УДК 004

Применение методов машинного обучения для анализа и прогнозирования ответов на опрос «В каком возрасте лучше заводить семью?»

Лапушкина Елена Павловна
Приамурский государственный университет им. Шолом-Алейхема
Студент

Языков Кирилл Александрович
Приамурский государственный университет им. Шолом-Алейхема
Студент

Трапезников Владислав Николаевич
Приамурский государственный университет им. Шолом-Алейхема
Студент

Аннотация
Целью исследования является применение различных методов машинного обучения для построения модели с целью прогнозирования ответа на вопрос «Как вы думаете, влияет ли ваш круг общения на возраст, в котором вы хотите создать семью?». В результате исследования выбрана модель и переменные, обеспечивающие лучшие результаты для текущей задачи. Сравнение моделей произведено на основании основных метрик: AUC, CA, F1, Precision, Recall.
Ключевые слова: Социологический опрос, Google форма, сведения, диаграмма, orange, модель машинного обучения, искусственный интеллект.

The use of machine learning methods to analyze and predict responses to the survey "At what age is it better to start a family?"

Lapushkina Elena Pavlovna
Sholom-Aleichem Priamursky State University
Student

Yazykov Kirill Alexandrovich
Sholom-Aleichem Priamursky State University
Student

Trapeznikov Vladislav Nikolaevich
Sholom-Aleichem Priamursky State University
Student

Abstract
The aim of the study is to use various machine learning methods to build a model in order to predict the answer to the question "Do you think your social circle affects the age at which you want to start a family?". As a result of the research, a model and variables were selected that provide the best results for the current task. The models were compared based on the main metrics: AUC, CA, F1, Precision Recall.
Keywords: Sociological survey, Google form, information, diagram, orange, machine learning model, artificial intelligence.

1 Введение

1.1 Актуальность

Используя машинное обучение, можно получить информированные прогнозы относительно изменений мнений о возрасте для основания семьи, а также лучше понять, какие факторы влияют на такие предпочтения. Это может стать полезным инструментом для социологов, демографов и психологов, а также для организаций, работающих в области планирования семьи и социальных услуг.

1.2 Обзор исследований

О.И. Китаева рассмотрела понятие интеллектуального анализа образовательных данных и выполнила анализ на основе данных учебной дисциплины ВУЗа [1]. А.А. Чернышев в своем исследовании описал такие методы оценки качества моделей машинного обучения, как оценка на обучающих данных, процентное разделение и кросс-валидация. Также рассмотрел важность начального значения рандомизации [2]. Н. Юсупов, А.Савельева, О.Г. Леонова рассмотрели использование методов классификации в программе orange на основе реальной базы данных [3]. Р.Э. Яхшибоев, Н.М. Апсилям, Л.Р. Шамсудинова посвятили исследование анализу современных подходов к моделированию механизмов ИИ, включая как традиционные методы машинного и глубокого обучения, так и новаторские подходы, основанные на последних достижениях в области нейронауки и когнитивной психологии [4]. К.С. Интинсон исследовал преимущества применения сервиса Google Forms, и дал краткую характеристику основных возможностей и ее функций.

1.3 Цель исследования

Целью исследования является применение различных методов машинного обучения для построения модели с целью прогнозирования «Как вы думаете, влияет ли ваш круг общения на возраст, в котором вы хотите создать семью?»

2 Материалы и методы

В данном опросе приняло участие сто шесть респондентов мужского и женского пола в возрасте от 25 лет. Для реализации создан опрос в google-форме. Вопросы сгенерированы с помощью сервиса chadgpt.

3 Результаты и дискуссии

Для того чтобы заполнить данные опроса, нужно зайти в Google аккаунт, нажав на значок приложения Google, выбрать «формы» (рис. 1).

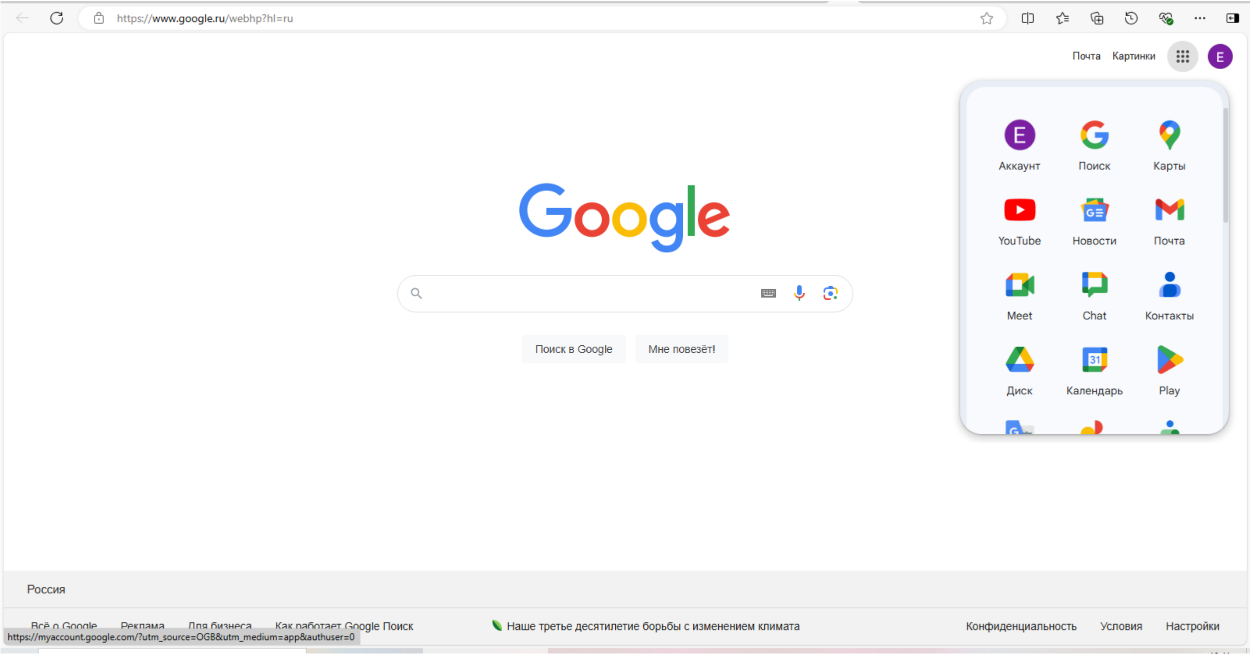


Рисунок 1 –Аккаунт Google

На открывшейся странице нужно создать форму (рис. 2).

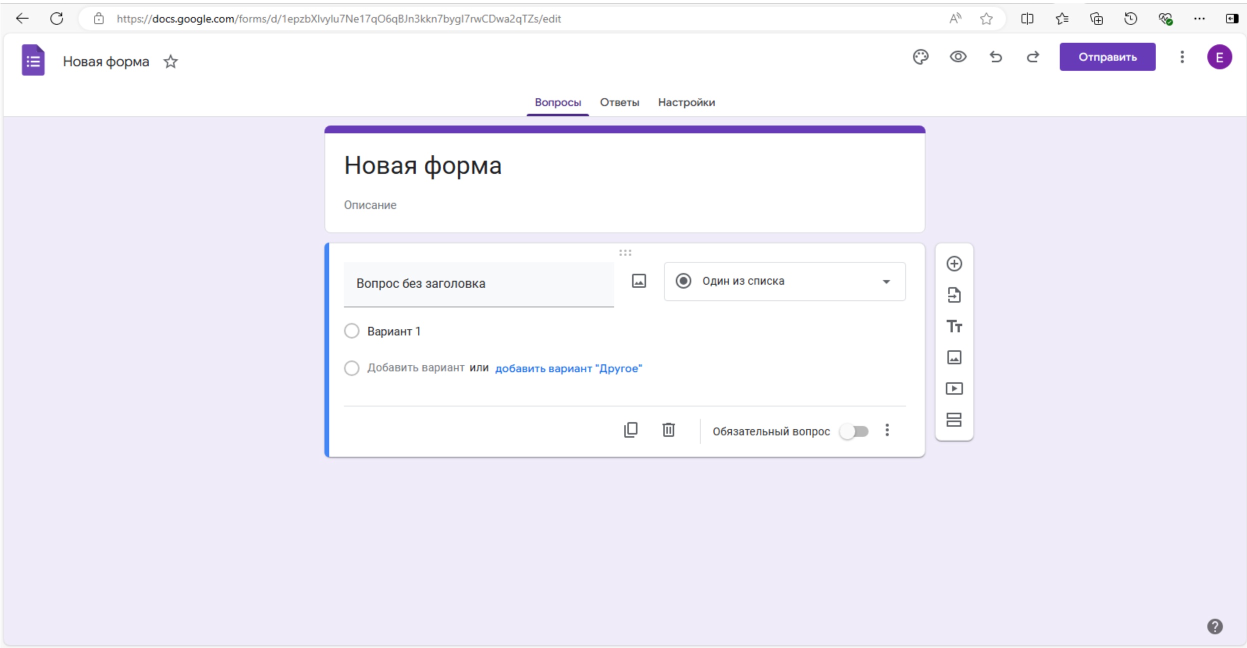


Рисунок 2- Создание новой формы

В новой форме необходимо прописать название опроса, далее заполнить поле вопрос, и при необходимости дополнить еще варианты ответа, ниже в поле нужно выбрать «обязательный вопрос» (рис. 3).

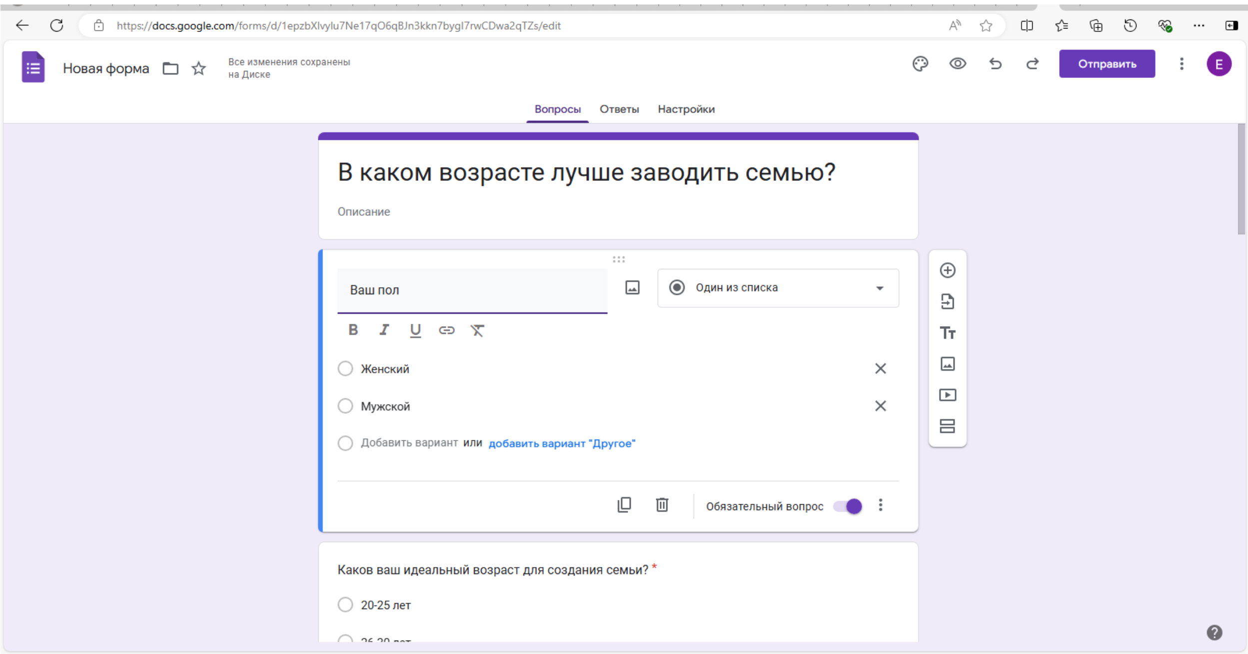


Рисунок 3- Заполнение форм

После того как вопросы в форме заполнены, его можно открыть в режиме просмотра, после вернуться назад для редактирования (рис. 4).

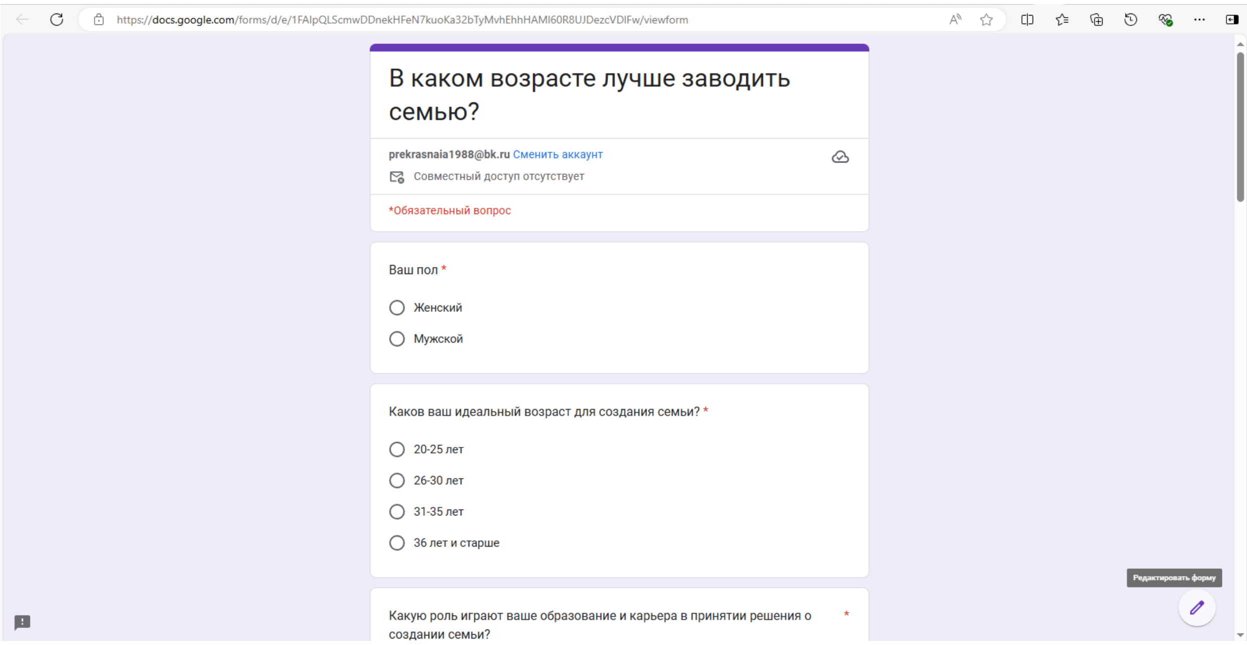


Рисунок 4- Ссылка для обзора опроса

Следующее что нужно сделать- это опубликовать форму для опроса участников, нажать ссылка, лучше выбрать галку- короткий URL, копировать, направить ссылку для прохождения опроса (рис. 5).

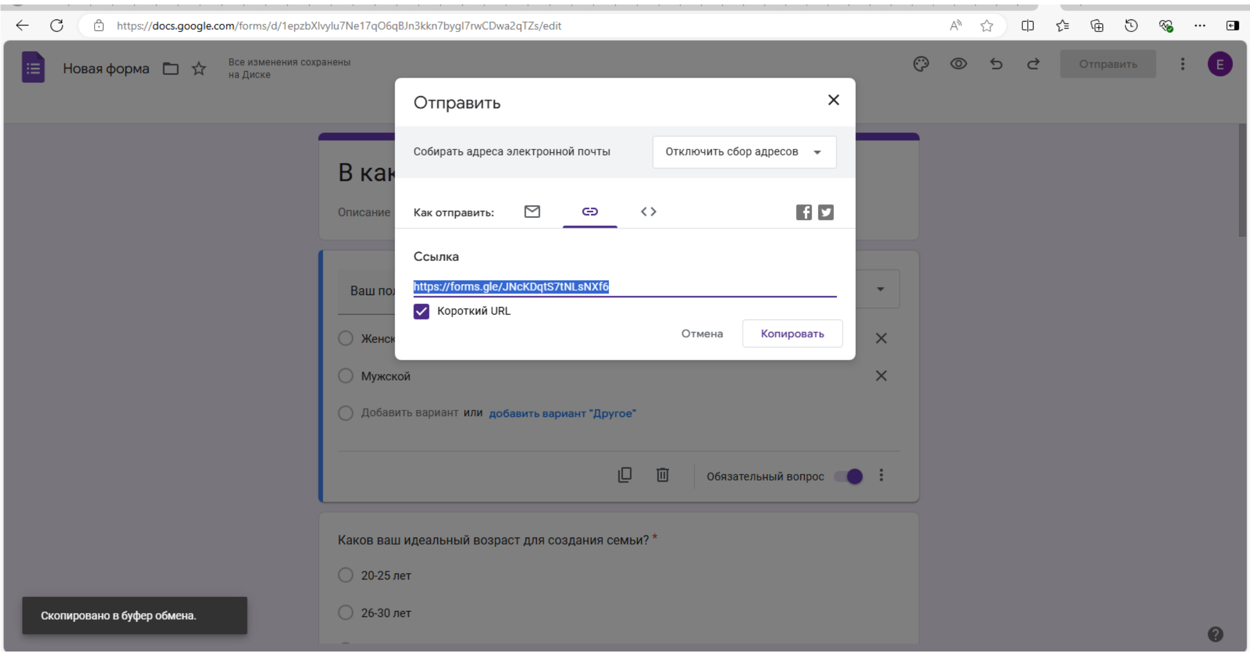


Рисунок 5- Создание ссылки

Для того чтобы сохранить диаграмму по опросу на любой из ответов, необходимо создать Google документ, выбрав вкладку документы (рис. 6).

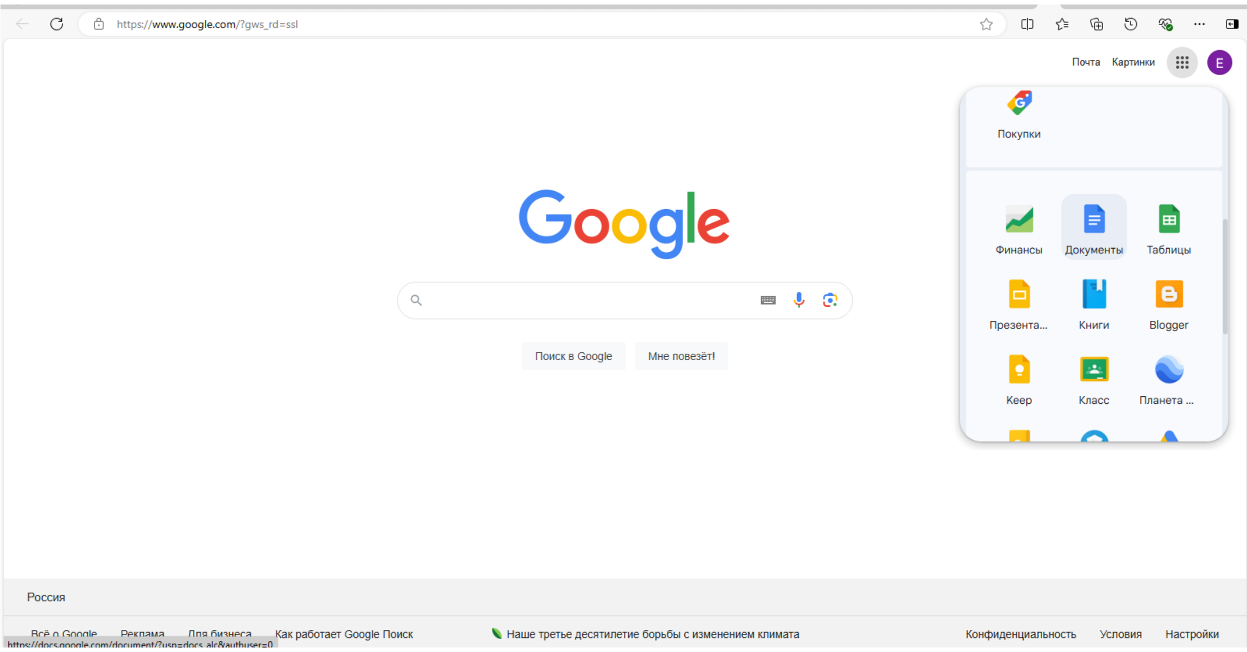


Рисунок 6- Создание нового документа

Далее в пустой документ вставить диаграмму (рис.7).

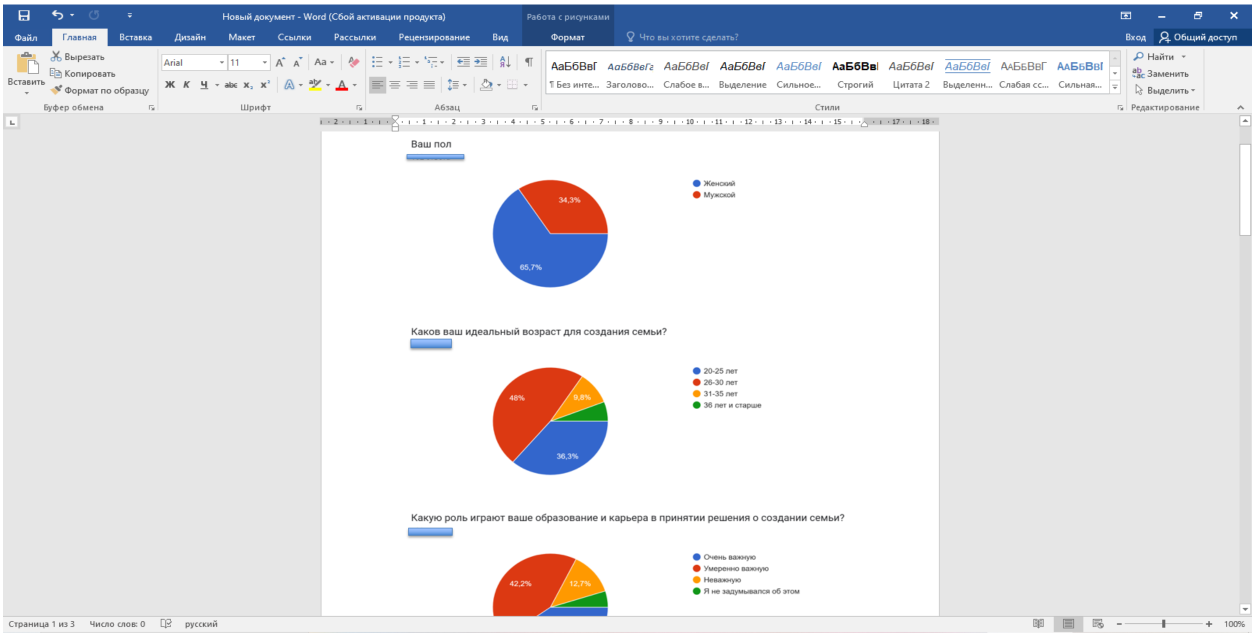
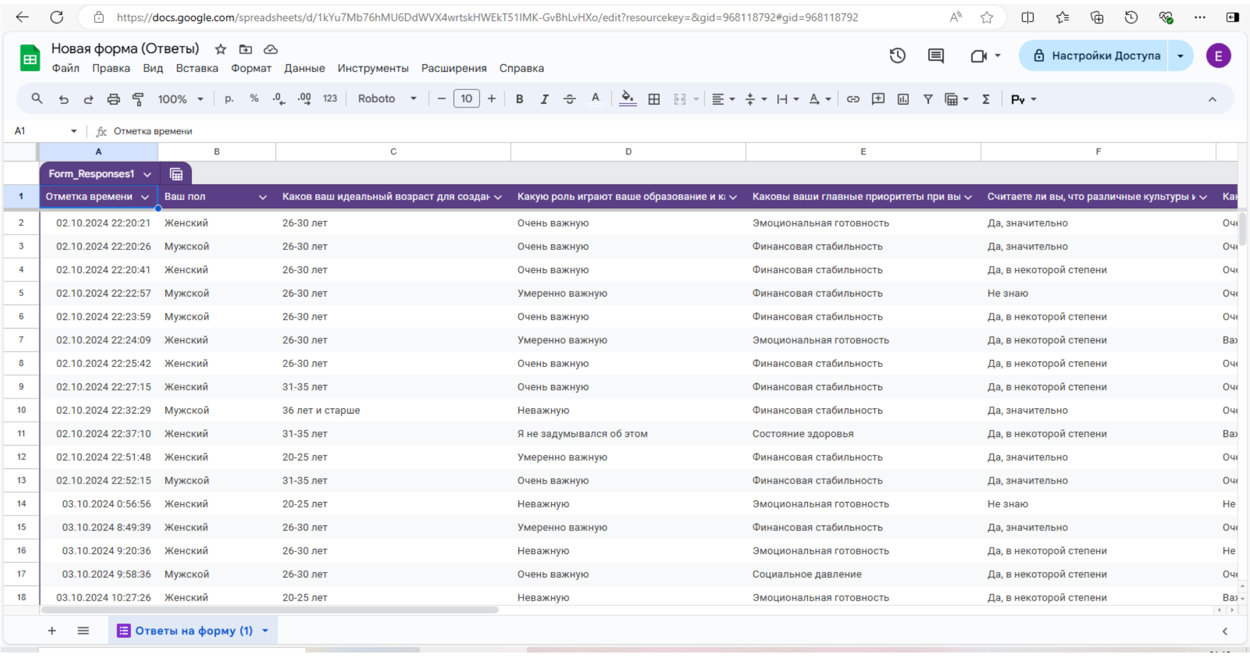


Рисунок 7- Вставка диаграммы

Ответы по форме в развернутом виде можно сохранить в таблице, для этого нажать вкладку, посмотреть в таблицах, и все данные автоматически будут сохранены в Google таблице (рис. 8).



	Отметка времени	Ваш пол	Каков ваш идеальный возраст для создан	Какую роль играют ваше образование и ка	Каковы ваши главные приоритеты при вы	Считаете ли вы, что различные культуры и	Ка
2	02.10.2024 22:20:21	Женский	26-30 лет	Очень важную	Эмоциональная готовность	Да, значительно	Оч
3	02.10.2024 22:20:26	Мужской	26-30 лет	Очень важную	Финансовая стабильность	Да, значительно	Оч
4	02.10.2024 22:20:41	Женский	26-30 лет	Очень важную	Финансовая стабильность	Да, в некоторой степени	Оч
5	02.10.2024 22:22:57	Мужской	26-30 лет	Умеренно важную	Финансовая стабильность	Не знаю	Оч
6	02.10.2024 22:23:59	Мужской	26-30 лет	Очень важную	Финансовая стабильность	Да, в некоторой степени	Оч
7	02.10.2024 22:24:09	Женский	26-30 лет	Умеренно важную	Эмоциональная готовность	Да, в некоторой степени	Ва
8	02.10.2024 22:25:42	Женский	26-30 лет	Очень важную	Финансовая стабильность	Да, в некоторой степени	Оч
9	02.10.2024 22:27:15	Женский	31-35 лет	Очень важную	Финансовая стабильность	Да, в некоторой степени	Оч
10	02.10.2024 22:32:29	Мужской	36 лет и старше	Неважную	Финансовая стабильность	Да, значительно	Оч
11	02.10.2024 22:37:10	Женский	31-35 лет	Я не задумывался об этом	Состояние здоровья	Да, в некоторой степени	Ва
12	02.10.2024 22:51:48	Женский	20-25 лет	Умеренно важную	Финансовая стабильность	Да, значительно	Оч
13	02.10.2024 22:52:15	Мужской	31-35 лет	Очень важную	Финансовая стабильность	Да, значительно	Оч
14	03.10.2024 0:56:56	Женский	20-25 лет	Неважную	Эмоциональная готовность	Не знаю	Не
15	03.10.2024 8:49:39	Женский	26-30 лет	Умеренно важную	Финансовая стабильность	Да, значительно	Оч
16	03.10.2024 9:20:36	Женский	26-30 лет	Неважную	Эмоциональная готовность	Да, в некоторой степени	Не
17	03.10.2024 9:58:36	Мужской	26-30 лет	Очень важную	Социальное давление	Да, в некоторой степени	Оч
18	03.10.2024 10:27:26	Женский	20-25 лет	Неважную	Эмоциональная готовность	Да, в некоторой степени	Ва

Рисунок 8- Сохранение ответов опроса в Google таблице

Для того чтобы построить диаграмму по каждому вопросу опроса, необходимо результаты опроса «закодировать», будет лучше если варианты ответа, как и сами вопросы «закодировать» в цифры и буквы. Для этого, выделить вопрос, нажать заменить, и прописать тот вариант ответа, который необходимо заменить на букву или цифру (рис. 9).

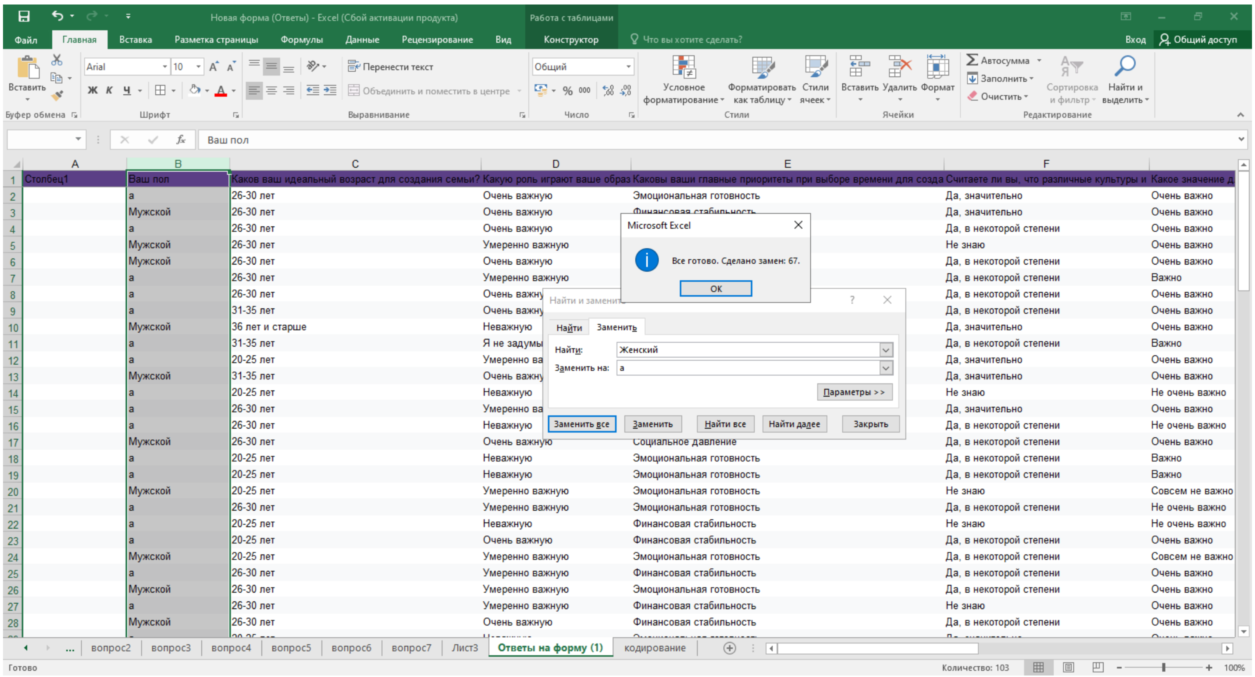


Рисунок 9- Обзор «Кодирования» данных

После того как данные опроса «закодированы», таблица выглядит следующим образом (рис. 10).

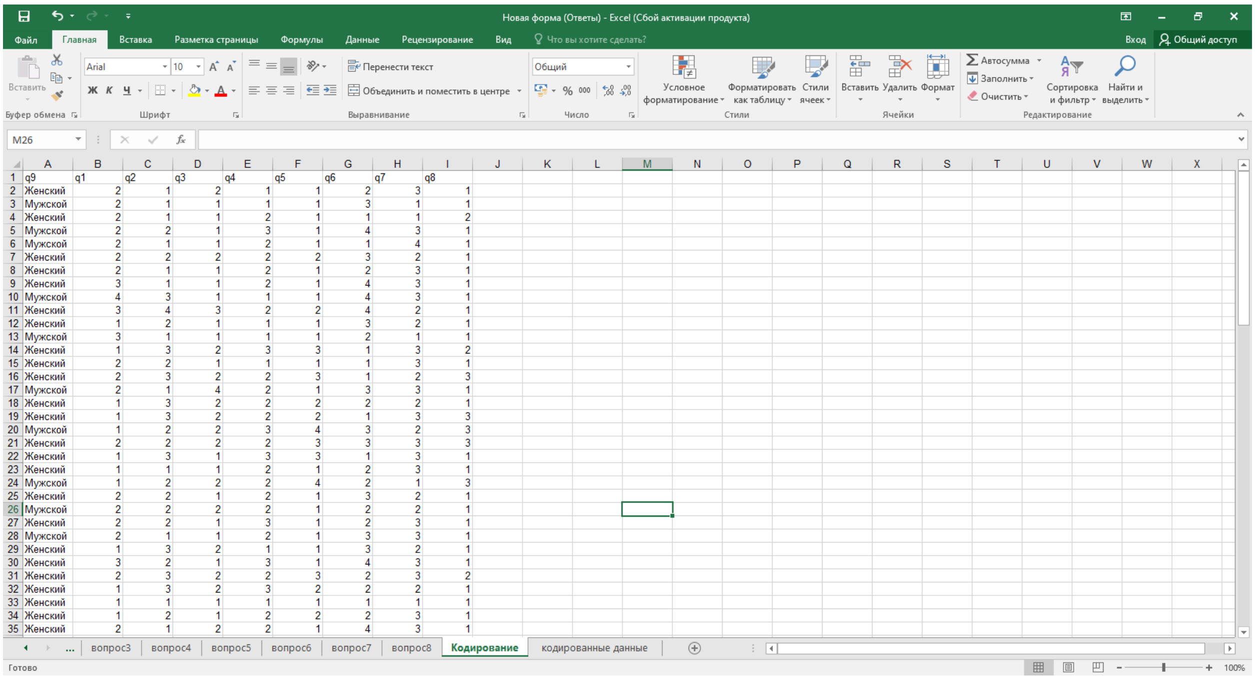


Рисунок 10- Готовый «закодированный» файл Excel

Для того чтобы увидеть ответы респондентов в более развернутом виде, нужно построить диаграмму на основании этих данных. Первая диаграмма показывает какое количество респондентов, считают идеальным возраст для создания семьи. Данные опроса показали, что большинство женщин считают создавать семью необходимо в 20-25 лет, количество опрошенных мужчин считают идеальным возрастом 26-30 лет (рис.11).

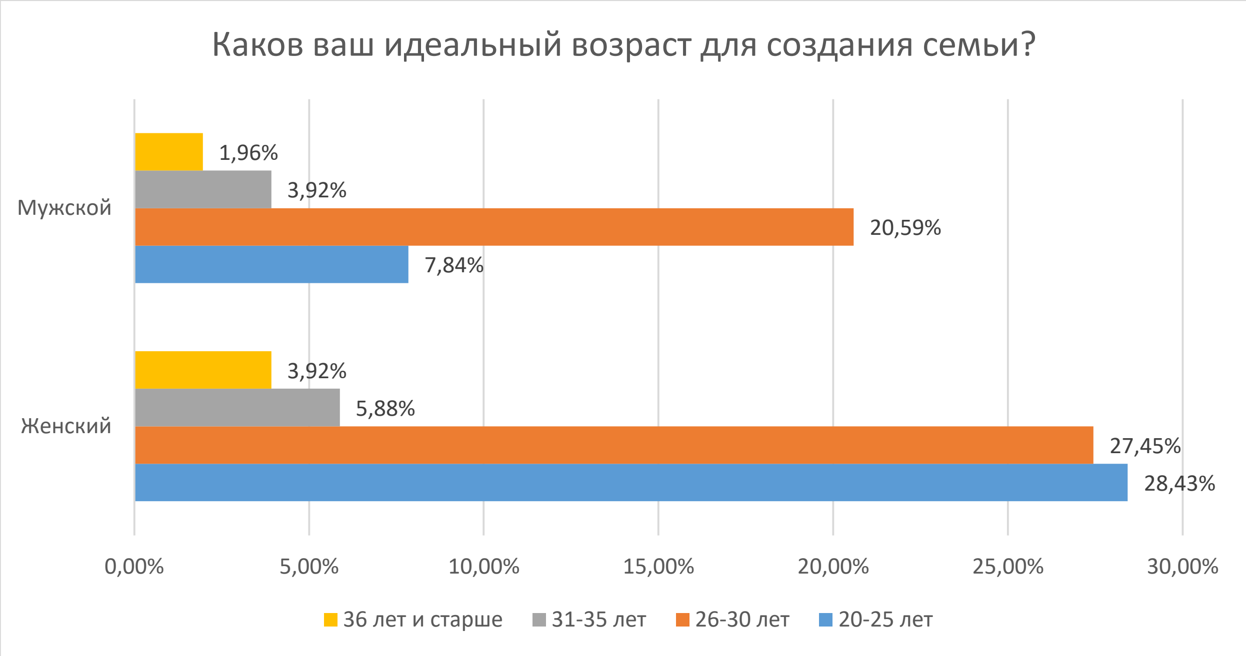


Рисунок 11- Обзор ответа на вопрос «Каков ваш идеальный возраст для создания семьи?»

В данной диаграмме показано как именно распределились ответы в процентном соотношении между респондентами в каждом варианте ответа (рис.12).

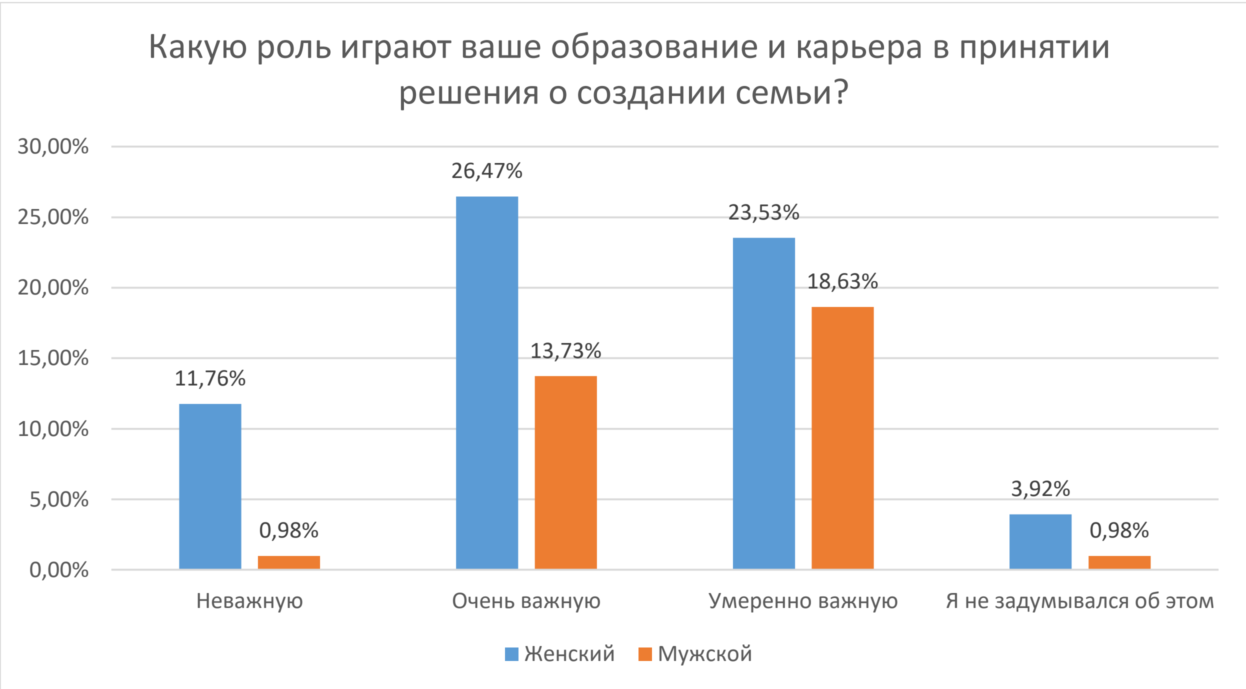


Рисунок 12- Обзор ответа на вопрос «Какую роль играют ваше образование и карьера в принятии решения о создании семьи?»

Для обзора ответов респондентов на третий вопрос, создана гистограмма с группировкой (рис. 13).

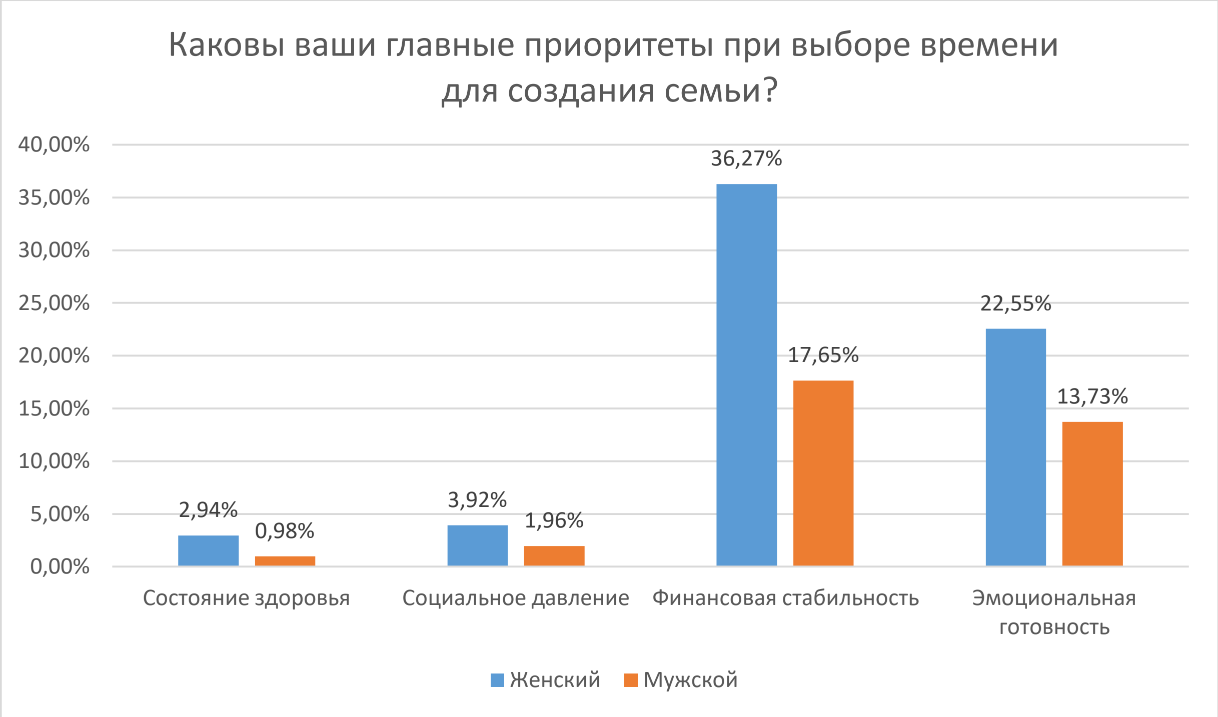


Рисунок 13- Обзор ответа на вопрос «Каковы ваши главные приоритеты при выборе времени для создания семьи?»

В построении диаграммы можно использовать объемную гистограмму, она позволяет максимально лучше увидеть результаты опроса за счет небольшого количества предложенных ответов (рис.14).

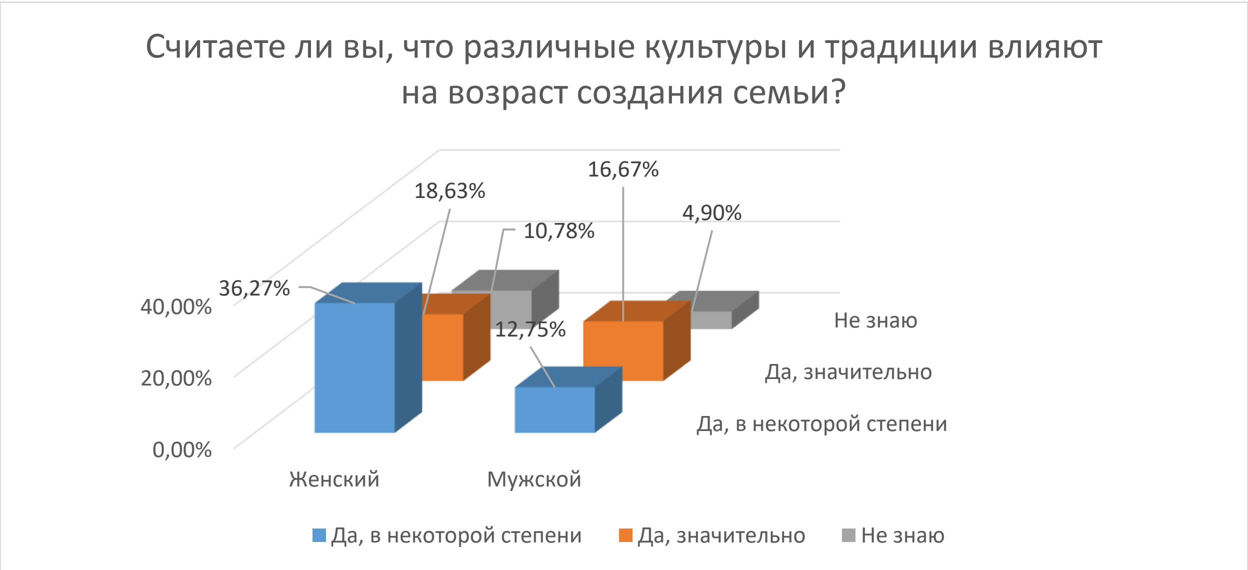


Рисунок 14- Обзор ответа на вопрос «Считаете ли вы, что различные культуры и традиции влияют на возраст создания семьи?»

При построении диаграммы можно менять ее цвет (рис.15).

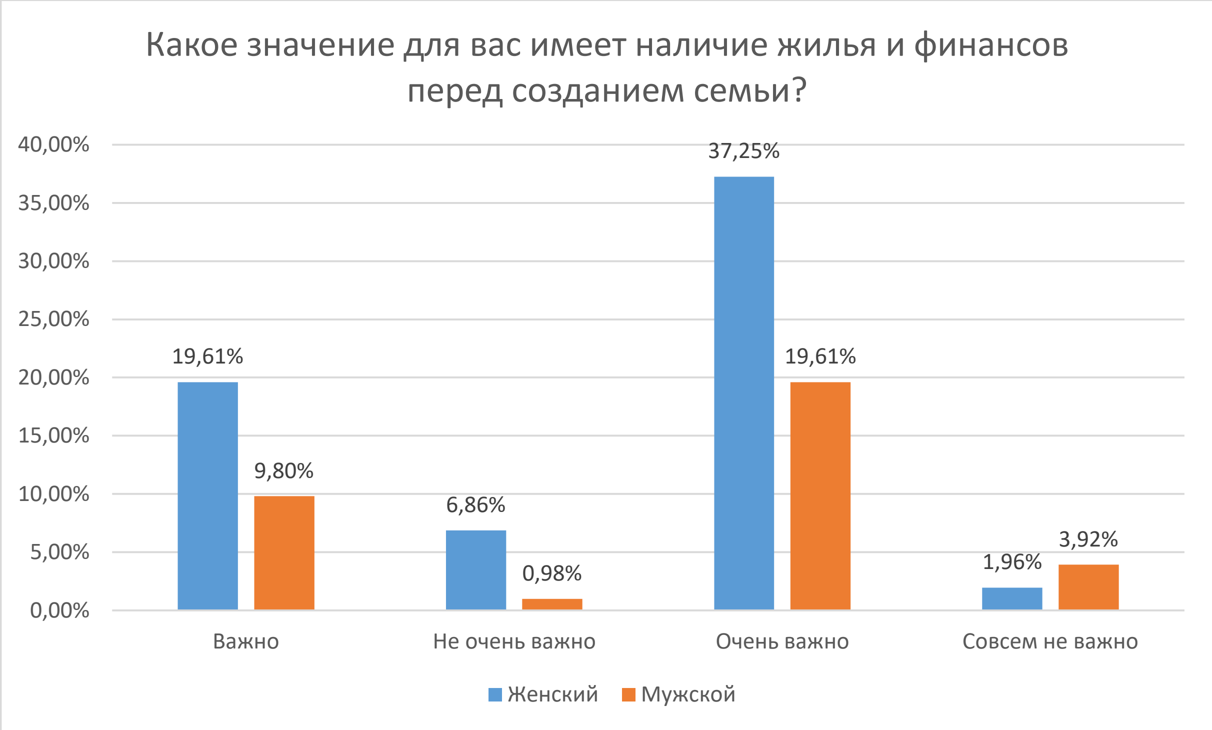


Рисунок 15- Обзор ответа на вопрос «Какое значение для вас имеет наличие жилья и финансов перед созданием семьи?»

Лучше всего данные опроса, в проценте соотношений, показывает объемная гистограмма с группировкой (рис.16).

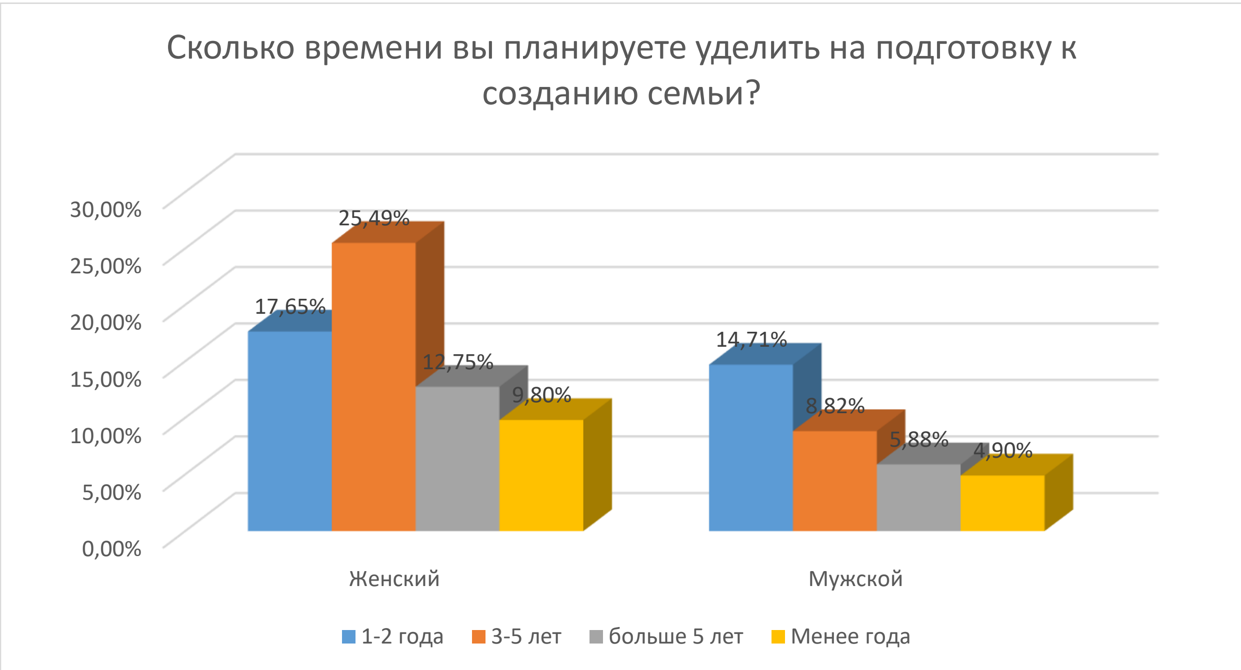


Рисунок 16 – Обзор ответа на вопрос «Сколько времени вы планируете уделить на подготовку к созданию семьи?»

В следующей диаграмме большинство женщин- это 35,29% ответили, что круг общения не влияет на возраст при создании семьи (рис.17).

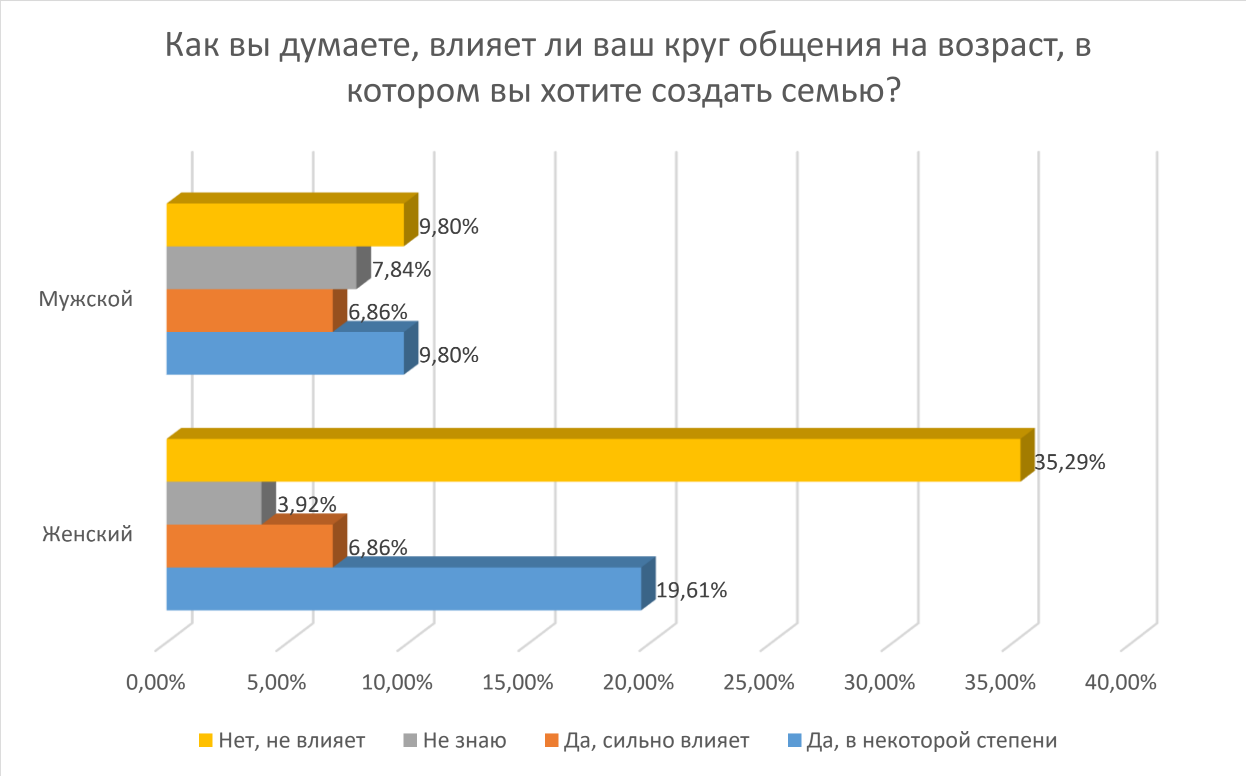


Рисунок 17- Обзор ответа на вопрос «Как вы думаете, влияет ли ваш круг общения на возраст, в котором вы хотите создать семью?»

В последней диаграмме тоже можно наблюдать, что большинство женщин считают романтические отношения важными составляющими для успешного создания семьи (рис.18).

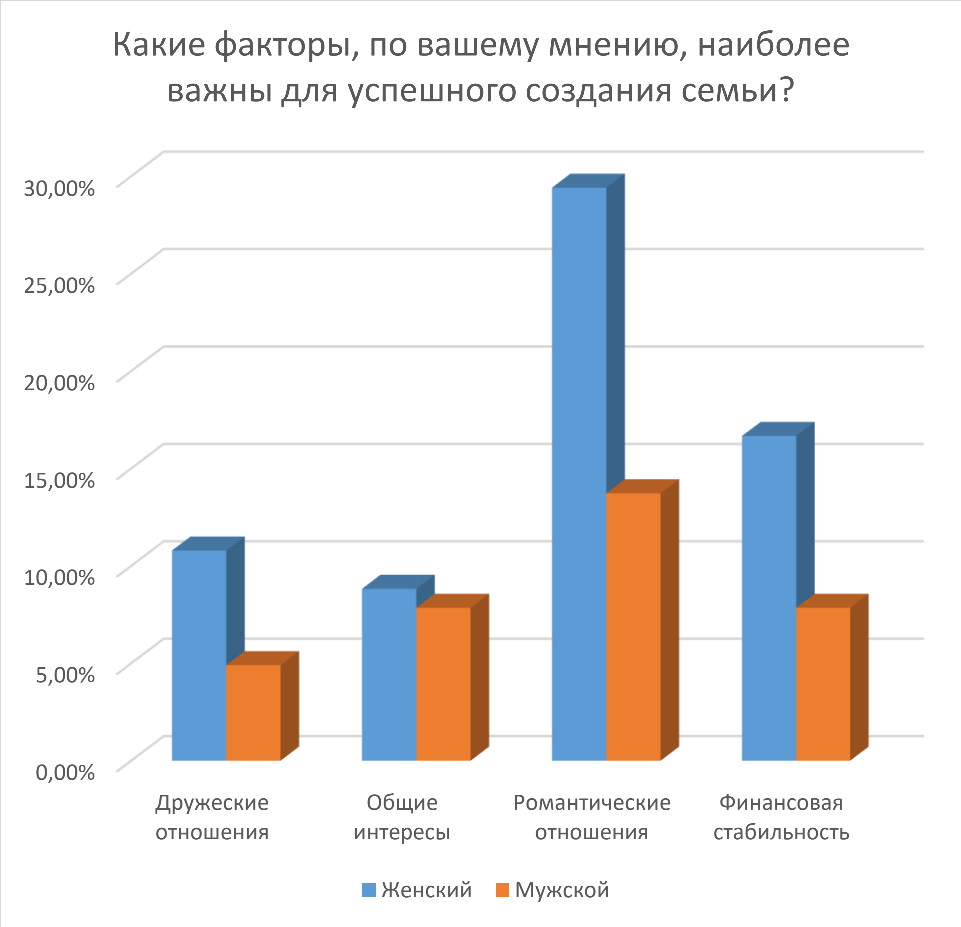


Рисунок 18- Обзор ответа на вопрос «Какие факторы, по вашему мнению, наиболее важны для успешного создания семьи?»

На основании данных опроса необходимо подсчитать корреляции в Orange, и построить модели машинного обучения. На сегодняшний день Orange - это мощная и интуитивно понятная платформа для визуального анализа данных и машинного обучения. Нужно создать новый файл (рис. 19).

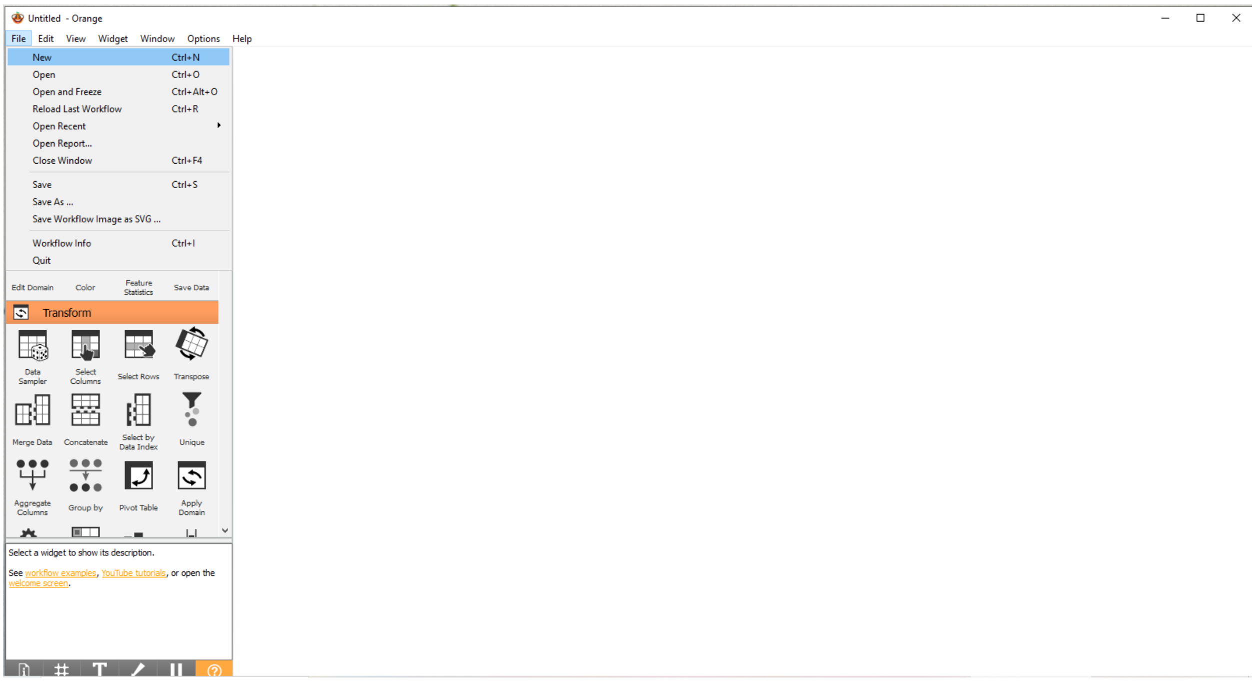


Рисунок 19- Создание нового файла в Orange

В новом открывшемся файле добавить файл данные которого уже предварительно «закодированы» в числа и буквы (рис.20).

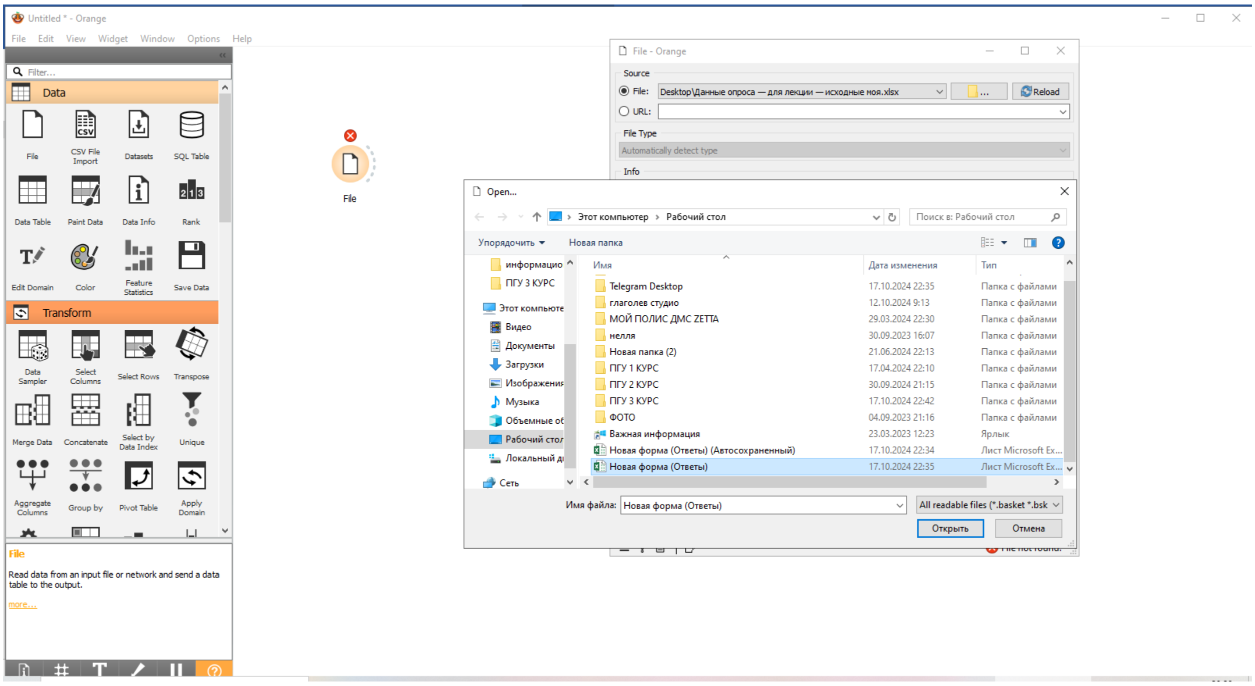


Рисунок 20- Добавление «закодированного» файла в Orange

После добавление файла, в открывшемся окне видно, что все вопросы представлены по типу категориального, вопрос по которому предстоит сделать прогнозирование имеет роль «target», все остальные вопросы имеют название роли «feature» (рис. 21).

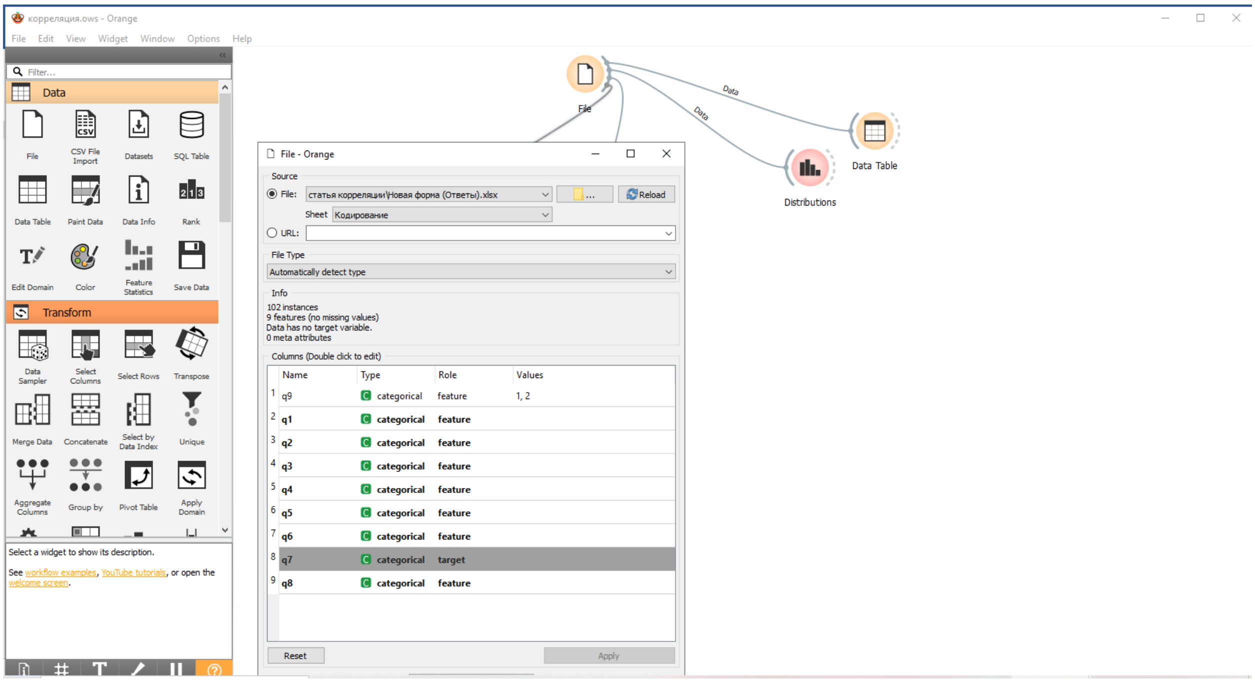


Рисунок 21- Обзор выбора переменной

Далее к новому файлу добавить модуль data table. В Orange модуль data table отвечает за отображение и манипуляцию данными в табличном формате (рис. 22).

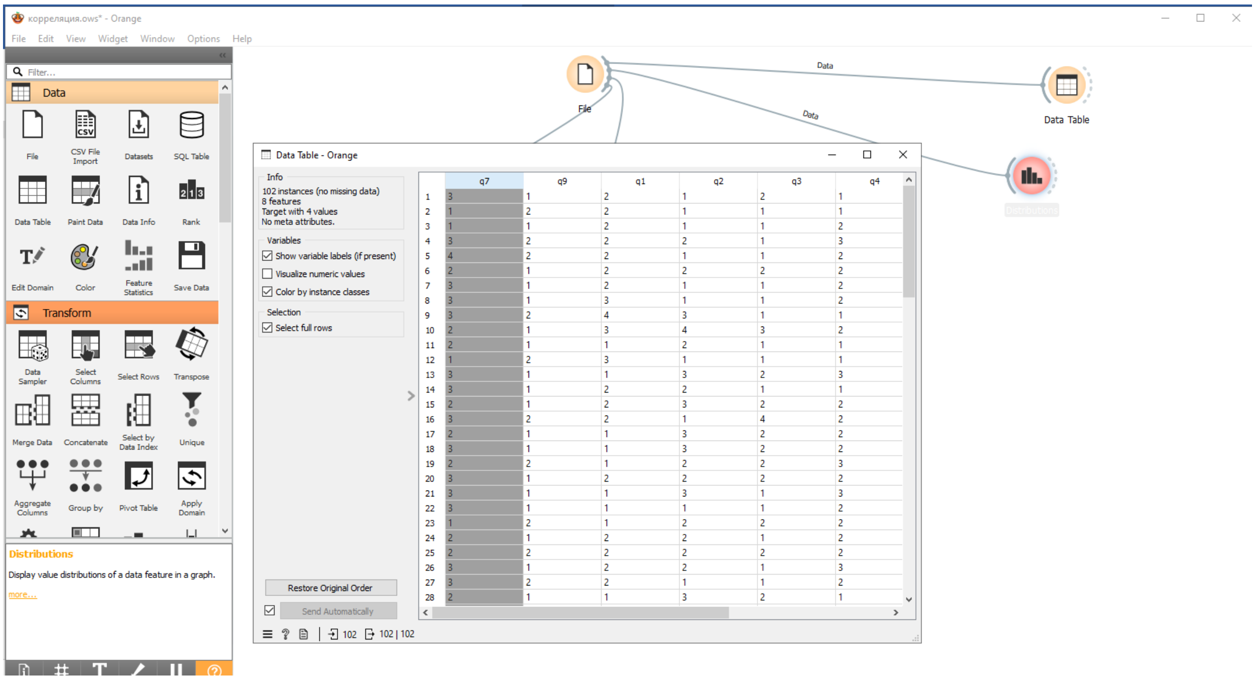


Рисунок 22- Обзор модуля data table

Далее добавить модуль distributions. В Orange, библиотеке для визуализации и анализа данных, модуль distributions отвечает за исследование распределений переменных в данных (рис.23).

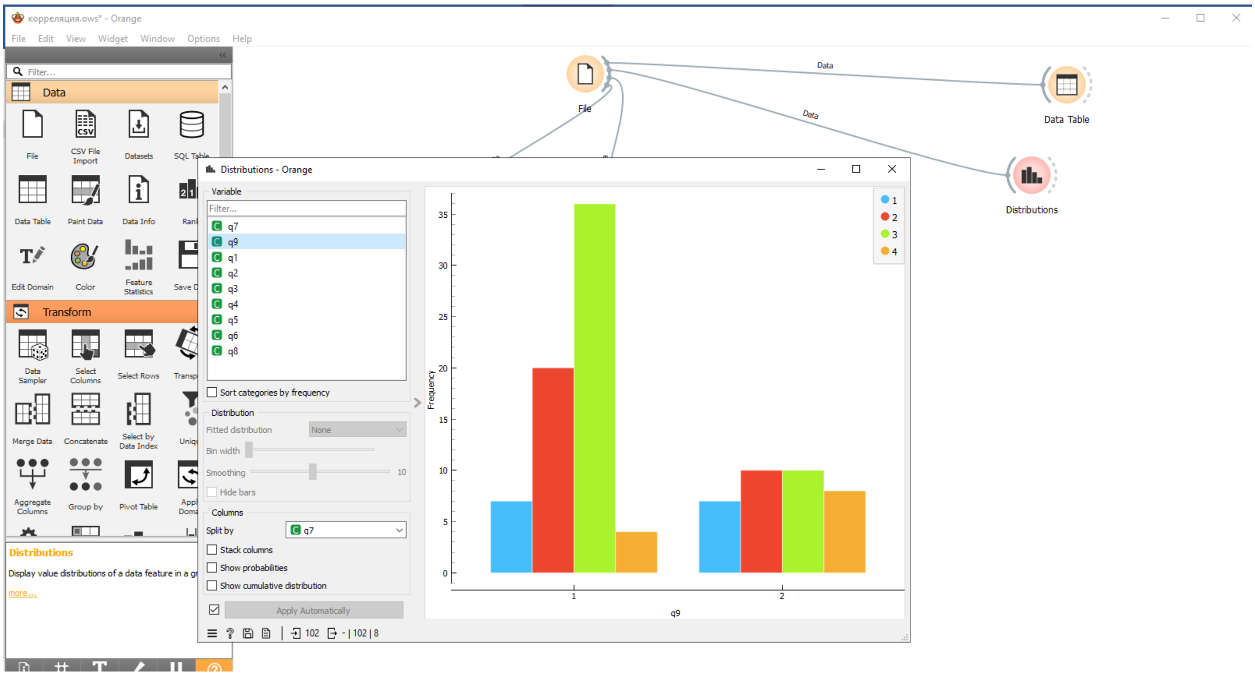


Рисунок 23- Обзор данных с помощью модуля distributions

Далее добавить модуль correlations. Этот модуль в Orange отвечает за анализ взаимосвязей между переменными в наборе данных, также модуль является полезным инструментом для оценки и визуализации взаимосвязей в данных, что может помочь в анализе и принятии решений (рис.24).

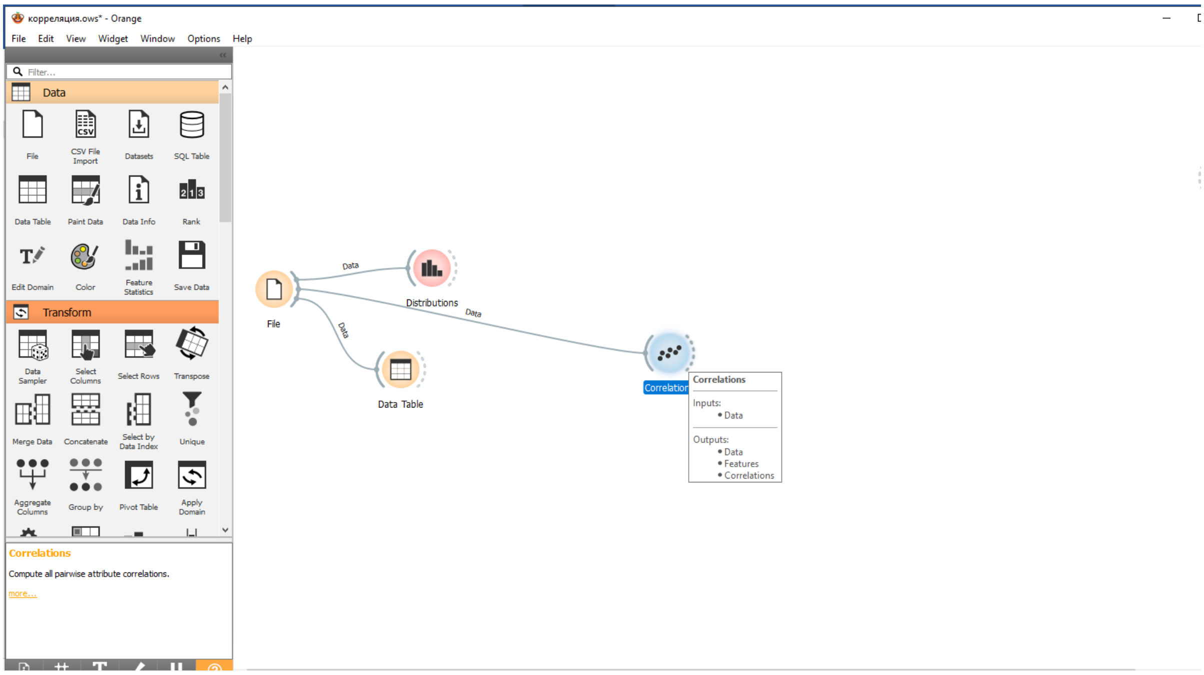


Рисунок 24- Обзор данных модуля correlations

Анализируя данные модуля correlations, можно отследить 2 метода корреляции, первый метод- spearman correlations и второй метод pearson correlations, метод коэффициента корреляции spearman отвечает за оценку степени и направления монотонной связи между двумя переменными, коэффициент корреляции pearson измеряет линейную зависимость между двумя непрерывными переменными, в этом случае данные любого метода указывают корреляционную зависимость между вопросами, для подсчета корреляции по методу spearman correlations необходимо все вопросы перевести по типу в числовой (numeric) вопрос по которому будет выполнятся прогнозирование оставить прежним (рис. 25) (рис. 26).

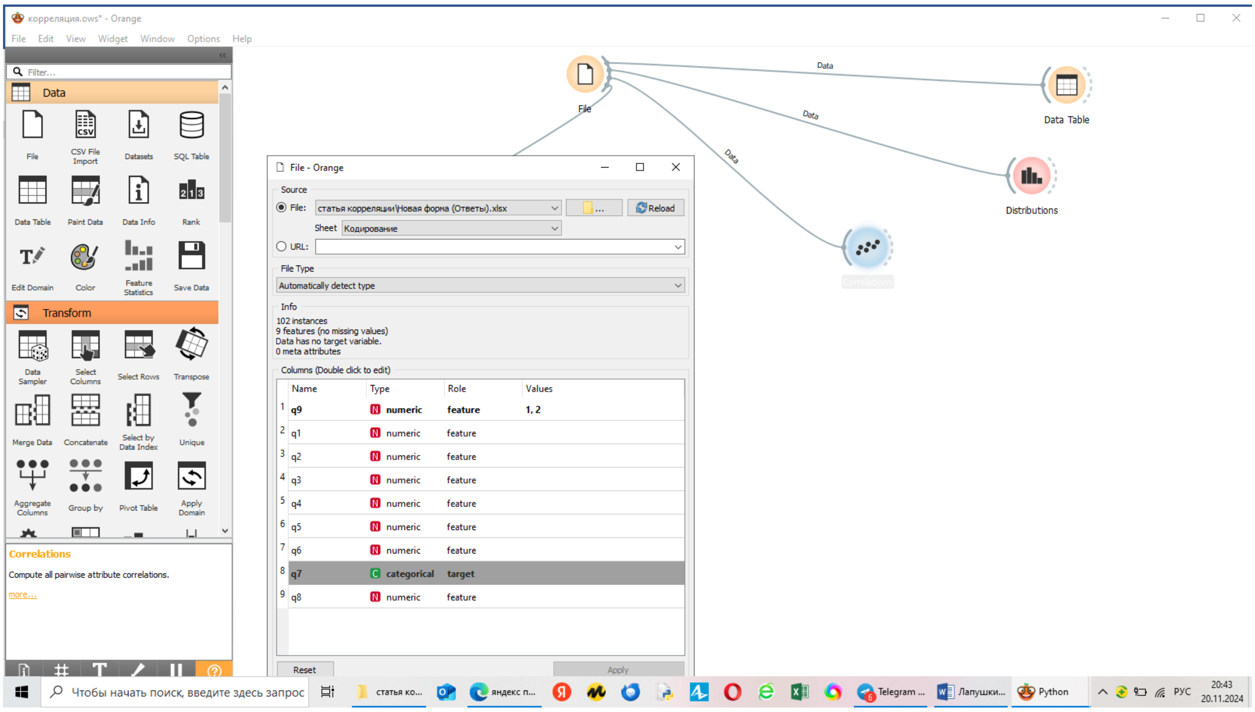


Рисунок 25- Обзор выбора переменной

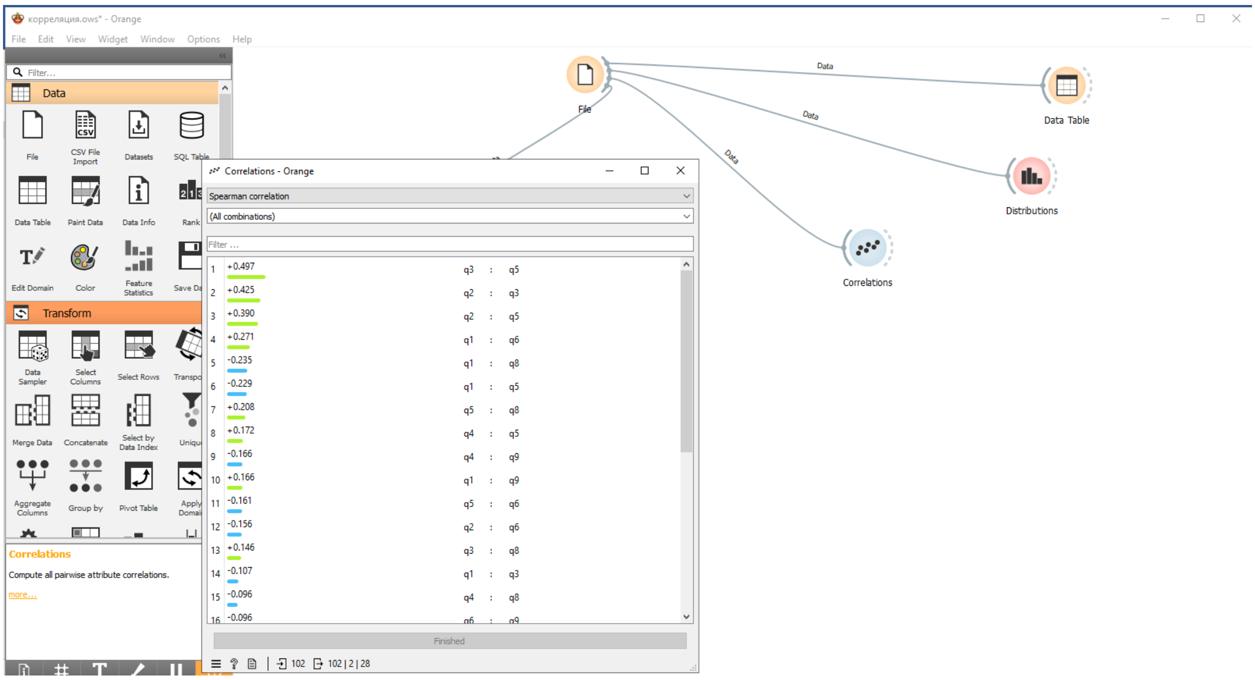


Рисунок 26- Обзор метода spearman correlations

Следующим для обзора будет построение классификационной модели для прогнозирования ответа на вопрос «Как вы думаете, влияет ли ваш круг общения на возраст, в котором вы хотите создать семью?»

Далее добавить модуль test and score в Orange данный модуль оценивает производительность моделей машинного обучения, позволяя выполнять кросс-валидацию, вычислять метрики (точность, полноту и др.), сравнивать модели и визуализировать результаты. (рис. 27).

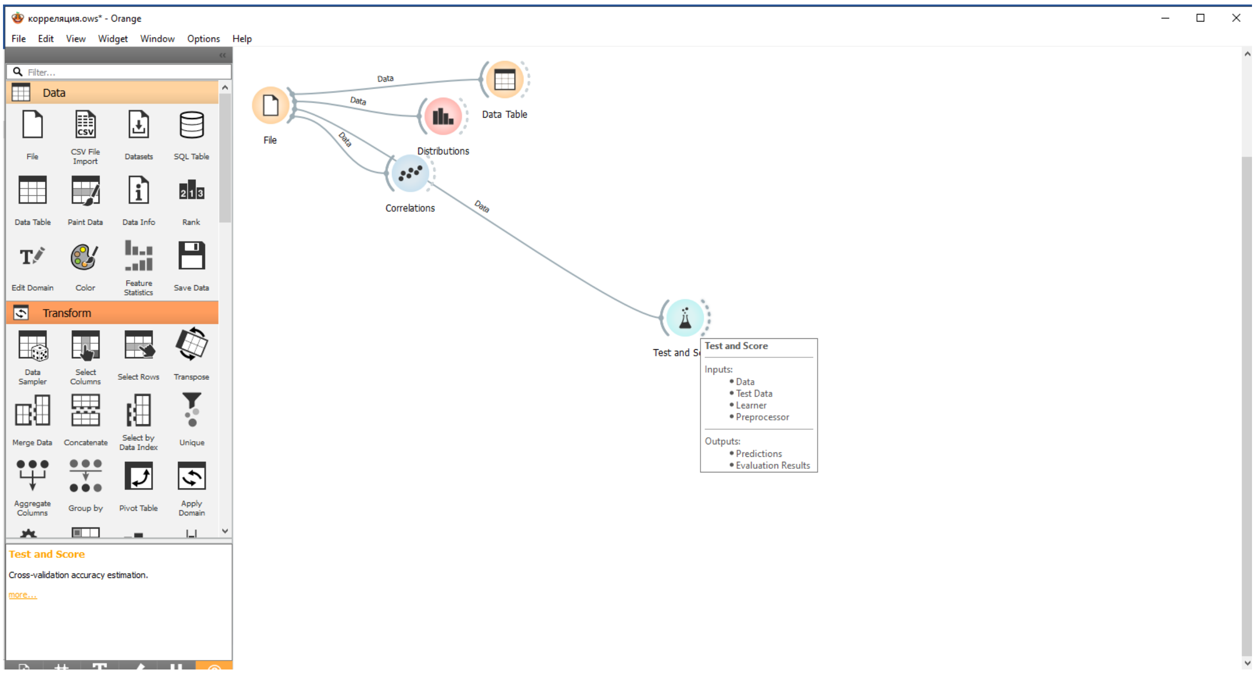


Рисунок 27- Обзор модуля test and score

При открытии модуля test and score можно увидеть данные по метрикам AUC, CA, F1, Prec, Recall, где AUC (Area Under the Curve): показывает способность модели различать классы. Значение от 0 до 1, где 1 — идеальная модель, а 0.5 — случайная. Чем ближе показания к единице, тем вероятность предсказания выше, CA (Classification Accuracy): доля правильно классифицированных примеров среди всех, F1 Score: гармоническое среднее между точностью (Precision) и полнотой (Recall), Precision (точность): доля правильно предсказанных положительных примеров среди всех предсказанных положительных, Recall (Полнота): доля правильно предсказанных положительных примеров среди всех реальных положительных (рис. 28).

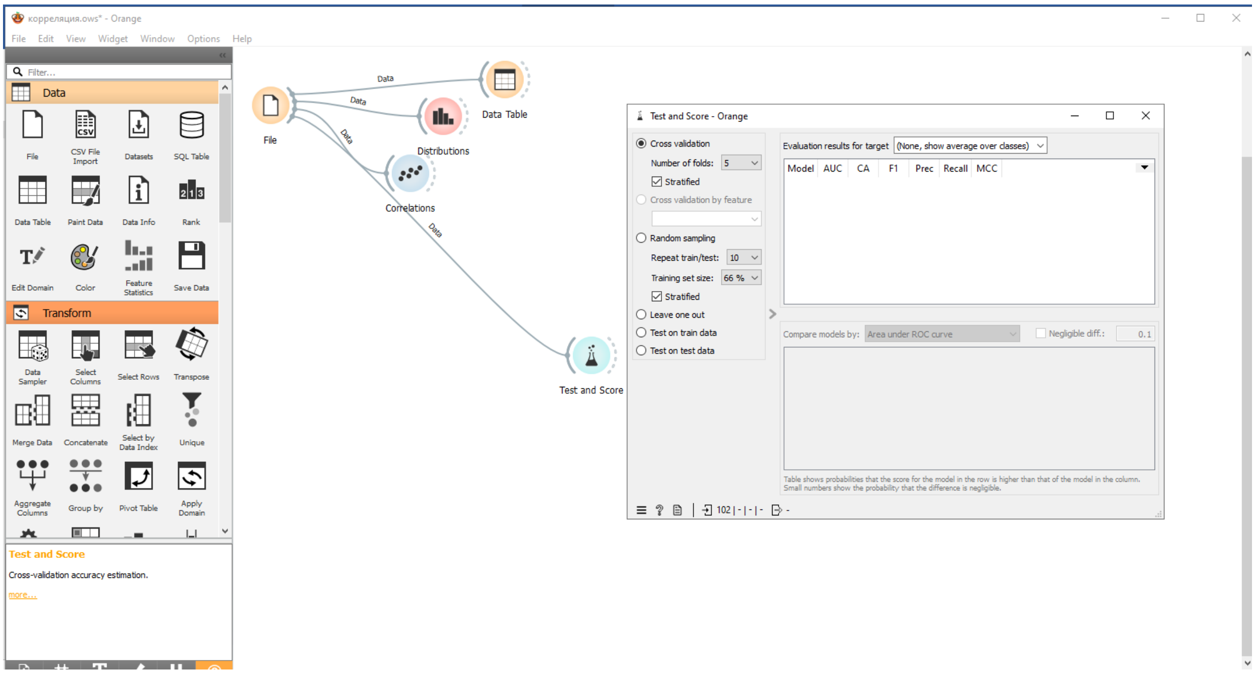


Рисунок 28- Обзор данных параметров AUC, CA, F1, Precision, Recall

Далее добавить модели для оценки данных- kNN (k ближайших соседей) в Orange — этот алгоритм предназначен для классификации и регрессии, основанный на принципе, что схожие объекты находятся близко друг к другу в пространстве признаков, модуль- Naïve Bayes, это простой вероятностный классификатор, основанный на теореме Байеса и предположении о независимости признаков, модель Logistic Regression в Orange- это инструмент предназначен для выполнения логистической регрессии, который используется в задачах классификации, он помогает в анализе данных и прогнозировании вероятностей принадлежности объектов к разным классам на основе имеющихся признаков, и модуль Tree-в Orange он представляет собой инструмент для построения деревьев решений, который используется для решения задач классификации и регрессии (рис. 29).

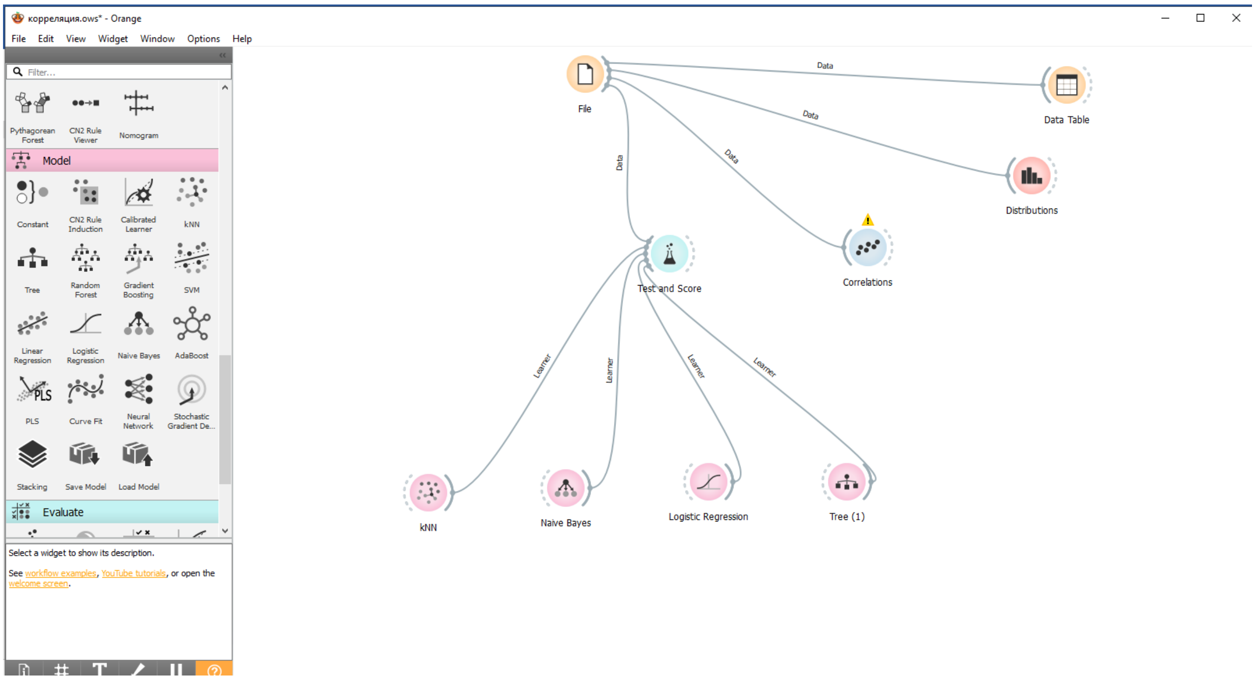


Рисунок 29- Добавление моделей kNN и Naïve Bayes, Logistic Regression, Tree

По данным кросс-валидации (cross- validation) которая используется для оценки производительности модели в машинном обучении, помогает избежать переобучения и дает более надежную оценку ее качества, весь диапазон данных поделен на 5 частей, по каждой метрики можно определить какая модель является обучающая, а какая тренировочная (рис. 30).

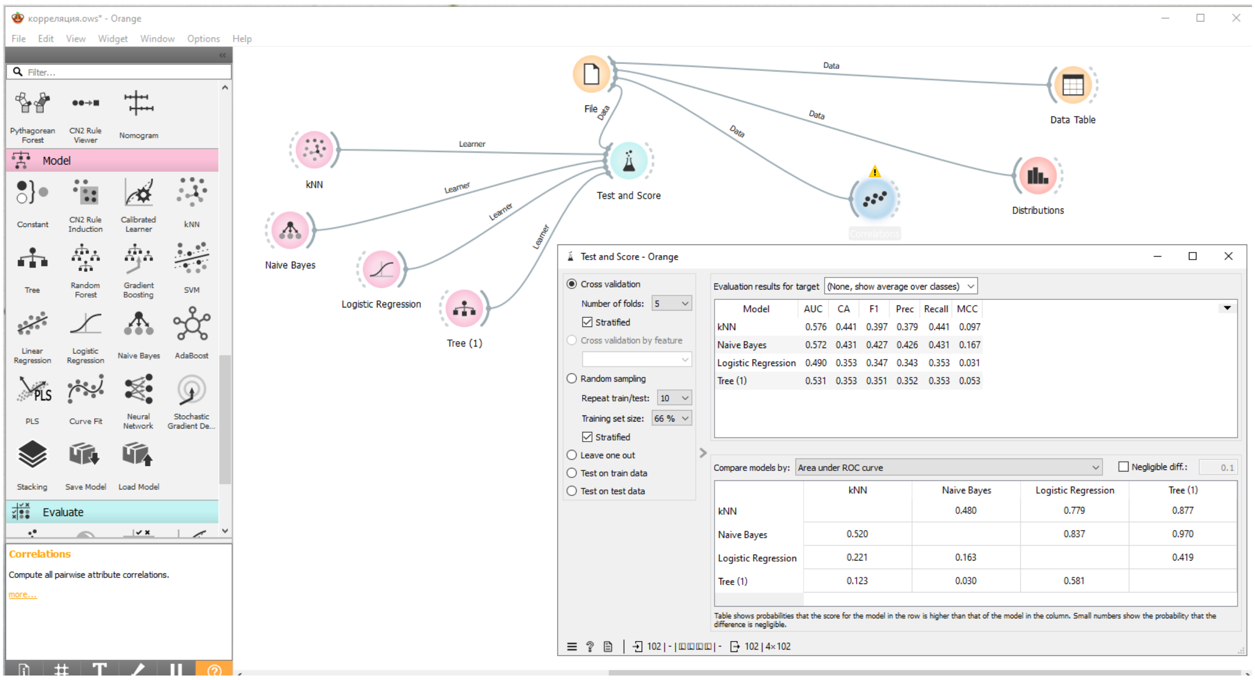


Рисунок 30- Обзор метрик моделей kNN, Naïve Bayes, Logistic Regression, Tree

Следующим этапом будет построение модели confusion matrix — это инструмент, который используется для оценки производительности классификационной модели. Она предоставляет наглядное представление о том, как хорошо модель справляется с задачей классификации, отображая количество предсказаний для моделей- kNN, Naïve Bayes, Logistic Regression, Tree (рис. 31), (рис. 32), (рис. 33), (рис. 34).

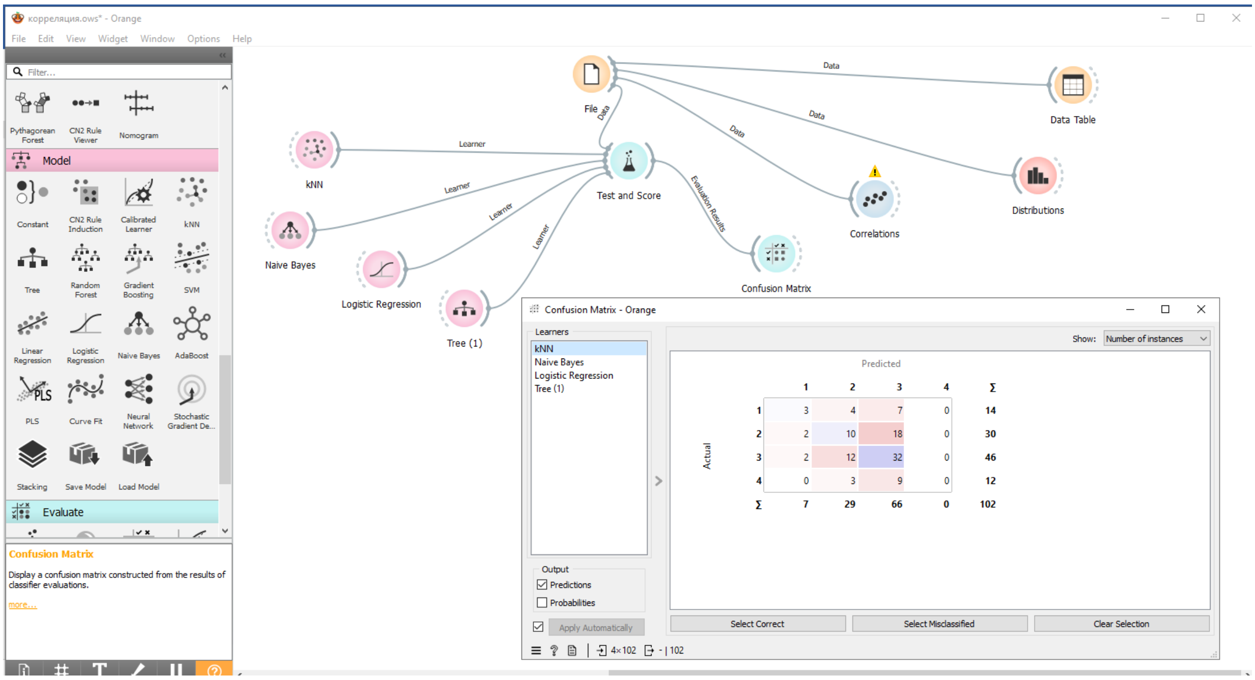


Рисунок 31- Обзор матрицы ошибок по данным модели kNN

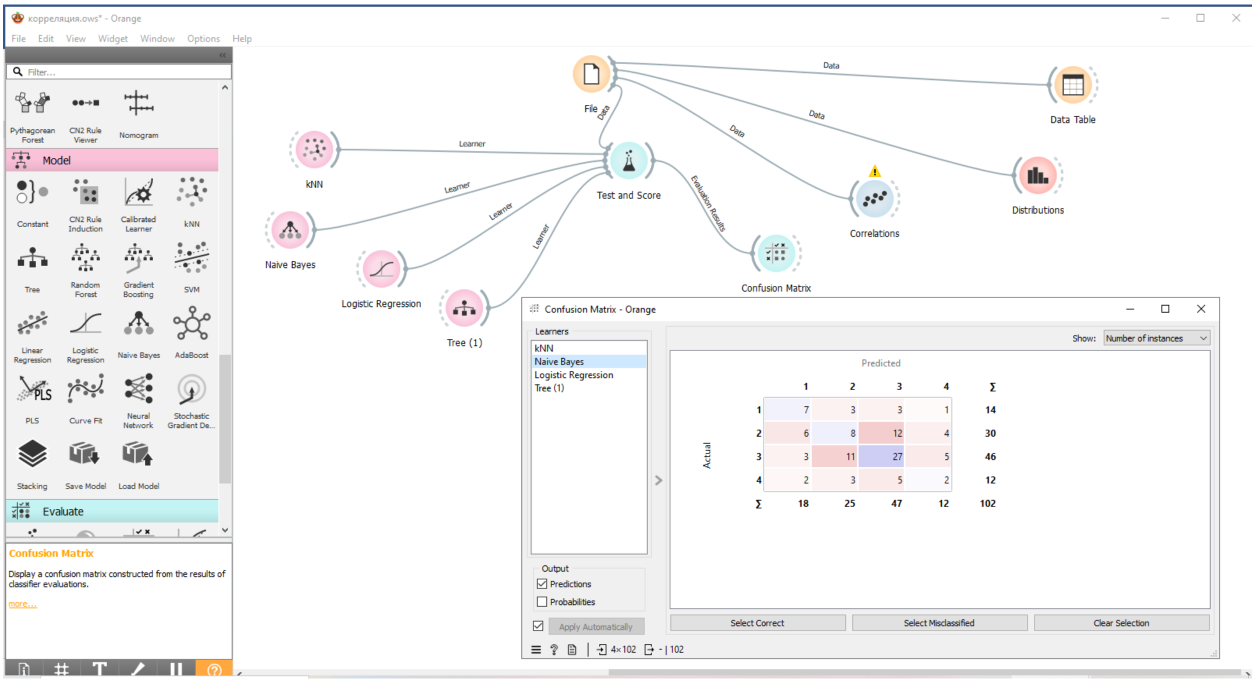


Рисунок 32- Обзор матрицы ошибок по данным модели Naïve Bayes

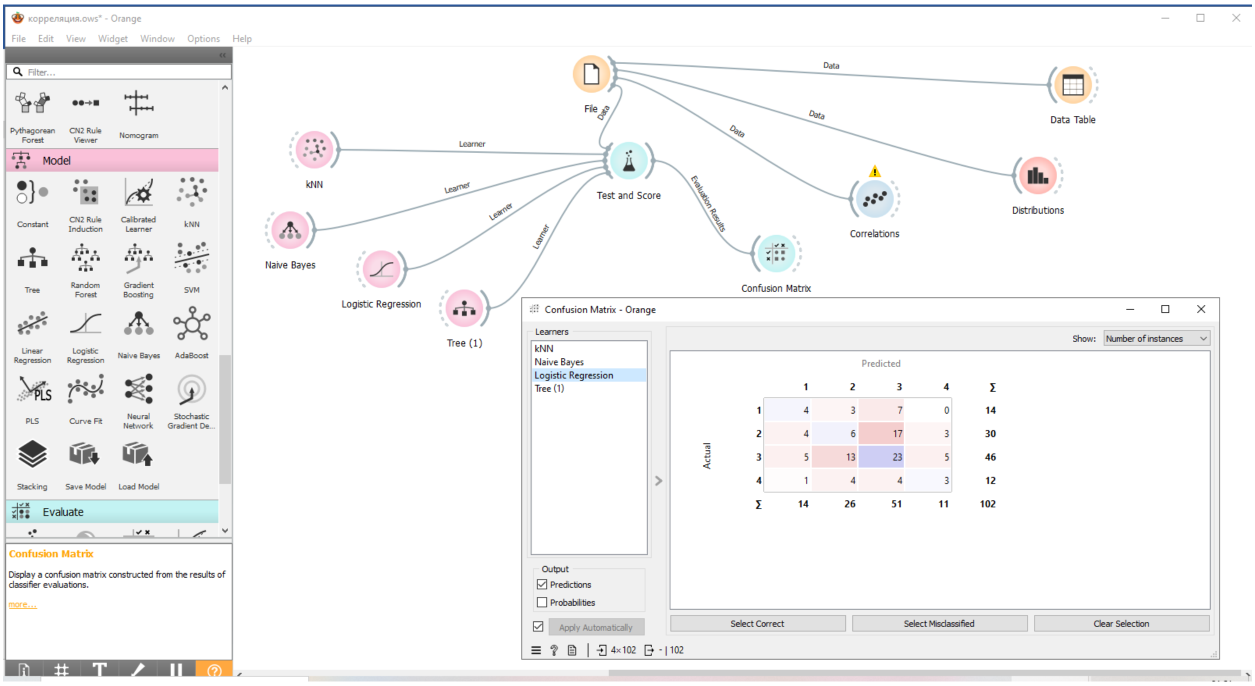


Рисунок 33- Обзор матрицы ошибок по данным модели Logistic Regression

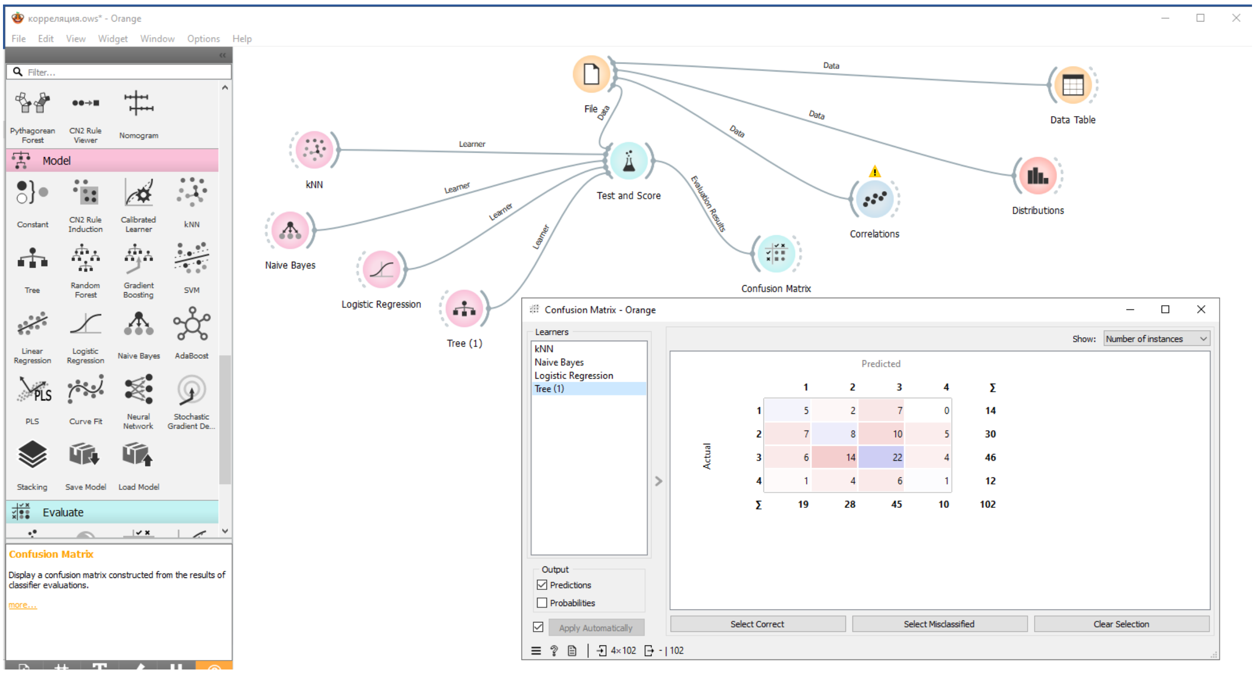


Рисунок 34- Обзор матрицы ошибок по данным модели Tree

Данные по метрикам вывести в итоговую таблицу (табл.1).

Таблица 1 - Сравнение точности предугадывания результата

Model	Area Under the Curve	Classification Accuracy	F1 Score	Precision	Recall
kNN	0,576	0,441	0,397	0,379	0,441
Naïve Bayes	0,572	0,431	0,427	0,426	0,431

Logistic Regression	0,490	0,353	0,347	0,343	0,353
Tree	0,531	0,353	0,351	0,352	0,353

4 Выводы

В данном исследовании была построена модель классификации на вопрос «Как вы думаете, влияет ли ваш круг общения на возраст, в котором вы хотите создать семью?» по представленным данным в модуле test and score, видно, что модель kNN 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 ля 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-distsipliny-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., Апсилям 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