

Применение методов машинного обучения для анализа и прогнозирования ответов на опрос «Как много времени люди проводят за ПК, консолях, телефонах?»

Неретин Андрей Сергеевич

Приамурский государственный университет им. Шолом-Алейхема

Студент

Никифорова Людмила Евгеньевна

Приамурский государственный университет им. Шолом-Алейхема

Студент

Акчулпанов Дмитрий Русланович

Приамурский государственный университет им. Шолом-Алейхема

Студент

Аннотация

Цель данного исследования заключается в использовании различных методов машинного обучения для создания модели, которая позволит прогнозировать, сколько времени люди проводят за гаджетами и как они к этому относятся. В ходе работы была выбрана наиболее эффективная модель и переменные, обеспечивающие наилучшие результаты для поставленной задачи. Сравнение моделей проводилось на основе ключевых метрик: AUC, точности (CA), F1, точности (Precision) и полноты (Recall).

Ключевые слова: Социологический опрос, Google форма, данные, диаграмма, Orange, модель машинного обучения, искусственный интеллект.

The use of machine learning methods to analyze and predict responses to the survey "How much time do people spend on PCs, consoles, and phones?"

Neretin Andrey Sergeevich

Sholom-Aleichem Priamursky State University

Student

Nikiforova Ludmila Evgenevna

Sholom-Aleichem Priamursky State University

Student

Akchulpanov Dmitry Ruslanovich

Sholom-Aleichem Priamursky State University

Student

Abstract

The purpose of this study is to use various machine learning methods to create a model that will allow predicting how much time people spend on gadgets and how they feel about it. In the course of the work, the most effective model and variables were selected that provide the best results for the task. The models were compared based on key metrics: AUC, Accuracy (CA), F1, Precision and Recall.

Keywords: Sociological survey, Google form, data, chart, Orange, machine learning model, artificial intelligence.

1 Введение**1.1 Актуальность**

С помощью машинного обучения можно сделать обоснованные прогнозы относительно изменений в восприятии возраста, необходимого для создания семьи, а также глубже понять факторы, влияющие на эти предпочтения. Это может стать ценным инструментом для социологов, демографов и психологов, а также для организаций, занимающихся вопросами планирования семьи и предоставления социальных услуг.

1.2 Обзор исследований

О.И. Китаева изучила концепцию интеллектуального анализа образовательных данных и провела анализ на основе информации, собранной по учебной дисциплине в высшем учебном заведении [1]. А.А. Чернышев в своем исследовании описал методы оценки качества моделей машинного обучения, включая оценку на обучающих данных, процентное разделение и кросс-валидацию, а также подчеркнул важность начального значения рандомизации [2]. Н. Юсупов, А. Савельева и О.Г. Леонова исследовали применение методов классификации в программе Orange, используя реальную базу данных [3]. Р.Э. Яхшибоев, Н.М. Апсилям и Л.Р. Шамсудинова посвятили свое исследование анализу современных подходов к моделированию механизмов искусственного интеллекта, включая как традиционные методы машинного и глубокого обучения, так и инновационные подходы, основанные на последних достижениях в нейронауке и когнитивной психологии [4].

1.3 Цель исследования

Целью исследования является применение различных методов машинного обучения для построения модели с целью прогнозирования «Как много времени люди проводят за ПК, консолях, телефонах?».

2. Материалы и методы

В данном опросе приняло участие сто пять респондентов мужского и женского пола в возрасте от 14 лет. Для реализации создан опрос в google форме.

3.Результаты и дискуссии

Для того чтобы увидеть ответы респондентов в более развернутом виде, нужно построить диаграмму на основании этих данных. Первая диаграмма показывает количество респондентов и их пол (рис.1).

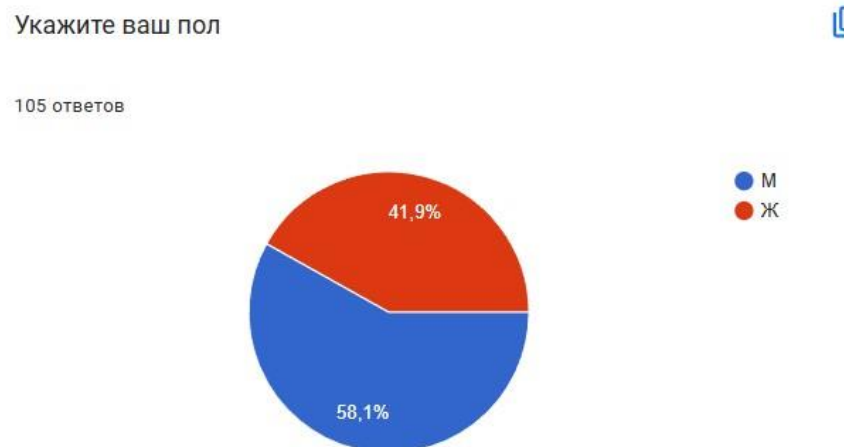


Рисунок 1. Количество респондентов и их пол

Вторая диаграмма показывает считают ли респонденты, что гаджеты стали неотъемлемой частью их жизни. По ней видно, что 92.2% людей дали положительный ответ (рис.2).

Считаете ли вы, что гаджеты стали неотъемлемой частью нашей жизни?

103 ответа

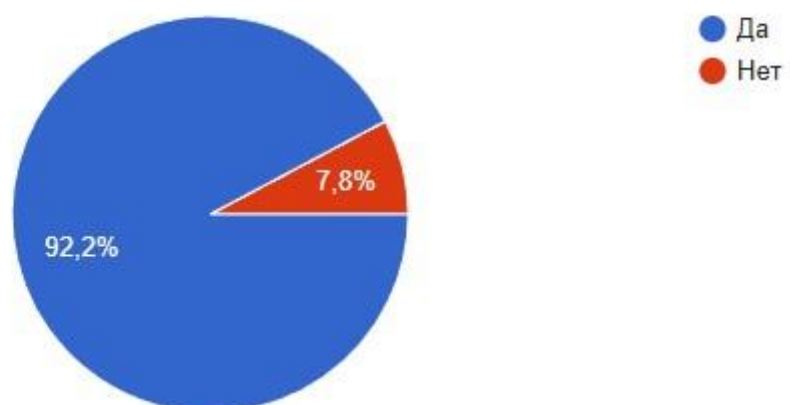


Рисунок 2. Респонденты, считающие гаджеты неотъемлемой частью их жизни

Третья диаграмма показывает количество времени, проводимого за гаджетами в день. По ней видно, что 41% людей проводят 4-6 часов за гаджетами, 30.5% 1-4 часа, 22.9% более 6 часов, 5.6% менее 1 часа(рис.3).

Сколько времени в день в среднем вы проводите за гаджетами?

 Копировать

105 ответов

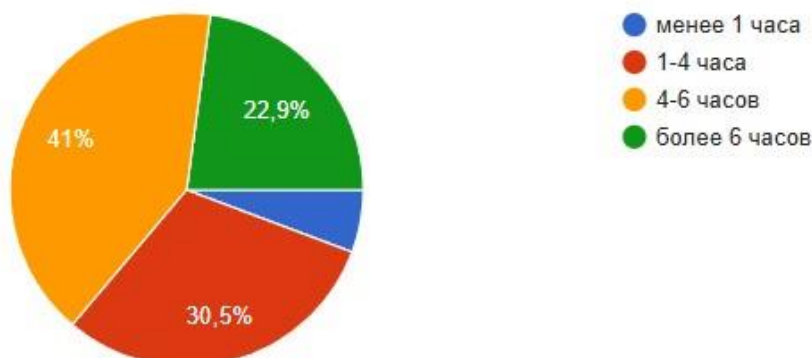


Рисунок 3. Количество времени, проводимого за гаджетами в день

Четвертая диаграмма показывает цель, с которой сидят в гаджетах. По ней видно, что 60% людей проводят за играми, 52.4% за работой, 37.1% за мессенджерами, 32.4% за чтением, 31.4% за просмотром фильмов и сериалов, 17.1% за поиском информации (рис.4).

С какой целью вы сидите в гаджетах?

 Копировать диаграмму

105 ответов

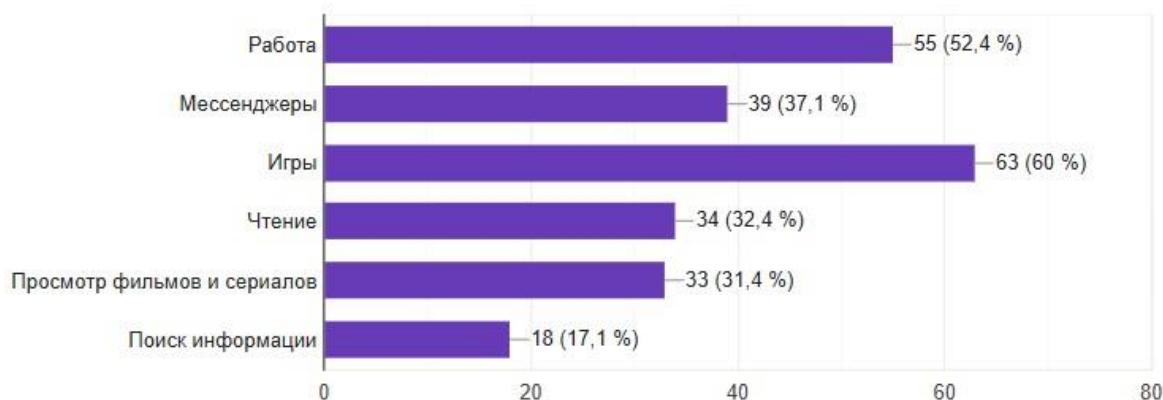


Рисунок 4. Цель, с которой сидят в гаджетах

Пятая диаграмма показывает хотят ли респонденты проводить меньше времени за гаджетами. По ней видно, что 86.5% не хотят этого (рис.5).

Хотели бы вы проводить меньше времени за гаджетами?



104 ответа

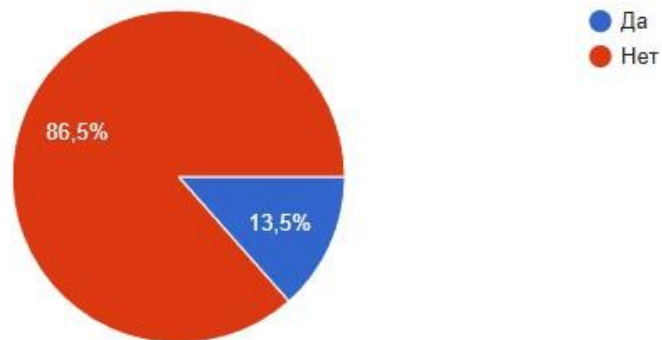


Рисунок 5. Хотят ли респонденты проводить меньше времени за гаджетами

Шестая диаграмма показывает мнение респондентов о том, влияют ли гаджеты на здоровье. По ней видно, что 78.8% людей отрицают это (рис.6).

Как вы думаете гаджеты влияют на ваше здоровье?

104 ответа

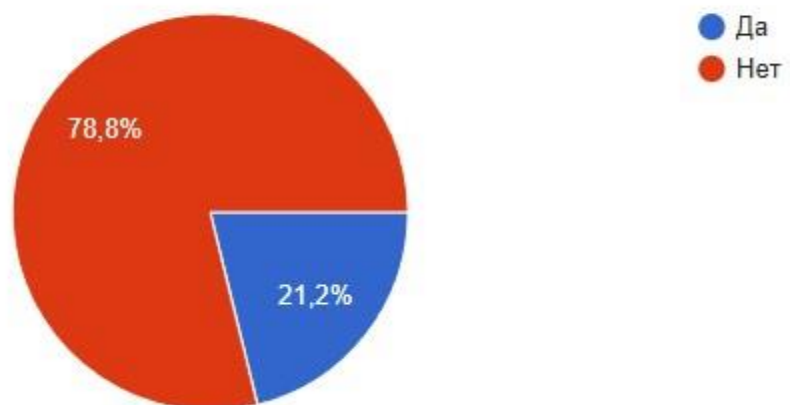


Рисунок 6. Мнение респондентов о том, влияют ли гаджеты на здоровье

На основе данных опроса необходимо рассчитать корреляции в Orange и создать модели машинного обучения. В настоящее время Orange представляет собой мощную и интуитивно понятную платформу для визуального анализа данных и машинного обучения. Также требуется создать новый файл (рис.7).

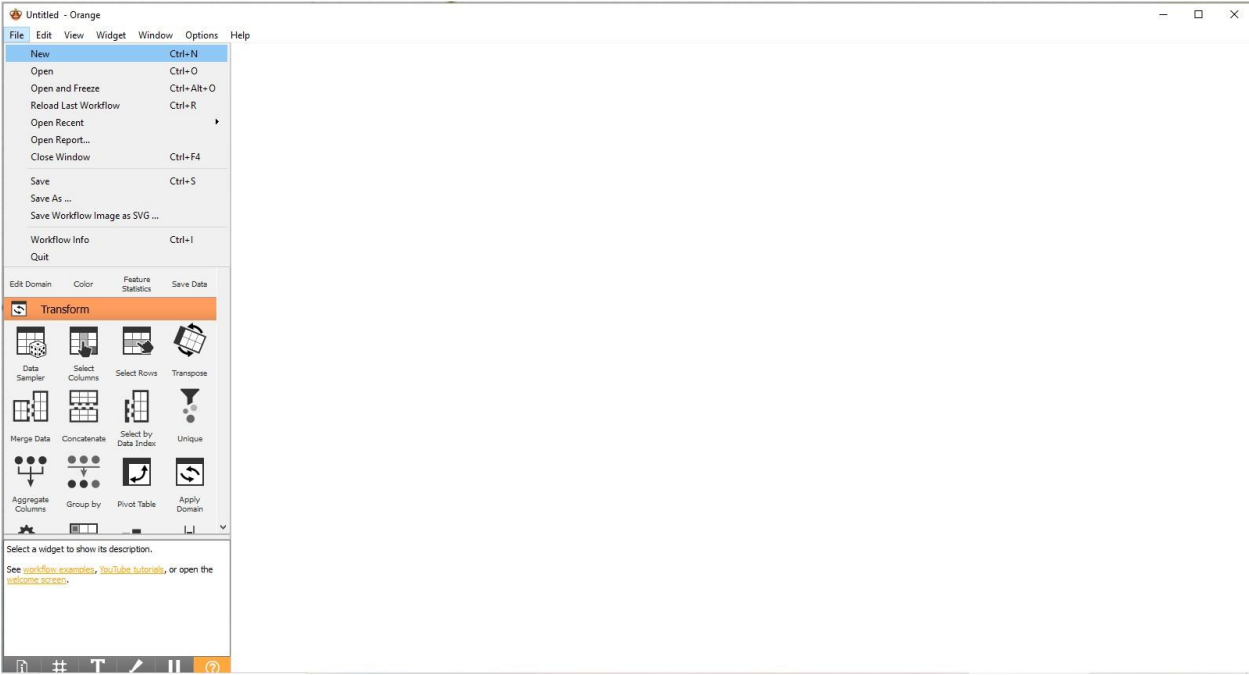


Рисунок 7. Создание нового файла модели в Orange

Чтобы построить диаграмму для каждого вопроса опроса, необходимо «закодировать» результаты. Будет более эффективно, если варианты ответов и сами вопросы будут представлены в виде чисел и букв (рис.8).

Укажите ваш пол	Считаете ли вы, что гаджеты стали неотъемлемой частью жизни?	Сколько времени в день в среднем вы сидите в гаджетах?	С какой целью вы сидите в гаджетах?	Хотели бы вы проводить меньше времени в гаджетах?	Как вы думаете гаджеты влияют на ваше здоровье?
Ж	Да	менее 1 часа	Мессенджеры	Нет	Да
Ж	Да	1-4 часа	Мессенджеры, Просмотр фильмов и сериалов	Нет	Да
Ж	Нет	1-4 часа	Чтение, Поиск информации	Нет	Да
Ж	Да	4-6 часов	Работа, Мессенджеры, Чтение	Нет	Да
Ж	Нет	4-6 часов	Работа, Мессенджеры	Нет	Нет
Ж	Да	менее 1 часа	Поиск информации	Нет	Нет
Ж	Да	4-6 часов	Игры	Нет	Да
Ж	Да	4-6 часов	Работа	Нет	Нет
Ж	Да	4-6 часов	Работа, Игры	Нет	Нет
Ж	Да	1-4 часа	Работа, Игры	Нет	Нет
Ж	Да	более 6 часов	Работа, Игры	Нет	Нет
Ж	Да	1-4 часа	Мессенджеры, Просмотр фильмов и сериалов	Нет	Нет
Ж	Да	1-4 часа	Работа	Нет	Нет

Рисунок 8. Незакодированные данные

После того как данные опроса «закодированы», таблица выглядит следующим образом (рис.9).

q1	q2	q3	q4	q5	q6
2	1	3	Работа, Мессенджеры	1	2
2	1	3	Мессенджеры, Чтение, Просмотр филь	2	2
2	1	4	Работа, Мессенджеры, Поиск информ	2	2
2	1	1	Мессенджеры	2	1
2	1	2	Мессенджеры, Просмотр фильмов и се	2	1
2	2	2	Чтение, Поиск информации	2	1
2	1	3	Работа, Мессенджеры, Чтение	2	1
2	2	3	Работа, Мессенджеры	2	2
2	1	1	Поиск информации	2	2
2	1	3	Игры	2	1
2	1	3	Работа	2	2

Рисунок 9. Закодированные данные

Расшифровка закодированных вопросов (таб.1).

Таблица 1. Расшифровка закодированных вопросов

Укажите ваш пол	q1
Считаете ли вы, что гаджеты стали неотъемлемой частью нашей жизни?	q2
Сколько времени в день в среднем вы проводите за гаджетами?	q3
С какой целью вы сидите в гаджетах?	q4
Хотели бы вы проводить меньше времени за гаджетами?	q5
Как вы думаете гаджеты влияют на ваше здоровье?	q6

Расшифровка закодированных ответов (таб.2).

Таблица 2. Расшифровка закодированных ответов

	1	2	3	4
q1	М	Ж	-	-
q2	Да	Нет	-	-
q3	менее 1 часа	1-4 часа	4-6 часов	более 6 часов
q5	Да	Нет	-	-
q6	Да	Нет	-	-

В новом открывшемся файле необходимо добавить данные, которые уже были предварительно «закодированы» в числовом формате.

После добавления файла в открывшемся окне видно, что все вопросы представлены в категориальном формате. Вопрос, по которому будет производиться прогнозирование, имеет роль «target», в то время как все остальные вопросы обозначены как «feature», один вопрос представлен как «text» (рис 10).

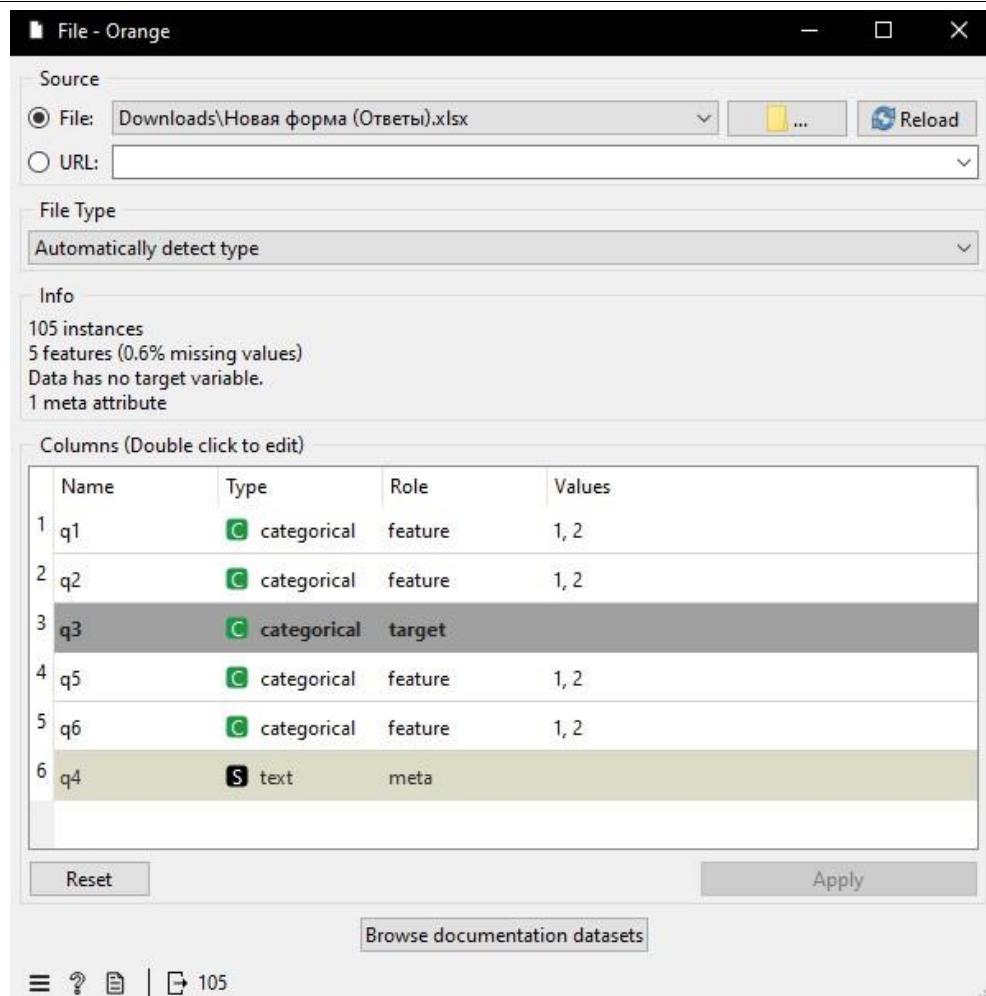


Рисунок 10. Вопросы, представленные в категориальном формате, один вопрос представлен как «text»

Далее к новому файлу добавить модуль data table. Модуль data table отвечает за отображение и манипуляцию данными в табличном формате (рис.11).

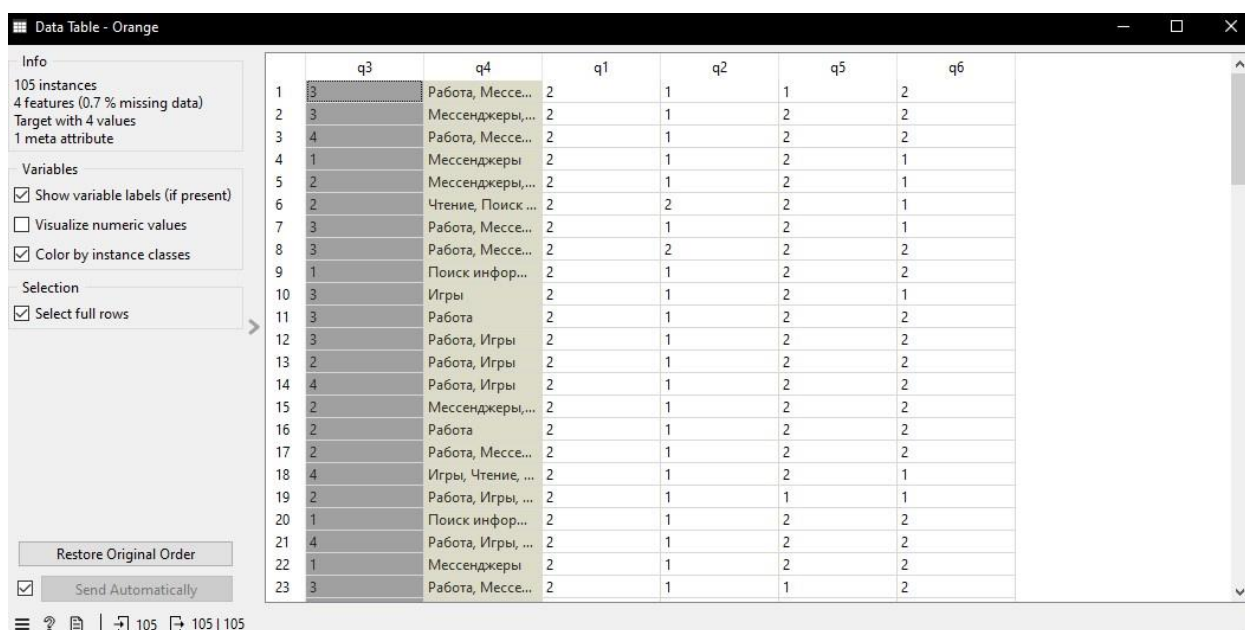


Рисунок 11. Отображение данных в табличном формате

Затем необходимо добавить модуль distributions. В Orange, библиотеке для визуализации и анализа данных, этот модуль предназначен для изучения распределений переменных в наборе данных (рис.10).

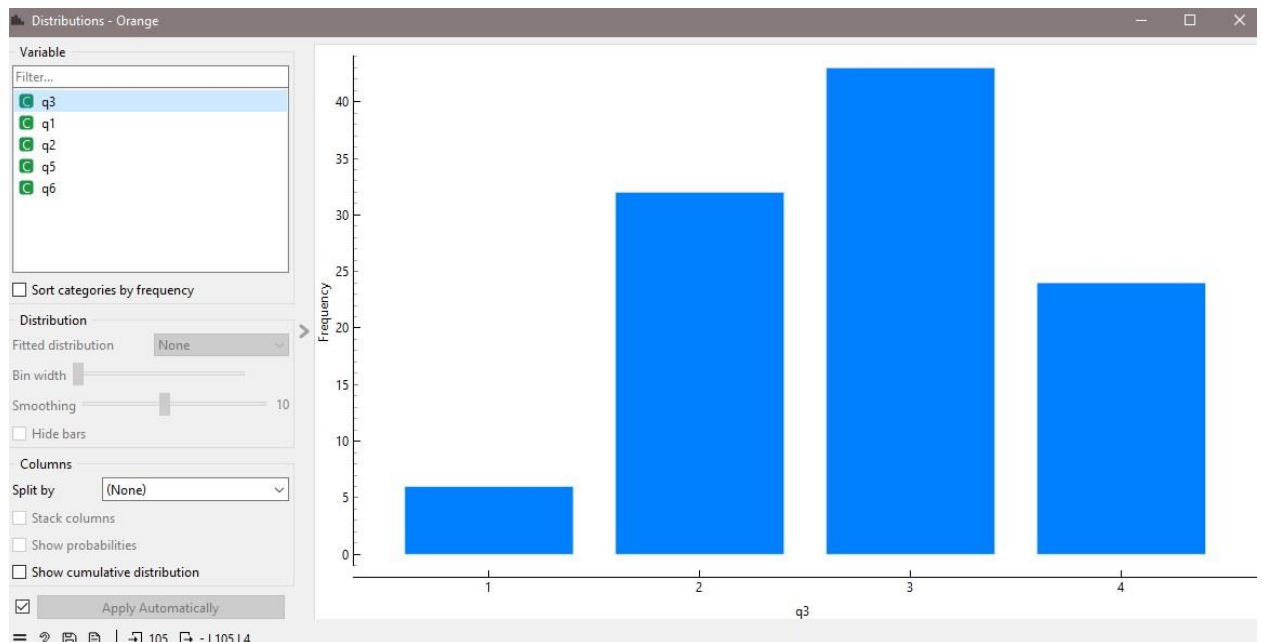


Рисунок 12. Изучение распределений переменных в наборе данных

Далее следует добавить модуль correlations. Этот модуль в Orange предназначен для анализа взаимосвязей между переменными в наборе данных. Он также является полезным инструментом для оценки и визуализации этих взаимосвязей, что может оказать помощь в анализе и принятии решений (рис.13).

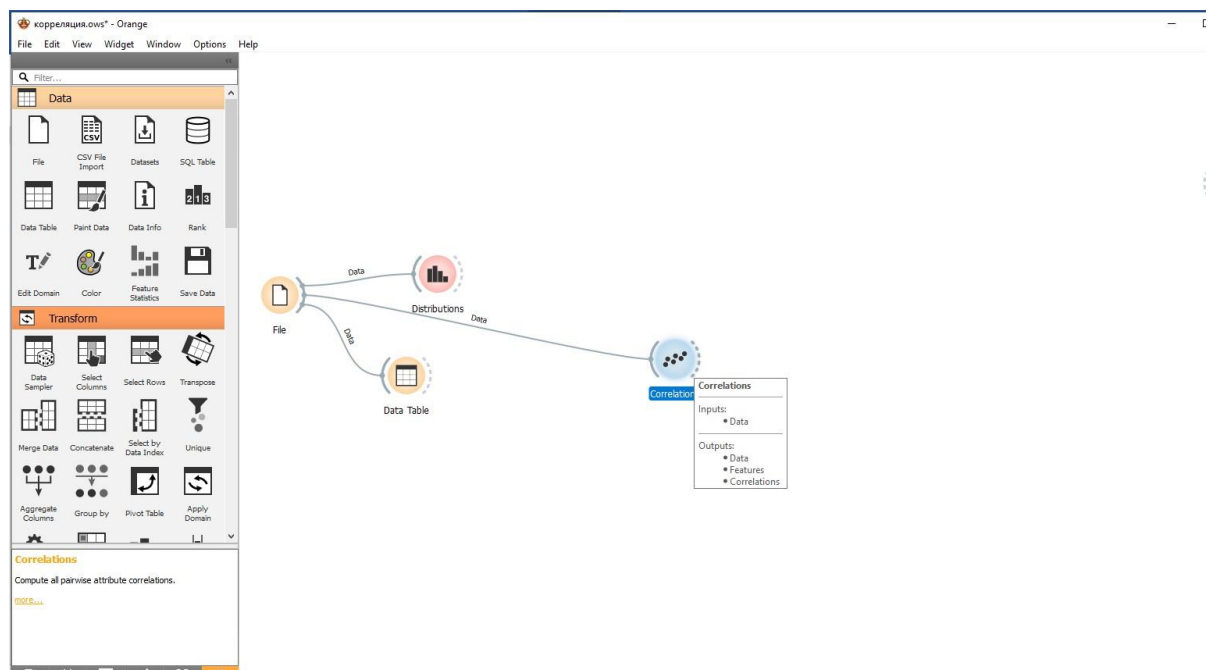
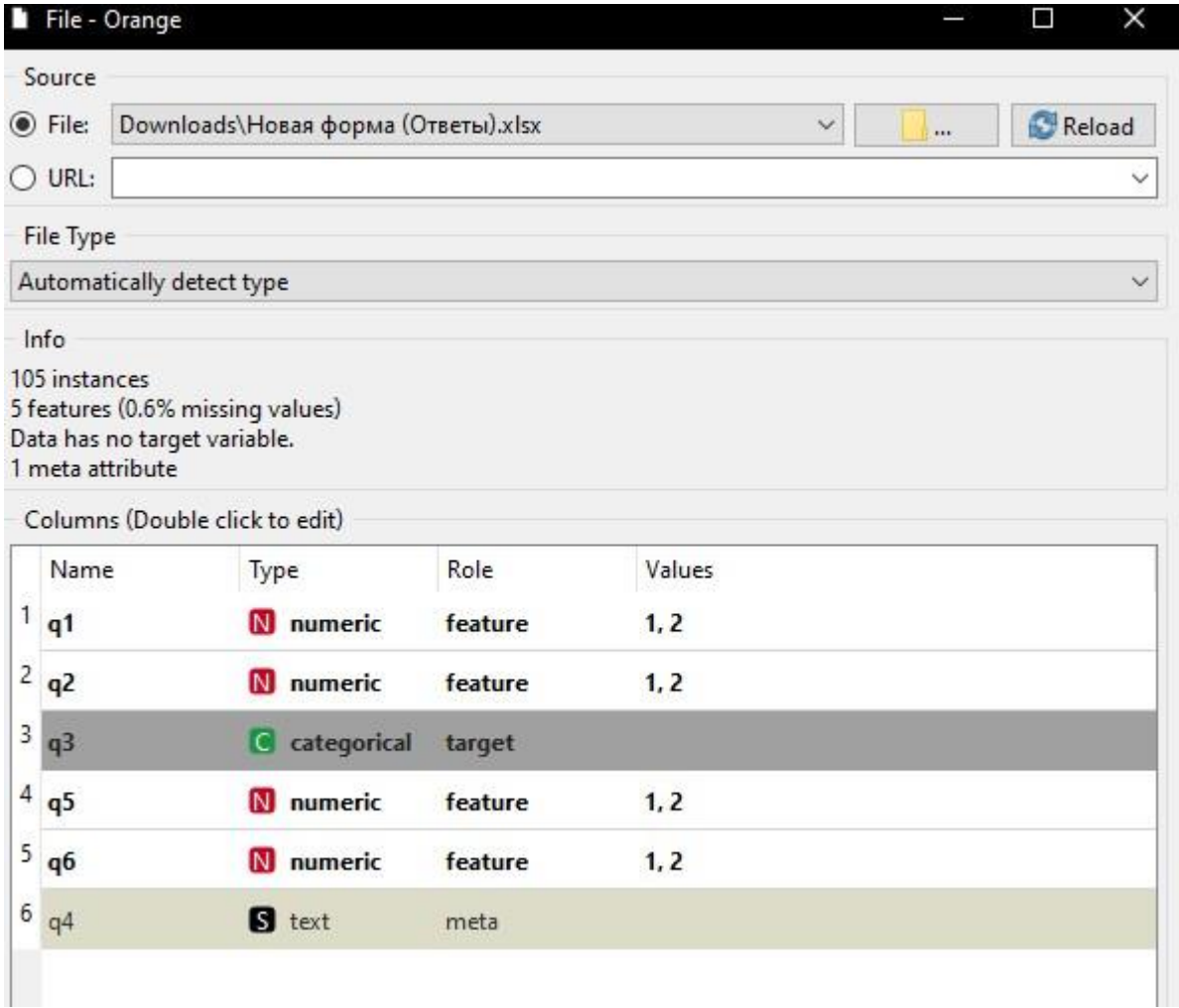


Рисунок 13. Анализ взаимосвязей между переменными в наборе данных

Анализируя данные модуля корреляций, можно выделить два метода: корреляцию Спирмена и корреляцию Пирсона. Метод коэффициента корреляции Спирмена оценивает степень и направление монотонной связи между двумя переменными, в то время как коэффициент корреляции Пирсона измеряет линейную зависимость между двумя непрерывными переменными. В данном случае данные, полученные с помощью любого из методов, указывают на корреляционную зависимость между вопросами. Для расчета корреляции по методу Спирмена необходимо преобразовать все вопросы в числовой формат (numeric), оставив прежним тот вопрос, по которому будет выполняться прогнозирование (рис. 14) (рис.15).



The screenshot shows the Orange data mining software interface. The 'Source' tab is selected, with a file path 'Downloads\Новая форма (Ответы).xlsx' entered. The 'File Type' is set to 'Automatically detect type'. The 'Info' section displays: '105 instances', '5 features (0.6% missing values)', 'Data has no target variable.', and '1 meta attribute'. The 'Columns' section shows a table with 6 columns: Name, Type, Role, and Values.

	Name	Type	Role	Values
1	q1	N numeric	feature	1, 2
2	q2	N numeric	feature	1, 2
3	q3	C categorical	target	
4	q5	N numeric	feature	1, 2
5	q6	N numeric	feature	1, 2
6	q4	S text	meta	

Рисунок 14. Преобразование всех вопросов в числовой формат, оставив прежним тот вопрос, по которому будет выполняться прогнозирование

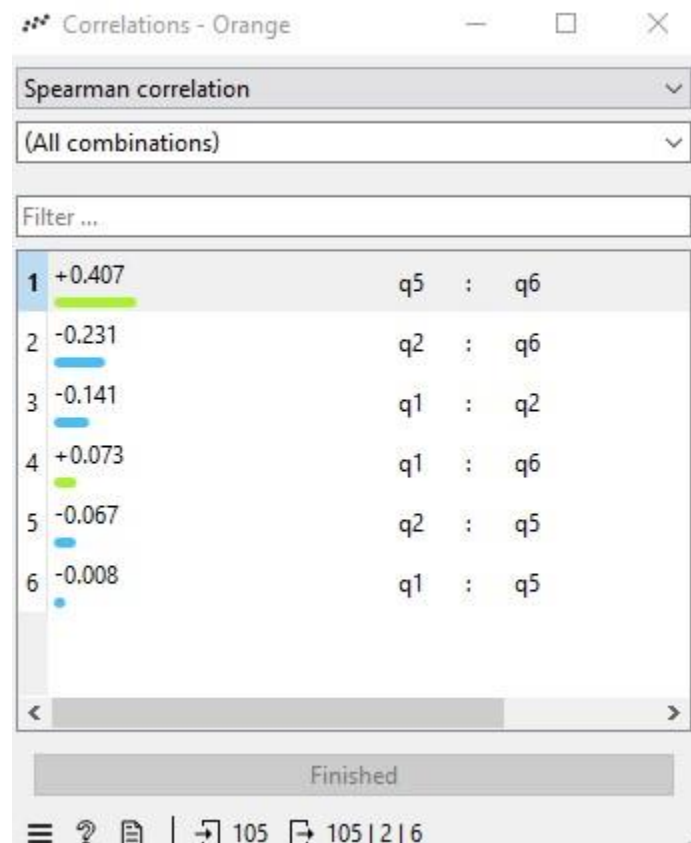


Рисунок 15. Корреляция по Спирману

Следующим шагом будет создание классификационной модели для прогнозирования ответа на вопрос: «Как много времени люди проводят за ПК, консолях, телефонах?». Затем, необходимо добавить модуль "test and score" в Orange. Этот модуль оценивает производительность моделей машинного обучения, позволяя проводить кросс-валидацию, вычислять метрики (такие как точность и полнота), сравнивать модели и визуализировать результаты (рис.16).

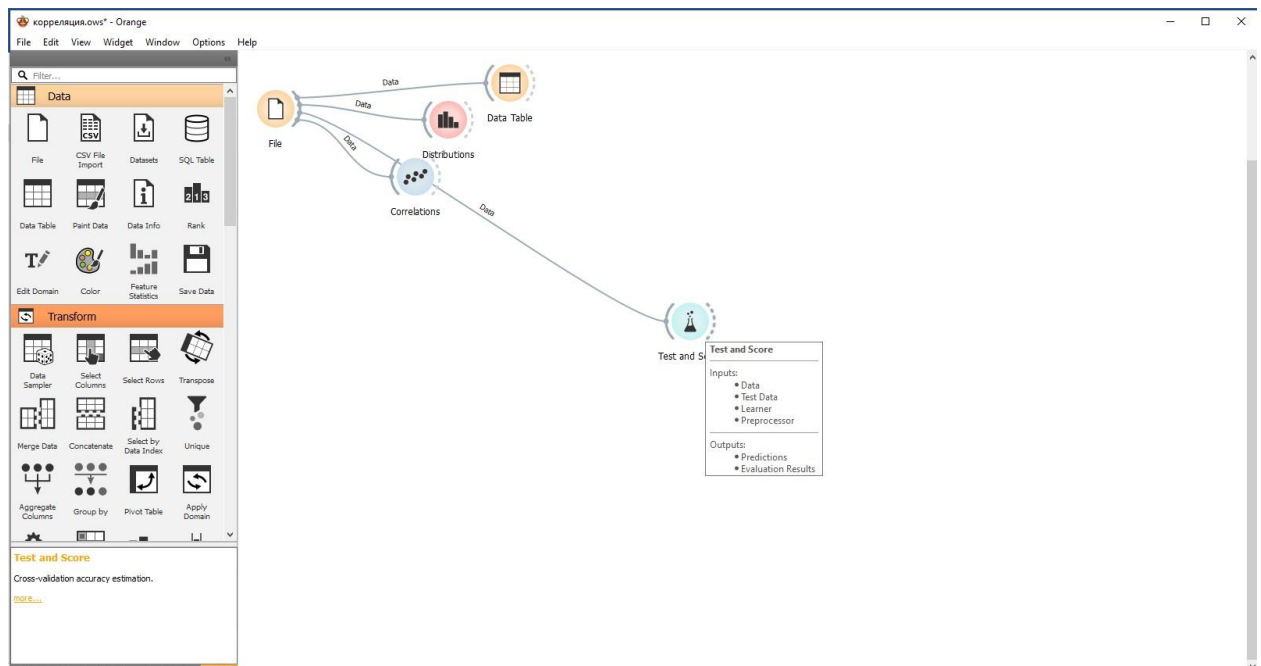
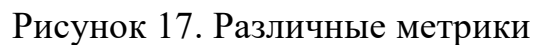


Рисунок 16. Модуль "test and score"

При открытии модуля "test and score" можно увидеть данные по различным метрикам, таким как AUC, CA, F1, Precision и Recall. AUC (Area Under the Curve) показывает, насколько хорошо модель различает классы. Значение AUC варьируется от 0 до 1, где 1 соответствует идеальной модели, а 0.5 — случайной. Чем ближе значение к единице, тем выше вероятность точного предсказания. CA (Classification Accuracy) представляет собой долю правильно классифицированных примеров среди всех. F1 Score — это гармоническое среднее между точностью (Precision) и полнотой (Recall). Precision (точность) измеряет долю правильно предсказанных положительных примеров среди всех предсказанных положительных, в то время как Recall (полнота) показывает долю правильно предсказанных положительных примеров среди всех реальных положительных (рис.17).



Затем необходимо добавить модели для оценки данных: алгоритм kNN (k ближайших соседей) в Orange, который предназначен для классификации и регрессии, основывается на принципе, что схожие объекты располагаются близко друг к другу в пространстве признаков. Также следует включить модуль Naïve Bayes, представляющий собой простой вероятностный классификатор, основанный на теореме Байеса и предположении о независимости признаков. Модель Logistic Regression в Orange служит инструментом для выполнения логистической регрессии, используемой в задачах классификации; она помогает в анализе данных и прогнозировании вероятностей принадлежности объектов к различным классам на основе имеющихся признаков. Наконец, модуль Tree в Orange представляет собой инструмент для построения деревьев решений, который применяется для решения задач классификации и регрессии, а также параметры моделей (рис.18) (рис.19).

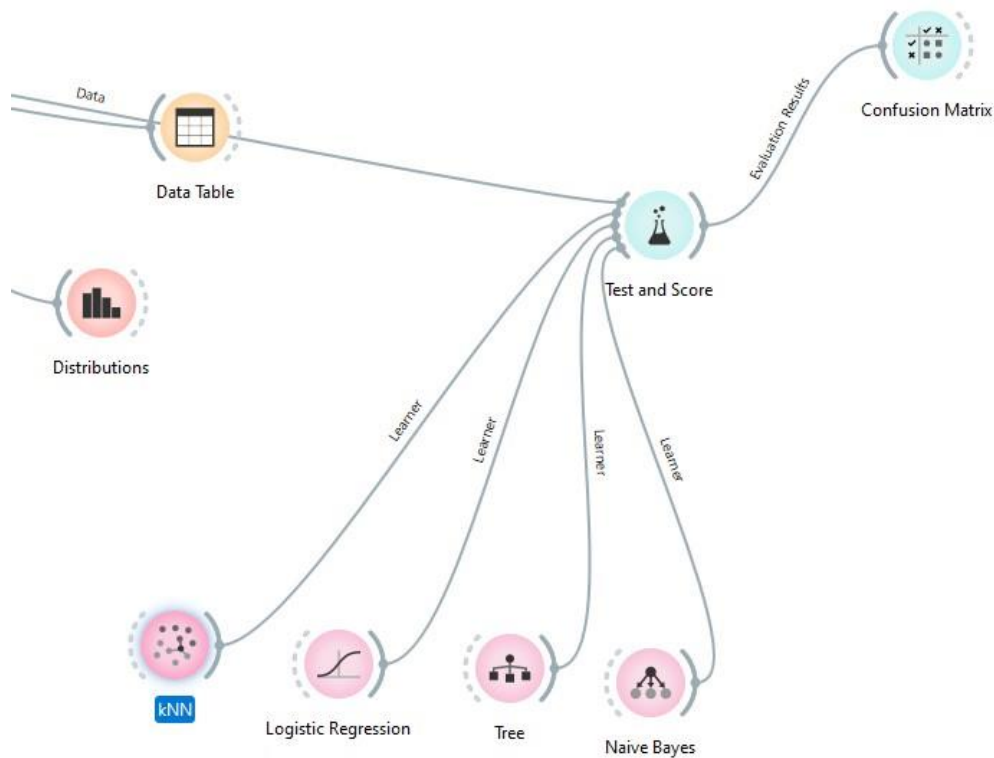


Рисунок 18. Модели для оценки данных

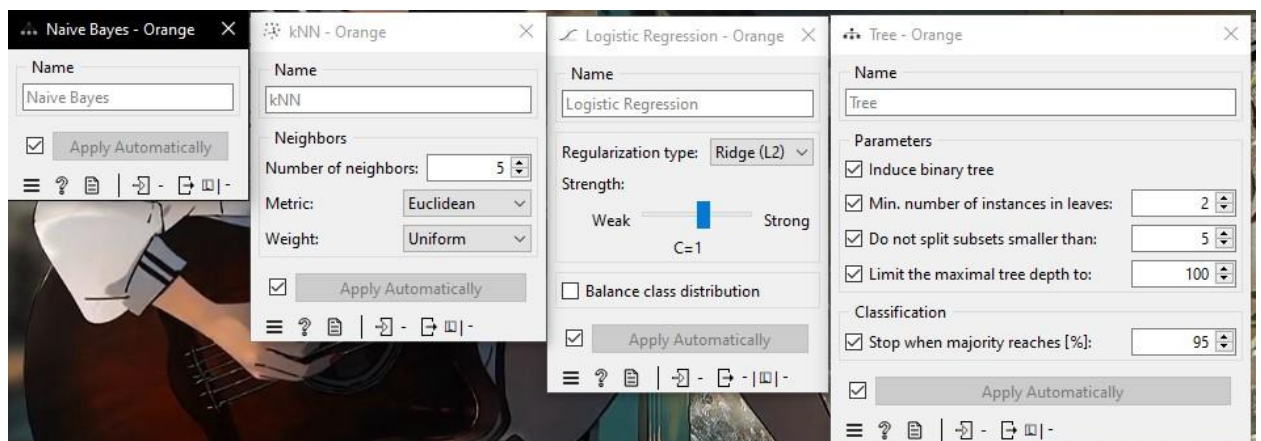


Рисунок 19. Параметры

Кросс-валидация — это метод, используемый для оценки производительности моделей в машинном обучении, который помогает избежать переобучения и предоставляет более надежную оценку качества модели. В данном случае весь набор данных делится на 5 частей, что позволяет по каждой метрике определить, какая модель является обучающей, а какая — тренировочной (рис.20).

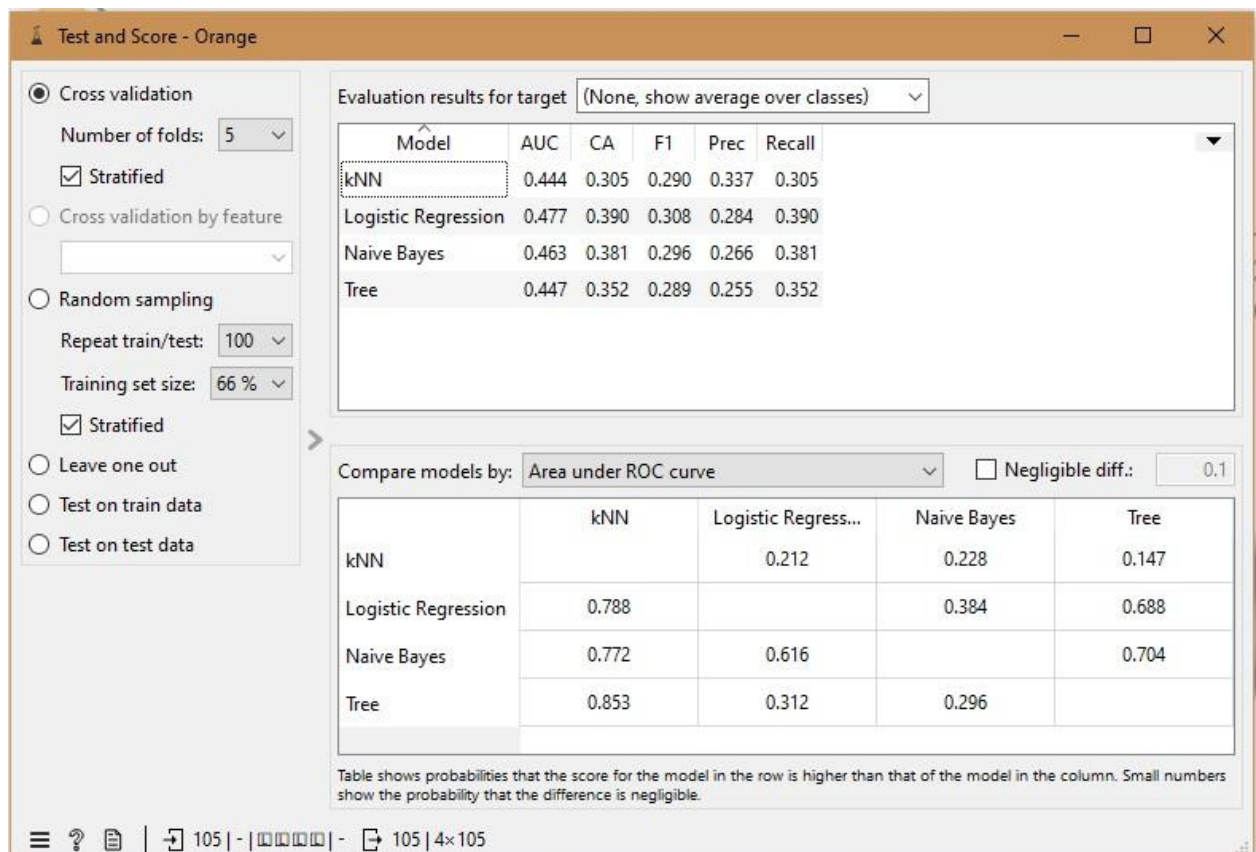


Рисунок 20. Модуль "test and score"

Следующим этапом будет создание модели с использованием матрицы ошибок (confusion matrix), которая служит инструментом для оценки производительности классификационной модели. Она предоставляет наглядное представление о том, насколько эффективно модель выполняет задачу классификации, отображая количество предсказаний для различных моделей, таких как kNN, Naïve Bayes, Logistic Regression и Tree, а также их параметры (рис. 21, рис. 22, рис. 23, рис. 24, рис. 25).

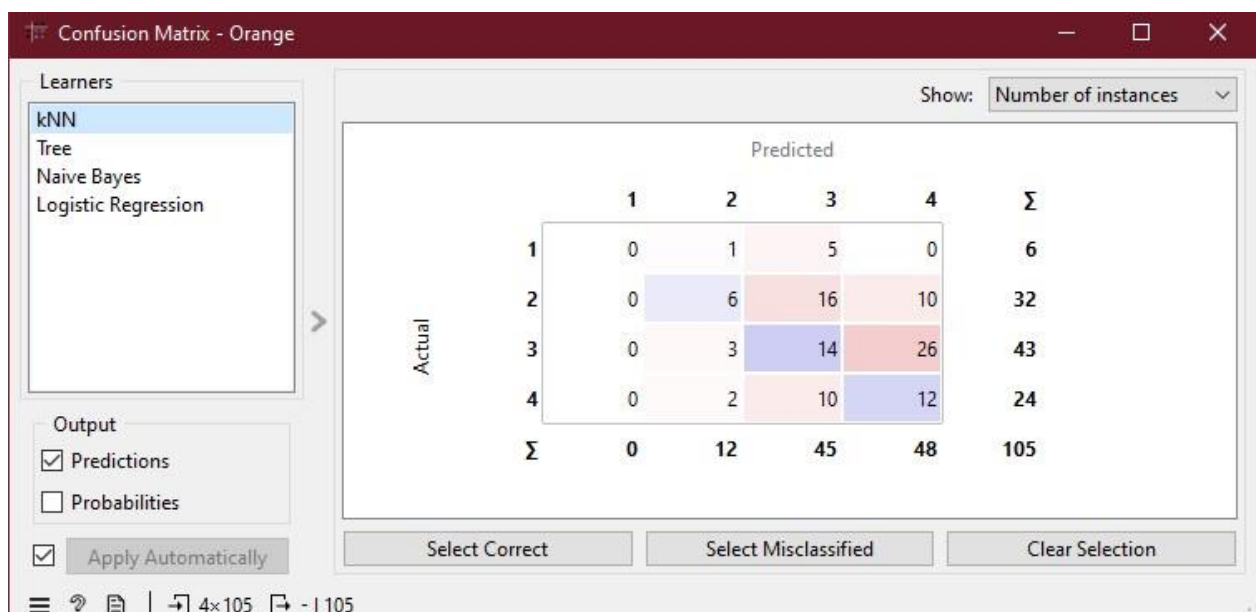


Рисунок 21. Модель kNN

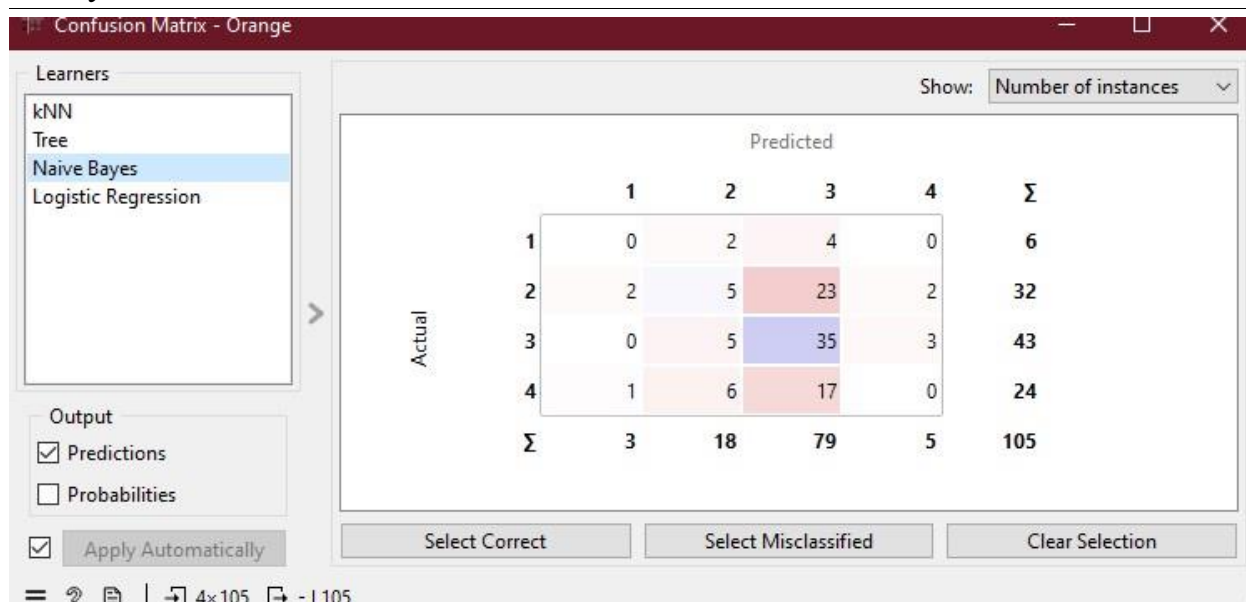


Рисунок 22. Модель Naive Bayes

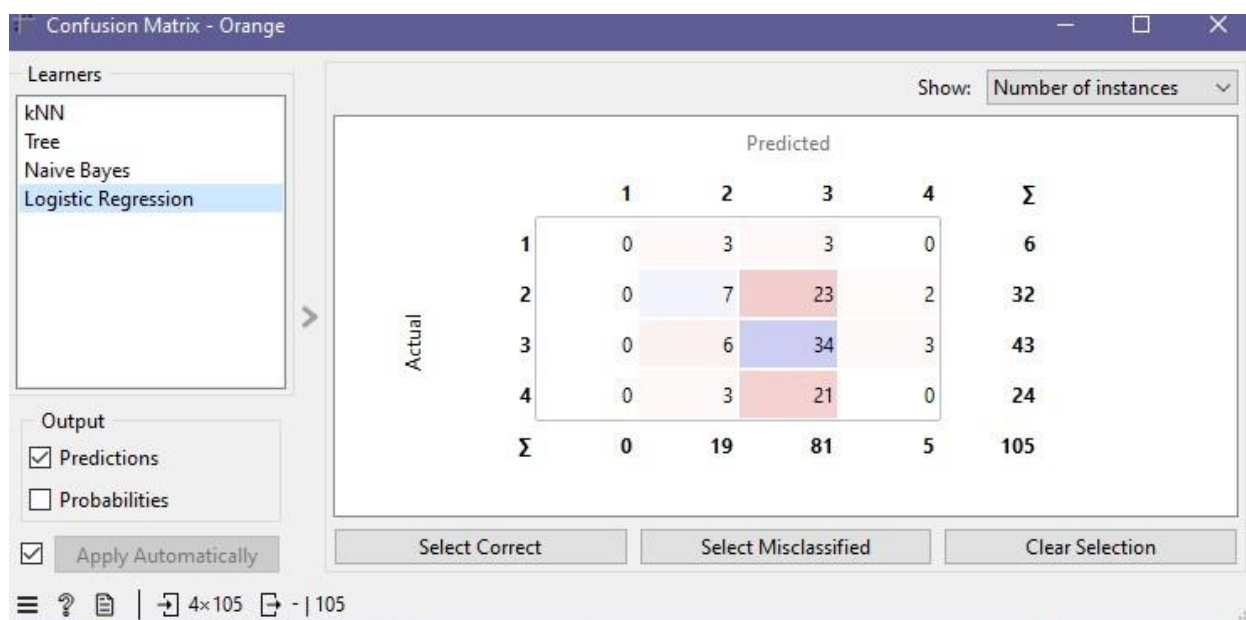


Рисунок 23. Модель Logistic Regression

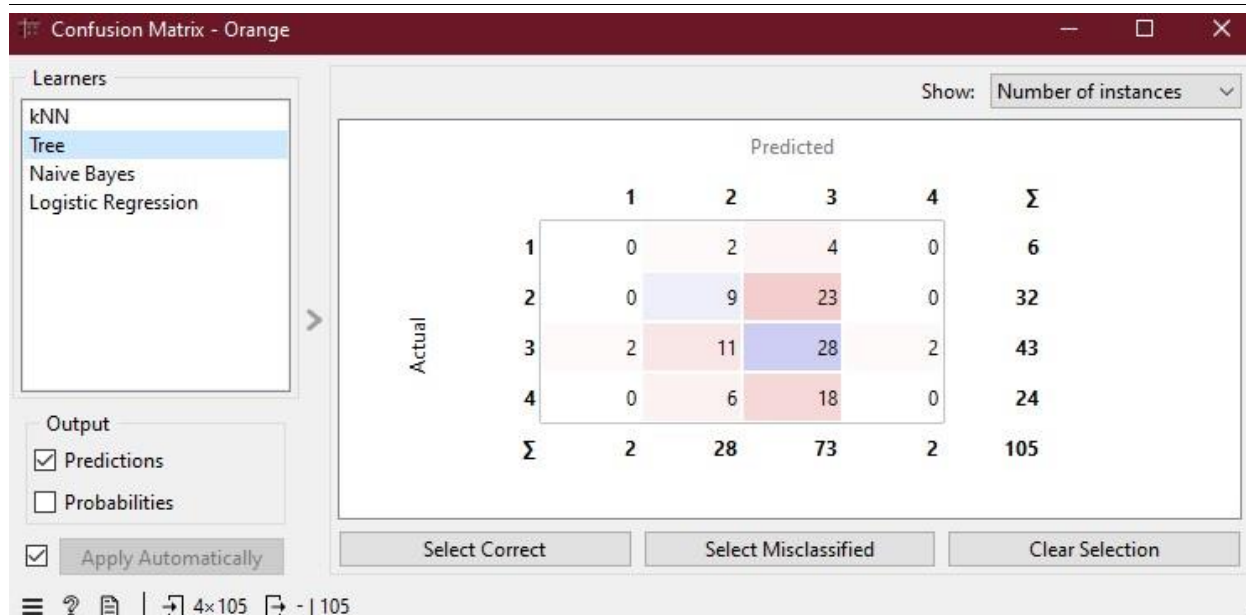


Рисунок 24. Модель Tree

4 Выводы

В данном исследовании была построена модель классификации на вопрос «Как много времени люди проводят за ПК, консолях, телефонах?». По представленным данным в модуле test and score, видно, что модель Logistic Regression по данным метрики Area Under the Curve (которая позволяет сравнивать модели и выбирать наиболее эффективные для решения задач классификации) оказалась более удачной, все практически ее значения ближе к 1, однако для улучшения модели возможно, потребуется дополнительное исследование и настройка параметров.

Библиографический список

1. Китаева О. И. Интеллектуальный анализ образовательных данных учебной дисциплины с использованием программы Orange // Информационные и математические технологии в науке и управлении. 2023. №1. С. 190-200. URL: <https://cyberleninka.ru/article/n/intellektualnyy-analiz-obrazovatelnyh-dannyh-uchebnoy-distipliny-s-ispolzovaniem-programmy-orange>
2. Чернышев А. А. Методы оценки качества моделей машинного обучения // Наука и просвещение. 2023. №34. С. 34-37. URL: <https://naukaip.ru/wp-content/uploads/2023/07/MK-1771-1.pdf#page=34>
3. Юсупов Н., Савельева А., Леонова О. Г. Исследование методов классификации в программе Orange // Молодежная школа-семинар по проблемам управления в технических системах имени АА Вавилова. 2020. Т.1. С. 27. URL: https://vavilovschool.etu.ru/assets/files/2020/sbornik_2020.pdf#page=28
4. Яхшибоев Р. Э., Апислям Н. М., Шамсудинова Л. Р. Моделирование

- механизмов искусственного интеллекта // Innovations in Science and Technologies. 2024. №1. С. 35-42. URL: <https://www.innoist.uz/index.php/ist/article/view/11>
5. Итинсон К. С. Возможности сервиса Google формы для организации образовательного процесса // Региональный вестник. 2020. №10. С. 47-48. URL: <https://www.elibrary.ru/item.asp?id=43090495>