**Министерство Образования и Исследований Молдовы**

**Технический Университет Молдовы**

**Факультет Вычислительной техники, Информатики и Микроэлектроники**

**Департамент Информатики и Системной Инженерии**

**Предмет:** Analiza Datelor

**Научная статья**

**На тему: «Анализ факторов, влияющих на успеваемость учеников»**

**Выполнила:** Costrici A. гр. MI-213

**Кишинёв 2023**

*Статистика и анализ данных в образовании*

**УДК:** 37.07

**JEL:** I21, I24, I2

**Анализ факторов, влияющих на успеваемость учеников.**

**Кострич Анастасия**

Студентка третьего курса, специальность Информационный Менеджмент, Технический Университет Молдова, Кишинев, Республика Молдова.

[anastasia.costrici@iis.utm.md](mailto:anastasia.costrici@iis.utm.md)

**Аннотация**

**Предмет.**

В настоящей статье проводится исследование, направленное на анализ факторов, оказывающих влияние на успеваемость учеников. Проблема влияния различных переменных на образовательные результаты является актуальной и требует системного подхода для выявления ключевых аспектов, формирующих учебные успехи, что позволит обеспечить идеальные условия для обучения и успешной сдачи экзаменов.

**Цели.**

Целью настоящего исследования является проведение анализа факторов, влияющих на успеваемость учеников, с использованием методов Exploratory Data Analysis и разработка прогностической модели оценок по математике. Основной упор делается на исследование влияния таких независимых переменных, как пол, образование родителей, этническая группа, уровень подготовки к тесту, тип обеда, а также предварительные оценки по чтению и письму.

**Методология.**

Для проведения анализа данных в работе были применены методы Exploratory Data Analysis с целью оценки распределения и взаимосвязей между переменными. В дополнение к этому, проведен статистический анализ, включающий корреляционный анализ и тесты гипотез, с целью выявления статистически значимых факторов. Для построения прогностической модели использовались методы множественной регрессии, а именно линейная регрессия, рандомный лес и метод опорных векторов (SVM).

**Результаты.**

В ходе реализации различные интересные зависимости. Так, было выявлено, что мальчики лучше справляются с точными науками, такими как математика. Показатель же девочек лидирует в гуманитарных предметах. Были выявлены закономерности между модготовкой к экзаменам и топом питания ученика, что может говорить о финансовом положении семьи. Следом были выявлены переменные, обладающие наибольшим влиянием на показатель оценки по математике у студентов.

Помимо этого, в ходе создания моделей с помощью различных методов были получены схожие результаты точности, которые объясняют 87.51% дисперсии зависимой переменной.

**Заключение.**

В проведенном исследовании были выделены значимые факторы, влияющие на успеваемость учеников в математике. Результаты анализа данных свидетельствуют о том, что пол, уровень образования родителей, этническая группа, уровень подготовки к тесту, тип обеда, а также предварительные оценки по чтению и письму оказывают существенное воздействие на учебные показатели. В целом, результаты исследования не только подчеркивают важность учета множества факторов при анализе успеваемости учеников, но и предоставляют основу для разработки эффективных стратегий и политик в образовательной сфере, направленных на поддержку студентов и создание условий для максимального обучения.

**Ключевые слова:** успеваемость, образовательные факторы, анализ данных, прогностическая модель

**Введение**

**1.1 Актуальность темы**

Сегодня, в эпоху стремительных технологических изменений и глобальных вызовов, необходимость в высококачественном образовании становится неотъемлемым элементом обеспечения устойчивого социально-экономического развития. Образовательные системы по всему миру находятся под постоянным давлением, вынужденные адаптироваться к новым требованиям и ожиданиям. В свете этих вызовов, анализ успеваемости учеников приобретает стратегическое значение, предоставляя важные инсайты для совершенствования образовательных стратегий и принятия обоснованных решений. Более того, с учетом динамичного характера современного общества и его постоянного развития, анализ успеваемости учеников становится инструментом для предвидения будущих требований рынка труда и формирования образовательных программ, отвечающих потребностям наукоемкой экономики. Успешная интеграция информационных технологий, инновационных методов обучения и персонализированных подходов к образованию зависит, в частности, от понимания, какие факторы способствуют или препятствуют достижению высоких результатов обучающимися.

**1.2 Цель и задачи**

В данной научной статье стоит цель систематизировать и выявить влияние различных переменных на итоговые оценки учащихся по математике, чтению и письму. В образовании решение задачи обеспечения эффективности обучения требует глубокого анализа, учета многогранных факторов и разработки системных подходов. Настоящее исследование направлено на выделение тех аспектов, которые играют ключевую роль в формировании успеха учащихся, с целью создания более благоприятных условий для развития образовательной среды. В контексте настоящей работы, необходимость анализа и выявления влияния факторов на успеваемость становится неотъемлемой частью стремления обеспечить более эффективное и инклюзивное образование для каждого обучающегося.

**Методология**

**2.1 Описание датасета**

В ходе исследования использовался обширный датасет, охватывающий информацию о 30,682 учениках, собранную с целью тщательного анализа влияния различных факторов на их успеваемость в области математики, чтения и письма. Этот массив данных предоставляет уникальную возможность лучше понять динамику образовательной среды и взаимосвязи между различными параметрами, влияющими на учебные результаты.

Для анализа использовались следующие данные:

* **Гендер** - В датасете учтены два основных гендера: мужчины и женщины. Этот параметр позволяет оценить, существуют ли гендерные различия в успеваемости и какие именно.
* **Этническая принадлежность** - Трехуровневая категоризация этнической принадлежности (A, B, C) позволяет провести анализ влияния культурных особенностей на образовательные результаты. Реальные названия групп не предоставлены во избежание конфликтов.
* **Образование родителей** – Многообразие вариантов уровня образования родителей (старшая школа, бакалавриат, колледж и др.) позволяет исследовать влияние образовательного фона семьи на успеваемость учеников.
* **Тип обеда** - Категоризация обедов на стандартные и бесплатные/льготные обеспечивает индикатор финансового положения семьи, что может оказывать влияние на успеваемость.
* **Подготовка к экзаменам** - Информация о подготовке к экзаменам (готовился/не готовился) помогает определить, насколько активно учащиеся подходили к подготовке, и как это сказывается на их результативности.
* **Результаты экзамена** - Итоговые оценки по трем предметам (математика, чтение, письмо) предоставляют ключевые данные для оценки общей успеваемости учеников.

**2.2 Источник данных**

Данные были доступны на платформе Kaggle, широко используемой для обмена исследовательскими наборами данных. За счет публичного доступа к датасету и использования его в образовательных целях, предоставляется возможность для широкого анализа и применения результатов исследования в практических аспектах образовательной сферы. Анализ этого датасета направлен на выявление ключевых факторов, воздействующих на успеваемость учеников, что позволит сформировать более глубокое понимание динамик и взаимосвязей в сфере образования. Полученные результаты могут быть использованы для разработки практических рекомендаций, способствующих оптимизации образовательных стратегий с целью повышения качества обучения и равенства образовательных возможностей.

**2.3 Инструменты и этапы исследования**

Первым этапом в работе над проектом стал разведочный анализ данных (EDA). Разведочный анализ данных (Exploratory Data Analysis) — предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками, а также сужения набора методов, используемых для создания модели Машинного обучения [1]. Для реализации данного процесса был использован язык R - язык программирования с открытым исходным кодом, который широко используется как статистическое программное обеспечение и инструмент анализа данных [2]. Данный язык обладает множеством удобных библиотек для анализа и визуализации данных. В данном проекте были использованы:

1. **Tidyverse** - библиотека для удобной работы с данными, которая активно используется в аналитике. С ее помощью можно более быстро получать описание датафрейма, сохранять полученные результаты, группировать наблюдения по определенному признаку, а также строить красивые графики [3].
2. **Openintro —** это пакет в языке R, созданный для поддержки использования в учебных целях в курсах по статистике и анализу данных. Пакет включает в себя набор данных, функции для их обработки и визуализации, а также учебные материалы [4].
3. **corrplot** — это пакет в языке R, предназначенный для визуализации матрицы корреляции. Матрица корреляции используется для оценки степени линейной взаимосвязи между переменными в данных [5].
4. **Rsample** - это пакет в языке программирования R, предназначенный для облегчения процесса разделения данных на обучающие и тестовые наборы, а также для проведения перекрестной проверки (cross-validation) [6].
5. **Knitr** - это пакет, который предоставляет инструменты для динамической генерации отчётов, документации и презентаций из исходного кода на R. Он обеспечивает возможность объединения кода, текста и результатов выполнения кода в одном документе, что делает процесс создания отчётов более автоматизированным и воспроизводимым [7].
6. **Caret** - это пакет, который предоставляет единый интерфейс для обучения и оценки моделей машинного обучения. Он создан для упрощения процесса поиска, обучения и оценки различных моделей, а также для автоматизации многих рутинных задач [8].
7. **Vip** - то R-пакет, предназначенный для визуализации важности переменных в моделях машинного обучения [9].
8. **Metrics** - это пакет в языке программирования R, предназначенный для оценки качества моделей машинного обучения. Он содержит различные функции для вычисления метрик, таких как среднеквадратичная ошибка (RMSE), средняя абсолютная ошибка (MAE), коэффициент детерминации (R-squared) и другие. [10]

Помимо описанных пакетов, также, были использованы и другие, например: dummy[11], randomForest и др.

Для начала, была произведена обработка данных, включая изменение типов данных, анализ пропущенных значений и обработку выбросов. Затем проведен разведочный анализ данных (EDA), в ходе которого изучены распределения переменных, взаимосвязи между ними, а также выявлены основные статистические характеристики.

Для более глубокого понимания влияния факторов на успеваемость, использованы различные методы визуализации. Построены гистограммы, ящичные диаграммы, линейные графики и тепловые карты корреляции. Эти визуализации позволили выявить тенденции и закономерности в данных.

Затем был использован метод one-hot encoding [12] для категориальных переменных, что позволило представить их в виде числовых признаков для использования в моделях машинного обучения.

Для построения и оценки моделей машинного обучения применены методы линейной регрессии [13], случайного леса (Random Forest) [14] и метод опорных векторов (SVM) [15]. Проведена оценка качества моделей с использованием различных метрик, таких как коэффициент детерминации, среднеквадратичная ошибка и средняя абсолютная ошибка. В конечном итоге, данная методология анализа факторов, влияющих на успеваемость учеников, позволяет выделить ключевые факторы, которые следует учитывать при разработке образовательных стратегий и политик. Результаты исследования могут быть использованы для оптимизации образовательного процесса и повышения эффективности обучения.

**Результаты**

**3.1 Разведочный анализ данных**

В ходе реализации данного этапа работы, были созданы различного рода графики, визуализирующие различные моменты дата сета для выявления интересных закономерностей и прочей всевозможной информации.

В начале анализа данных была построена круговая диаграмма (Pie Chart) для визуального сравнения количества мальчиков и девочек в датасете (см. Приложение 1, Рисунок 1). Этот график подчеркнул, что соотношение полов практически равномерно, что составляет 50.34% девочек и 49.66% мальчиков. Это является важным стартовым пунктом, поскольку равномерное распределение полов может влиять на результаты исследования и статистические выводы.

Далее был проведен обзор уровня образования родителей учеников с использованием столбчатой диаграммы (см. Приложение 1, Рисунок 2). На этом графике визуализированы типы образования родителей в зависимости от пола учащихся. В результате анализа можно отметить, что большая часть родителей окончила колледж, в то время как меньшая часть получила степень магистра. Важно отметить, что для всех уровней образования родителей нет явного преобладания какого-либо из полов учащихся.

Рассмотрим следующий показатель, а именно этнические группы учеников. Как было сказано ранее, реальные наименования этнических групп заменены буквенным кодом. Для данного исследования также была использована столбчатая диаграмма (см. Приложение 1, Рисунок 3). Можем отметить, что ни в одной из этнических групп не присутствует преобладания какого-либо из полов. Также диаграмма говорит нам, что этническая группа С сильно преобладает над остальными, самой же редкой является этническая группа А.

Следующий этап, а именно рассмотрение оценок учащихся дает нам интересную информацию. Так на линейной диаграмме сравнения оценок по математике у мальчиков и девочек (см. Приложение 1, Рисунок 4) можно увидеть, что мальчики больше преуспели в данной дисциплине, так как получивших 100 балов за тест больше среди представителей данного пола. Гуманитарные же предметы такие как письмо (см. Приложение 1, Рисунок 5) и чтение (см. Приложение 1, Рисунок 6), лучше даются девочкам. Количество девочек получивших 100 балов по эти предметам больше, чем мальчиков, при этом отрыв в количестве значительнее, чем в показателях по математике.

Теперь сравним средние показатели по предметам, для этого я использовала ящик с усами (box plot). В данном случае также будем сравнивать оценки по каждому из предметов. Так, диаграмма оценок по математике (см. Приложение 1, Рисунок 7) показывает, что средняя оценка у мальчиков выше, чем у девочек. Более того, минимальный бал, полученный мальчиком выше, чем минимальный бал у девочек. Перейдем к результатам по чтению (см. Приложение 1, Рисунок 8). В данном случае уже лидируют девочки даже с большим отрывом. Минимальный бал теперь выше у представительниц женского пола, мальчики плошают. Показатель по письму (см. Приложение 1, Рисунок 9) даёт схожую информацию, однако различие средней оценки даже больше, чем в чтении.

С предпочтениями в предметах определились, теперь предлагаю рассмотреть типы обеда и то, как они связаны со сдачей тренировочного теста (см. Приложение 1, Рисунок 10). Так мы видим, что большая часть учеников не сдала тренировочный тест, что могло повлиять на итоговые результаты. Также большая часть учеников питается стандартным обедом, в том числе и среди тех, кто сдал тренировочный тест. Также мы можем увидеть, как мальчики и девочки готовятся к экзаменам (см. Приложение 1, Рисунок 11). График говорит, что все же большее количество девочек готовятся к экзаменам, чем мальчики, хотя в числе неподготовленных девочки также лидируют.

Таким образом в ходе разведочного анализа мы выяснили, что мальчики лучше в точных науках, девочки же в гуманитарных. Большая часть учеников не готовится к экзаменам, однако среди тех, кто тренировочный экзамен сдал преобладают девочки, питающиеся стандартным обедом. Также родители большей части учеников закончили колледж. На данном этапе это все, что удалось выяснить. Переходим к следующей части проекта.

**3.2 Построение модели**

Для данной части проекта используем оценки по трем предметам. В ходе работы было выяснено, что для построения модели нужно выбрать лишь один предмет для получения хороших результатов. Таким образом целевой переменной стала оценка по математике, оценки же по чтению и письму мы включаем как признаки (независимые переменные). Однако на тепловой карте корреляции (см. Приложение 1, Рисунок 12) мы видим, что оценки по другим предметам очень сильно влияют на оценку по математике, что является плохой практикой. По этой причине в модель мы будем включать только один из предметов, а именно результаты по письму. При избавлении же от двух этих показателей точность модели очень сильно падает, а именно до 0,2 для R^2.

Для прогнозирования оценок по математике была использована модель линейной регрессии, которую я упоминала ранее. Линейная регрессия — это статистический метод, используемый для моделирования отношений между зависимой переменной (или целевой переменной) и одной или несколькими независимыми переменными (предикторами или факторами) [13]. Полученная модель дает хорошие показатели точности (см. Приложение 1, Рисунок 13), мы можем видеть это по показателю R^2 равному 0.8751. Это значение показывает, насколько хорошо модель соответствует данным. В данном случае, модель объясняет 87.51% дисперсии зависимой переменной, что является хорошим показателем. Также можем отметить и другие показатели, например RSE (Стандартная ошибка остатков), которая равна 5.375. Этот показатель измеряет, насколько точно модель соответствует данным. В нашем случае это достаточно хороший показатель, так как чем меньше это число, тем лучше соответствие модели данным. Помимо этого, был получен показатель Adjusted R-squared равный 0.875, который учитывает количество переменных в модели и позволяет оценивать, насколько хорошо эти переменные объясняют изменение в зависимой переменной. Он может принимать значения от 0 до 1, где 1 означает, что модель полностью объясняет изменчивость зависимой переменной. Это означает, что около 87.5% изменчивости зависимой переменной объясняется моделью, принимая во внимание количество предикторов. Значение p-value, < 2.2e-16, указывает на то, что модель в целом является значимой. Также были выявлены следующие показатели: MAE = 4,312076; RMSE = 5,374167 Таким образом была построена хорошая регрессионная модель, результаты которой также можно просмотреть и графически (см. Приложение 1, Рисунок 14).

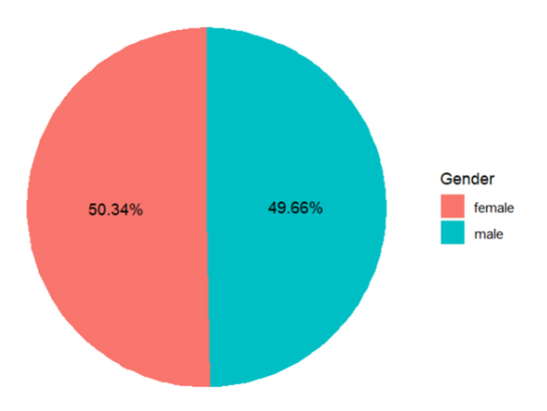
Однако, помимо данного метода построения модели, были использованы еще два, а именно Random Forest и SVM. Это было сделано для проверки показателей основной модели. Таким образом были выявлены следующие показатели для Random Forest: R^2 = 0,8743732; MAE = 4,312076; RMSE = 5,384144. А также для SVM: R^2 = 0,8736061; MAE = 4,32243; RMSE = 5,40049. Показатели трех методов очень близки по значению, однако не идентичны, это говорит о том, что построенная модель является точной и вероятно не содержит ошибок в работе.

Также теперь мы можем выяснить какие показатели наибольшим образом влияют на успеваемость ученика по предмету математика (см. Приложение 1, Рисунок 15). На графике показан топ наиболее важных показателей определяющие успех ребенка в учебе. Из этого вытекают интересные выводы. Можем прийти к выводу, что наиболее успешным по этому предмету будет мальчик, у которого неплохие оценки по другим предметам, а именно по письму, питающийся стандартным обедом и относящийся к этнической группе E. Что удивило, так это отсутствие хорошей подготовки к тесту, но, вероятно, хорошему ученику это не обязательно. Также было выяснено, что образование родителя практически не влияет на результаты ребенка, главное — это его собственное старание и хорошая обстановка в семье.

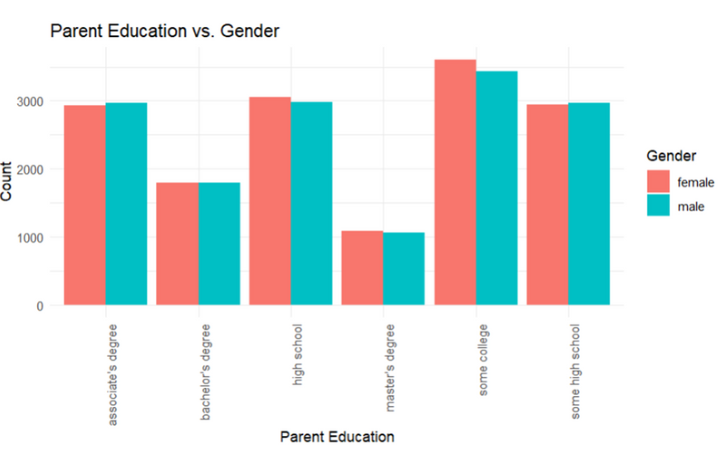
**Вывод**

В ходе реализации проекта была проделана обширная работа по выявлению факторов, наиболее благоприятно влияющих на успеваемость ученика. Были использованы различные методы для достижения результата, такие как разведочный анализ данных, а также построение разного рода предсказывающих моделей. Мы получили интересные, и возможно неожиданные результаты. Как оказалось на успеваемость ребенка влияет его собственное рвение к учебе, а также благополучие в семье. К сожалению, не смогли разобрать каждый из возможных предметов, однако проделанная работа с показателями по математике заслуживает своего внимания. В ходе выполнения были получены высокие результаты точности модели, а также были исключены возможные ошибки в реализации благодаря сравнению с другими методами. Цель проекта была достигнута, а результаты могут помочь организовать учебный процесс ребенка наиболее правильным и комфортным для него образом.

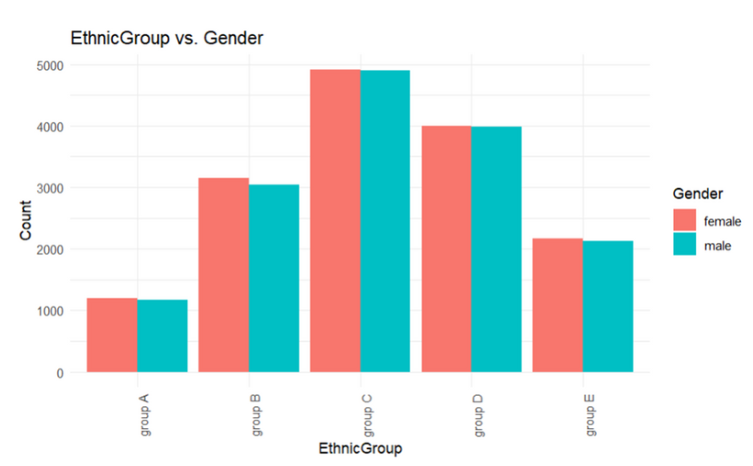
**Приложение 1:**

****

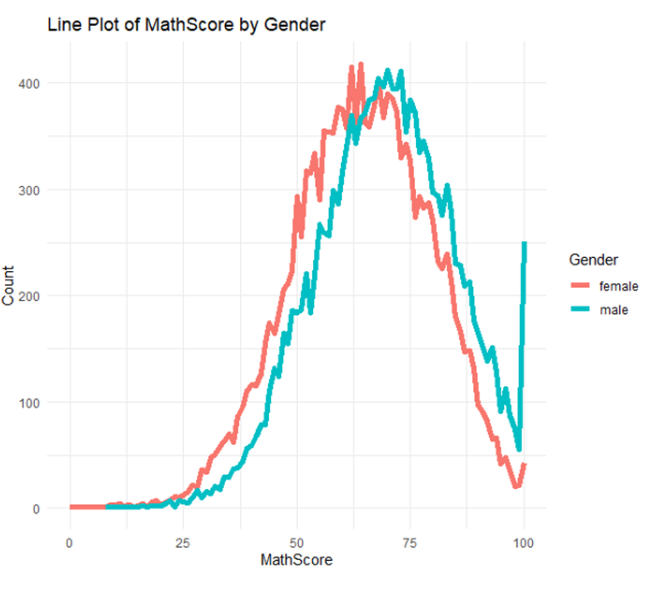
**Рисунок 1** Соотношение полов в датасете



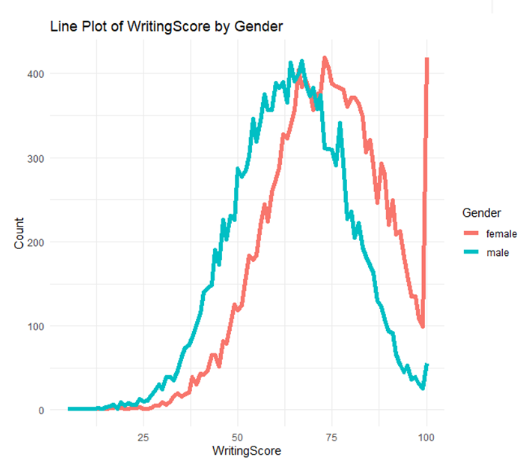
**Рисунок 2** Уровень образования родителей учеников



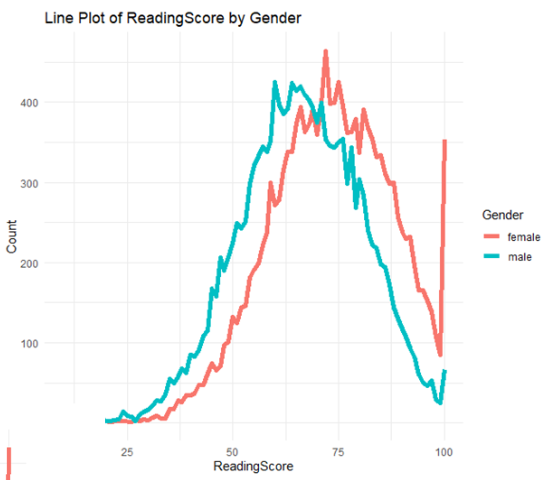
**Рисунок 3** Соотношение этнических групп



**Рисунок 4** Разброс оценок по математике



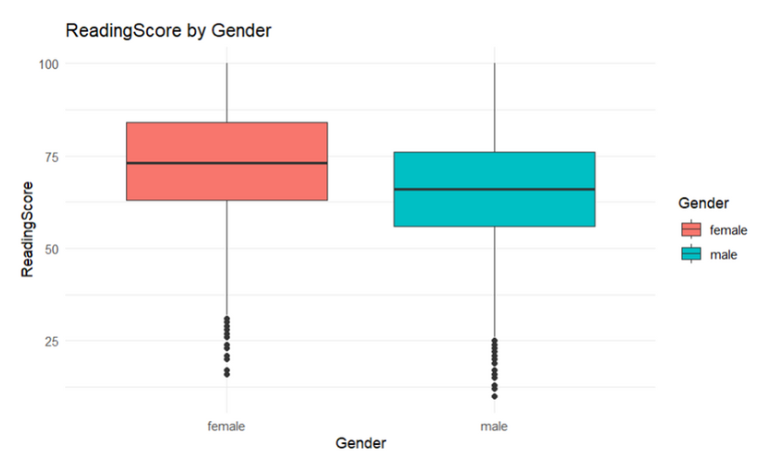
**Рисунок 5** Разброс оценок по письму



**Рисунок 6** Разброс оценок по чтению



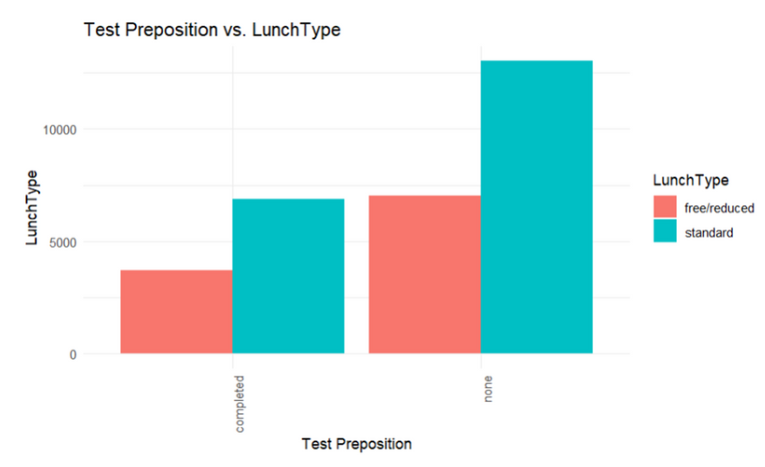
**Рисунок 7** Сравнение средней оценки по математике



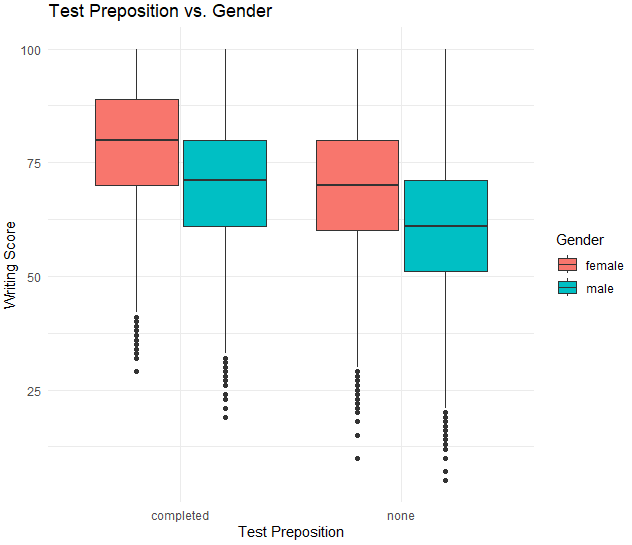
**Рисунок 8** Сравнение средней оценки по чтению



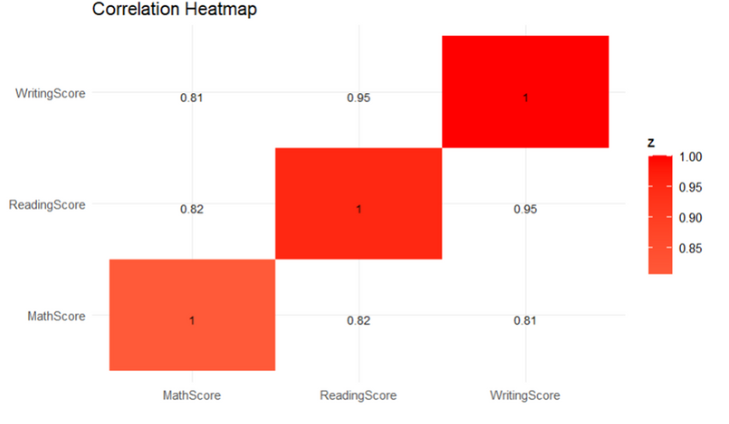
**Рисунок 9** Сравнение средней оценки по письму



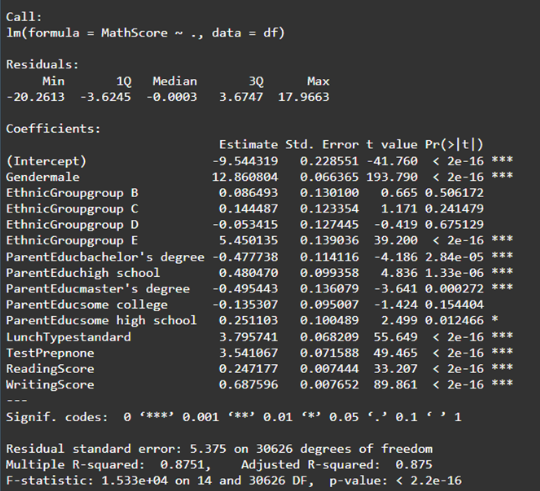
**Рисунок 10** Зависимость типа обеда и подготовки к экзаменам



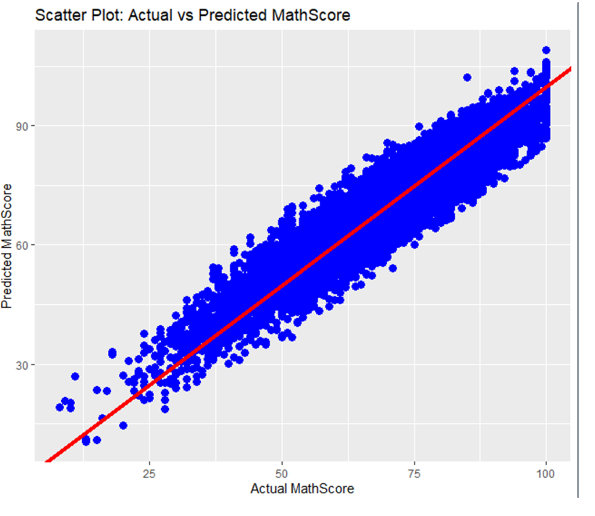
**Рисунок 11 Уровень** подготовки



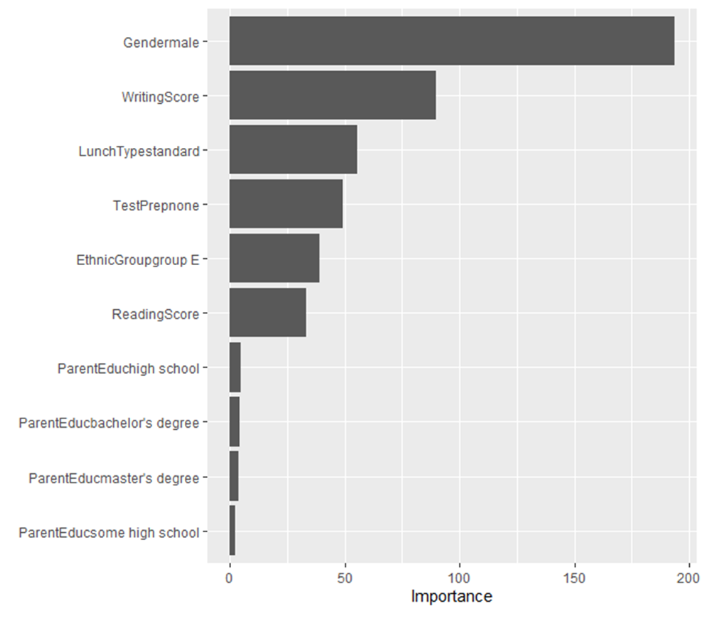
**Рисунок 12** Корреляция



**Рисунок 13** Регрессионная модель



**Рисунок 14** График линейной регрессии



**Рисунок 15** Важность переменных

**Библиография**

1. Kapatsa, H. (2023, 15 декабря). Разведочный анализ данных (EDA), ч. 1. Взято из <https://mlaccessible.medium.com/eda-в-машинном-обучении-простыми-словами-ч-1-663fd5231196>
2. Акод (2023, 15 декабря). Введение в язык программирования R. Взято из <https://ravesli.com/introduction-r/>
3. Алла Тамбовцева (2023, 15 декабря). Занятие 04: библиотека tidyverse. Взято из <https://api.rpubs.com/AllaT/r-tidyverse>
4. Berkan Cetinkaya (2023, 15 декабря). Statistics with R – Roadmap. Взято из <https://medium.com/@berkancetinkaya/statistics-with-r-roadmap-b49792e50d30>
5. Saurav Das (2023, 15 декабря). Correlation plots in R. Взято из <https://medium.com/@saurav12das/correlation-plots-in-r-b392056a2ce>
6. Hannah Frick (2023, 15 декабря). Rsample. Взято из <https://rsample.tidymodels.org/>
7. Donald Knuth (2023, 15 декабря). Introduction to knitr. Взято из <https://sachsmc.github.io/knit-git-markr-guide/knitr/knit.html>
8. Yu-En Hsu (2023, 15 декабря). Create Predictive Classification Models in R with Caret. Взято из <https://yuenhsu.medium.com/create-predictive-classification-models-in-r-with-caret-19a83c1b742>
9. Brandon M. Greenwell, Brad Boehmke (2023, 15 декабря). vip: Variable Importance Plots. Взято из <https://koalaverse.github.io/vip/>
10. Said Bleik, Shaheen Gauher (2023, 15 декабря). Computing Classification Evaluation Metrics in R. Взято из <https://blog.revolutionanalytics.com/2016/03/com_class_eval_metrics_r.html>
11. Prashant Nair (2023, 15 декабря). Creating Dummy Variables for Categorical Data in R programming. Взято из <https://medium.com/@prashant.nair2050/creating-dummy-variables-for-categorical-data-in-r-programming-c2afe5d1a4ac>
12. Martinqiu (2023, 15 декабря). Easy Implementation of Dummy Coding/One-Hot Coding in R. Взято из <https://medium.com/codex/easy-implementation-of-dummy-coding-one-hot-coding-in-r-42c0486f8996>
13. Gennady Nemchin (2023, 15 декабря). Как работает линейная регрессия. Взято из <https://gennadynemchin.medium.com/how-to-linear-regression-works-2d383c3875f>
14. Shubhang Agrawal (2023, 15 декабря). RANDOM FOREST. Взято из <https://medium.com/swlh/random-forest-ac5227dabb08>
15. Indriani Sitorus (2023, 15 декабря). Support Vector Machine (SVM) and Kernels Trick. Взято из <https://medium.com/analytics-vidhya/introduction-to-svm-and-kernel-trick-part-1-theory-d990e2872ace>