис.1 - Результаты сравнения роста спортсменов спортивной борьбы и дзюдо с использованием t-критерия Стьюдента

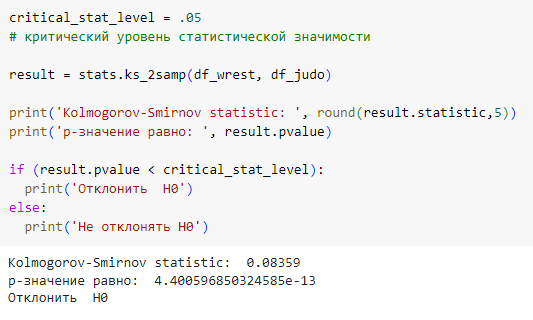


Рис.1 - Результаты сравнения роста спортсменов спортивной борьбы и дзюдо с использованием критерия Колмогорова-Смирнова

Для сравнения результатов спортивная борьба и дзюдо подверглись исследованию с помощью критерия Колмогорова-Смирнова (рис.1) и Z-теста (рис.1)

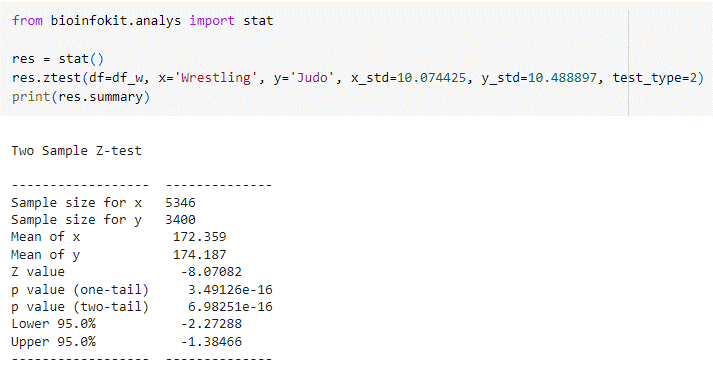


Рис.1 - Результаты сравнения роста спортсменов спортивной борьбы и дзюдо с использованием Z-теста

Результаты всех трёх исследований были аналогичны и рост спортсменов спортивной борьбы и дзюдо оказался достоверно различным (гипотезу о равенстве среднего роста пришлось отвергнуть). При этом результаты (p-value) Z-тест и t-критерия Стьюдента гораздо ближе друг к другу чем к критерию Колмогорова-Смирнова. Результаты же сравнения одновыборочного теста Стьюдента и Z-теста показывают, что значимость (p-value) отличается на порядок. Это конечно различие, но для точности в физической культуре и спорте различие в 194-ом знаке после запятой не является определяющей и как в случае одной выборки и двух независимых можно сделать заключение о том, что действительно между этими тестами есть различие, которое растёт с увеличением выборки, но при этом устойчивость критерия Стьюдента достаточна для целей и задач физической культуры и спорта на текущем объёме данных (300000 наблюдений), что превышает объём данных обычного исследования.

Далее были использованы непараметрические критерии. Первым был использован критерий Вилкоксона (Wilcoxon Т-test) Этот тест используется вместо t-критерия Стьюдента для парных выборок. Тест проводится путем ранжирования абсолютных различий между парными наблюдениями с учетом их знаков. Затем вычисляется сумма рангов положительных разностей и сравнивается с суммой отрицательных разностей. Затем статистика теста рассчитывается как меньшая из этих двух сумм. Таким образом, он проверяет, является ли распределение разностей x - y симметричным относительно нуля. Тем самым определяет происходят ли две связанные выборки из популяций с одинаковой медианой. Результаты исследования представлены на рис.1

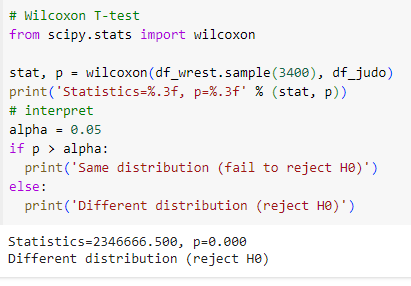


Рис.1 - Результаты сравнения роста спортсменов спортивной борьбы и дзюдо с использованием критерия Вилкоксона

Далее был использован критерий Краскалла-Уоллеса (Kruskal-Wallis H-test) который проверяет нулевую гипотезу о том, что медиана значений всех групп равна. Это непараметрическая версия ANOVA. Тест работает на двух и более независимых выборках, которые могут иметь разные размеры в отличии от критерия Вилкоксона, который требует равенства наблюдений в группах. Отказ от нулевой гипотезы не позволяет уточнить как именно отличаются выбранные признаки, поэтому нужны апостериорные сравнения. Ввиду предположения, что критерий Краскалла-Уоллеса имеет распределение хи-квадрат, количество выборок в каждой группе не должно быть слишком маленьким. Типичное правило состоит в том, что каждая группа должна содержать не менее 5 измерений, но в выбранном направлении исследования размер данных напротив стремиться к увеличению и это делает оценки более резкими, как в приведённом случае на рис.1

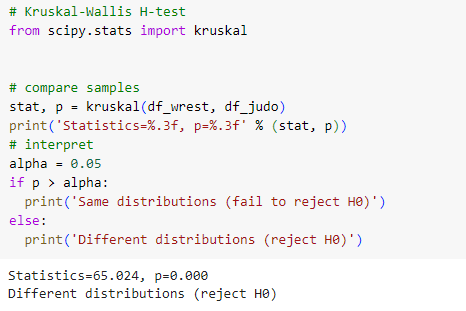


Рис.1 - Результаты сравнения роста спортсменов спортивной борьбы и дзюдо с использованием критерия Краскалла-Уоллеса

Далее был использован тест Фридмана проверяет нулевую гипотезу о том, что повторяющиеся выборки одного и того же набора данных имеют одинаковое распределение. Его часто используют для проверки согласованности образцов, полученных разными способами. Например, если к одной и той же группе лиц используются два метода выборки, можно использовать критерий Фридмана, чтобы определить, являются ли эти два метода выборки согласованными. С помощью этого теста было проведено моделирование с целью тестирования согласованности методов выборки, применявшихся ранее. Модель представлена на рис.1

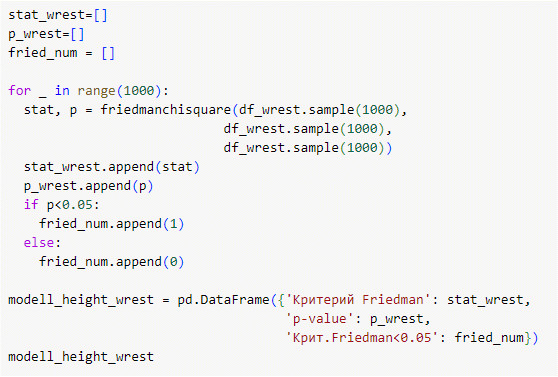


Рис.1 – Моделирование согласованности метода выборки из набора данных значений роста спортсменов спортивной борьбы

В результате моделирования из 1000 случаев в 34 было выявлено различие между тремя подвыборками с р≤0,05 и 4 случая с р≤0,01. Таким образом метод выборки используемый в исследовании вполне удовлетворяет требованиям и проводимые исследования вполне надёжны.

Из проведённых исследований можно сделать несколько выводов:

Конечно же без средств автоматизации исследования такая работа не была бы возможной.

Если выборка имеет достаточно большое количество наблюдений (n≥1000) и её распределение стремится к нормальному, то использование t-критерия Стьюдента для исследования вполне продуктивно, хотя Z-тест будет более продуктивен при дальнейшем росте количества наблюдений (n≥300000).

Получено подтверждение, что при компьютерном и математическом моделировании даже в одной группе спорта распределение значений роста спортсменов будет отличаться и на современном этапе развития информационных технологий необходимо учитывать это в исследованиях.

# 4. Регрессионный и классификационный анализ

В математическом моделировании регрессионный анализ представляет собой набор методов для оценки статистических взаимосвязей между зависимой переменной (называемой также «результат» или «ответ») и одной или несколькими независимыми переменными (называемыми также «предикторами», «объясняющими переменными» или «признаками»)[[250]](#footnote-250). Наиболее распространенным методом регрессионного анализа является простая линейная регрессия, при которой находятся коэффициенты перед каждой из независимых переменных, которые наиболее точно соответствуют данным в соответствии с выбранным математическим критерием оценки. В случае одной объясняющей (независимой) переменной называется простой линейной регрессией; для более чем одного этот процесс называется множественной линейной регрессией. Наиболее часто используемый метод для построения регрессионной модели - это метод наименьших квадратов который определяет прямую (или кривую, в общем же случае гиперплоскость) минимизирующую сумму квадратов разностей между истинными данными и результатами решения уравнения описывающих эту гиперплоскость (прямую или кривую). Такой подход позволяет исследователю оценить условное ожидание зависимой переменной, когда независимые переменные принимают заданный набор значений. Менее распространенные формы регрессии используют немного другие процедуры для оценки альтернативных параметров местоположения (например, квантильная регрессия или Necessary Condition Analysis (NCA)) или оценки условного ожидания в более широком наборе нелинейных моделей (например, непараметрическая регрессия).

Как и все другие формы регрессионного анализа, линейная регрессия фокусируется на условном распределении вероятностей ответа (значения зависимой переменной) с учетом значений предикторов (значений независимых переменных), а не на совместном распределении вероятностей всех этих переменных, что является областью многомерного анализа.

Линейная регрессия была первым типом регрессионного анализа, который был тщательно изучен и широко использовался в практических приложениях. Это связано с тем, что модели, которые линейно зависят от неизвестных параметров, легче подобрать и они более интуитивно понятны, чем модели, которые нелинейно связаны с их параметрами, и статистические свойства полученных оценок легче определить.

Линейная регрессия имеет множество практических применений. Большинство приложений попадают в одну из следующих двух широких категорий: прогнозирование будущего изменения зависимой переменной при получении новых данных и объяснение связи зависимой переменной от независимых в актуальной ситуации.

В прогнозе целью является ошибка, то есть уменьшение различия прогноза от значений, которые будут получены при использовании на новых наборах данных. Таким образом можно использовать линейную регрессию для сближения прогностической модели к наблюдаемому набору данных, значений результирующей и объясняющих переменных. Если после разработки такой модели собираются дополнительные значения объясняющих переменных без сопровождающего значения результирующей переменной, подобранную модель можно использовать для прогнозирования нового результата или совершенствовать имеющуюся модель если результирующая переменная будет известна.

Если цель состоит в том, чтобы объяснить изменение результирующей переменной, которое можно соотнести с изменениями объясняющих переменных. Здесь можно применить линейный регрессионный анализ для количественной оценки силы связи между результирующей переменной и объясняющими переменными и, в частности, для определения того, имеют ли некоторые объясняющие переменные линейную связи с ответом или определить, какие подмножества объясняющих переменных содержат полезную или напротив избыточную информацию о результирующей переменной.

# 4.1. Линейная регрессия

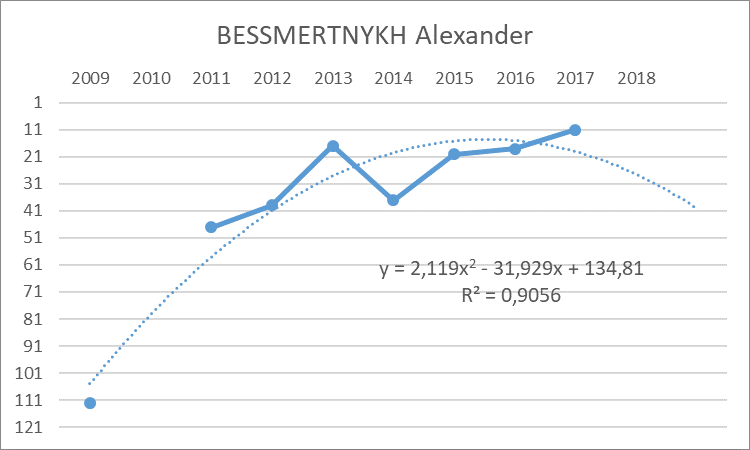
Прогнозирование является достаточно востребованным методом в физической культуре спорта и одно из направлений – это прогнозирование спортивного результата. Действительно важно в ходе процесса многолетней спортивной подготовки иметь знания об ожидаемом результате выполнения того или иного упражнения или занятия той или иной позиции в рейтинге. Это нужно и для корректной организации учебно-тренировочного процесса, и для обоснованного формирования спортивного коллектива, для определения тренда в спортивной карьере. Прогнозирование необходимо для поддержки принятия решения и технологии искусственного интеллекта, в данном случае регрессионный анализ с использованием линейной регрессии позволяют сделать этот процесс быстрым, эффективным и доступным для использования в практике.

Ранее нами были проверены исследования с использованием регрессионных моделей с использованием скользящей средней на спортсменах зимних видов спорта 31 спортсмен (мужчины) фристайл (лыжная акробатика), 22 спортсменки конькобежного спорта (женщины) и 31 спортсмен лыжник (мужчины) по результатам их участия в кубке мира при завершении годичного цикла[[251]](#footnote-251). Была выявлена возможность построения математической модели га основе полинома второй (первая модель) и полинома второй степени (вторая модель). Результаты представлены в табл.1

Таблица 1. - Продуктивность прогнозирования спортивной успешности с помощью регрессионной модели основанной наскользящей средней (Ермаков А.В., Мякинченко П.Е., 2021).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Вид спорта** | **Фристайл (Лыжная акробатика)** | **Лыжные гонки** | **Конькобежный спорт** | **Итого** |
| **Количество исследуемых спортивных карьер** | | | | |
| Спортсмены (чел.) | 31 | 22 | 31 | 84 |
| Спортсмены с достоверностью аппроксимации R²≥8 (чел./%) | 14  (41,18%) | 5  (22,73%) | 15  (45,16%) | 34  (40,45%) |
| Спортсмены относящиеся к первой модели (чел./%) | 7  (50%) | 4  (80%) | 8  (53,33%) | 19  (55,88%) |
| Спортсмены относящиеся ко второй модели (чел./%) | 7  (50%) | 1  (20%) | 7  (46,67%) | 15  (44,12%) |
| **Точность прогнозирования тренда** | | | | |
| Точность прогнозирования тренда для первой модели (%) | 85,71% | 75% | 75% | 78,95%. |
| Точность прогнозирования тренда второй модели (%) | 57,14% | 100% | 71,43% | 66,67%. |
| Точность прогнозирования тренда средняя (%) | 71,43%, | 80% | 73,33% | 73,53% |

Для более наглядного отображения можно привести иллюстрацию этого процесса на графике (рис.1)



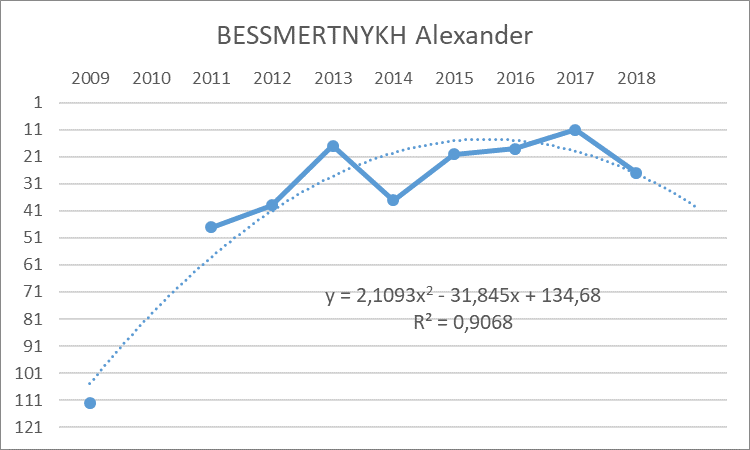


Рис.1 – Результаты прогноза карьеры спортсмена полиномом второй степени (Ермаков А.В., Мякинченко П.Е., 2021)

В качестве примера построения простой линейной регрессии с целью объяснения целевой переменной можно выбрать связь упражнений в ходе спортивной подготовки. Например, в течении 8 учебно-тренировочных занятий спортсмен выполнял комплекс упражнений в котором одно упражнение было целевым (соревновательным), а два других должны были способствовать его совершенствованию (рис.1)

Таблица1. – Результаты наблюдений за тренировочной деятельностью

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Занятие/  Упражнение | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Первое | 8 | 9 | 16 | 15 | 15 | 17 | 16 | 18 |
| Второе | 20 | 20 | 25 | 40 | 35 | 30 | 29 | 27 |
| Целевое | 25 | 30 | 50 | 60 | 63 | 62 | 48 | 55 |

Как можно убедиться на рис.1 между упражнениями если и есть связь, то она относительно сложна для описания и возникает задача создания модели для этого учебно-тренировочного процесса и цифрового двойника спортсмена.

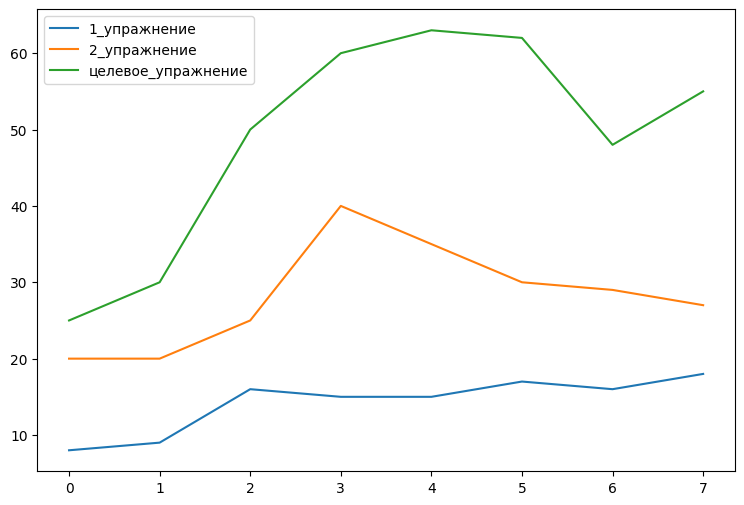


Рис.1 – Результативность выполнения упражнений в 8 занятиях

Используя диаграмму Boxplot можно предположить, что эти упражнения имеют разное распределение и используемыми ранее методами сложно получить полезные исследовательские данные (рис.1).

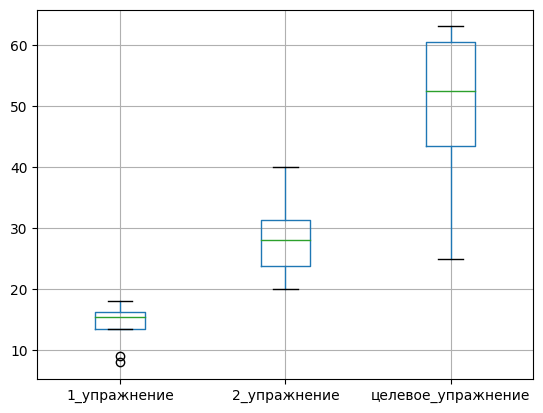


Рис.1 – Распределения выполнения упражнений на 8 занятиях

Данная гипотеза была проверена с помощью дисперсионного анализа, однофакторного теста ANOVA (рис.1)

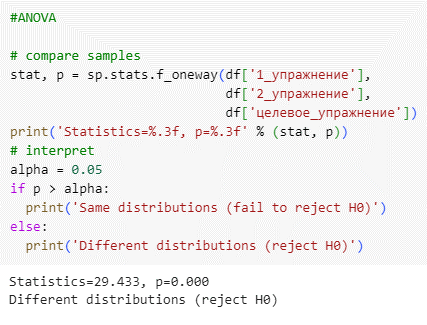


Рис.1 – Результаты использования однофакторного дисперсионного анализа ANOVA

Для более углубленного анализа был применён критерий Тьюки, который позволяет увидеть и попарное сравнение убедившись, что распределения всех признаков различны (рис.1).



Рис.1 – Результаты использования критерия Тьюки

При визуальной оценке можно заметить, что, хотя распределения различны, но зависимость всё же есть как между целевым упражнением и первым (рис.1), так и между целевым упражнением и вторым (рис.1)

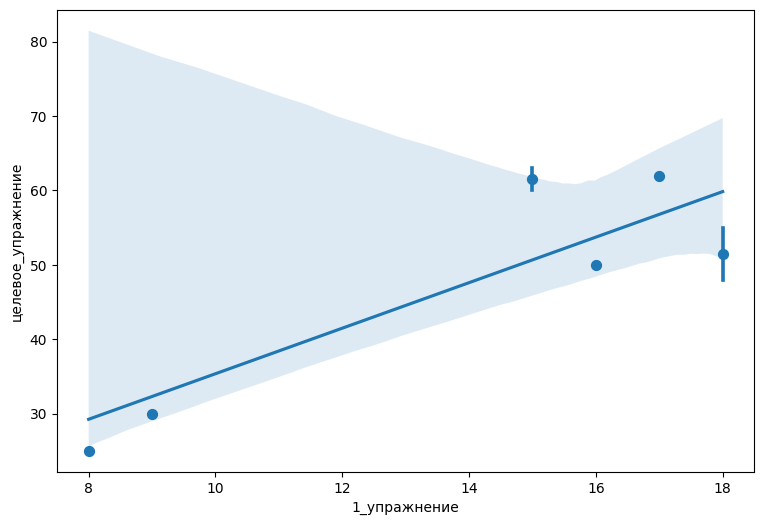


Рис.1 – Линейная зависимость целевого упражнения от первого

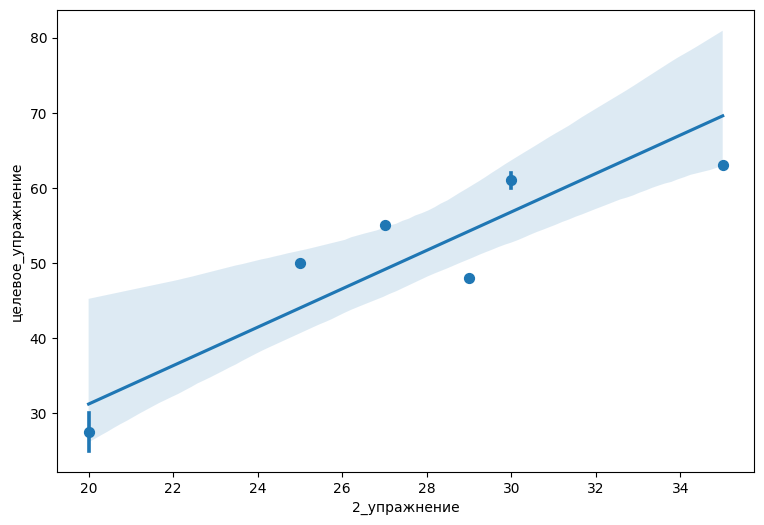


Рис.1 – Линейная зависимость целевого упражнения от второго

После определения неравенства распределений возникает вопрос о поиске иной зависимости между переменными и возникает необходимость проведения корреляционного анализа. Его результаты на рис.1

Значение коэффициента корреляции показывает о достаточно высокой статистической взаимосвязи между всеми упражнениями, но превращение этой информации в модель затруднительно. Поэтому следующим шагом явился переход к регрессионной модели. Здесь необходимо учитывать, что существует связь первого и второго упражнения с целевым (r= и r= соответственно) и связь между первым и вторым упражнением с достаточно высоким значением коэффициента корреляции 0,72, то есть можно говорить о наличии гетероскедастичности.

Существование гетероскедастичности является серьезной проблемой в регрессионном анализе и дисперсионном анализе, поскольку оно делает недействительными статистические тесты значимости, которые предполагают, что все ошибки моделирования имеют одинаковую дисперсию или можно сказать, что это нарушение предпосылок теоремы Гаусса-Маркова[[252]](#footnote-252).

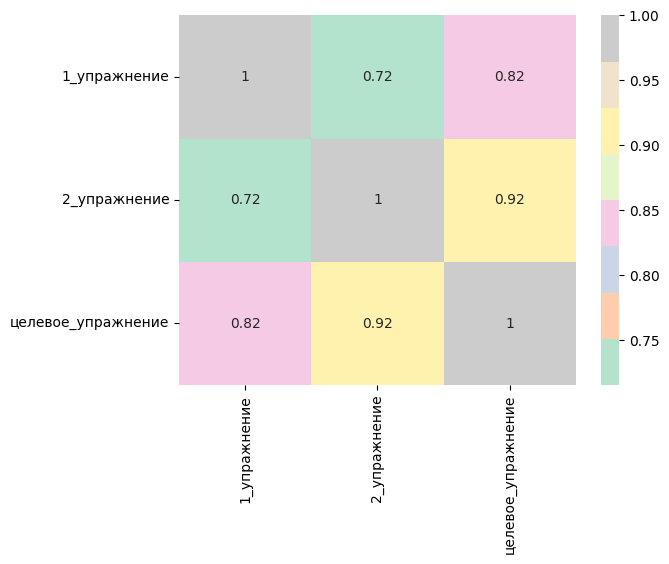


Рис.1 – Корреляционная матрица исследуемого набора данных

Хотя обычная оценка методом наименьших квадратов все еще является несмещенной при наличии гетероскедастичности, она не всегда эффективна, а вывод, основанный на предположении о гомоскедастичности может вводить в заблуждение. В этом случае в прошлом часто использовался метод обобщенных наименьших квадратов (GLS)[[253]](#footnote-253). В настоящее время одним из способов решения проблемы в данной ситуации является включение стандартных ошибок, согласованных с гетероскедастичностью, вместо использования GLS, поскольку GLS может демонстрировать сильную предвзятость в небольших выборках. Применяется взвешенный метод наименьших квадратов и существует подход преобразующий зависимую переменной (например, использование логарифма от зависимой переменной).

Сначала необходимо объективно удостовериться (несмотря на очевидность), что гетероскедастичность в данном случае присутствует и для этого был проведён White test — это статистический тест, который устанавливает, является ли дисперсия ошибок в регрессионной модели постоянной, то есть является проверкой на гетероскедастичность (рис.1).

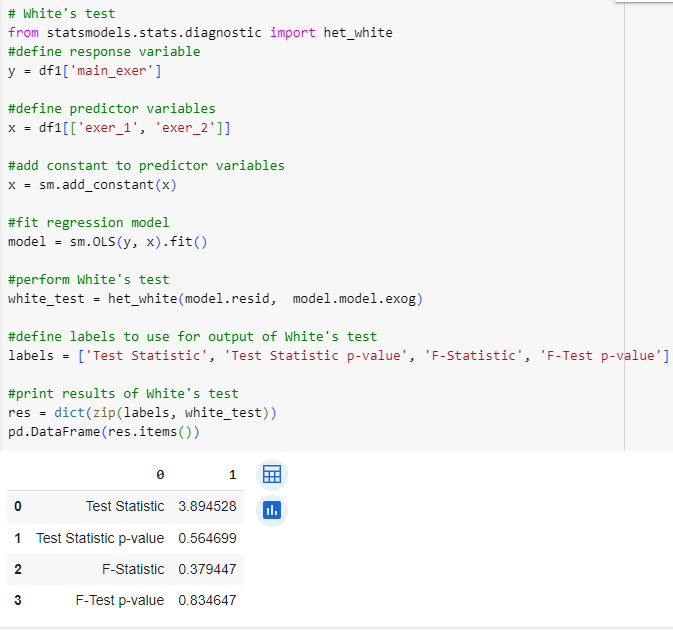


Рис.1 – Результаты проверки на гетероскедастичность (White test)

Как можно заметить, что в исследуемом наборе данных гомоскедастичность присутствует и это может уменьшить продуктивность модели, но согласно особенностям сферы физической культуры и спорта и конкретно спортивной тренировки, где редко присутствуют абсолютно изолируемые друг от друга средства и методы и большую значимость имеет их комплексность и взаимопроникновение была выдвинута гипотеза о допустимости в данном случае наличия гетероскедастичности и слабого её влияния на продуктивность модели.

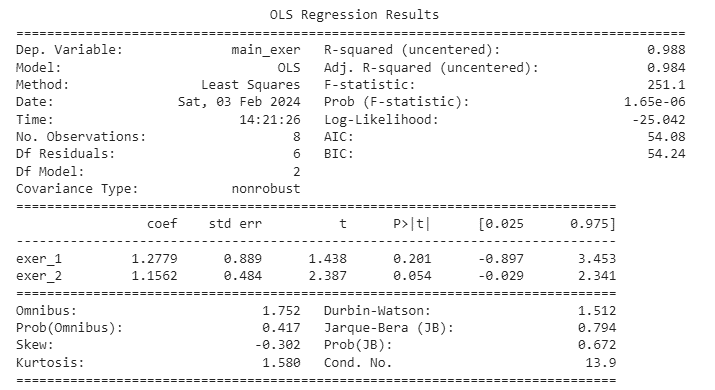


Рис.1 – Регрессионная модель построенная на основе метода наименьших квадратов.

После построения регрессионной модели может оказаться важным подтвердить соответствие модели и статистическую значимость оцененных параметров. Обычно используемые проверки согласия включают R-квадрат, анализ структуры остатков и проверку гипотез. Статистическую значимость можно проверить с помощью F-критерия общего соответствия, за которым следуют t-критерии отдельных параметров, сами модели можно сравнить с помощью информационных критериев и коэффициента подобия. Здесь можно видеть следующие оценки:

R-squared - коэффициент (аппроксимации) детерминации R², представляет собой долю дисперсии зависимой переменной, которая объяснена независимыми переменными в регрессионной модели. R-квадрат объясняет, в какой степени дисперсия одной переменной объясняет дисперсию второй переменной. Таким образом, если R-квадрат составляет 0,50, то примерно половину наблюдаемых изменений можно объяснить входными данными модели.

Adj. R-squared - скорректированный R-квадрат используют, чтобы сравнить качество регрессионных моделей, содержащих разное количество независимых переменных. При сравнении модели с большим количеством независимыми переменными и модели с меньшим, более высокий R-квадрат будет у модели с большим. Так как возможно переобучение и значение R-квадрата может быть неоправданно высокое. В свою очередь скорректированный R-квадрат используется для определения того, насколько надежна корреляция и насколько она определяется добавлением независимых переменных. Штраф за количество увеличивается при добавлении каждой новой переменной. Таким образом скорректированный R-квадрат адаптируется к количеству признаков в модели. R-квадрат, в отличие от скорректированного R-квадрата, используется для указания того, насколько продуктивно модель прогнозирует ответы при анализе новых наблюдений. Одно из ошибочных представлений о регрессионном анализе состоит в том, что низкое значение R-квадрата будет негативным фактором при оценке модели.

t-критерий Стьюдента оценивающий остатки, его большое значение говорит, что коэффициент описывает зависимость лучше (чем меньше стандартное отклонение, тем лучше измерен коэффициент). t-критерий позволяет оценить важность независимых переменных (предикторов), предполагая, что остатки модели нормально распределены около нуля. Если остатки не ведут себя таким образом, то это говорит о наличии некоторой нелинейности между переменными и о том, что их t-тесты не следует использовать для оценки важности отдельных предикторов.

Доверительные интервалы коэффициентов показывают уверенность в значении оцениваемого коэффициента, зависимость здесь обратная, при меньшем доверительном интервале увеличивается уверенность в значении оцениваемого коэффициента.

F-statistic – это критерий Фишера который можно использовать для определения общей значимости для нескольких независимых переменных. F-statistic используется для сравнения способности разных моделей объяснять дисперсию зависимой переменной. Если нет взаимосвязи между зависимой и независимыми переменными то F-statistic принимает значение, близкое к 1, иначе число больше. Значение F-statistic представляет собой отношение суммы квадратов средней регрессии к сумме квадратов средней ошибки. Его значение будет варьироваться от нуля до сколь угодно большого числа.

Prob (F-statistic) — это вероятность того, что нулевая гипотеза для полной модели верна (т. е. что все коэффициенты регрессии равны нулю). Например, если Prob(F-statistic) имеет значение 0,01, то есть 1 шанс из 100, что все параметры регрессии равны нулю. Значение будет означать, что по крайней мере некоторые из параметров регрессии отличны от нуля и что уравнение регрессии имеет некоторую достоверность при построении модели на конкретных данных (т. е. независимые переменные не являются случайными по отношению к зависимой переменной).

Durbin-Watson Statistic (тест на автокорреляцию) определяет вероятность того, что значения отклонения (ошибки) для регрессии имеют компонент авторегрессии первого порядка. В моделях регрессии предполагается, что отклонения ошибок некоррелированы. Небольшие значения статистики Durbin-Watson указывают на наличие автокорреляции.

Omnibus описывает нормальность распределения остатков при прогнозе (остатки необъяснённой дисперсии), используя в качестве измерений смещение (skew) (коэффициент симметрии распределения) и эксцесс (kurtosis) (мера остроты случайной величины в вершине графика). 0 означает полную нормальность.

Prob (Omnibus) - это статистический тест, измеряющий вероятность нормального распределения остатков. Значение 1 означает что распределение соответствует нормальному.

Skew (Смещение) - это мера симметрии распределения остатков, где 0 означает идеальную симметрию.

Kurtosis (Эксцесс) - измеряет остроту распределения остатков или его концентрацию около 0 на нормальной кривой. Более высокий эксцесс означает меньшее количество выбросов.

Jarque-Bera (JB) проверка нормальности распределения остатков по критерию Jarque-Bera и его расчетный уровень значимости Prob(JB). Критерий Харке-Бера является асимптотическим, расчетное значение имеет распределение хи-квадрат, поэтому данный критерий рекомендуют применять только для больших выборок

Cond. No Число обусловленности используется для проверки мультиколлинеарности (считается, что мультиколлинеарность вероятна, если значение Cond. No > 30)

Log-Likelihood или логарифмическое значение правдоподобия регрессионной модели является одним из способов измерить продуктивность модели. Чем выше значение логарифмической вероятности, тем лучше модель описывает набор данных. Значение логарифмического правдоподобия для данной модели может варьироваться от отрицательной бесконечности до положительной бесконечности. Фактическое значение логарифмического правдоподобия для данной модели в большинстве случаев не имеет смысла, но оно полезно для сравнения двух или более моделей. Можно заметить, что в более сложной модели произошло незначительное улучшение этой оценки.

Качество статистических методов из теории информации может быть измерено информационными критериями (IC). Таким образом, это относится к методам выбора модели, основанным на функциях правдоподобия. Самый низкий балл получает лучшая модель.

AIC (информационный критерий Акаике) предназначен для поиска модели, которая объясняет наибольшую изменчивость данных, и штрафует модели, использующие чрезмерное количество параметров. В нашем случае если модель имеет меньший AIC, значит лучший показатель соответствия набору данных. AIC не зависит от размера выборки, но хуже работает с малыми наборами данных. Основные ошибки AIC связаны с переобучением, и он представляет достаточно сложную модель, больше подходящую для оценки прогностических моделей.

BIC (байесовский информационный критерий) из байесовской вероятности чувствителен к размеру выборки, добавляет больший штраф, более простая модель и подходит для задач оценки моделей, объясняющих связь между зависимой и независимыми переменными.

Интерпретация этих диагностических тестов во многом зависит от предположений модели. Хотя анализ остатков может использоваться для признания модели недействительной, результаты t-теста или F-теста иногда труднее интерпретировать, если предположения модели нарушаются. Например, если член ошибки не имеет нормального распределения, в небольших выборках оцененные параметры не будут соответствовать нормальному распределению, что усложнит вывод. Однако при относительно больших выборках можно применить центральную предельную теорему, чтобы проверка гипотез могла продолжаться с использованием асимптотических приближений.

Модели линейной регрессии подбираются не только с использованием метода наименьших квадратов, но их также можно подбирать и другими способами, например, путем минимизации «несоответствия» какой-либо другой норме (как в случае с регрессией наименьших абсолютных отклонений) или гребневой регрессии (штраф за норму L2) и лассо регрессии (штраф за норму L1). Использование среднеквадратической ошибки (MSE) в качестве стоимости набора данных, который имеет много крупных выбросов, может привести к тому, что модель будет соответствовать выбросам больше, чем данным из доверительного интервала, из-за более высокой важности, придаваемой MSE величине ошибки. Таким образом, если в наборе данных много выбросов больших значений, следует использовать функции измерения ошибки более устойчивые к выбросам. Иногда метод наименьших квадратов можно использовать для подбора моделей, которые не являются линейными моделями. Таким образом, хотя термины «наименьшие квадраты» и «линейная модель» тесно связаны, они не являются синонимами.

Поэтому была построена регрессионная модель, построенная на основе взвешенного метода наименьших квадратов для сравнения с простым методом наименьших квадратов (рис.1)

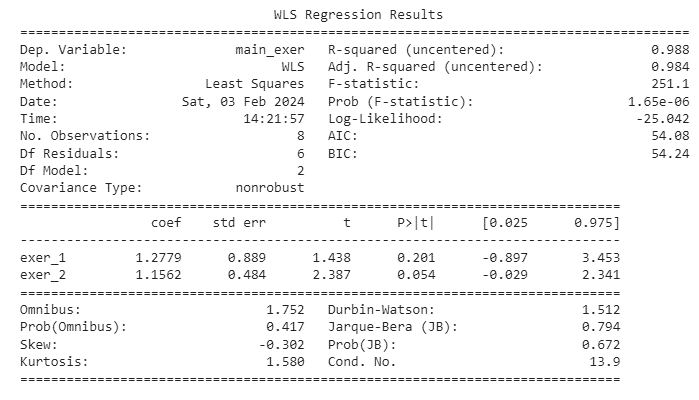


Рис.1 – Регрессионная модель, построенная на основе взвешенного метода наименьших квадратов.

Сравнивая эти две модели можно увидеть, что они и соответствующие оценки абсолютно идентичны. Как и предполагалось наличие гетероскедастичности оказало слабое влияние на продуктивность модели. Согласно проводившимся наблюдениям, действительно упражнения в ходе спортивной подготовки связаны друг с другом, но несмотря на это линейная регрессия методом наименьших квадратов достаточно хорошо справляется с моделированием учебно-тренировочного процесса. В исследуемом случае возникает формула линейной зависимости:

1.2779\***1упражнение** + 1.1562\***2упражнение**=**целевое упражнение**

Зная данную математическую модель, мы можем понять, как воздействовать на целевое упражнение и более корректно строить учебно-тренировочный процесс. Первое упражнение имеет чуть больший коэффициент и можно предположить, что именно оно больше влияет на целевое упражнение, однако по оценке распределения остатков t-критерием Стьюдента, доверительным интервалам и меньшей среднеквадратичной ошибки можно сказать, что второе упражнение в большей степени отражает характер целевого упражнения. Учитывая это можно ставить тренировочную цель и подбирать нагрузку исходя из имеющейся модели.

После получения понимания причинно-следственных связей выбранных упражнений и получения математической модели комплекса этих упражнений при необходимости можно перейти к прогнозированию. Для этого новые данные (значения независимых переменных) вводятся в модель и получается прогноз значения зависимой переменной (рис.1).

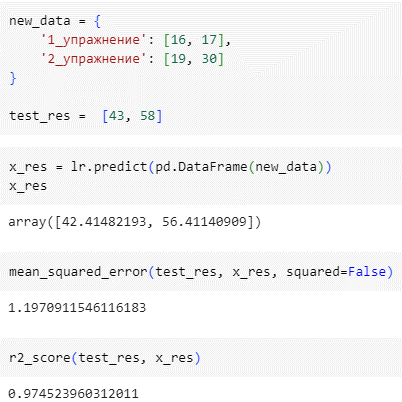


Рис.1 – Прогноз и результаты оценки продуктивности модели

В результате можно убедиться, что при введении ранее неизвестных модели данных её продуктивность изменилась не сильно. Коэффициент детерминации ранее был 0,988, стал 0,975, ошибка ранее была 0.889, а стала 1,197. Таким образом модель оказалась достаточно надёжной и может приносить пользу при использовании. Конечно в условиях спортивной подготовки модели должны строиться и изменяться постоянно, согласно с актуальными требованиями к учебно-тренировочному процессу.

В данном случае были использованы средства и методы построения модели, относящиеся к так называемому Classical Machine Learning («Классическое машинное обучение»). Однако наиболее популярным сейчас является Deep Learning («Глубокое обучение») использующий нейросети. Поэтому имеющийся набор данных был исследован с помощью, созданной простейшей нейросети с одним слоем, так задача нахождения линейной модели не требует большего (рис.1)

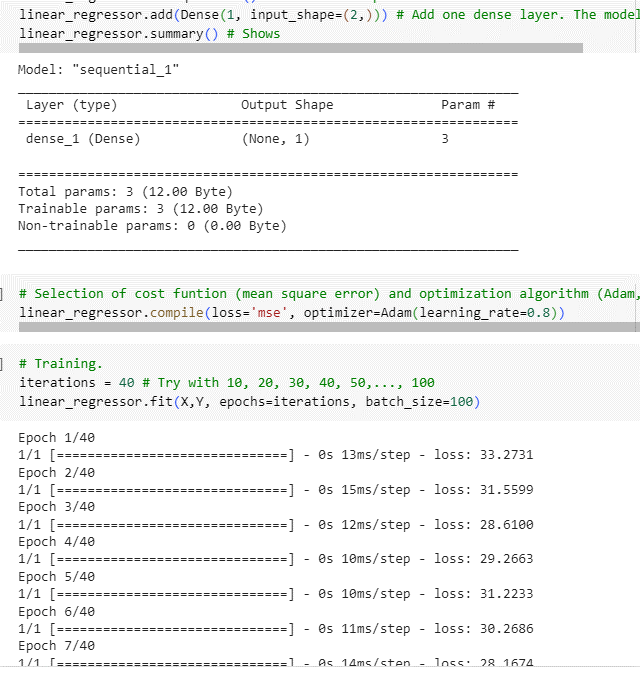


Рис.1 – Создание и обучение нейросети для создания линейной регрессионной модели

Однако обратившись к оценке модели можно увидеть, что значения коэффициента детерминации и среднеквадратичной ошибки хуже чем для метода наименьших квадратов (рис.1)

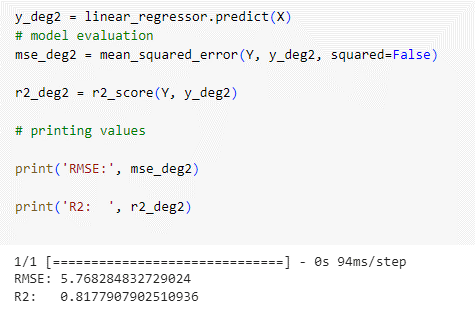


Рис.1 – Определение продуктивности нейросети для создания линейной регрессионной модели

Таким образом можно утверждать, что в случае относительно простых задач, которыми изобилует сфера физической культуры и спорта более простые решения могут быть продуктивнее сложных и оперативная оценка, и оперативное же планирование нагрузки в процессе спортивной подготовки вполне реализуемо на основании простых методов, что делает их применимыми в ежедневной спортивной практике.

Исследование с помощью регрессионных моделей может быть достаточно сложным. Например, даже простые регрессионные модели, используемые совместно в исследовании спортивной деятельности способны обеспечить более глубокое понимание спортивной деятельности там где на первый взгляд всё понятно и очевидно. Например при исследовании набора данных «UFC-Fight historical data from 1993 to 2021»[[254]](#footnote-254) где были собраны данные о профессиональных поединках в ММА в размере 6012 записей по 160 признакам были проведены различные оценки, но продуктивность их была невысокой несмотря на то, что использовались в том числе и методы глубокого обучения (нейросети). Простые парные корреляции показали очевидное (рис.1)

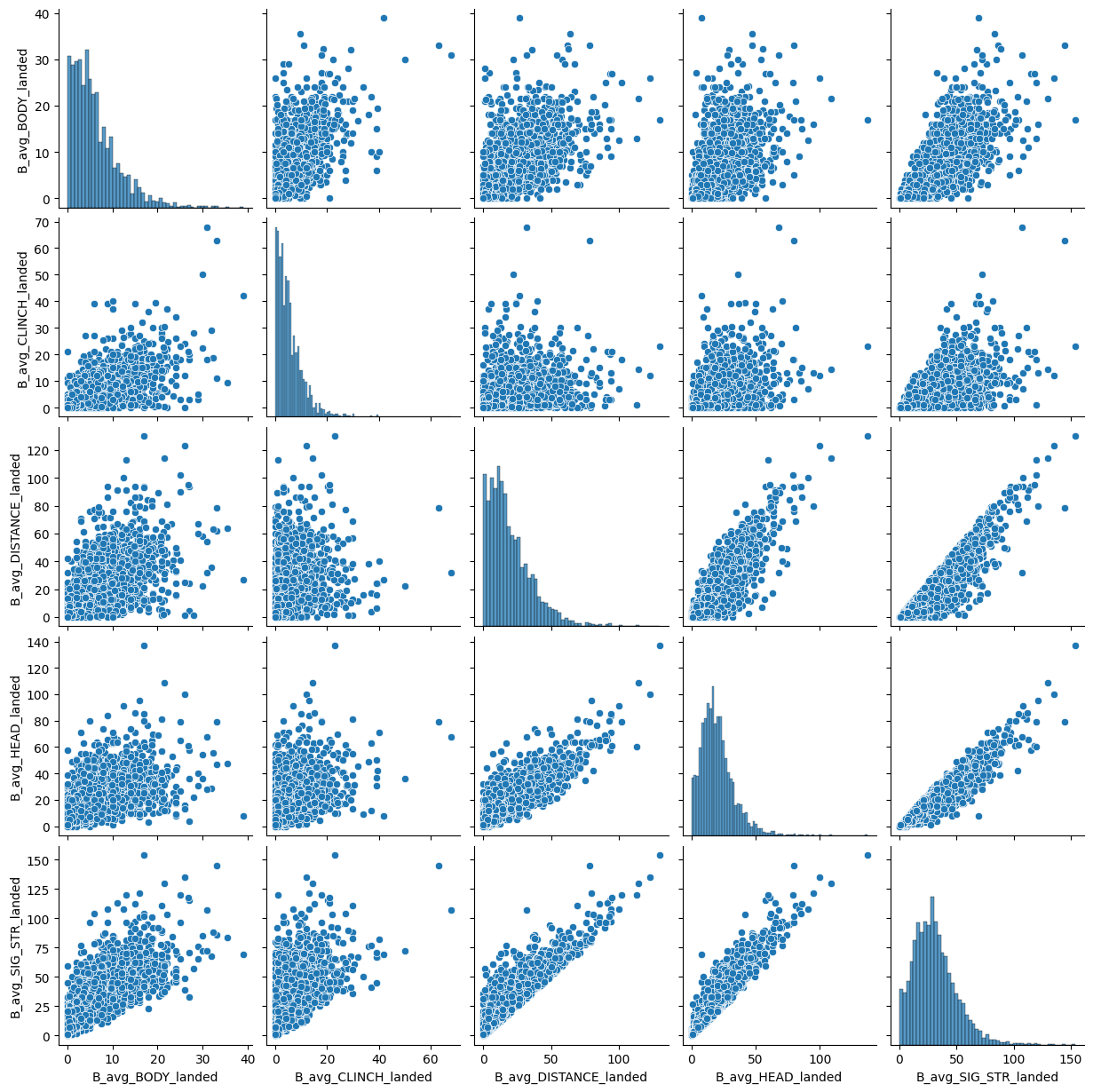


Рис.1 –Парные корреляции различных дошедших до цели атак ударом в ММА

Действительно удары, которые судьи посчитали «значительными», а не просто дошедшими до цели должны быть статистически связанными с количеством ударов, дошедших до цели, но это с одной стороны тривиально, а с другой модель будет явно некорректной (рис.1)

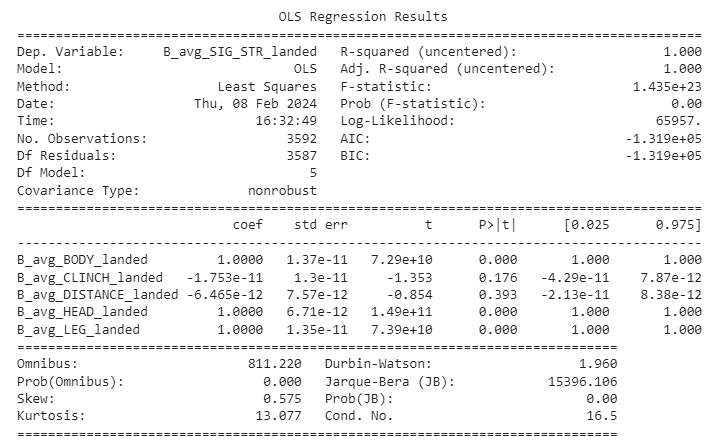


Рис.1 – Некорректно построенная модель зависимости «значительных» ударов от количества ударов дошедших до цели в разных условиях

Однако, тут можно сделать шаг в глубь данной информации и узнать, например, какую значимость имеет коэффициент эффективности ударов для различных целей – голова и туловище (рис.1)

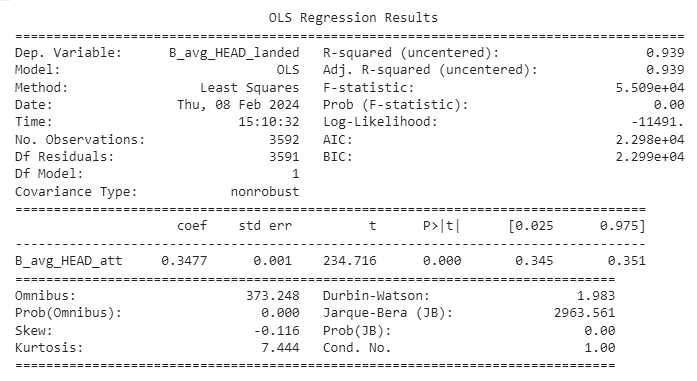




Рис.1 Модель зависимости ударов, достигших цели от общего количества нанесённых ударов (вверху - в голову, внизу – в туловище)

Сравнивая полученные модели можно заметить, что с одной стороны коэффициент логарифмического правдоподобия и оба информационных критерия свидетельствуют о том, что модель эффективности ударов по туловищу более продуктивна. С другой стороны, коэффициент при независимой также больше в случае ударов по туловищу почти в два раза. Это говорит о том, что удары по туловищу чаще достигают цели, попадания по туловища и больше связаны с общим количеством попыток их нанесения сем удары в область головы. Причины этого могут быть разнообразны. Возможно играет роль дистанция, так как удары по туловища чаще производятся с более близкой дистанции и соответственно это оставляет меньше времени для организации защиты, чем при ударах в область головы, дистанция для которых чаще несколько больше. Возможно большая подвижность головы относительно подвижности туловища обеспечивает лучшую защиту чем менее подвижное туловище. Также имеется вероятность, что психологически удары в область головы воспринимаются как более «опасные», тем самым спортсмены могут просто уделять больше внимания защите головы, чем защите туловища. Поэтому необходимо проверить не просто отношение количества ударов, достигших цели к количеству общего числа нанесённых ударов, а каково влияние количества нанесённых ударов вне зависимости от их успешности и места нанесения на количество ударов определённых как «значимые». Построенная модель отражена на рис.1

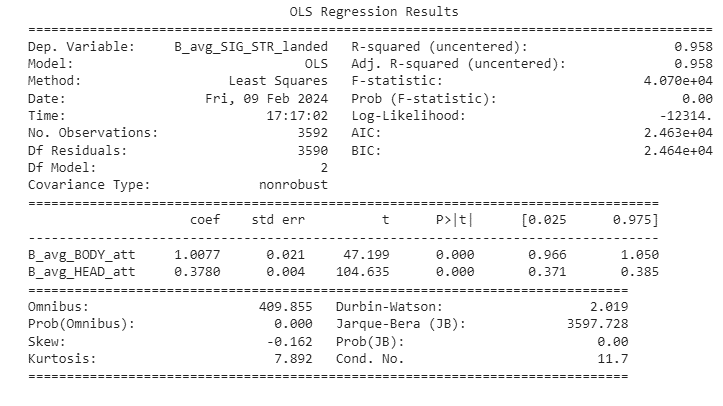


Рис.1 - Модель зависимости значимых ударов от общего количества нанесённых ударов в область головы и в туловище

Согласно оценки построенной модели можно заметить, что её продуктивность достаточно высока. При этом общее количество ударов по туловищу более значимо для количества «значимых» ударов, и эта значимость приблизительно в три раза выше чем подобная же оценка для ударов, нанесённых в область головы. Отсюда можно сделать несколько умозаключений оказывающих влияние как на учебно-тренировочный процесс спортсменов данного вида спорта, так и для тактики, избираемой на конкретную схватку. Например, при организации учебно-тренировочного процесса можно делать ставку на увеличение общего количества ударов по туловищу, а для ударов в область головы избрать средства и методы для повышения точности этих атак.

Обратившись ко всему изложенному в этом разделе работы материалу можно сделать несколько выводов:

1. Регрессионные модели хорошо подходят для прогнозирования и объяснения закономерностей спортивной подготовки и соревновательной деятельности, когда речь идёт о зависимой переменной представленной в виде числовой последовательности.
2. Регрессионные модели достаточно продуктивны и одновременно просты в использовании, оценке и позволяют получить математические модели высокой наглядности в следствии связанности их с традиционными для сферы спорта коэффициентами эффективности.
3. При возможности желательно выбирать самую простую модель в следствии её большей устойчивости. Так если линейная модель не удовлетворяет требованиям точности можно переходить к полиномам второй степени, а если и в этом случае продуктивность модели недостаточна, то тогда можно увеличивать степень полинома (не допуская переобучения модели). Использование же сложных методов сразу, как в примере с построением, обучением и применением нейросети, не обязательно даст лучший или сопоставимый результат.
4. Регрессионные модели желательно использовать в комплексе переходя от одной гипотезе к другой последовательно увеличивая качество получаемых знаний из собранных данных.
5. Применение технологий искусственного интеллекта позволяет создавать, применять и оценивать регрессионные модели быстро и эффективно, что позволяет использовать их даже при работе со спортсменами массовых разрядов в ходе ежедневной спортивной практики.

# 4.2. Логистическая регрессия

Модель логистической регрессии — это метод классификации обучения с учителем, который используется для прогнозирования наблюдений для дискретного набора признаков[[255]](#footnote-255). На практике он используется для классификации сделанных наблюдений по различным категориям. Следовательно, результат имеет ограниченный дискретный характер. Логистическая регрессия - это один из самых простых, понятных и универсальных алгоритмов классификации, который используется для решения задач классификации. Алгоритм логистической регрессии работает путем реализации линейного уравнения с независимыми или независимыми переменными для прогнозирования значения ответа, как и в рассматриваемом ранее случае линейной и полиномиальной регрессионных моделей. Что отличает модель логистической регрессии от модели линейной регрессии, так это то, что переменная результата в логистической регрессии является бинарной или дихотомической. Эта разница между логистической и линейной регрессией отражается как в форме модели, так и в ее предположениях. Как только эта разница учтена, методы, используемые при анализе с использованием логистической регрессии, более или менее следуют тем же общим принципам, которые используются в линейной регрессии.

Поэтому данную модель используют для бинарной классификации. В сфере физической культурны спорта это может быть результатом спортивного состязания в качественном выражении – победа или поражение. Подобным же образом наличие или отсутствие определённого состояния логистическая регрессия широко используется как в биологии, медицине, так и в социальных исследованиях[[256]](#footnote-256)[[257]](#footnote-257). Формально в бинарной логистической регрессии существует одна бинарная зависимая переменная, кодируемая индикаторной переменной, где два значения помечены «0» и «1», в то время как каждая из независимых переменных может быть двоичной переменной (два класса, которые кодирует индикаторная переменная) или непрерывная переменная (принимает любое действительное значение). Соответствующая вероятность значения, помеченного «1», может варьироваться от 0 до 1, отсюда и подобная маркировка[[258]](#footnote-258). Функция, которая преобразует логарифм шансов в вероятность – это и есть логистическая функция, отсюда и происходит название алгоритма. Единица измерения шкалы логарифмических шансов называется логитом, от логистической единицы. Зависимость вероятности наступления события от уравнения регрессии геометрически представляет собой логистическую кривую (рис.1)

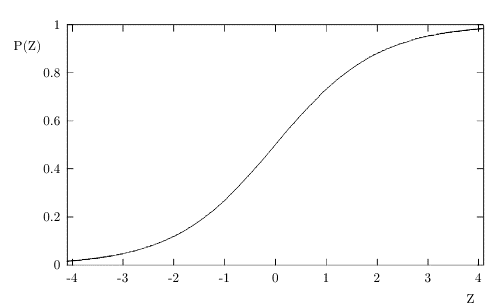


Рис.1 – Логистическая кривая (сигмоида) зависимости вероятности наступления события (Р) от уравнения регрессии (Z)[[259]](#footnote-259)

Описываемая зависимость описывается формулой: 

Таким образом получается возможность разделить набор данных по определённому признаку с достаточно высокой точностью основываясь на вероятности его появления.

Поэтому модель логистической регрессии просто моделирует вероятность результата в соответствии со значениями входных данных, но не выполняет статистическую классификацию, хотя ее можно использовать для создания классификатора выбрав пороговое значение и классифицируя входные данные с вероятностью. больше порогового значения для одного класса и ниже порогового значения для другого. Именно это является распространенным способом создания бинарного классификатора.

Также могут использоваться аналогичные линейные модели для двоичных переменных с другой сигмовидной функцией вместо логистической функции (для преобразования линейной комбинации в вероятность), например пробит-модель. Определяющей характеристикой логистической модели является то, что увеличение одной из независимых переменных масштабирует шансы данного результата с постоянной скоростью, при этом каждая независимая переменная имеет свой собственный параметр для двоичной зависимой переменной. Логистическая функция является естественным параметром распределения Бернулли и в этом смысле является «простейшим» способом преобразования действительного числа в вероятность. В частности, она максимизирует энтропию (минимизирует добавленную информацию).

Для исследования был избран набор данных по соревновательной компьютерной игре League of Legends отражающий параметры первых 10 минут игры[[260]](#footnote-260). Соответственно формулируется гипотеза о возможности успешного прогноза победителя в матче двух команд. Для каждой из команд в наборе данных присутствуют количество уничтоженных персонажей соперника и собственные потери, значения показателей взаимопомощи игроков для каждой из команд, количество уничтоженных неиграбельных персонажей (NPC) в результате чего нанёсший последнюю атаку получает внутриигровую валюту (Minions, Jungle Minions, Elite Monsters) или другие преимущества для команды (Dragon, Herald), а также установка своих и уничтожение вражеских объектов предназначенных для увеличения видимой части карты (Ward) и обеспечивающие оборонные задачи (Tower). Всего набор данных содержит записи о 9879 матчах игроков с высоким рейтингом (от Diamond I до Master). Для этого набора данных определены 40 признаков по 19 для каждой команды.

Первоначально был проведён корреляционный анализ для определения статистических связей между изучаемыми признаками. Так как количество признаков велико (40), то вместо приведения таблицы здесь приводится визуальное отображение корреляционной матрицы в виде тепловой карты (рис.1). На данном рисунке можно видеть, что присутствуют признаки как с очень большой положительной, так и сочень большой отрицательной корреляцией и это закономерно так как некоторые признаки являются идеальными противоположностями друг друга например «кто пролил первую кровь» (blueFirstBlood), количество уничтоженных персонажей соперника должно быть равно количеству персонажей, которые соперник потерял (blueDeaths и blueKills).

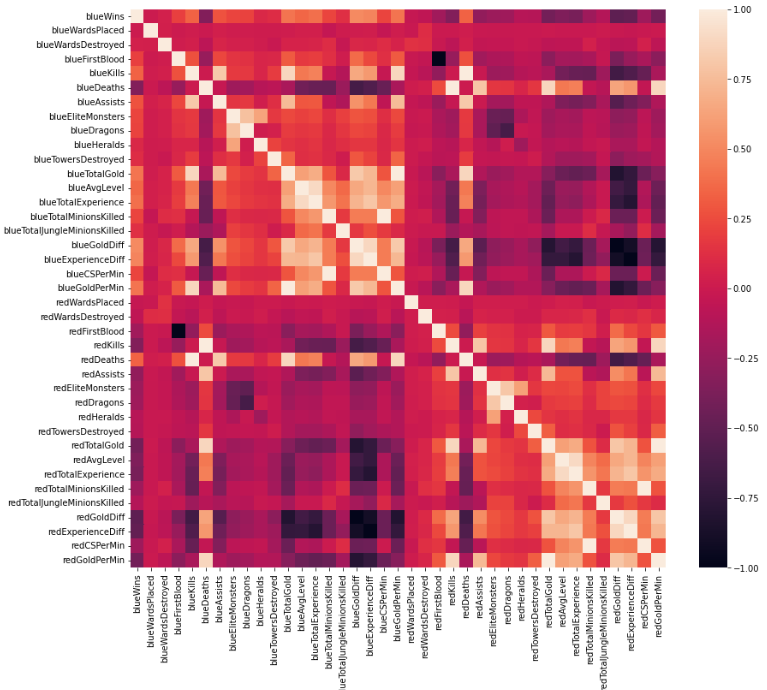


Рис.1 – Корреляционная матрица полного набора данных

Даже при простом визуальном анализе легко можно заметить, что в корреляционной матрице есть две очень похожие друг на друга участка практически дублирующие друг друга. Поэтому некоторое количество признаков, которые являются зеркальным отображением друг друга могут быть удалены из набора данных без потери точности модели. После удаления был получен скорректированный набор данных, который содержит уже двадцать три признака и согласно правилам подготовки, данных должен позволять строить более продуктивные модели. Визуальное представление скорректированного набора данных можно увидеть на рис.1

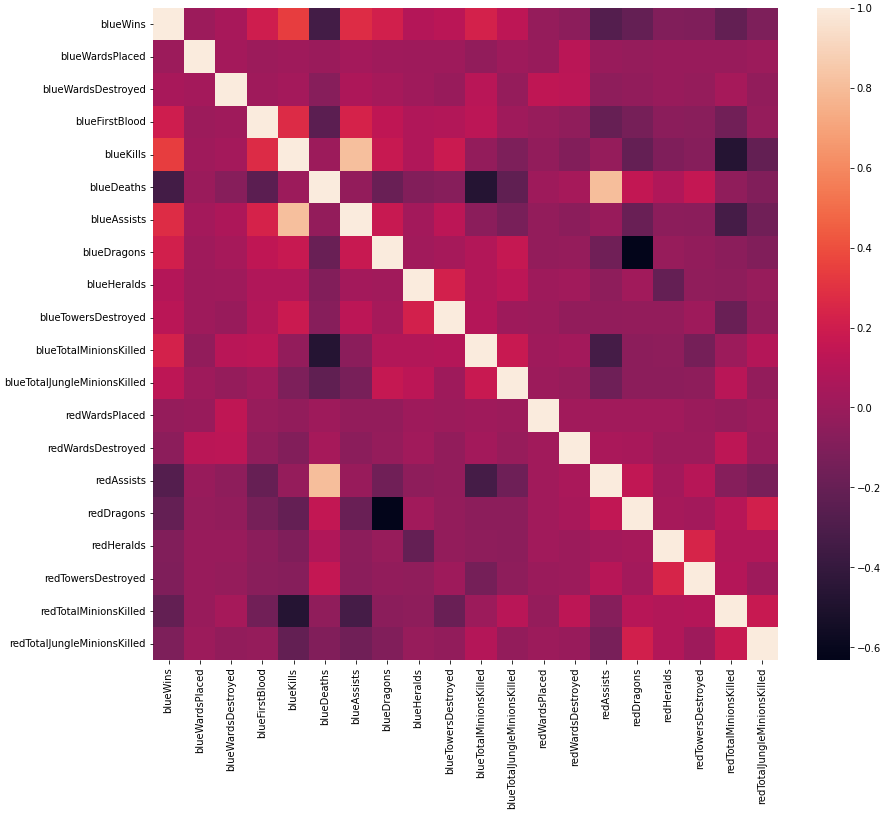


Рис.1 – Корреляционная матрица скорректированного набора данных

Анализируя полученную матрицу можно обратить внимание на то, что присутствует завышенная корреляция между показателями Assists и Deaths –Kills, но избавляться от части этих данных было неразумно, так как можно было исказить важную информацию и основываясь на соревновательной практике было принято решение эти признаки в скорректированном наборе данных.

Создав модель биномиальной классификации с использованием логистической регрессии на скорректированном наборе данных прежде всего необходимо оценить продуктивность созданной модели. Для подобной оценки было использовано несколько методов. В начале результативность была визуализирована с помощью матрицы ошибок и кривой AUC-ROC (рис.1)

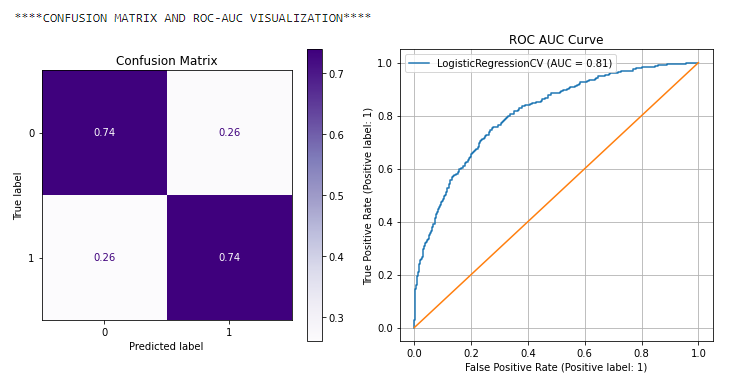


Рис.1 – Confusion matrix (Матрица ошибок) и кривая AUC-ROC для модели логистической регрессии на скорректированном наборе данных.

Confusion matrix (матрица ошибок) представляет собой таблицу, которая позволяет визуализировать производительность алгоритма[[261]](#footnote-261). Каждая строка матрицы представляет экземпляры реального класса, а каждый столбец представляет экземпляры прогнозируемого класса, или наоборот[[262]](#footnote-262). Название отражает факт того, что оно позволяет легко увидеть, как система может ошибиться при выборе класса (т. е. обычно ошибочно маркирует один как другой). Это особый вид таблицы сопряженности с двумя измерениями («фактическое» и «прогнозируемое») и идентичными наборами «классов» в обоих измерениях (каждая комбинация измерения и класса является переменной в таблице сопряженности). Согласно матрицы определяется несколько важных значений:

Количество положительных случаев (condition positive, P).

Количество отрицательных случаев (condition negative, N).

Истинно положительный (true positive, TP) результат теста, который корректно указывает на наличие состояния или характеристики.

Истинно отрицательный (true negative, TN) Результат теста, который корректно указывает на отсутствие состояния или характеристики.

Ложноположительный результат (false positive, FP), ошибка I рода, который ошибочно указывает на наличие определенного условия или атрибута.

Ложноотрицательный результат (false negative, FN), ошибка II рода, который ошибочно указывает на отсутствие определенного условия или атрибута.

Чувствительность, отзыв, частота совпадений или истинно положительный результат (sensitivity, recall, hit rate, true positive rate TPR)



Специфичность, селективность или истинно отрицательный показатель (specificity, selectivity, true negative rate, TNR)



Precision или положительное прогностическое значение (precision, positive predictive value, PPV)



Отрицательное прогностическое значение (negative predictive value, NPV)



Частота ложноотрицательных результатов (miss rate or false negative rate, FNR)



Частота ложноположительных результатов (fall-out or false positive rate, FPR)



Частота ложных обнаружений (false discovery rate, FDR)



Частота ложных пропусков (false omission rate, FOR)



Отношение правдоподобия положительного результата (positive likelihood ratio, LR+)



Отношение правдоподобия отрицательного результата (negative likelihood ratio, LR-)



Оценка F1 среднее гармоническое значение положительного прогностического значения и истинно положительного результата:



В исследуемом случае значения получились симметричные, что бывает не всегда. Значения нормированы и истинно положительный и истинно отрицательные результаты имеют значения 0,74 (иными словами в 74% случаев алгоритм точно указал победу или поражение), а ложно положительный и ложноотрицательный результат – 0,26

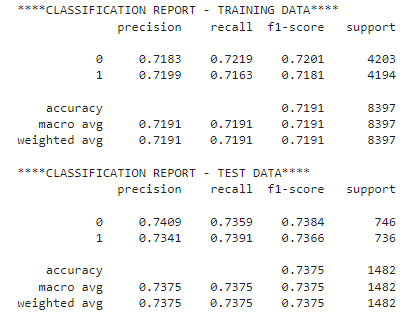
Кривая AUC-ROC (Area Under Curve - Receiver Operating Characteristics) или в прямом переводе это звучит как площадь под кривой рабочей характеристики приемника представляет собой графическое представление производительности модели бинарной классификации при различных пороговых значениях классификации. Название пришло из времён второй мировой войны и относилось к качеству распознавания самолётов противника при радиолокации[[263]](#footnote-263).

ROC означает «Рабочие характеристики приемника», а кривая ROC — это графическое представление эффективности модели двоичной классификации. Он отображает соотношение истинно положительного результата (TPR) и уровня ложноположительного результата (FPR) при различных порогах классификации.

AUC представляет собой площадь под кривой ROC. Она измеряет общую производительность модели бинарной классификации. Поскольку и истинно положительный результат (TPR), и ложноположительный результат (FPR) нормированы и варьируются от 0 до 1, площадь всегда будет находиться в диапазоне от 0 до 1. Диагональ, то есть значение 0,5 означает случайный выбор, с которым и идёт сравнение классификатора. Большее значение AUC означает лучшую производительность модели. Главная цель модели — максимизировать эту область, чтобы иметь самый высокий TPR и самый низкий FPR при заданном пороге. AUC измеряет вероятность того, что модель присвоит случайно выбранному положительному экземпляру более высокую прогнозируемую вероятность по сравнению со случайно выбранным отрицательным экземпляром. В данном случае показатель достаточно высок - 0,81.

Другая система оценок продуктивности представлена в таб.1

Таблица 1. – Оценка продуктивности модели логистической регрессии на скорректированном наборе данных



В данной таблице представлены результаты оценки как на тренировочном (на котором было произведено обучение) и на тестовом (новом неизвестном при обучении модели) наборах данных. В таблице представлены следующие оценки:

Precision - это количество корректно предсказанных положительных результатов (истинно положительный результат), разделенное на все предсказанные положительные результаты (истинный положительный результат + ложный положительный результат).



Чем precision выше, тем лучше модель машинного обучения минимизирует количество ложных срабатываний. Когда точность высока, это означает, что большинство положительных прогнозов модели являются хорошо предсказанными положительными результатами.

Recall - это количество корректно предсказанных положительных результатов (истинно положительный результат), разделенное на общее количество положительных результатов (истинный положительный результат + ложный отрицательный результат).



Чем recall выше, тем лучше модель машинного обучения максимизирует количество истинных положительных результатов.

Кроме описанного ранее критерия F1 (среднее гармоническое значение положительного прогностического значения и истинно положительного результата) здесь представлена и другая интегральная оценка. Оценка F1 — хороший способ сравнить производительность нескольких классификаторов. При выборе между несколькими моделями, каждая из которых имеет разные значения precision и/или recall, ее можно использовать, чтобы определить, какая из моделей продуктивнее. Тем не менее, оценка F1 не идеальна и подходит не для всех задач классификации. Она полностью игнорирует истинные отрицательные значения в своих расчетах, поэтому это создаёт несбалансированность. Оценка F1 придает равную важность precision и recall.

Accuracy — это то, насколько близок данный набор значений (наблюдений) к их истинному значению, а precision — насколько близки значения друг к другу. Обе эти оценки переводятся как «точность», поэтому они и приводятся без перевода. Другими словами, precision — это описание случайных ошибок, мера статистической изменчивости. Accuracy - это описание только систематических ошибок, мера статистической погрешности данной меры центральной тенденции; низкая accuracy приводит к отличию результата от истинного значения; ISO называет это правдивостью (trueness).



Accuracy означает, насколько близко значение к истинному или ожидаемому значению. С другой стороны, precision означает, насколько последовательным или воспроизводимым является измерение. Другими словами, точность — это правильность, а точность — это последовательность. Проще говоря, при наличии статистической выборки или набора точек данных, полученных в результате повторных измерений одной и той же величины, можно сказать, что accuracy для выборки высока, если среднее значение близко к истинному значению измеряемой величины, в то время как набор может можно сказать, что значение precision высоко, если стандартное отклонение относительно невелико.

Кроме прочего таблица содержит среднее значения. Macro-avg — это среднее значение для всех классов (точность класса 0 + точность класса 1)/2. В то время как weighted avg (взвешенное среднее) - это общее количество TP (истинное положительное значение для всех классов)/общее количество объектов во всех классах. пример на основе вашей модели. В исследуемом случае все значения находятся на уровне достаточным, чтобы признать модель рабочей и использовать её в практике. Коэффициенты полученной модели можно увидеть на рис.1

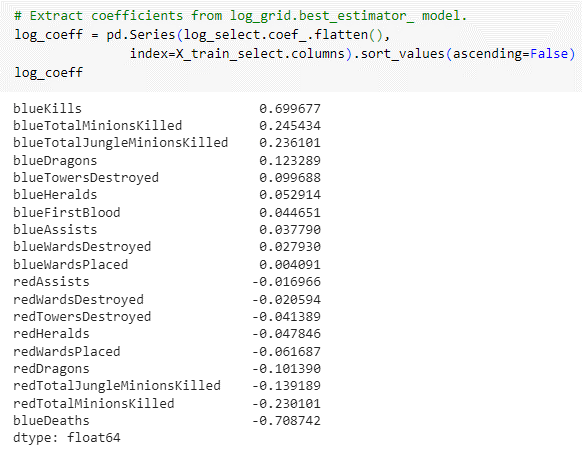


Рис.1 – Коэффициенты полученной модели логистической регрессии

Как можно заметить наивысшее значение имеют количество уничтоженных юнитов соперника и потерянных собственных и это очевидно. Менее очевидным фактом является почти сопоставимый суммарный вклад от уничтожения неиграбельных персонажей NPC. В случае этой игры это миньоны (Minions) и джунглевые миньоны. Визуально вклад значений коэффициентов в достижение победы отображён на рис.1

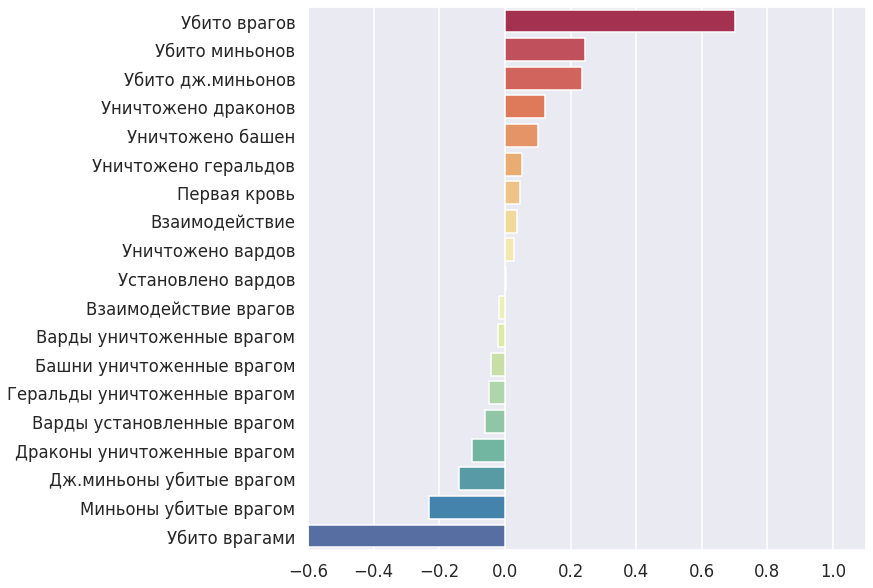


Рис.1 – Вклад отдельных параметров деятельности в достижение победы согласно коэффициентам логистической регрессии

Даже при визуальном анализе можно убедиться, что суммарный вклад от уничтожения NPC очень высок и это может оказать влияние на выбор тактики ведения спортивной борьбы и возможные пути обеспечения преимущества на первых десяти минутах игры. Так как уничтожение NPC приносит внутриигровую валюту, то это подчёркивает значимость экономического фактора на начальном этапе игры.

Следуя используемого ранее подхода было использовано глубокое обучение (deep learning). Так как данные имеют более сложный по сравнению линейной регрессии вид, то и созданная нейросеть имела более сложную структуру (рис.1)

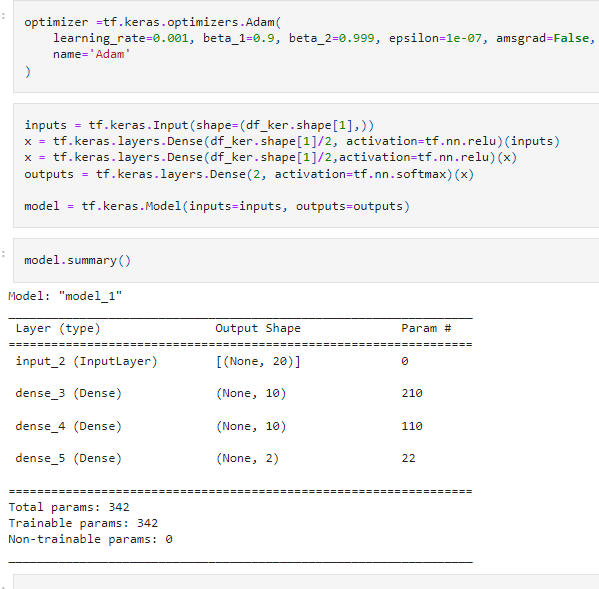
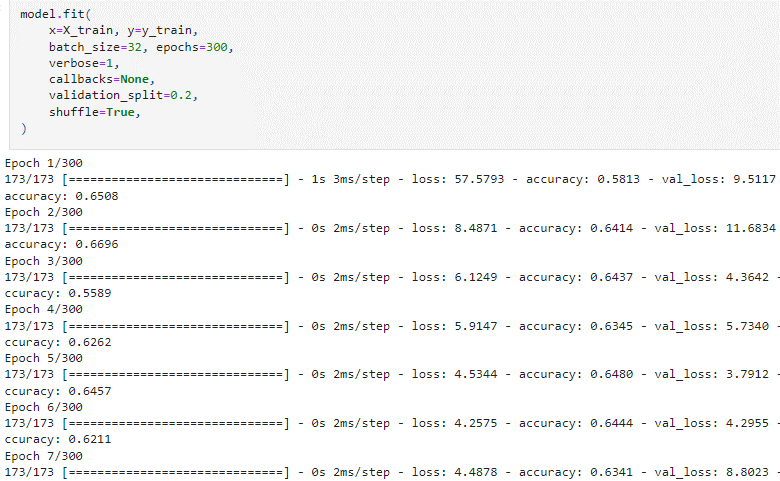


Рис.1 – Организация и структура нейросети для создания модели логистической регрессии на скорректированном наборе данных

В этом случае было использовано больше слоёв и соответственно больше параметров. В качестве оптимизатора был использован Оптимизатор Adam — это метод стохастического градиентного спуска, основанный на адаптивной оценке моментов первого и второго порядка. Считают что, этот метод «вычислительно эффективен, требует мало памяти, инвариантен к диагональному масштабированию градиентов и хорошо подходит для задач, больших по объему данных/параметров»[[264]](#footnote-264). Результаты обучения модели представлены на рис.1



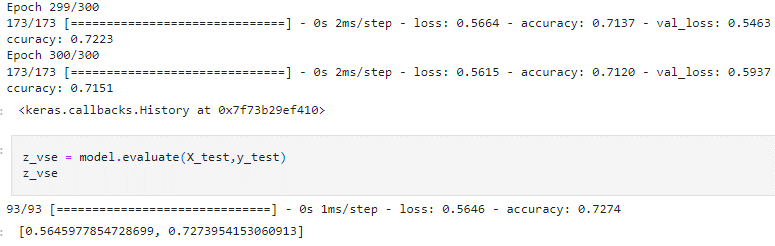


Рис.1 - Результаты обучения регрессионной модели с использованием нейросети

Для обучения использовалось большее количество слоёв и 300 эпох обучения. В результате получена модель, не превосходящая обычную логистическую регрессию. С другой стороны оценки расходятся только, начиная с тысячной и это показывает, что при увеличении количества данных и сложности задачи значимость глубокого обучения будет возрастать.

В результате исследования можно придти к следующим выводам:

Регрессионный анализ в основном используется для двух концептуально различных целей. Во-первых, регрессионный анализ широко используется для прогнозирования. Во-вторых, регрессионный анализ можно использовать для вывода причинно-следственных связей между независимыми и зависимыми переменными.

Важно отметить, что регрессии сами по себе выявляют только отношения между зависимой переменной и набором независимых переменных в фиксированном наборе данных. Чтобы использовать регрессии для прогнозирования или вывода причинно-следственных связей, соответственно, исследователь должен тщательно обосновать, почему существующие взаимосвязи обладают предсказательной силой для нового контекста или почему связь между двумя переменными имеет причинно-следственную интерпретацию.

На тех объёмах данных, сложности их и решаемых задач в сфере спорта эффективность простых решений выше и именно они могут быть встроены в ежедневный процесс спортивной подготовки.

# 4.3. Ансамблевые методы улучшения производительности математических и компьютерных моделей в сфере спорта

Предварительно было выяснено, что более сложные модели не могут обеспечить более высокую производительность на тех объёмах и сложности данных, которые сейчас распространены в сфере физической культуры и спорта. Закономерно возникает вопрос о возможности не использовать более сложный метод, а создать для решения задачи комплекс более простых методов.

В статистическом анализе и машинном обучении ансамблевые методы используют несколько алгоритмов для получения более высокой производительности модели, чем можно было бы получить с помощью любого из составляющих её алгоритмов по отдельности[[265]](#footnote-265)[[266]](#footnote-266).

Оценка прогноза ансамбля обычно требует больше вычислений, чем оценка прогноза одной модели. В каком-то смысле ансамблевое обучение можно рассматривать как способ компенсировать плохие алгоритмы обучения за счет выполнения большого количества дополнительных вычислений. Ансамблевая система может быть более эффективной в улучшении общей точности при том же увеличении вычислительных ресурсов, ресурсов хранения или связи за счет использования этого увеличения для двух или более методов, чем было бы улучшено за счет увеличения использования ресурсов для одного метода. Быстрые алгоритмы, такие как деревья решений, обычно используются в ансамблевых методах (например, случайные леса), хотя и более медленные алгоритмы также могут выиграть от ансамблевых методов. В начале было рассмотрено применение алгоритма решающего дерева. Дерево решений — это иерархическая модель, применяемая поддержки принятия решений, в которой используется древовидная модель, включающая решения и их возможные последствия[[267]](#footnote-267). Деревья решений обычно используются в исследованиях операций, чтобы помочь определить стратегию, которая с наибольшей вероятностью достигнет цели, а также являются популярным инструментом в машинном обучении.

Дерево решений — это классификатор, выраженный как рекурсивное разделение данных. Дерево решений состоит из узлов, которые образуют корневое дерево, то есть это направленный граф с узлом, называемым «корнем», который не имеет входящих ребер. Все остальные узлы имеют ровно одно входящее ребро. Узел с исходящими ребрами называется внутренним узлом. Все остальные узлы называются листьями (также известными как терминальные узлы или узлы принятия решений). В дереве решений каждый внутренний узел разбивает пространство экземпляров на два или более подпространства в соответствии с определенной дискретной функцией значений входных атрибутов. Для примера был рассмотрен случай с набором данных о боях ММА (рис.1)

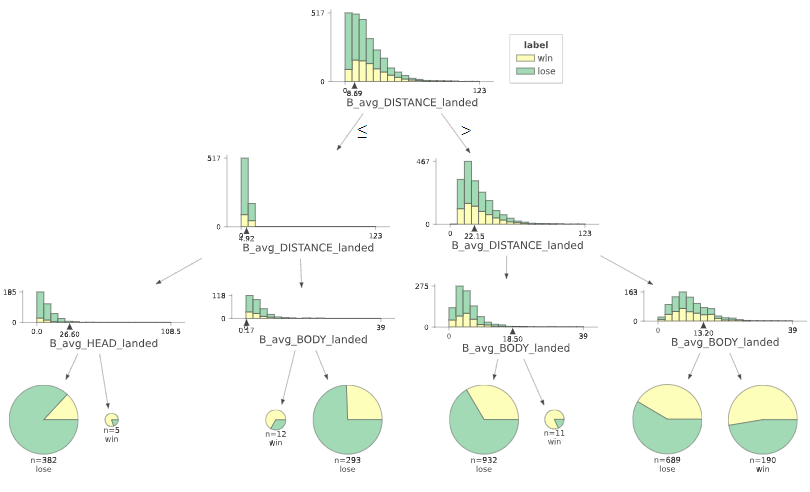


Рис.1 – Классификация алгоритмом Decision Tree результатов матча в Ultimate Fighting Championship (UFC)

В итоге была получена структуру, подобная блок-схеме, в которой каждый внутренний узел представляет собой «тест» на атрибут (была зафиксирована победа или поражение в этом поединке), каждая ветвь представляет результат теста, а каждый листовой узел представляет собой метку класса (решение, принятое после вычисления всех атрибутов). Пути от корня к листу представляют правила классификации. В случае логистической регрессии построенные модели для Ultimate Fighting Championship (UFC) были непродуктивны и их не было смысла использовать для принятия решений. В случае же применения Decision Tree можно получить некоторые интуиции, а именно если победу прогнозировать оказалось сложнее, то определить минимальное количество точных ударов в область головы и туловища соперника меньше которых вероятность проигрыша составит более 75% с помощью этого алгоритма оказалась возможной в отличии от регрессионных моделей.

Для более глубокого анализа был использован набор данных League of Legends Diamond Ranked Games (10 min) на котором была ранее построена модель логистической регрессии. Результата оценки на скорректированном наборе данных представлены на рис.1.

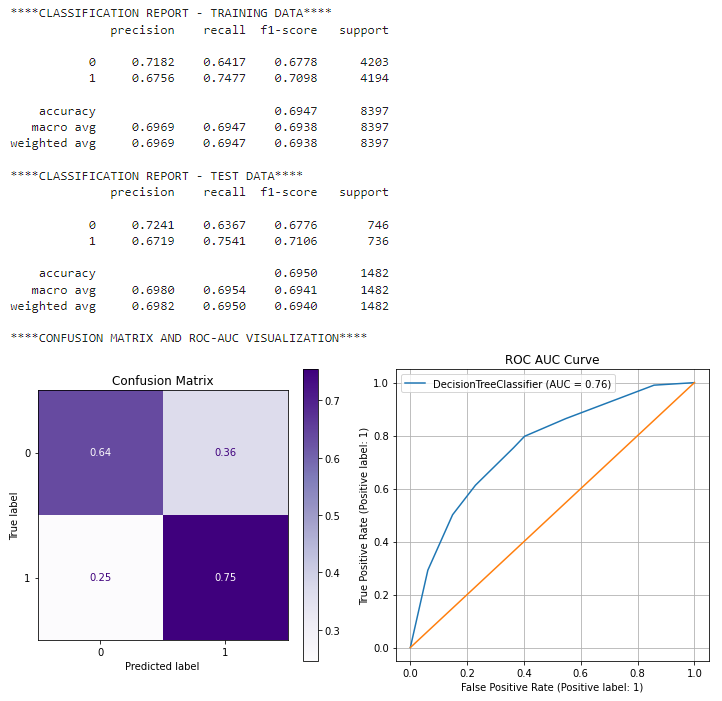


Рис.1 – Продуктивность алгоритма Decision Tree на скорректированном наборе данных League of Legends

Можно обратить внимание, что практически по всем критериям модель логистической регрессии оказалась продуктивнее. Однако Decision Tree не обладает теми же ограничениями как логистическая регрессия и можно обучить модель на полном наборе данных (рис.1)

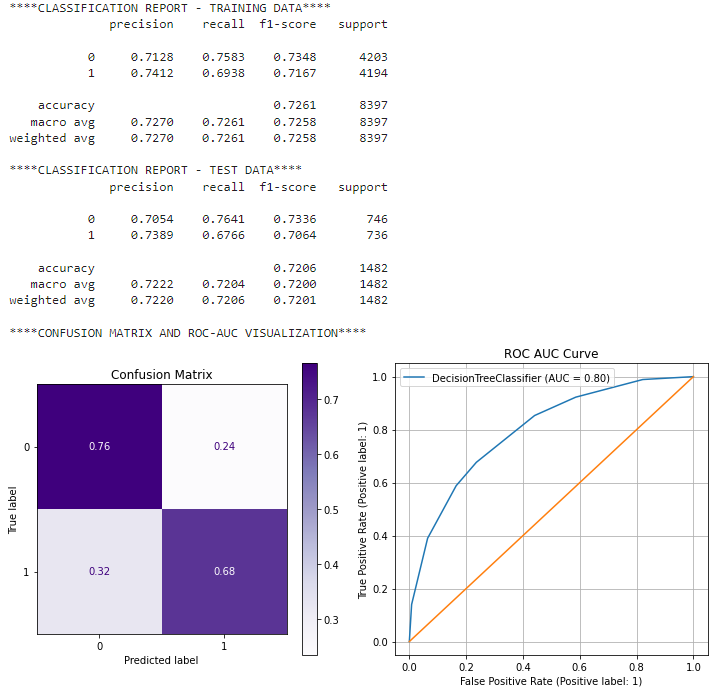


Рис.1 – Продуктивность алгоритма Decision Tree на полном наборе данных League of Legends

В результате обучения на полном наборе данных продуктивность модели улучшилась и приблизилась к продуктивности логистической регрессии, но не достигло её. Поэтому было принято решение об использовании ансамблевого алгоритма Random forest.

Random Forest (случайный лес) — это метод ансамблевого обучения модели для классификации, регрессии и других задач. Он состоит в оценки результатов работы множества деревьев решений во время обучения. Для задач классификации выходными данными случайного леса является класс, выбранный большинством деревьев[[268]](#footnote-268)[[269]](#footnote-269).

Исследователи считают метод случайного леса наиболее близко соответствующим требованиям, предъявляемым к использованию в качестве готовой процедуры интеллектуального анализа данных так как он толерантен к различным размерам данных и наличию гетероскедастичности и создает проверяемые модели. Однако точность его не всегда высока и хотя и в меньшей степени чем деревья решений, но всё же случайный лес также склонен к переобучению[[270]](#footnote-270).

Самым большим недостатком случайных лесов является их вычислительная сложность. Данный метод работает медленнее логистической регрессии, поскольку для прогнозирования используется большое количество деревьев решений. Все деревья в лесу должны произвести анализ всего набора данных, а затем провести голосование. Поэтому с ростом количества деревьев в лесу вырастает не только продуктивность модели, но и вычислительная сложность алгоритма.

Модель, построенная методом случайного леса сложно интерпретировать по сравнению с одним деревом решений, для дерева решений объяснительная способность гораздо выше и в следствии интерпретируемости дерево решений может быть преобразовано в правила, которые и будут служить основой поддержки для принятия решений.

Согласно устоявшейся последовательности алгоритм случайного леса первоначально был использован для скорректированного набора данных (рис.1). И несмотря на большую устойчивость по сравнению с алгоритмом дерева решений к переобучению именно в этом случае для тестовой выборки модель показала традиционные признаки переобучения. При этом на тестовой выборке модель отработала вполне корректно и показатели продуктивности были выше чем для алгоритма дерева решений, но при этом меньше чем для ситуации с использованием алгоритма логистической регрессии. Поэтому в данном случае лучшим будет сочетание алгоритма логистической регрессии для получения параметров математической модели и алгоритма дерева решений для поиска интуиций, неочевидных закономерностей.

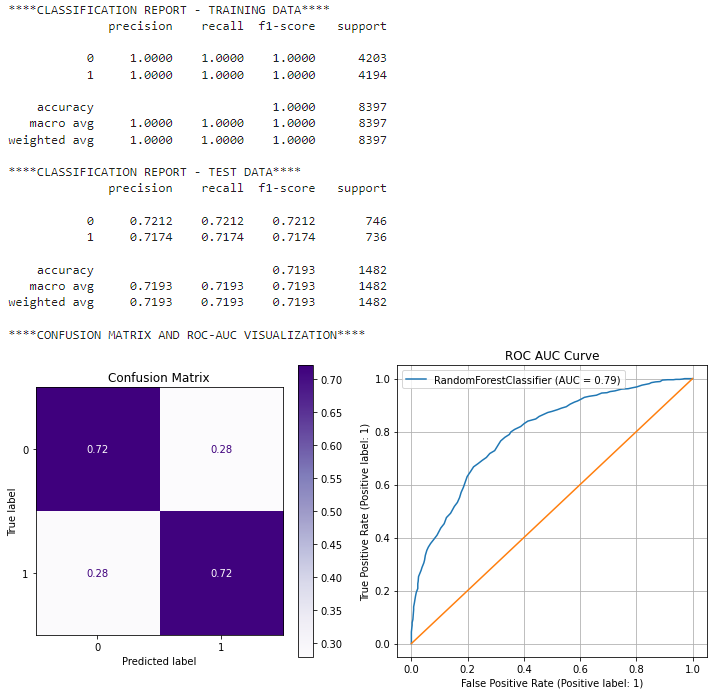


Рис.1 – Продуктивность алгоритма Random Forest на скорректированном наборе данных League of Legends

После скорректированного набора данных алгоритм случайного леса был использован для построения модели на полном наборе данных (рис.1).

При оценке результатов можно заметить, что и в этом случае модель максимально близка к переобучению на тестовой выборке, но также, как и в ситуации для скорректированного набора данных, для тестовой выборки модель отработала вполне корректно.

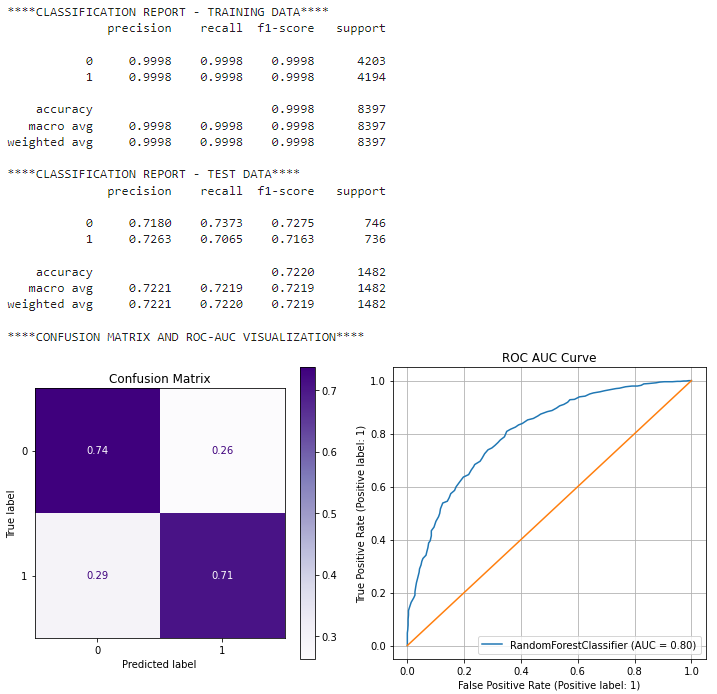


Рис.1 – Продуктивность алгоритма Random Forest на полном наборе данных League of Legends

На полном наборе данных модель так же показывает ситуацию схожую с переобучением на тренировочной выборке, а на тестовой продуктивность несколько хуже чем для модели логистической регрессии на скорректированном наборе данных.

Для дальнейшего совершенствования ансамбля было применено усложнение модели случайного лесса за счёт градиентного бустинга (XGBoost: Random Forest)

XGBoost и Random Forest —представляют собой древовидные модели и демонстрируют хорошую производительность при решении задачи обнаружения сложных закономерностей в данных. Случайный лес — это ансамблевая модель так называемого беггинга (bagging), которая обучает несколько деревьев параллельно, а конечный результат — это то, что результат полученный большинством деревьев. Ансамбль так называемого бустинга (boosting) XGBoost создает последовательный ансамбль древовидных моделей, каждая из которых улучшает результаты полученные предыдущей и так до получения конечного результата. Каждая модель ансамбля учится именно на ошибках предшественницы, пытаясь их исправить. Этот алгоритм можно использовать для создания и классификатора, и регрессора.

Обе модели различаются по производительности для разных случаев использовании. Они эффективны при обучении благодаря возможности использовать несколько ядер центрального процессора компьютера для параллельных операций. Скорость обучения во многом зависит от значения наборов гиперпараметров, то есть количества, глубины дерева и т. д., а также размера обучающих данных. В сфере физической культуры и спорта пока объёмы данных незначительны для того чтобы скорость работы модели была критическим фактором, упитывающимся исследователями. После своего появления Random Forest быстро стал популярным благодаря своей высокой точности. Однако высокая точность может достигаться за счет переобучения, как это было в произошедшей ситуации, из-за чего модель может демонтировать худшую производительность при исследовании новых, ранее неизвестных, наборов данных. Разработчики могут противостоять этому путем предварительной обработки и очистки данных, но проблема остается для сценариев использования в реальном времени.

Алгоритм XGBoost: Random Forest в свою очередь не строит деревья на полную глубину. Он обрезает деревья в определенной точке в зависимости от показателя сходства узла. Он рассматривает «Прирост» узла как разницу между показателем сходства узла и его дочерних элементов и прекращает построение дерева, если показатель считается минимальным. Меньшие деревья решают проблему переобучения и позволяют модели лучше обобщать. Модели XGBoost: Random Forest демонстрируют более высокую точность тестовых данных, что делает их более подходящими для решения реальных задач.

XGBoost: Random Forest представляет собой значительно более сложный алгоритм по сравнению с Random Forest и включает в себя параметры, которые могут значительно изменить производительность. При этом гиперпараметры для XGBoost: Random Forest устанавливаются только для первого дерева. Остальные деревья корректируются на каждой итерации, используя потерю предыдущего дерева и осуществляют поиск с использованием алгоритма градиентного спуска. Это значимо в случаях, когда входные данные поступают в режиме реального времени и демонстрируют высокую вариативность. Модель, параметры которой настраиваются итеративно (XGBoost: Random Forest), лучше обучается на потоковых данных, чем модель с фиксированным набором параметров для всего ансамбля (Random Forest). На рис.1 представлены результаты продуктивности алгоритма Градиентный бустинг случайного леса XGBoost: Random Forest на скорректированном наборе данных League of Legends.

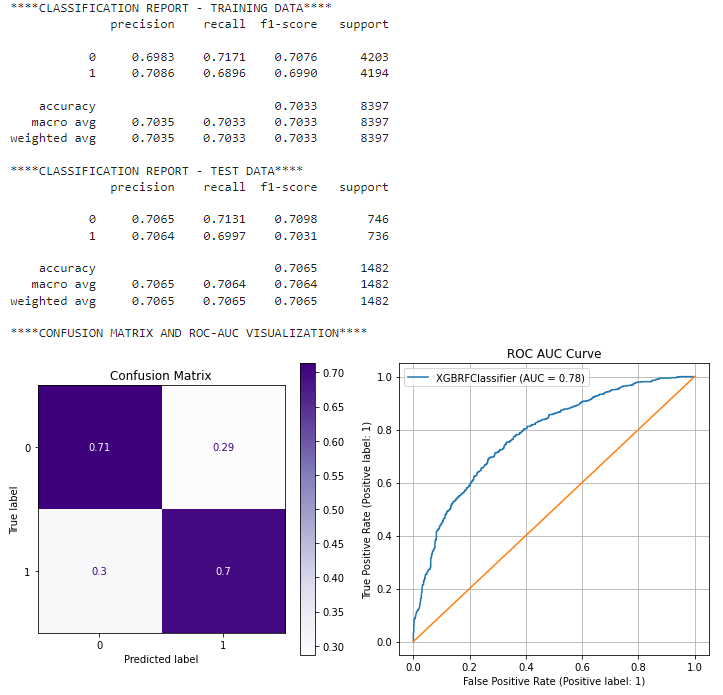


Рис.1 – Продуктивность алгоритма Градиентный бустинг случайного леса XGBoost: Random Forest на скорректированном наборе данных League of Legends

Результаты работы модели на скорректированном наборе данных League of Legends показывает, что в этом случае переобучения нен произошло и можно продолжить исследование уже на полном наборе данных (рис.1).

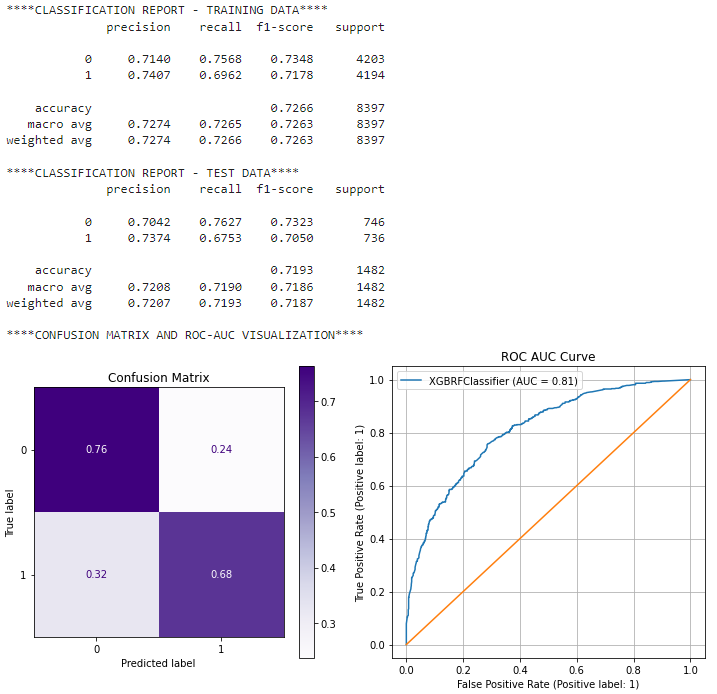


Рис.1 – Продуктивность алгоритма градиентный бустинг случайного леса (XGBoost: Random Forest) на полном наборе данных League of Legends

Модель XGBoost: Random Forest работает лучше, когда присутствует дисбаланс классов. Модель бустинга итеративно учится на ошибках предыдущего дерева. Таким образом, если дереву не удается предсказать класс (скорее всего, несбалансированный класс), следующее дерево придаст больше веса этому образцу. Интуитивно понятно, что это попытка уравнять модель, придав больший вес менее представленным категориям. В Random Forest соответственно нет такого механизма для устранения дисбаланса данных. В исследуемом случае произошло именно таким образом. Продуктивность модели на полном наборе данных выросла и ещё более приблизилась к продуктивности логистической регрессии. Дополнительно возможно выполнить упоминаемый выше подбор гиперпараметров модели (GridSearch CV) для попытки дополнительного улучшения продуктивности модели.

Согласно методу GridSearch сначала определяются желаемые к исследованию гиперпараметры модели и их возможные значения: решите, указывается диапазон или список значений для каждого гиперпараметра. Далее генерируются все возможные комбинации значений гиперпараметров. Это формирует так называемую сеть (Grid), где каждая точка соответствует уникальной комбинации гиперпараметров.

Потом происходит обучение, и оценка моделей для каждой комбинации гиперпараметров в созданной сети и используя данные обучения оценивается производительность каждой из полученных моделей, используя валидационный набор данных или метод перекрёстной валидации (cross-validation).

Выбор лучшей комбинации есть определение того, какая комбинация гиперпараметров обеспечивает наилучшую производительность модели на валидационном или тестовой выборке данных. Это дает оценку того, насколько хорошо ваша модель может работать на новых, ранее неизвестных данных.

GridSearch — это простой в реализации метод, но он может оказаться достаточно затратным в вычислительном отношении, особенно если модель имеет много гиперпараметров и широкий диапазон возможных значений для каждого. Чтобы избежать подобного ресурсоемкого процесса, исследователи часто используют такие методы, как случайный поиск (Random Search), при котором данные из пространства гиперпараметров выбираются случайным образом, или более продвинутые методы оптимизации, такие как байесовская оптимизация (Bayesian optimization). В нашем случае был применён двухступенчатый поиск, на первом этапе были определены тенденции в улучшении гиперпараметров за счёт создания «грубой» сети, а потом уточняющий поиск с учётом выявленных тенденций.

Так как наилучшие результаты показала ансамблевая модель с использованием бустинга, алгоритмом XGBoost: Random Forest реализуемая на полном наборе данных, то и подбор оптимальных гиперпараметров был произведён именно на ней, без использования скорректированного набора данных. Продуктивность модели XGBoost: Random Forest с подобранными гиперпараметрами можно видеть на рис.1

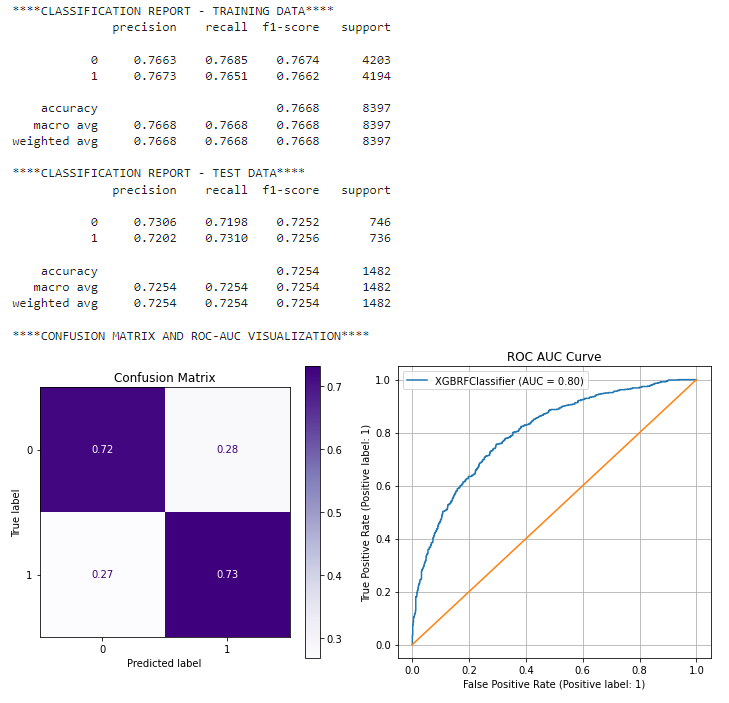


Рис.1 – Увеличение продуктивности алгоритма градиентный бустинг случайного леса (XGBoost: Random Forest) подбором гиперпараметров модели методом GridSearch CV

В результате подбора гиперпараметров модели XGBoost: Random Forest она продемонстрировала лучшую продуктивность среди ансамблей древовидных моделей и вплотную приблизилась к модели бинарной классификации на основе алгоритма логистической регрессии, но не превзошла её. Это стало ещё одним дополнительным свидетельством, что на тех наборах данных которые сейчас распространены в сфере физической культуры и спорта более простые решения на основе регрессионных моделей более предпочтительны. Более же сложные решения подходят для выявления сложных закономерностей и получения дополнительной информации влияющей на принятия решения. Так полученная модель может также дать достаточно информации для выбора оптимальной тактики на первые десять минут игры. Вклад отдельных переменных в победу одной из команд (в данном исследовании рассматривается победа команды синих) можно увидеть на рис.1

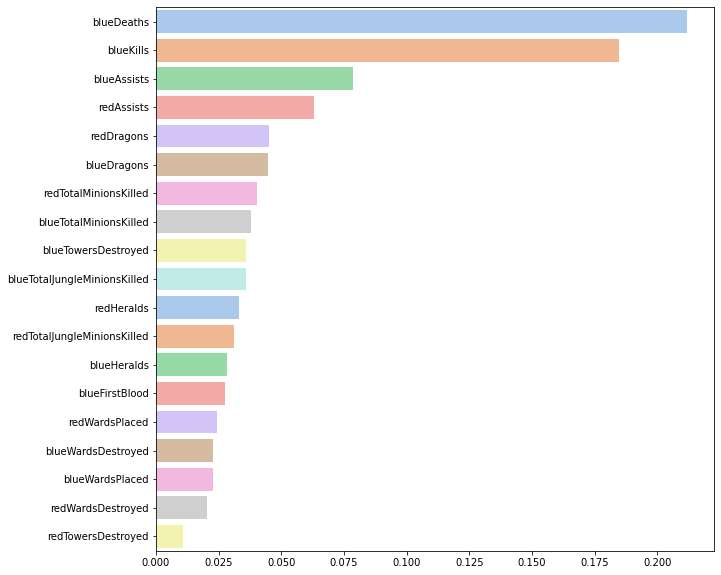


Рис.1 – Вклад отдельных переменных в победу команды синих (Blue) в модели на основе алгоритма градиентный бустинг случайного леса (XGBoost: Random Forest) на полном наборе данных League of Legends

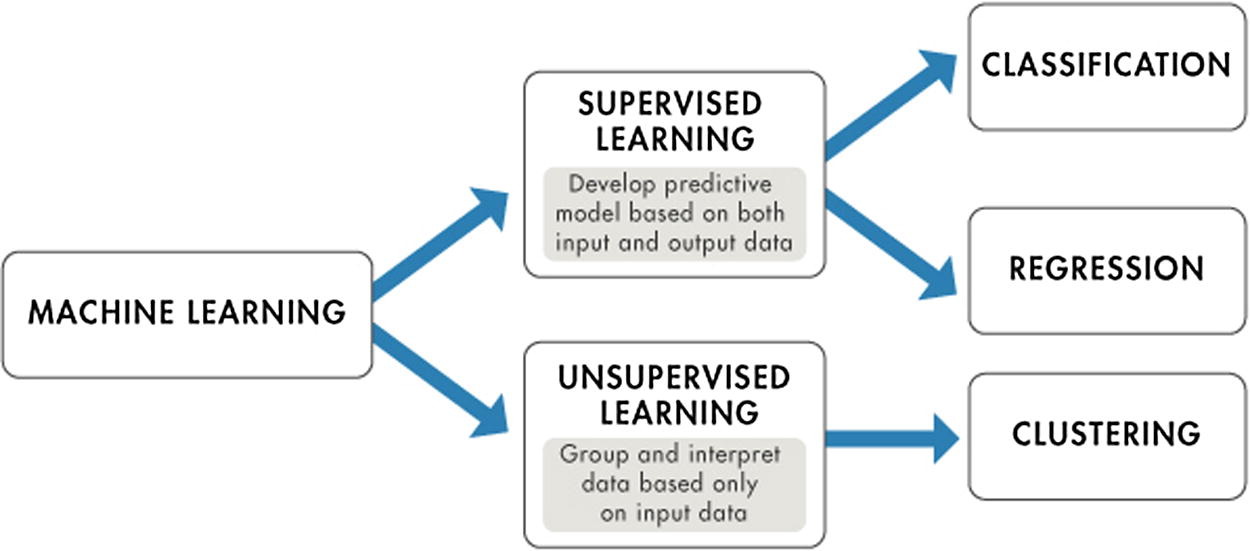
При рассмотрении результатов оценки вклада отдельных независимых переменных в определении победы команды синих (Blue) после первых десяти минут соревновательной компьютерной игры League of Legends можно заметить, что количество потерянных играбельных персонажей оказывает большее влияние чем уничтожение персонажей соперника, что предполагает большую эффективность более осторожной тактики в начале игры. Примерно такие же результаты можно наблюдать и с взаимопомощью во время уничтожения персонажей соперника, относительно собственных потерь при взаимодействии персонажей соперника. Подобные наблюдения помогают сформировать общую модель тактики ведения соревновательной борьбы на определённом этапе спортивного мастерства и позволяет сформировать как общую модель оптимальной тактики, так и создать, а затем постоянно совершенствовать индивидуальные тактические паттерны сравнивая их как с общей моделью, так и с построенными моделями тактики соперников.

В результате исследования ансамблевых моделей можно сделать следующие выводы:

Современное развитие информационных технологий в целом и искусственного интеллекта в частности позволяет использовать их для создания и совершенствования отдельных аспектов спортивной деятельности в ходе ежедневной практики для всех уровней спортивной подготовки.

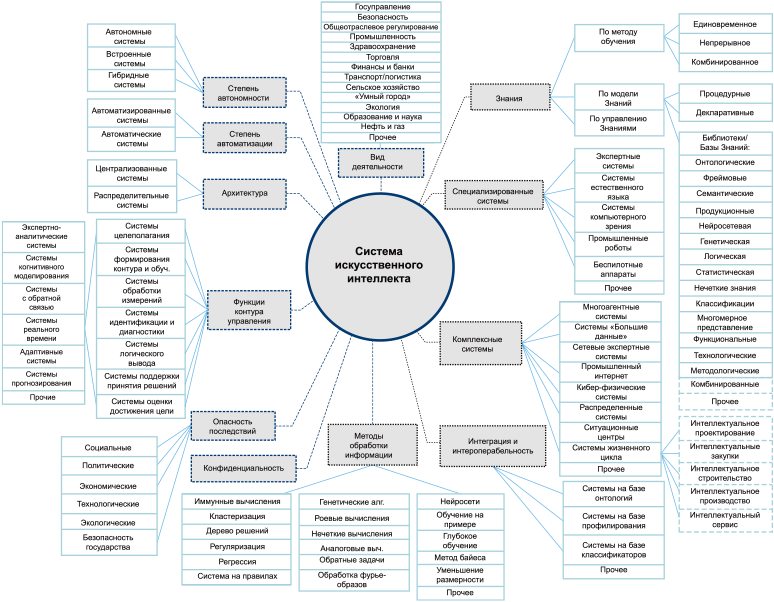
Ансамблевые методы построения модели соревновательной деятельности позволяют значительно улучшить изначальное качество алгоритмов искусственного интеллекта, но при этом относительно простые регрессионные модели зачастую оказываются более продуктивными, что связано с качественным и количественным составом данных наиболее часто используемых на современном этапе развития спортивной науки.

# Заключение



Bunker R. P., Thabtah F. A machine learning framework for sport result prediction //Applied computing and informatics. – 2019. – Т. 15. – №. 1. – С. 27-33.

<https://www.youtube.com/watch?v=qPzKWceRcmA>



ГОСТ Р 59277-2020 Системы искусственного интеллекта. Классификация систем искусственного интеллекта https://allgosts.ru/35/020/gost\_r\_59277-2020

1. Baumol W. J. Productivity growth, convergence, and welfare: what the long-run data show //The american economic review. – 1986. – С. 1072-1085. [↑](#footnote-ref-1)
2. Sala-i-Martin X. The Classical Approach to Convergence Analysis // The Economic Journal. — 1996. — Т. 106. — С. 1019—1036 [↑](#footnote-ref-2)
3. Ip B. Technological, content, and market convergence in the games industry //Games and Culture. – 2008. – Т. 3. – №. 2. – С. 199-224. [↑](#footnote-ref-3)
4. Jenkins H., Deuze M. Convergence culture //Convergence. – 2008. – Т. 14. – №. 1. – С. 5-12. [↑](#footnote-ref-4)
5. Jenkins H. Fans, bloggers, and gamers: Exploring participatory culture. – NYU Press, 2006 – 279 р. [↑](#footnote-ref-5)
6. Jenkins H., Ford S., Green J. Spreadable media //Spreadable media. – New York University Press, 2013. - 352 р. [↑](#footnote-ref-6)
7. Carter D. M. Money games: profiting from the convergence of sports and entertainment. – Stanford University Press, 2010. - 304 p. [↑](#footnote-ref-7)
8. McLeod J., Shilbury D., Zeimers G. An institutional framework for governance convergence in sport: The case of India //Journal of sport management. – 2020. – Т. 35. – №. 2. – С. 144-157. [↑](#footnote-ref-8)
9. Pu H., Kim J., Daprano C. Can esports substitute traditional sports? The convergence of sports and video gaming during the pandemic and beyond //Societies. – 2021. – Т. 11. – №. 4. – С. 129. [↑](#footnote-ref-9)
10. Pichersky E., Lewinsohn E. Convergent evolution in plant specialized metabolism //Annual review of plant biology. – 2011. – Т. 62. – р. 549-566. [↑](#footnote-ref-10)
11. Stern D. L. The genetic causes of convergent evolution //Nature Reviews Genetics. – 2013. – Т. 14. – №. 11. – С. 751-764. [↑](#footnote-ref-11)
12. Roco M. C. Nanotechnology: convergence with modern biology and medicine //Current opinion in biotechnology. – 2003. – Т. 14. – №. 3. – С. 337-346. [↑](#footnote-ref-12)
13. Perestrelo A. R. et al. Microfluidic organ/body-on-a-chip devices at the convergence of biology and microengineering //Sensors. – 2015. – Т. 15. – №. 12. – С. 31142-31170. [↑](#footnote-ref-13)
14. Roco M. C., Bainbridge W. S. Converging technologies for improving human performance: Integrating from the nanoscale //Journal of nanoparticle research. – 2002. – Т. 4. – С. 281-295. [↑](#footnote-ref-14)
15. Tarca A. L. et al. Machine learning and its applications to biology //PLoS computational biology. – 2007. – Т. 3. – №. 6. – С. e116. [↑](#footnote-ref-15)
16. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain //Psychological review. – 1958. – Т. 65. – №. 6. – С. 386. [↑](#footnote-ref-16)
17. Stormo G. D. et al. Use of the ‘Perceptron’algorithm to distinguish translational initiation sites in E. coli //Nucleic acids research. – 1982. – Т. 10. – №. 9. – С. 2997-3011. [↑](#footnote-ref-17)
18. Carpenter G. A., Grossberg S. The ART of adaptive pattern recognition by a self-organizing neural network //Computer. – 1988. – Т. 21. – №. 3. – С. 77-88. [↑](#footnote-ref-18)
19. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position //Biological cybernetics. – 1980. – Т. 36. – №. 4. – С. 193-202. [↑](#footnote-ref-19)
20. Kelly D. The evolutionary ecology of mast seeding //Trends in ecology & evolution. – 1994. – Т. 9. – №. 12. – С. 465-470. [↑](#footnote-ref-20)
21. Ciomaga B. Convergence challenges in sport-related applied disciplines: The case of sport management //Quest. – 2015. – Т. 67. – №. 3. – С. 300-316. [↑](#footnote-ref-21)
22. Houlihan B. Sport policy convergence: a framework for analysis //European sport management quarterly. – 2012. – Т. 12. – №. 2. – С. 111-135. [↑](#footnote-ref-22)
23. De Bosscher V. et al. Convergence and divergence of elite sport policies: is there a one-size-fits-all model to develop international sporting success? //Journal of Global Sport Management. – 2016. – Т. 1. – №. 3-4. – С. 70-89. [↑](#footnote-ref-23)
24. McCree R. Sport Policy and the new public management in the Caribbean: Convergence or resurgence? //Public management review. – 2009. – Т. 11. – №. 4. – С. 461-476. [↑](#footnote-ref-24)
25. Jackson S. J., Haigh S. (ed.). Sport and foreign policy in a globalizing world. – Routledge, 2013. – 176 р. [↑](#footnote-ref-25)
26. Stoddart B. Convergence: sport on the information superhighway //Journal of Sport and Social Issues. – 1997. – Т. 21. – №. 1. – С. 93-102. [↑](#footnote-ref-26)
27. Hutchins B., Rowe D. Sport beyond television: The internet, digital media and the rise of networked media sport. – Routledge, 2012. – Т. 40. – 237 р. [↑](#footnote-ref-27)
28. Miah A. Sport 2.0: Transforming sports for a digital world. – MIT Press, 2017. – 288 р. [↑](#footnote-ref-28)
29. Fritsch J. et al. The role of emotional convergence in sport: A scoping review //Sport, Exercise, and Performance Psychology. – 2023. Advance online publication. <https://doi.org/10.1037/spy0000330> [↑](#footnote-ref-29)
30. Green K. et al. Girls, young women and sport in Norway: a case of sporting convergence amid favourable socio-economic circumstances //International Journal of Sport policy and politics. – 2015. – Т. 7. – №. 4. – С. 531-550. [↑](#footnote-ref-30)
31. Pu H., Kim J., Daprano C. Can Esports Substitute Traditional Sports? The Convergence of Sports and Video Gaming during the Pandemic and Beyond. Societies 2021, 11, 129. – 2021. [↑](#footnote-ref-31)
32. Pizzo A. D., Funk D. C. Convergence of Sport and Esports //The Routledge Handbook of Digital Sport Management. – Routledge, 2022. – С. 293-307. [↑](#footnote-ref-32)
33. Buns M. T., Thomas K. T. Convergent validity between a sport video game and real sport performance //Sports Technology. – 2011. – Т. 4. – №. 1-2. – С. 77-87. [↑](#footnote-ref-33)
34. Miah A. Rethinking enhancement in sport //Annals of the New York Academy of Sciences. – 2006. – Т. 1093. – №. 1. – С. 301-320. [↑](#footnote-ref-34)
35. Kim M. et al. A new paradigm for the spread sport leisure culture focusing on the IT-based convergence interactive system //Information Science and Applications (ICISA) 2016. – Springer Singapore, 2016. – С. 1477-1485. [↑](#footnote-ref-35)
36. Kim J. Y. A Study on the Convergence of Sport Commitment and Smartphone Addiction in adolescents //Journal of the Korea Convergence Society. – 2021. – Т. 12. – №. 4. – С. 237-248. [↑](#footnote-ref-36)
37. Sandvoss C. Technological evolution or revolution? Sport online live internet commentary as postmodern cultural form //Convergence. – 2004. – Т. 10. – №. 3. – С. 39-54. [↑](#footnote-ref-37)
38. Malcolm D. Sport, medicine and health: The medicalization of sport?. – Taylor & Francis, 2016.-222p. [↑](#footnote-ref-38)
39. Kretschmann R., Von Grabowiecki U. Sport and Art: a View on the Worldwide and Connecting Area from the Perspective of Sports Science–or: about the Convergence of Two Worlds //Physical Culture and Sport. Studies and Research. – 2012. – Т. 55. – №. 1. – С. 5-16. [↑](#footnote-ref-39)
40. Кибернетика и спорт, научная конференция (1965; Москва). Материалы Научной конференции "Кибернетика и спорт". (Ноябрь, 1965 г.) [Текст] / Центр. совет Союза спортивных обществ и организаций СССР. Гос. ордена Ленина ин-т физ. культуры. - Москва : [б. и.], 1965. - 120 с. [↑](#footnote-ref-40)
41. Всесоюзный научно-методический семинар по проблеме "Кибернетика и спорт" (1967; Ленинград).[Тезисы докладов] [Текст] / Центр. совет Союза спортивных о-в и организаций СССР. Науч.-метод. совет. Гос. ордена Ленина и ордена Красного Знамени ин-т физ. культуры им. П. Ф. Лесгафта. Всесоюз. науч.-метод. семинар по проблеме "Кибернетика и спорт". - Ленинград : [б. и.], 1967. - 25 с [↑](#footnote-ref-41)
42. Винер, Н. Кибернетика, или Управление и связь в животном и машине / Пер. с англ. И. В. Соловьева. Под ред. Г. Н. Поварова. - М.: Сов. Радио, 1958. - 215 с [↑](#footnote-ref-42)
43. Винер, Н. Творец и робот : Обсуждение некоторых проблем, в которых кибернетика сталкивается с религией / Норберт Винер ; пер. с англ. М.Н. Аронэ, Р.А. Фесенко. - 1966. - 102 с. [↑](#footnote-ref-43)
44. "Кибернетика и управление движениями в спорте", всесоюзная научная конференция (1971; Москва).Тезисы Всесоюзной научной конференции "Кибернетика и управление движениями в спорте". (7-8 декабря 1971 г.) [Текст] / Ком. по физ. культуре и спорту при Совете Министров СССР. Гос. Центр. ордена Ленина ин-т физ. культуры. - Москва : [б. и.], 1971. - 59 с. [↑](#footnote-ref-44)
45. Бионика и биолмедикибернетика-85 : Биол. кибернетика. Кибернетика спорта, труда и отдыха : Тез. докл. всесоюз. конф. / [Редкол.: В. М. Ахутин (пред.) и др.]. - Ленинград : Науч. совет по комплекс. пробл. "Кибернетика" АН СССР, 1986. - 176 с. [↑](#footnote-ref-45)
46. Методы биокибернетического анализа функционального состояния спортсменов-подростков [Текст] : Сборник науч. трудов / Ин-т гигиены детей и подростков, Всесоюз. науч.-исслед. ин-т физ. культуры ; [Под ред. Г.Н. Сердюковой и П.И. Гуменера]. - Москва : Ин-т гигиены детей и подростков, 1977. - 146 с [↑](#footnote-ref-46)
47. Применение электронно-вычислительной техники в науке о спорте [Текст] : (Кибернетика и спорт) : Тезисы докладов Конференции. 10-12 сентября 1968 г. / Центр. совет Союза спортивных о-в и организаций СССР. Науч.-метод. совет. Гос. центр. ордена Ленина ин-т физ. культуры. Метод. кабинет. - Москва : [б. и.], 1968. - 139 с. [↑](#footnote-ref-47)
48. Бернштейн, Н.А. Избранные труды по биомеханике и кибернетике : Учеб. пособие для студентов высш. и сред. спец. учеб. заведен. / Н.А. Бернштейн. - М.: СпортАкадемПресс, 2001. - 295 с. [↑](#footnote-ref-48)
49. Бернштейн, Н.А. Новые линии развития в физиологии и их соотношение с кибернетикой. - М.: Теория и практика физической культуры, 1996. - 52 с. [↑](#footnote-ref-49)
50. Бернштейн, Н. А. Очерки по физиологии движений и физиологии активности [Текст] / Н. А. Бернштейн ; Акад. мед. наук СССР. - Москва : Медицина, 1966. - 349 с. [↑](#footnote-ref-50)
51. Чхаидзе, Л.В. Об управлении движениями человека [Текст] / Л. В. Чхаидзе, д-р биол. наук проф. засл. деят. физкультуры и спорта ГССР. - Москва : Физкультура и спорт, 1970. - 136 с. : [↑](#footnote-ref-51)
52. Козлов, А. Е.Кибернетика и спорт [Текст] / А. Е. Козлов, В. Б. Морозов, П. Н. Хломенок ; М-во сел. хоз-ва СССР. Кафедра физ. воспитания Кирг. с.-х. ин-та им. К. И. Скрябина. - Фрунзе : Мектеп, 1968. - 68 с. [↑](#footnote-ref-52)
53. Петровский, В.В. Кибернетика и спорт [Текст]. - Киев : Здоров'я, 1973. - 111 с [↑](#footnote-ref-53)
54. Морозов, В. Б. Использование некоторых приборов и методов кибернетики в практике физической культуры и спорта : диссертация ... кандидата педагогических наук : 13.00.00. - Тарту, 1970. - 156 с. [↑](#footnote-ref-54)
55. Садовский, Л.Е. Математика и спорт : (Мат. моделирование в задачах спорта) / Л. Е. Садовский, А. Л. Садовский. - Москва : Знание, 1990. - 45 с. [↑](#footnote-ref-55)
56. Суслаков, Б.А. Разработка математического обеспечения для задач многомерного статистического анализа : с приложениями к биомеханике спорта : диссертация ... кандидата технических наук : 05.13.01. - Москва, 1976. - 106 с. [↑](#footnote-ref-56)
57. Волков, Н.И. Математическое моделирование процессов энергетического обмена у человека при мышечной деятельности // Теория и практика физ. культуры. - 1966. - № 5. - С. 37-43. [↑](#footnote-ref-57)
58. Зациорский, В.М. Кибернетика, математика, спорт [Текст] : (Применение матем. и кибернет. методов в науке о спорте и в спортивной практике). - Москва : Физкультура и спорт, 1969. - 199 с. [↑](#footnote-ref-58)
59. Автоматизация биологических исследований / Отв. ред. В. В. Тихомиров. - М.: Наука, 1968. - 158 с. [↑](#footnote-ref-59)
60. Годик М.А., Тимошкин В.Н. Исследование двигательной активности студентов с помощью компьютерной экспертизы //Теор. и практ. физ. культ., 1990.- № 2.- С. 32-33 [↑](#footnote-ref-60)
61. Жбанков О. В., Лебяжьев А. Н. Компьютеризованная система как средство управления психофизическим состоянием спортсмена //Теория и практика физ. культуры. – 1994. – №. 67. – С. 46-48. [↑](#footnote-ref-61)
62. Жбанков О. В., Соловьев Е. В. Информационно-методическая система как средство индивидуализации процесса физического воспитания в вузе //Теор. и практ. физ. культ. – 1995. – Т. 11. – С. 51-53. [↑](#footnote-ref-62)
63. Аверкин, А.Н. Нечеткая семиотическая модель управления на основе отношения моделирования / Аверкин А.Н., Белов С.В. // Соврем. олимп. спорт и спорт для всех : 7 Междунар. науч. конгр. : Материалы конф., 24-27 мая 2003 г. - М., 2003. - Т. 3. - С. 363-364. [↑](#footnote-ref-63)
64. Балантер, Б.И. Вероятностные модели в физиологии : К анализу механизмов функционирования физиол. систем и их нарушений / Б. И. Балантер ; [АН СССР, Науч. совет по комплексной проблеме "Кибернетика]. - М.: Наука, 1977. - 251 с. [↑](#footnote-ref-64)
65. Вероятностное прогнозирование в деятельности человека / под ред.: И.М. Фейгенберга, Г.Е. Журавлева ; Акад. наук СССР [и др.]. - М.: Наука, 1977. - 391 с. [↑](#footnote-ref-65)
66. Ботвинник, М.М. О решении неточных переборных задач / Ботвинник Михаил Моисеевич. - М.: Сов. Радио, 1979. - 149 с. [↑](#footnote-ref-66)
67. Поспелов, Г.С.Искусственный интеллект - основа новой информационной технологии / Г. С. Поспелов; АН СССР. - Москва : Наука, 1988. – 278 с. [↑](#footnote-ref-67)
68. Шестаков, М. П. Теоретико-методическое обоснование процессов управления технической подготовкой спортсменов на основе компьютерного моделирования : диссертация ... доктора педагогических наук : 01.02.08. - Москва, 1997. - 317 с. [↑](#footnote-ref-68)
69. Фураев, А.Н. Система оперативного контроля и коррекции спортивной техники на основе современных информационных технологий : диссертация ... доктора педагогических наук : 01.02.08 / Фураев Александр Николаевич; [Место защиты: ФГБОУ ВО ‹‹Российский университет спорта ‹‹ГЦОЛИФК››]. - Малаховка, 2022. - 445 с. [↑](#footnote-ref-69)
70. Кузин В. В., Кутепов М. Е., Сейранов С. Г. Компьютерное тестирование как способ отбора студентов специализации" Спортивный менеджмент" и контроля их подготовленности //Теория и практика физической культуры. – 1997. – №. 1. – С. 48-51. [↑](#footnote-ref-70)
71. Гзовский, Б.М. Состояние и перспективы механико-математического и кибернетического образования в институтах физической культуры / Гзовский Б.М., Гинзбург Г.И. // Теория и практика физ. культуры. - 1976. - № 6. - С. 50-53. [↑](#footnote-ref-71)
72. Жилина М. Я. Применение ЭВМ в учебном процессе институтов физической культуры //Теор. и практ. физ. культ. – 1989. – №. 3. – С. 10-12. [↑](#footnote-ref-72)
73. Богданов С. Н., Чубаров М. M., Жуковский Ю. Т. Обучение предмету «физическое воспитание» с использованием персонального компьютера //Теор. и практ. физ. культ. – 1990. – №. 7. – С. 46-50. [↑](#footnote-ref-73)
74. Киршев С. П. Компьютерные технологии обучения упражнениям на уроке физической культуры //Теор. и практ. физ. культ. – 1993. – №. 5. – С. 38-40. [↑](#footnote-ref-74)
75. Самсонова А. В., Козлов И. М., Таймазов В. А. От ЭВМ-к информационным технологиям //Теория и практика физической культуры. – 2000. – №. 11. – С. 9-15. [↑](#footnote-ref-75)
76. Распоряжение Правительства РФ от 28 июля 2017 г. N 1632-р Об утверждении программы "Цифровая экономика Российской Федерации" URL: <https://base.garant.ru/71734878/> (дата обращения: 11.12.2023). [↑](#footnote-ref-76)
77. Указ Президента РФ от 9 мая 2017 г. N 203 "О Стратегии развития информационного общества в Российской Федерации на 2017 - 2030 годы" URL: <https://base.garant.ru/71670570/> (дата обращения: 11.12.2023). [↑](#footnote-ref-77)
78. Ермаков Алексей Валерьевич, Облог Кристина Андреевна, Щёголева Анастасия Павловна, Зиборова Светлана Александровна Оценка интенсивности физической нагрузки в демонстрационном экзамене по компетенции "Физическая культура, спорт и фитнес" по стандартам WORLDSKILLS RUSSIA при помощи умных браслетов ONETRAK C 320 PULSE // Вестник спортивной науки. 2020. №4. URL: <https://cyberleninka.ru/article/n/otsenka-intensivnosti-fizicheskoy-nagruzki-v-demonstratsionnom-ekzamene-po-kompetentsii-fizicheskaya-kultura-sport-i-fitnes-po> (дата обращения: 11.12.2023). [↑](#footnote-ref-78)
79. Ермаков Алексей Валерьевич, Облог Кристина Андреевна, Щёголева Анастасия Павловна, Зиборова Светлана Александровна Разработка модуля "С" - привлечение разных возрастных групп населения к здоровьесбережению - для компетенции "Физическая культура, спорт и фитнес" по стандартам “WORLDSKILLS RUSSIA” // Вестник спортивной науки. 2020. №3. URL: https://cyberleninka.ru/article/n/razrabotka-modulya-s-privlechenie-raznyh-vozrastnyh-grupp-naseleniya-k-zdoroviesberezheniyu-dlya-kompetentsii-fizicheskaya-kultura (дата обращения: 11.12.2023). [↑](#footnote-ref-79)
80. Ермаков, А. В. Результаты мониторингадвигательной активностинаселения с использованиеманализа "больших данных"компании ONETRAK / А. В. Ермаков // Спортивно-педагогическое образование: сетевое издание. – 2020. – № 2. – С. 18-22. [↑](#footnote-ref-80)
81. Статистика населения [сайт]. All-Populations, 2020 [обновлено 14.04.2020; процитировано 14.04.2020]. Доступно: https://all-populations.com/ru/ru/list-of-cities-in-russia-by-population.html. [↑](#footnote-ref-81)
82. Паспорт национального проекта Национальная программа "Цифровая экономика Российской Федерации" (утв. протоколом заседания президиума Совета при Президенте Российской Федерации по стратегическому развитию и национальным проектам от 4 июня 2019 г. N 7) . URL: <https://base.garant.ru/72296050/> (дата обращения: 11.12.2023) [↑](#footnote-ref-82)
83. Указ Президента РФ от 10 октября 2019 г. № 490 "О развитии искусственного интеллекта в Российской Федерации URL: <https://www.garant.ru/products/ipo/prime/doc/72738946/> (дата обращения: 11.12.2023) [↑](#footnote-ref-83)
84. Дорожная карта развития «сквозной» цифровой технологии «Технологии виртуальной и дополненной реальности» URL: <https://digital.gov.ru/ru/documents/6654/> (дата обращения: 11.12.2023) [↑](#footnote-ref-84)
85. Указ Президента Российской Федерации от 15.02.2024 № 124

    "О внесении изменений в Указ Президента Российской Федерации от 10 октября 2019 г. № 490 "О развитии искусственного интеллекта в Российской Федерации" и в Национальную стратегию, утвержденную этим Указом" <http://publication.pravo.gov.ru/document/0001202402150063> (дата обращения: 21.02.2024) [↑](#footnote-ref-85)
86. Дорожная карта развития «сквозной» цифровой технологии «Нейротехнологии и искусственный интеллект» URL: <https://digital.gov.ru/uploaded/files/07102019ii.pdf> (дата обращения: 11.12.2023) [↑](#footnote-ref-86)
87. Ермаков, А. В. Цифровые технологии в спорте как инструмент воспитания / А. В. Ермаков, Е. Н. Скаржинская // Физическая культура и спорт в XXI веке: актуальные проблемы и их решения : Сборник материалов Всероссийской с международным участием научно-практической конференции, Волгоград, 21–22 октября 2020 года. Том 2. – Волгоград: Волгоградская государственная академия физической культуры, 2020. – С. 189-193 [↑](#footnote-ref-87)
88. "Паспорт федерального проекта "Кадры для цифровой экономики" (утв. президиумом Правительственной комиссии по цифровому развитию, использованию информационных технологий для улучшения качества жизни и условий ведения предпринимательской деятельности, протокол от 28.05.2019 N 9) URL <https://www.consultant.ru/document/cons_doc_LAW_328933/> (дата обращения: 11.12.2023) [↑](#footnote-ref-88)
89. Реализуемые направления подготовки URL <https://gtsolifk.ru/obrazovanie/fakultet_magisterskoy_podgotovki/realizuemie_napravleniya_podgotovki> (дата обращения: 11.12.2023) [↑](#footnote-ref-89)
90. Brannigan K. J. et al. An Empirical Investigation of the Variables Influencing Contributions in NCAA Division I Athletics: A Quantitative Analysis //The Sport Journal. – 2020. – Т. 21. URL <https://thesportjournal.org/article/an-empirical-investigation-of-the-variables-influencing-contributions-in-ncaa-division-i-athletics-a-quantitative-analysis/> (дата обращения: 11.12.2023) [↑](#footnote-ref-90)
91. Aoki R. Y. S., Assuncao R. M., Vaz de Melo P. O. S. Luck is hard to beat: The difficulty of sports prediction //Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2017. – С. 1367-1376. [↑](#footnote-ref-91)
92. Куряев, Р. В. Прогнозирование результатов соревнований спортсменов-биатлонистов с использованием методов математического моделирования / Р. В. Куряев, С. А. Фирсова, О. А. Нелин // Тенденции развития науки и образования. – 2022. – № 89-1. – С. 16-22. [↑](#footnote-ref-92)
93. Севдалев, С. В. Прогноз достижений ведущих пятиборок мира на основе математического моделирования / С. В. Севдалев, В. Н. Старченко // Мир спорта. – 2023. – № 3(92). – С. 39-43 [↑](#footnote-ref-93)
94. Dobravec S. Predicting sports results using latent features: A case study //2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). – IEEE, 2015. – С. 1267-1272. [↑](#footnote-ref-94)
95. Ермаков, А. В. Использование моделей машинного обучения (логистической регрессии, случайного леса) для прогнозирования результатов матча League of Legends по оценке его первой части / А. В. Ермаков // Компьютерный спорт (киберспорт): проблемы и перспективы развития : Материалы Всероссийской научно-практической конференции, Москва, 08 декабря 2022 года / Под редакцией М.А. Новоселова. – Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский университет спорта "ГЦОЛИФК", 2022. – С. 60-68. [↑](#footnote-ref-95)
96. Xu J. Prediction and Planning of Sports Competition Based on Deep Neural Network //Computational Intelligence and Neuroscience. – 2022. – Т. 2022. [↑](#footnote-ref-96)
97. Дмитриев, Р. М. Использование методов дата майнинг для анализа больших данных на примере спортсменов, выступающих по дзюдо на олимпийских играх / Р. М. Дмитриев, Д. В. Попов // Интеграция науки и спортивной практики в единоборствах : материалы XXII Всероссийской с международным участием научно-практической конференции молодых ученых, посвящённой памяти заслуженного мастера спорта СССР, заслуженного тренера СССР, профессора Евгения Михайловича Чумакова, Москва, 16 февраля 2023 года. – Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский университет спорта "ГЦОЛИФК", 2023. – С. 96-102 [↑](#footnote-ref-97)
98. Javed A. et al. A decision tree framework for shot classification of field sports videos //The Journal of Supercomputing. – 2020. – Т. 76. – С. 7242-7267. [↑](#footnote-ref-98)
99. Serrien B., Goossens M., Baeyens J. P. Issues in using self-organizing maps in human movement and sport science //Journal homepage: http://iacss. org/index. php? id. – 2017. – Т. 16. – №. 1 [↑](#footnote-ref-99)
100. Akhanli S. E., Hennig C. Clustering of football players based on performance data and aggregated clustering validity indexes //Journal of Quantitative Analysis in Sports. – 2023. – Т. 19. – №. 2. – С. 103-123. [↑](#footnote-ref-100)
101. Barker J. B. et al. The effectiveness of psychological skills training and behavioral interventions in sport using single-case designs: A meta regression analysis of the peer-reviewed studies //Psychology of Sport and Exercise. – 2020. – Т. 51. – С. 101746. [↑](#footnote-ref-101)
102. Hodge K., Petlichkoff L. Goal profiles in sport motivation: A cluster analysis //Journal of Sport and Exercise Psychology. – 2000. – Т. 22. – №. 3. – С. 256-272. [↑](#footnote-ref-102)
103. Gaudreau P., Blondin J. P. Different athletes cope differently during a sport competition: A cluster analysis of coping //Personality and Individual differences. – 2004. – Т. 36. – №. 8. – С. 1865-1877. [↑](#footnote-ref-103)
104. Anıl Duman E., Sennaroğlu B., Tuzkaya G. A cluster analysis of basketball players for each of the five traditionally defined positions //Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology. – 2021. – С. 175 [↑](#footnote-ref-104)
105. González D. E. L. Technical-tactical performance in Greco-Roman wrestling: analysis of 2013 senior World Championships through multivariate analysis //International Journal of Wrestling Science. – 2014. – Т. 4. – №. 1. – С. 95-130. [↑](#footnote-ref-105)
106. Akay M. F. et al. Prediction of upper body power of cross-country skiers using support vector machines //Arabian Journal for Science and Engineering. – 2015. – Т. 40. – №. 4. – С. 1045-1055. [↑](#footnote-ref-106)
107. Ганичев, А. М. Использование методов машинного обучения для анализа спортивной деятельности в футболе и интерактивном футболе на примере ПФК ЦСКА в сезоне 2021/2022 РПЛ / А. М. Ганичев, А. В. Ермаков // Компьютерный спорт (киберспорт): проблемы и перспективы развития : Материалы Всероссийской научно-практической конференции, Москва, 08 декабря 2022 года / Под редакцией М.А. Новоселова. – Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский университет спорта "ГЦОЛИФК", 2022. – С. 39-47 [↑](#footnote-ref-107)
108. Carter N. et al. Clustering algorithms to increase fairness in collegiate wrestling //Journal of Quantitative Analysis in Sports. – 2022. – Т. 18. – №. 2. – С. 113-125 [↑](#footnote-ref-108)
109. Ardianto R. et al. Sentiment analysis on E-sports for education curriculum using naive Bayes and support vector machine //Jurnal Ilmu Komputer dan Informasi. – 2020. – Т. 13. – №. 2. – С. 109-122. [↑](#footnote-ref-109)
110. Liu P., Yue N., Chen J. A machine-learning-based medical imaging fast recognition of injury mechanism for athletes of winter sports //Frontiers in public health. – 2022. – Т. 10. – С. 842452. [↑](#footnote-ref-110)
111. Dobrota M., Delibašić B., Delias P. A skiing trace clustering model for injury risk assessment //International Journal of Decision Support System Technology (IJDSST). – 2016. – Т. 8. – №. 1. – С. 56-68. [↑](#footnote-ref-111)
112. Лапаева, А. Г. Верификация методики измерения Х-фактора при выполнении бросков с поворотом в самбо / А. Г. Лапаева, С. Е. Табаков, А. В. Ермаков // Вестник спортивной науки. – 2023. – № 3. – С. 82-87 [↑](#footnote-ref-112)
113. Сунгатуллина А.Х., Фонарев Д.В. Применение искусственного интеллекта в подготовке гимнастов. В сборнике: SMART SPORT & TOURISM: цифровая трансформация в сфере физической культуры, спорта и туризма. Материалы II Всероссийской научно-практической конференции с международным участием, прошедшей в рамках Международного форума KAZAN DIGITAL WEEK-2023 в онлайн-формате. Казань, 2023. С. 107-110. [↑](#footnote-ref-113)
114. Host K., Ivašić-Kos M. An overview of Human Action Recognition in sports based on Computer Vision //Heliyon. – 2022. . URL <https://www.cell.com/heliyon/pdf/S2405-8440(22)00921-5.pdf> (дата обращения: 11.12.2023) [↑](#footnote-ref-114)
115. Naik B. T., Hashmi M. F., Bokde N. D. A comprehensive review of computer vision in sports: Open issues, future trends and research directions //Applied Sciences. – 2022. – Т. 12. – №. 9. – С. 4429. [↑](#footnote-ref-115)
116. Свидетельство о государственной регистрации программы для ЭВМ № 2022613196 Российская Федерация. Программа по визуальному определению амплитуды движений в основных суставах человека (военнослужащего, спортсмена) при выполнении двигательных действий (спортивных и прикладных упражнений) с использованием искусственного интеллекта : № 2022612313 : заявл. 18.02.2022 : опубл. 01.03.2022 / А. В. Ермаков, А. Г. Белов, Е. Н. Скаржинская, М. А. Новоселов ; заявитель ФГКВОУВО "ВОЕННЫЙ УЧЕБНО-НАУЧНЫЙ ЦЕНТР СУХОПУТНЫХ ВОЙСК "ОБЩЕВОЙСКОВАЯ ОРДЕНА ЖУКОВА АКАДЕМИЯ ВООРУЖЕННЫХ СИЛ РОССИЙСКОЙ ФЕДЕРАЦИИ" URL <https://www.elibrary.ru/item.asp?id=48138477> (дата обращения: 11.12.2023) [↑](#footnote-ref-116)
117. Pandas URL <https://pandas.pydata.org/> (дата обращения: 11.12.2023) [↑](#footnote-ref-117)
118. Opencv URL: <http://opencv.org/> (дата обращения: 11.12.2023) [↑](#footnote-ref-118)
119. Mediapipe URL: https://google.github.io/mediapipe/ (дата обращения: 11.12.2023) [↑](#footnote-ref-119)
120. Pose Landmark Model (BlazePose GHUM 3D) <https://chuoling.github.io/mediapipe/solutions/pose.html#pose-landmark-model-blazepose-ghum-3d> (дата обращения: 11.12.2023) [↑](#footnote-ref-120)
121. Xu H. et al. Ghum & ghuml: Generative 3d human shape and articulated pose models //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2020. – С. 6184-6193. [↑](#footnote-ref-121)
122. Model card <https://mediapipe.page.link/blazepose-mc> (дата обращения: 11.12.2023) [↑](#footnote-ref-122)
123. Bazarevsky V. et al. Blazepose: On-device real-time body pose tracking //arXiv preprint arXiv:2006.10204. – 2020. <https://arxiv.org/pdf/2006.10204.pdf> (дата обращения: 11.12.2023) [↑](#footnote-ref-123)
124. COCO 2020 Keypoint Detection Task <https://cocodataset.org/#keypoints-2020> (дата обращения: 11.12.2023) [↑](#footnote-ref-124)
125. Pose Estimation Quality <https://chuoling.github.io/mediapipe/solutions/pose.html#pose-estimation-quality> (дата обращения: 11.12.2023) [↑](#footnote-ref-125)
126. Lafayette, T.B.G., Kunst, V.H.L., Melo, P.V.S., et al. Validation of Angle Estimation Based on Body Tracking Data from RGB-D and RGB Cameras for Biomechanical Assessment // Sensors (Basel). - 2022. - No. 23 (1). -P. 3. -. DOI: 10.3390/s23010003 [↑](#footnote-ref-126)
127. NumPy <https://numpy.org/> (дата обращения: 11.12.2023) [↑](#footnote-ref-127)
128. ГОСТ Р 59277-2020 Системы искусственного интеллекта. Классификация систем искусственного интеллекта <https://allgosts.ru/35/020/gost_r_59277-2020> (дата обращения: 11.12.2023) [↑](#footnote-ref-128)
129. Bunker R. P., Thabtah F. A machine learning framework for sport result prediction //Applied computing and informatics. – 2019. – Т. 15. – №. 1. – С. 27-33. [↑](#footnote-ref-129)
130. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of machine learning. – MIT press, 2018. – Р.427 [↑](#footnote-ref-130)
131. ГОСТ Р ИСО/МЭК 20546-2021 Информационные технологии (ИТ). Большие данные. Обзор и словарь URL: https://docs.cntd.ru/document/1200180276 (дата обращения: 27.12.2023) [↑](#footnote-ref-131)
132. Chen, H., Hsinchun, R., Chiang, R.H.L. and Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36, (2012)1165-1188. URL:https://www.semanticscholar.org/paper/Business-Intelligence-and-Analytics%3A-From-Big-Data-Chen-Chiang/f117c6f12d067bd66dad40996b3931c069daa2da (дата обращения: 27.12.2023) [↑](#footnote-ref-132)
133. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics //International journal of information management. – 2015. – Т. 35. – №. 2. – С. 137-144. URL: https://cebma.org/wp-content/uploads/Gandomi-2015.pdf (дата обращения: 27.12.2023) [↑](#footnote-ref-133)
134. De Mauro A., Greco M., Grimaldi M. What is big data? A consensual definition and a review of key research topics //AIP conference proceedings. – American Institute of Physics, 2015. – Т. 1644. – №. 1. – С. 97-104. URL: https://big-data-fr.com/wp-content/uploads/2015/02/aip-scitation-what-is-bigdata.pdf (дата обращения: 27.12.2023) [↑](#footnote-ref-134)
135. Big Data: как большие данные меняют бизнес https://psm7.com/analytics/big-data-kak-bolshie-dannye-menyayut-biznes.html (дата обращения: 27.12.2023) [↑](#footnote-ref-135)
136. David Kanter URL: https://twitter.com/TheKanter/status/559034352474914816 (дата обращения: 27.12.2023) [↑](#footnote-ref-136)
137. Big Data: 20 Mind-Boggling Facts Everyone Must Read URL: https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/?sh=60dba9dd17b1 (дата обращения: 27.12.2023) [↑](#footnote-ref-137)
138. Big Data URL:https://it-enterprise.com/knowledge-base/technology-innovation/big-data-bolshie-dannye (дата обращения: 27.12.2023) [↑](#footnote-ref-138)
139. Big Data: как большие данные меняют бизнес https://psm7.com/analytics/big-data-kak-bolshie-dannye-menyayut-biznes.html (дата обращения: 27.12.2023) [↑](#footnote-ref-139)
140. George A. Miller, “The magical number seven, plus or minus two: Some limits on our capacity for processing information,” Psychological Review, Volume 63(2), March 1956: 81–97 [↑](#footnote-ref-140)
141. Sreenivasan R. R. Characteristics of big data-a Delphi study : дис. – Memorial University of Newfoundland, 2017. URL:https://research.library.mun.ca/13080/1/thesis.pdf (дата обращения: 27.12.2023) [↑](#footnote-ref-141)
142. Laney D. et al. 3D data management: Controlling data volume, velocity and variety //META group research note. – 2001. – Т. 6. – №. 70. – С. 1. [↑](#footnote-ref-142)
143. Chen M. et al. Big data applications //Big Data: Related Technologies, Challenges and Future Prospects. – 2014. – С. 59-79. [↑](#footnote-ref-143)
144. Thiyagarajan V. S., Venkatachalapathy K. Isolating values from big data with the help of four V’S //International Journal of Research in Engineering and Technology. – 2014. – Т. 4. – №. 1. – С. 132-135. [↑](#footnote-ref-144)
145. Anuradha J. et al. A brief introduction on Big Data 5Vs characteristics and Hadoop technology //Procedia computer science. – 2015. – Т. 48. – С. 319-324. [↑](#footnote-ref-145)
146. Hadi H. J. et al. Big data and five V’s characteristics //International Journal of Advances in Electronics and Computer Science. – 2015. – Т. 2. – №. 1. – С. 16-23. [↑](#footnote-ref-146)
147. Yassin A. T., Hamoudy E. Analyzing 6Vs of big data using system dynamics //journal of kerbala university. – 2014. – №. 2 – P. 75-83. [↑](#footnote-ref-147)
148. Hamoudy E. et al. Analyzing 6Vs of Big Data using System Dynamics //journal of kerbala university. – 2011. – Т. 7. – №. 4. – С. 75-83. [↑](#footnote-ref-148)
149. Uddin M. F. et al. Seven V's of Big Data understanding Big Data to extract value //Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education. – IEEE, 2014. – С. 1-5. [↑](#footnote-ref-149)
150. Owais S. S., Hussein N. S. Extract five categories CPIVW from the 9V’s characteristics of the big data //International Journal of Advanced Computer Science and Applications. – 2016. – Т. 7. – №. 7(3) – р.254–258. [↑](#footnote-ref-150)
151. De Mauro A., Greco M., Grimaldi M. What is big data? A consensual definition and a review of key research topics //AIP conference proceedings. – American Institute of Physics, 2015. – Т. 1644. – №. 1. – С. 97-104 [↑](#footnote-ref-151)
152. M. A. Beyer and D. Laney, The Importance of “Big Data”: A Definition, Gartner report – 2012 - p. 1–9 [↑](#footnote-ref-152)
153. Dijcks J. P. Oracle: Big data for the enterprise //Oracle white paper. – 2012. – Т. 16. [↑](#footnote-ref-153)
154. Peer Research, Big Data Analytics. Intel’s IT Manager Survey on How Organizations Are Using Big Data, Intel report .URL:https://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf (дата обращения: 27.12.2023) [↑](#footnote-ref-154)
155. Suthaharan S. Big data classification: Problems and challenges in network intrusion prediction with machine learning //ACM SIGMETRICS Performance Evaluation Review. – 2014. – Т. 41. – №. 4. – С. 70-73 [↑](#footnote-ref-155)
156. Schroeck M. et al. Analytics: The real-world use of big data //IBM Global Business Services. – 2012. – Т. 12. – №. 2012. – С. 1-20 [↑](#footnote-ref-156)
157. Framework D. N. B. D. I. NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements //NIST Special Publication. – 2014. – Т. 1500. – С. 3 [↑](#footnote-ref-157)
158. Ward J. S., Barker A. Undefined by data: a survey of big data definitions //arXiv preprint arXiv:1309.5821. – 2013. URL: https://arxiv.org/abs/1309.5821 (дата обращения: 27.12.2023) [↑](#footnote-ref-158)
159. Customers Rapidly Adopting Big Data Solutions — Driven By Marketing, Sales and More — Reports New Microsoft Research URL:https://news.microsoft.com/2013/02/11/customers-rapidly-adopting-big-data-solutions-driven-by-marketing-sales-and-more-reports-new-microsoft-research/ (дата обращения: 27.12.2023) [↑](#footnote-ref-159)
160. Dumbill E. Making sense of big data //Big Data. – 2013. – Т. 1. – №. 1. – С. 1-2. [↑](#footnote-ref-160)
161. Fisher D. et al. Interactions with big data analytics //interactions. – 2012. – Т. 19. – №. 3. – С. 50-59 [↑](#footnote-ref-161)
162. Shneiderman B., in Proc. 2008 ACM SIGMOD Int. Conf. Manag. Data -2008- pp. 3–12 [↑](#footnote-ref-162)
163. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H.. Big data: The next frontier for innovation, competition. – and productivity. Technical report, McKinsey Global Institute, 2011. – Т. 5. – №. 6. URL:https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi\_big\_data\_exec\_summary.pdf (дата обращения: 27.12.2023) [↑](#footnote-ref-163)
164. Chen, H., Hsinchun, R., Chiang, R.H.L. and Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36, -2012-р.1165-1188 [↑](#footnote-ref-164)
165. Boyd D., Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon //Information, communication & society. – 2012. – Т. 15. – №. 5. – С. 662-679 [↑](#footnote-ref-165)
166. Mayer-Schönberger V., Cukier K. Big data: A revolution that will transform how we live, work, and think. – Houghton Mifflin Harcourt, 2013. URL: https://www.semanticscholar.org/paper/Big-Data%3A-A-Revolution-That-Will-Transform-How-We-Mayer-Schnberger-Cukier/2e76fa107c50485985487b03a601eb0c11894193 (дата обращения: 27.12.2023) [↑](#footnote-ref-166)
167. Технические характеристики и ограничения Excel URL:https://support.microsoft.com/ru-ru/office/%D1%82%D0%B5%D1%85%D0%BD%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B5-%D1%85%D0%B0%D1%80%D0%B0%D0%BA%D1%82%D0%B5%D1%80%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B8-%D0%B8-%D0%BE%D0%B3%D1%80%D0%B0%D0%BD%D0%B8%D1%87%D0%B5%D0%BD%D0%B8%D1%8F-excel-1672b34d-7043-467e-8e27-269d656771c3#ID0EDBD=Newer\_versions (дата обращения: 27.12.2023) [↑](#footnote-ref-167)
168. Ермаков А. В. ПРОБЛЕМА ОПРЕДЕЛЕНИЯ ПОНЯТИЯ «БОЛЬШИЕ ДАННЫЕ» В ОТРАСЛИ ФИЗИЧЕСКАЯ КУЛЬТУРА И СПОРТ //Цифровая трансформация физической культуры и спорта: теория, практика, подготовка кадров. – 2022. – С. 25-39. [↑](#footnote-ref-168)
169. Lyakh Y, Ermakov A, Sarafanova E. The Use Of Information Technologies In Fitness Activities Under The Covid-19 Restrictions //European Proceedings of Social and Behavioural Sciences. – 2021 – р. 727-735 [↑](#footnote-ref-169)
170. James D. Wearable Technology in sport, a convergence of trends //Journal of advanced sport technology. – 2017. – Т. 1. – №. 1. – С. 1-4. [↑](#footnote-ref-170)
171. Duffield R. et al. Accuracy and reliability of GPS devices for measurement of movement patterns in confined spaces for court-based sports //Journal of science and medicine in sport. – 2010. – Т. 13. – №. 5. – С. 523-525. [↑](#footnote-ref-171)
172. Vickery W. M. et al. Accuracy and reliability of GPS devices for measurement of sports-specific movement patterns related to cricket, tennis, and field-based team sports //The Journal of Strength & Conditioning Research. – 2014. – Т. 28. – №. 6. – С. 1697-1705. [↑](#footnote-ref-172)
173. Herren R. et al. The prediction of speed and incline in outdoor running in humans using accelerometry //Medicine and science in sports and exercise. – 1999. – Т. 31. – №. 7. – С. 1053-1059. [↑](#footnote-ref-173)
174. Porter M. E. et al. How smart, connected products are transforming competition //Harvard business review. – 2014. – Т. 92. – №. 11. – С. 64-88. [↑](#footnote-ref-174)
175. Jones R. L., Wallace M. Another bad day at the training ground: Coping with ambiguity in the coaching context //Sport, education and society. – 2005. – Т. 10. – №. 1. – С. 119-134. [↑](#footnote-ref-175)
176. Krüger A., Edelmann‐Nusser J. Biomechanical analysis in freestyle snowboarding: application of a full‐body inertial measurement system and a bilateral insole measurement system //Sports Technology. – 2009. – Т. 2. – №. 1-2. – С. 17-23. [↑](#footnote-ref-176)
177. Bächlin M., Tröster G. Swimming performance and technique evaluation with wearable acceleration sensors //Pervasive and mobile computing. – 2012. – Т. 8. – №. 1. – С. 68-81. [↑](#footnote-ref-177)
178. Li R. et al. A wearable biofeedback control system based body area network for freestyle swimming //2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). – IEEE, 2016. – С. 1866-1869. [↑](#footnote-ref-178)
179. Chardonnens J. et al. Measurement of the dynamics in ski jumping using a wearable inertial sensor-based system //Journal of sports sciences. – 2014. – Т. 32. – №. 6. – С. 591-600. [↑](#footnote-ref-179)
180. Chardonnens J. et al. Characterization of lower-limbs inter-segment coordination during the take-off extension in ski jumping //Human movement science. – 2013. – Т. 32. – №. 4. – С. 741-752. [↑](#footnote-ref-180)
181. Chardonnens J. et al. A system to measure the kinematics during the entire ski jump sequence using inertial sensors //Journal of biomechanics. – 2013. – Т. 46. – №. 1. – С. 56-62. [↑](#footnote-ref-181)
182. Blair S. et al. Concurrent validation of an inertial measurement system to quantify kicking biomechanics in four football codes //Journal of biomechanics. – 2018. – Т. 73. – С. 24-32. [↑](#footnote-ref-182)
183. Alonge F. et al. The use of accelerometers and gyroscopes to estimate hip and knee angles on gait analysis //Sensors. – 2014. – Т. 14. – №. 5. – С. 8430-8446. [↑](#footnote-ref-183)
184. Papi E. et al. Smart sensing system for combined activity classification and estimation of knee range of motion //IEEE Sensors Journal. – 2015. – Т. 15. – №. 10. – С. 5535-5544. [↑](#footnote-ref-184)
185. Reenalda J. et al. Continuous three dimensional analysis of running mechanics during a marathon by means of inertial magnetic measurement units to objectify changes in running mechanics //Journal of biomechanics. – 2016. – Т. 49. – №. 14. – С. 3362-3367. [↑](#footnote-ref-185)
186. Wisbey B., Pyne D., Rattray B. Quantifying changes in AFL player game demands using GPS tracking //FitSense Australia. – 2010. – р.36 [↑](#footnote-ref-186)
187. Siegmund G. P. et al. Laboratory validation of two wearable sensor systems for measuring head impact severity in football players //Annals of biomedical engineering. – 2016. – Т. 44. – С. 1257-1274. [↑](#footnote-ref-187)
188. Meamarbash A. A novel inertial technique to measure very high linear and rotational movements in sports, part I: The hardware //Journal of Applied Sciences. – 2009. – Т. 9. – №. 9. – С. 1746-1751. [↑](#footnote-ref-188)
189. Cutmore T. R. H., James D. A. Sensors and sensor systems for psychophysiological monitoring: A review of current trends //Journal of Psychophysiology. – 2007. – Т. 21. – №. 1. – С. 51-71. [↑](#footnote-ref-189)
190. Coughlin M. J., Cutmore T. R. H., Hine T. J. Automated eye tracking system calibration using artificial neural networks //Computer Methods and Programs in Biomedicine. – 2004. – Т. 76. – №. 3. – С. 207-220. [↑](#footnote-ref-190)
191. Qaisar S. et al. A hidden markov model for detection and classification of arm action in cricket using wearable sensors //Journal of Mobile Multimedia. – 2013. – С. 128-144. [↑](#footnote-ref-191)
192. Kenneally-Dabrowski C. J. B., Serpell B. G., Spratford W. Are accelerometers a valid tool for measuring overground sprinting symmetry? //International Journal of Sports Science & Coaching. – 2018. – Т. 13. – №. 2. – С. 270-277. [↑](#footnote-ref-192)
193. Lee J. B. et al. Inertial sensor, 3D and 2D assessment of stroke phases in freestyle swimming //Procedia Engineering. – 2011. – Т. 13. – С. 148-153. [↑](#footnote-ref-193)
194. Wixted A. et al. Wearable sensors for on field near real time detection of illegal bowling actions //1 of 1-Conference of Science, Medicine & Coaching in Cricket 2010. – 2010. – С. 165. [↑](#footnote-ref-194)
195. Salman M., Qaisar S., Qamar A. M. Classification and legality analysis of bowling action in the game of cricket //Data Mining and Knowledge Discovery. – 2017. – Т. 31. – С. 1706-1734. [↑](#footnote-ref-195)
196. Spratford W. et al. Illegal bowling actions contribute to performance in cricket finger‐spin bowlers //Scandinavian Journal of Medicine & Science in Sports. – 2018. – Т. 28. – №. 6. – С. 1691-1699. [↑](#footnote-ref-196)
197. Azevedo A., Santos M. F. KDD, SEMMA and CRISP-DM: a parallel overview //IADS-DM. – 2008. [↑](#footnote-ref-197)
198. Fayyad U., Piatetsky-Shapiro G., Smyth P. The KDD process for extracting useful knowledge from volumes of data //Communications of the ACM. – 1996. – Т. 39. – №. 11. – С. 27-34. [↑](#footnote-ref-198)
199. Fayyad U. M., Haussler D., Stolorz P. E. KDD for Science Data Analysis: Issues and Examples //KDD. – 1996. – Р. 50-56. [↑](#footnote-ref-199)
200. Benoît, G.,. Data Mining. Annual Review of Information Science and Technology, 2002 -Vol. 36, No. 1 - pp 265-310. [↑](#footnote-ref-200)
201. Chen M. S., Han J., Yu P. S. Data mining: an overview from a database perspective //IEEE Transactions on Knowledge and data Engineering. – 1996. – Т. 8. – №. 6. – С. 866-883. [↑](#footnote-ref-201)
202. Simoudis E. Reality check for data mining //IEEE Intelligent Systems. – 1996. – Т. 11. – №. 05. – С. 26-33. [↑](#footnote-ref-202)
203. Džeroski S. Towards a general framework for data mining //International Workshop on Knowledge Discovery in Inductive Databases. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2006. – С. 259-300. [↑](#footnote-ref-203)
204. Schröer C., Kruse F., Gómez J. M. A systematic literature review on applying CRISP-DM process model //Procedia Computer Science. – 2021. – Т. 181. – С. 526-534. [↑](#footnote-ref-204)
205. Thukral E., Ratten V. Sport companies using analytics and statistics //Statistical Modelling and Sports Business Analytics. – Routledge, 2020. – С. 48-53. [↑](#footnote-ref-205)
206. Mincev S. Analysing Data Mining Methods in Sports Analytics: a Case Study in NHL Player Salary Prediction : дис. – Universidade NOVA de Lisboa (Portugal), 2021. [↑](#footnote-ref-206)
207. Singh N., Kanthwal A., Bidhuri P. Soccer competitiveness using shots on target: data mining approach //HCI in Business, Government and Organizations. Information Systems and Analytics: 6th International Conference, HCIBGO 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part II 21. – Springer International Publishing, 2019. – С. 141-150. [↑](#footnote-ref-207)
208. Wirth R., Hipp J. CRISP-DM: Towards a standard process model for data mining //Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. – 2000. – Т. 1. – С. 29-39. [↑](#footnote-ref-208)
209. Albert J. et al. (ed.). Handbook of statistical methods and analyses in sports. – CRC Press, 2017.-520р. [↑](#footnote-ref-209)
210. Figueiredo P. et al. A review of the statistical methods used in sports and exercise sciences PhD theses-a case study in a single post-graduate program //R Bras Ci Mov. – 2018. – Т. 26. – №. 4. – С. 191-200. [↑](#footnote-ref-210)
211. Kendall M. G., Stuart A. The Advanced Theory of Statistic: Distribution Theory; Charles Griffin & Comp //Lim.: London, UK. – 1958.-676р. [↑](#footnote-ref-211)
212. Hand D. J. et al. Report of the Council for 2009 //Journal of the Royal Statistical Society. Series A (Statistics in Society). – 2010. – Т. 173. – №. 4. – С. 827-914. [↑](#footnote-ref-212)
213. De Smith M. J. Statistical analysis handbook. – The Winchelsea Press, 2018.-629р. [↑](#footnote-ref-213)
214. American Statistical Association URL: https://ww2.amstat.org/consultantdirectory/index.cfm?fuseaction=searchresults (дата обращения: 25.10.2023) [↑](#footnote-ref-214)
215. MacKay R. J., Oldford R. W. Scientific method, statistical method and the speed of light //Statistical Science. – 2000. – С. 254-278. [↑](#footnote-ref-215)
216. Olympedia <https://www.olympedia.org/> (дата обращения: 25.10.2023) [↑](#footnote-ref-216)
217. Olympics <https://olympics.com/> (дата обращения: 25.10.2023) [↑](#footnote-ref-217)
218. Sports Reference <http://www.sports-reference.com/> (дата обращения: 25.10.2023) [↑](#footnote-ref-218)
219. 120 years of Olympic history: athletes and results <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results> (дата обращения: 25.10.2023) [↑](#footnote-ref-219)
220. Ray H. L. The Sport Sculpture of R. Tait Mckenzie //Sport History Review. – 1993. – Т. 24. – №. 1. – С. 84-86. [↑](#footnote-ref-220)
221. Kramer B. In search of the lost champions of the Olympic art contests //Revue Olympique. – 1912. – Т. 12. – №. 7. – С. 105. [↑](#footnote-ref-221)
222. Crossman E. B. Shooting //Journal of Health, Physical Education, Recreation. – 1972. – Т. 43. – №. 2. – С. 42-47. [↑](#footnote-ref-222)
223. Rosin, P.; Rammler, E., "The Laws Governing the Fineness of Powdered Coal", Journal of the Institute of Fuel, 7 -1933 - рр.29–36 [↑](#footnote-ref-223)
224. Hesse R. Triangle distribution: Mathematica link for Excel //Decision Line. – 2000. – Т. 31. – №. 3. – С. 12-14. [↑](#footnote-ref-224)
225. Silverman M. P. Exact Statistical Distribution and Correlation of Human Height and Weight: Analysis and Experimental Confirmation //Open Journal of Statistics. – 2022. – Т. 12. – №. 5. - 743-787рр. [↑](#footnote-ref-225)
226. A’hearn B., Peracchi F., Vecchi G. Height and the normal distribution: Evidence from Italian military data //Demography. – 2009. – Т. 46. – №. 1. – С. 1-25. [↑](#footnote-ref-226)
227. Johnson N. L., Kotz S., Balakrishnan N. Normal distributions //Continuous univariate distributions. – 1987. – Т. 1. – С. 156-157. [↑](#footnote-ref-227)
228. ГОСТ Р 50779.21-2004 Статистические методы. Правила определения и методы расчета статистических характеристик по выборочным данным. Часть 1. Нормальное распределение (дата обращения: 11.12.2023) [↑](#footnote-ref-228)
229. Casella G., Berger R. L. Statistical Inference, volume Second Edition. – 2001. – 686 р. [↑](#footnote-ref-229)
230. Lyon A. Why are normal distributions normal? //The British Journal for the Philosophy of Science. – (65)2014.- 621–649 срр [↑](#footnote-ref-230)
231. Wright S. J. Numerical optimization. – 2006. – 249 р [↑](#footnote-ref-231)
232. Hosking J. R. M., Wallis J. R. Regional frequency analysis. – 1997. – С. 240. [↑](#footnote-ref-232)
233. ГОСТ Р 50779.27-2017 Статистические методы. Распределение Вейбулла. Анализ данных <https://docs.cntd.ru/document/1200146523?marker=65A0IQ&section=text> (дата обращения: 11.12.2023) [↑](#footnote-ref-233)
234. Zadik Z., Dovev R. ARE HEIGHTS AND HEIGHTS NORMALLY DISTRIBUTED? //Pediatric Research. – 1993. – Т. 33. – №. 5. – С. S54-S54. [↑](#footnote-ref-234)
235. Kim K., Shevlyakov G. Why gaussianity? //IEEE Signal Processing Magazine. – 2008. – Т. 25. – №. 2. – С. 102-113. [↑](#footnote-ref-235)
236. Kozubowski T., Podgorski K. A log-Laplace growth rate model //Mathematical Scientist. – 2003. – Т. 28. – С. 49-60. [↑](#footnote-ref-236)
237. Kozubowski T. J., Podgórski K. A multivariate and asymmetric generalization of Laplace distribution //Computational Statistics. – 2000. – Т. 15. – С. 531-540. [↑](#footnote-ref-237)
238. McDonald J., Michelfelder R. A., Theodossiou P. Robust regression estimation methods and intercept bias: A capital asset pricing model application //Multinational Finance Journal. – 2009. – Т. 13. – №. 3/4. – С. 293-321. [↑](#footnote-ref-238)
239. McDonald J. B., Michelfelder R. A., Theodossiou P. Robust estimation with flexible parametric distributions: estimation of utility stock betas //Quantitative Finance. – 2010. – Т. 10. – №. 4. – С. 375-387. [↑](#footnote-ref-239)
240. Clauset A., Shalizi C. R., Newman M. E. J. Power-law distributions in empirical data //SIAM review. – 2009. – Т. 51. – №. 4. – С. 661-703. [↑](#footnote-ref-240)
241. Mønness E. N. Height curves based on the bivariate Power-Normal and the bivariate Johnson’s System bounded distribution. – Rapport;04/2013 –17 р. [↑](#footnote-ref-241)
242. Blair Fix Visualizing Power-Law Distributions https://economicsfromthetopdown.com/2019/04/25/visualizing-power-law-distributions/(дата обращения: 11.12.2023). [↑](#footnote-ref-242)
243. Nakagami M. The m-distribution—A general formula of intensity distribution of rapid fading //Statistical methods in radio wave propagation. – Pergamon, 1960. – С. 3-36. [↑](#footnote-ref-243)
244. Gavagnin E. et al. The invasion speed of cell migration models with realistic cell cycle time distributions //Journal of Theoretical Biology. – 2019. – Т. 481. – С. 91-99. [↑](#footnote-ref-244)
245. Yates C. A., Ford M. J., Mort R. L. A multi-stage representation of cell proliferation as a Markov process //Bulletin of mathematical biology. – 2017. – Т. 79. – С. 2905-2928. [↑](#footnote-ref-245)
246. Hogg R. V., McKean J. W., Craig A. T. Introduction to mathematical statistics. – Pearson, 2019.- 762р. [↑](#footnote-ref-246)
247. NCD Risk Factor Collaboration (NCD-RisC) <https://www.ncdrisc.org/country-profile.html> (дата обращения: 25.10.2023) [↑](#footnote-ref-247)
248. Takiar R. The Validity of t-test, Mann-Whitney test and Z test for Testing Significant differences between two Sample Means When Sample size is 10 or below //Bulletin of Mathematics and Statistics Research. – 2023. – Т. 11. – №. 2. – С. 1-15. [↑](#footnote-ref-248)
249. Bradley J. V. Nonrobustness in one-sample Z and t tests: A large-scale sampling study //Bulletin of the Psychonomic Society. – 1980. – Т. 15. – №. 1. – С. 29-32. [↑](#footnote-ref-249)
250. Freedman D. A. Statistical models: theory and practice. – Cambridge university press, 2009.- 444р [↑](#footnote-ref-250)
251. Ермаков А.В., Мякинченко П.Е. ПРОГНОЗИРОВАНИЕ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ В СПОРТЕ ВЫСШИХ ДОСТИЖЕНИЙ НА ПРИМЕРЕ ЗИМНИХ ВИДОВ СПОРТА // ТиПФК. 2021. №2. URL: https://cyberleninka.ru/article/n/prognozirovanie-s-ispolzovaniem-metodov-matematicheskogo-modelirovaniya-v-sporte-vysshih-dostizheniy-na-primere-zimnih-vidov-sporta (дата обращения: 03.02.2024). [↑](#footnote-ref-251)
252. Gujarati D. N., Porter D. C. Basic econometrics. – McGraw-hill, 2009.- 400 р. [↑](#footnote-ref-252)
253. Goldberger, Arthur S.. Econometric Theory. New York: John Wiley & Sons. - 1964 - pp. 238–243 [↑](#footnote-ref-253)
254. UFC-Fight historical data from 1993 to 2021 <https://www.kaggle.com/datasets/rajeevw/ufcdata?resource=download> (дата обращения: 25.10.2023) [↑](#footnote-ref-254)
255. Tolles J., Meurer W. J. Logistic regression: relating patient characteristics to outcomes //Jama. – 2016. – Т. 316. – №. 5. – С. 533-534. [↑](#footnote-ref-255)
256. King G., Zeng L. Logistic regression in rare events data //Political analysis. – 2001. – Т. 9. – №. 2. – С. 137-163. [↑](#footnote-ref-256)
257. Boateng E. Y., Abaye D. A. A review of the logistic regression model with emphasis on medical research //Journal of data analysis and information processing. – 2019. – Т. 7. – №. 4. – С. 190-207. [↑](#footnote-ref-257)
258. Hosmer Jr D. W., Lemeshow S., Sturdivant R. X. Applied logistic regression. – John Wiley & Sons, 2013. – С. 398 [↑](#footnote-ref-258)
259. Cramer, J. S. The origins of logistic regression (PDF) (Technical report). Vol. 119. Tinbergen Institute. – 2002 - pp. 167–178 [↑](#footnote-ref-259)
260. League of Legends Diamond Ranked Games (10 min) <https://www.kaggle.com/datasets/bobbyscience/league-of-legends-diamond-ranked-games-10-min> (дата обращения: 25.10.2023) [↑](#footnote-ref-260)
261. Stehman, Stephen V. "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. 62 (1): 1997 – С.77–89. [↑](#footnote-ref-261)
262. Powers, David M. W. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 2011 – С.37–63. [↑](#footnote-ref-262)
263. Green D. M. et al. Signal detection theory and psychophysics. – New York : Wiley, 1988. – Т. 1. – С. 505 [↑](#footnote-ref-263)
264. Kingma D. P., Ba J. Adam: A method for stochastic optimization //arXiv preprint arXiv:1412.6980. – 2014. [↑](#footnote-ref-264)
265. Opitz, D.; Maclin, R. "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research. 11 – 1999 – р.169–198. [↑](#footnote-ref-265)
266. Polikar, R. "Ensemble based systems in decision making". IEEE Circuits and Systems Magazine. 6 (3) – 2006 – р.21–45. [↑](#footnote-ref-266)
267. Winterfeldt, Detlof; Edwards, Ward. "Decision trees". Decision Analysis and Behavioral Research. Cambridge University Press. - 1986 - pp. 63–89 [↑](#footnote-ref-267)
268. Ho T. K. The random subspace method for constructing decision forests //IEEE transactions on pattern analysis and machine intelligence. – 1998. – Т. 20. – №. 8. – С. 832-844. [↑](#footnote-ref-268)
269. Ho T. K. Random decision forests //Proceedings of 3rd international conference on document analysis and recognition. – IEEE, 1995. – Т. 1. – С. 278-282. [↑](#footnote-ref-269)
270. Hastie T. et al. The elements of statistical learning: data mining, inference, and prediction. – New York : springer, 2017. –С. 764. [↑](#footnote-ref-270)