УДК 004

Решение задачи классификации набора данных состоящий из слов названий различных продуктов и их категорий с помощью программного пакета визуального программирования Orange

Голубева Евгения Павловна
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
Цель данной статьи – решить задачу классификации набора данных состоящий из слов названий различных продуктов и их категорий. Для решения задачи классификации был использован программный пакет визуального программирования на основе компонентов для визуализации данных Orange и набор данных различный слов. С помощью средств визуализации Orange решили задачу классификации набора данных состоящий из слов названий различных продуктов и их категорий.
Ключевые слова: Orange, виджет, слова, классификация.

Solving the problem of classifying a dataset consisting of words, names of various products and their categories using the Orange visual programming software package

Golubeva Evgeniya Pavlovna
Sholom-Aleichem Priamursky State University
Student

Abstract
The purpose of this article is to solve the problem of classifying a dataset consisting of various words. To solve the classification problem, a visual programming software package based on Orange data visualization components and a dataset of various words were used. With the help of Orange visualization tools, we solved the problem of classifying a data set consisting of various words.
Keywords: Orange, widget, words, classification.

1 Введение

1.1 Актуальность

Классификация слов имеет широкий спектр применений, включая системы рекомендаций, поисковые системы, анализ социальных сетей, автоматическое аннотирование текстов и многое другое.

Классификация слов может стать основой для разработки новых методов анализа текста, таких как автоматическое определение тематики текста, выявление связей между словами и т.д.

Программный пакет визуального программирования Orange предоставляет удобный и интуитивно понятный интерфейс для решения задач анализа данных, включая классификацию. Благодаря этому пакету исследователи и специалисты в области анализа данных могут легко применять методы классификации к наборам данных, в том числе и к наборам из различных слов.

1.2 Обзор исследований

С.А. Костырева, И.С. Курьян, Д.В. Негина рассмотрели решение задачи классификации примере классической задачи о пассажирах Титаника с использованием визуального программирования в программном пакете Orange [1]. Показали использование методов классификации в программе Orange на основе реальной базы данных Н. Юсупов, А. Савельева, О.Г.Леонова [2]. Д. В. Гринченков, Ф. Х. Нгуен, Т. Т. Нгуен, Д. А. Горбушин выполнили краткий обзор и сравнительный анализ возможностей алгоритмов, используемых для интеллектуального анализа данных [3]. В статье рассматривала и описывала один из алгоритмов Data Mining, предназначенных для решения задач классификации и прогнозирования - метод деревьев решений А.А. Мифтахова [4]. К.А. Малышенко, В.А.Малышенко, М.В. Анашкина показали прогрессивные возможности применения программного продукта с открытым кодом “Orange”, для реализации комплекса операций исследовательского, классификационного характера на основе данных о химическом составе вина [5].

1.3 Цель исследования

Цель исследования - решить задачу классификации набора данных состоящий из слов названий различных продуктов и их категорий.

2 Материалы и методы

Для решения задачи классификации используется программа Orange. Работа будет происходить на готовом наборе данных состоящий из слов названий различных продуктов и их категорий, скачать которые можно по ссылке:

https://docs.google.com/spreadsheets/d/1psK2WHN3xVUW0d3FO3_2Xh6Qlsl2NCRU/edit?usp=sharing&ouid=104272149632818699735&rtpof=true&sd=true

3 Результаты и обсуждения

Перед началом работы требуется установить Orange с официального сайта и установить.

Создадим новый файл (см.рис.1).

Рисунок-1 Создание нового файла

Для решения задачи классификации необходимо установить дополнение Text. Для того, чтобы скачать дополнение, необходимо перейти в Options, далее в Add-ons, в появившимся окне выбираем Text (см.рис.2).

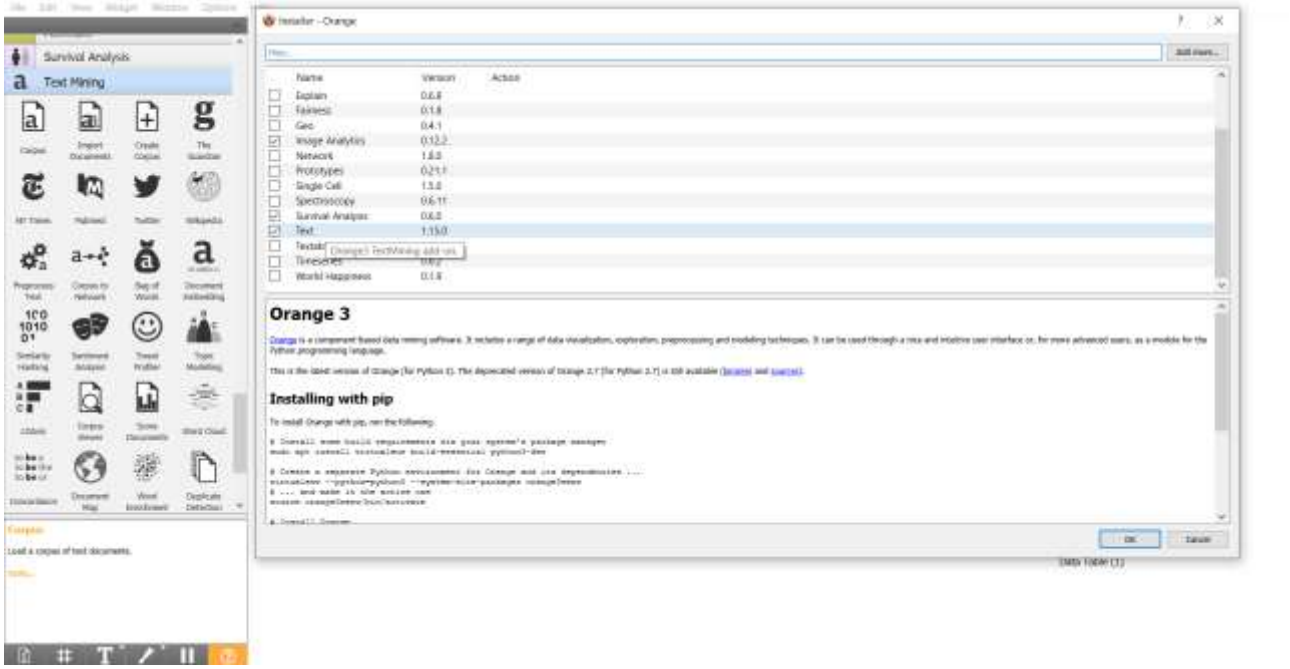

Рисунок-2 Установка дополнения Text

Для того, чтобы загрузить набор данных состоящий из названий различных продуктов и их категорий, необходимо из раздела Data выбрать виджет File на холст и перенести его на холст (см.рис.3).

Рисунок- 3 Добавление виджета File на холст

Открываем виджет File и добавляем набор данных words-food.xlsx. Набор данных words-food.xlsx содержит 108 данных. Для колонки category выбираем атрибут target (см.рис.4).

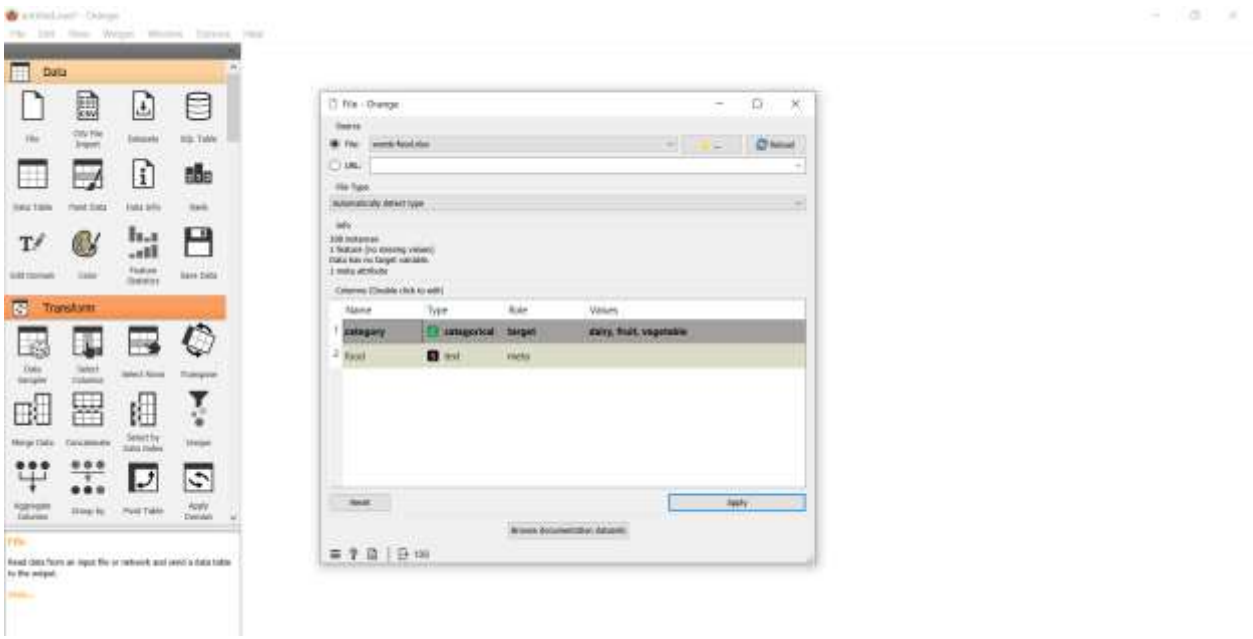

Рисунок-4 Добавление набора данных words - food.xlsx

Далее добавляем виджет Corpus на холст, и соединяем с виджетом File (см.рис.5).

Рисунок-5 Добавление виджета Corpus на холст

Открываем окно виджета Corpus. В открывшемся окне выбираем язык English (см.рис.6)

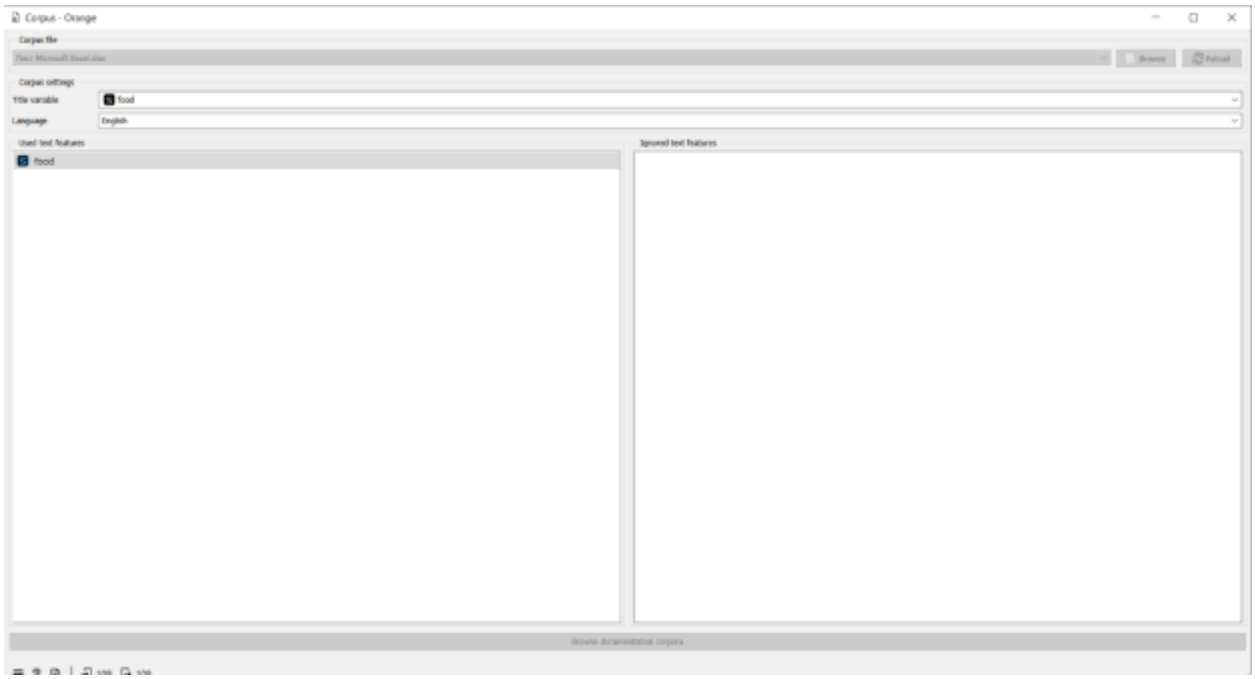

Рисунок-6 Просмотр набора данных

Далее добавляем виджет Document Embedding на холст, и соединяем с виджетом Corpus. Виджет Document Embedding представляет слова в многомерном пространстве таким образом, что слова со схожими значениями имеют сходное вложение. Это означает, что каждое слово сопоставляется с вектором вещественных чисел, представляющих слово (см.рис.7).

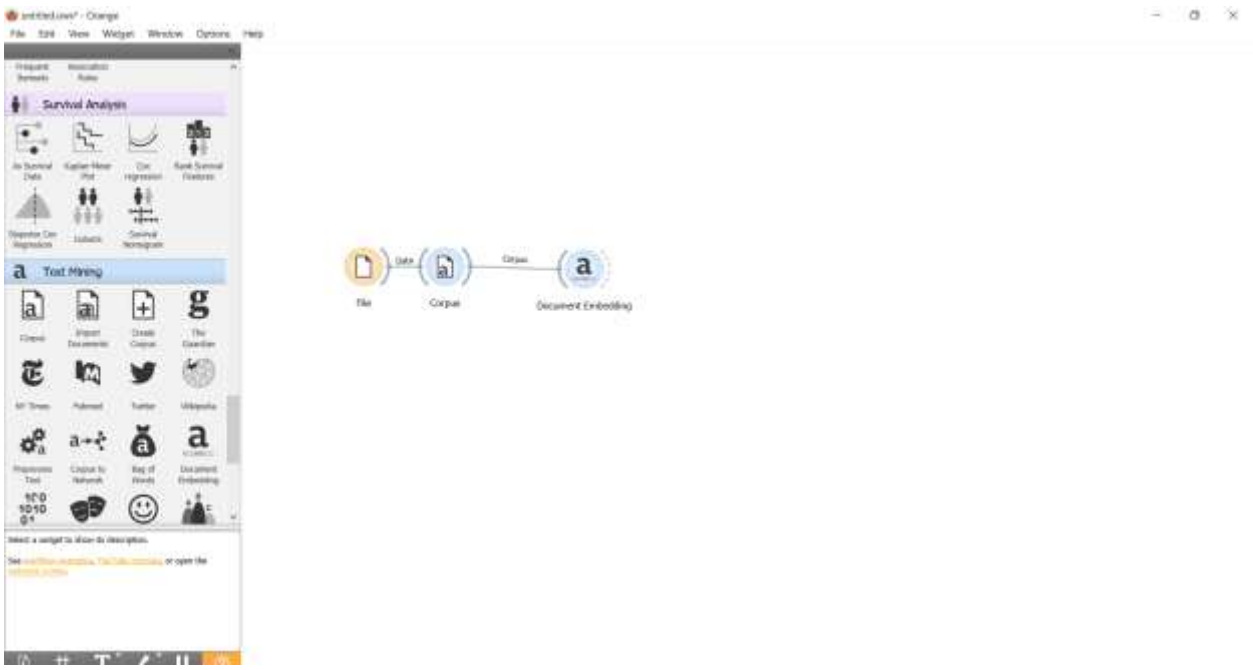

Рисунок- 7 Добавление виджета Document Embedding на холст

Открываем виджет Document Embedding, и в появившемся окне выбираем fastText (см.рис.8).

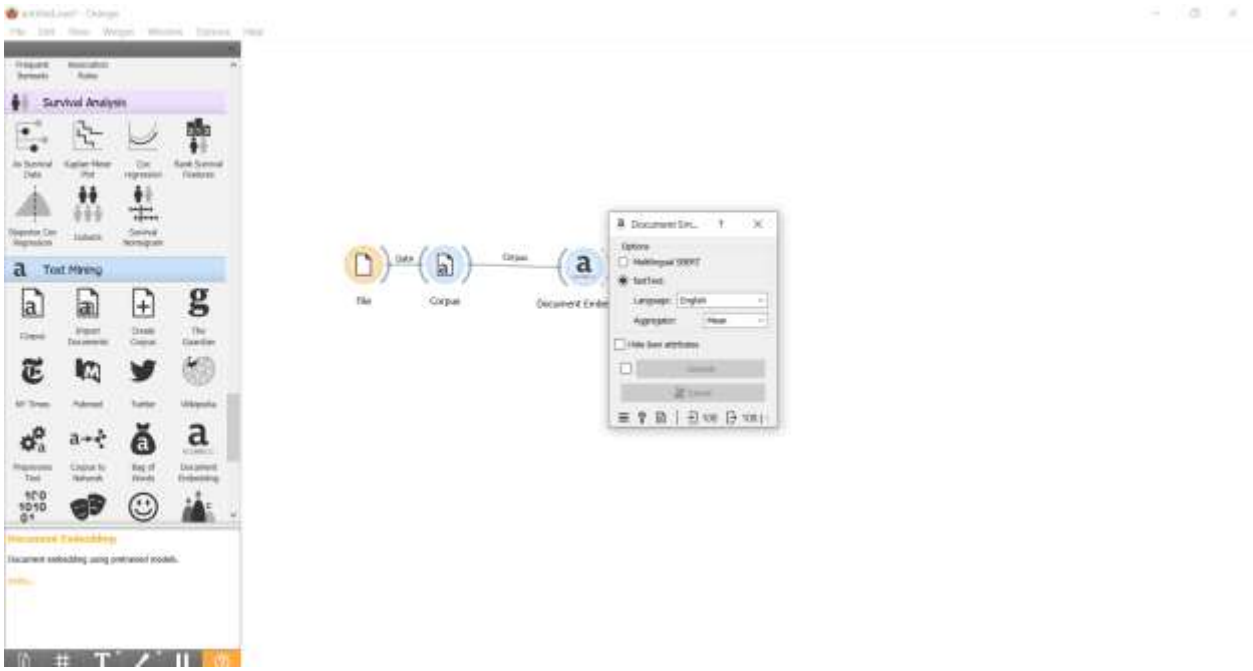

Рисунок- 8 Изменение настроек виджета Document Embedding

Добавим виджет Data Table на холст, и соединим с виджетом Document Embedding, для того чтобы посмотреть данные виджета Document Embedding (см.рис.9).

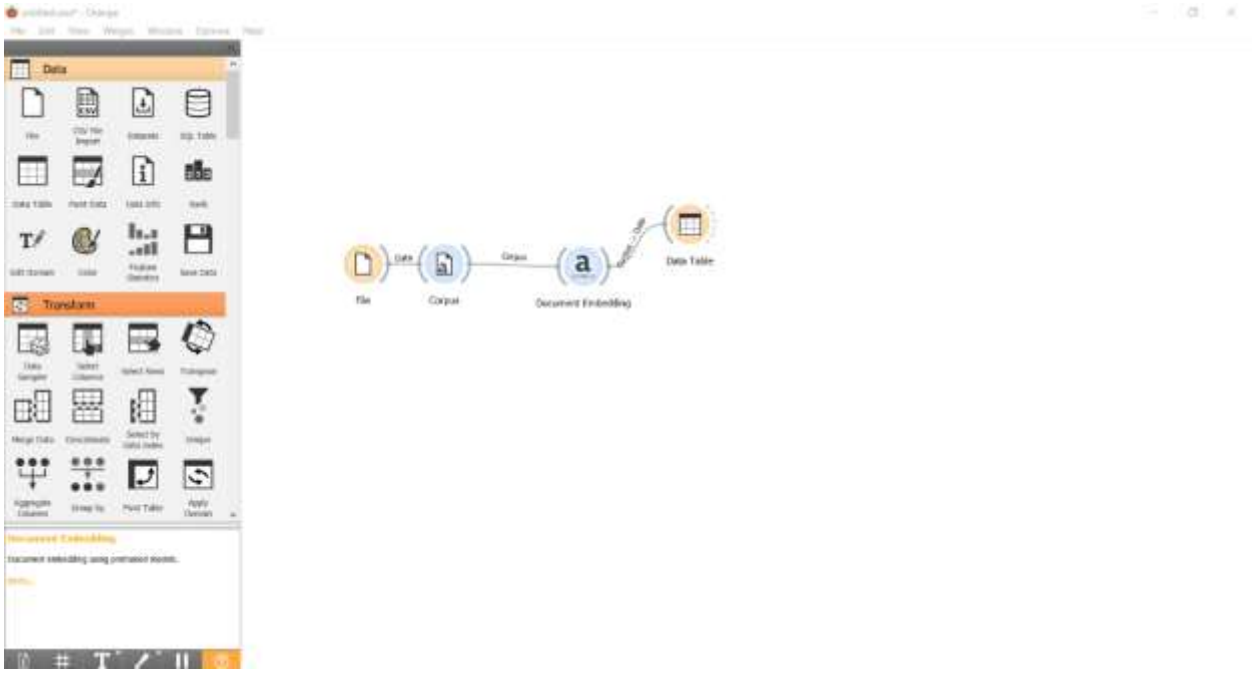

Рисунок-9 Добавление виджета Data Table

Открываем виджет Data Table. С помощью таблицы можем увидеть, что для каждого слова добавились 300 дополнительных признаков (см.рис.10).

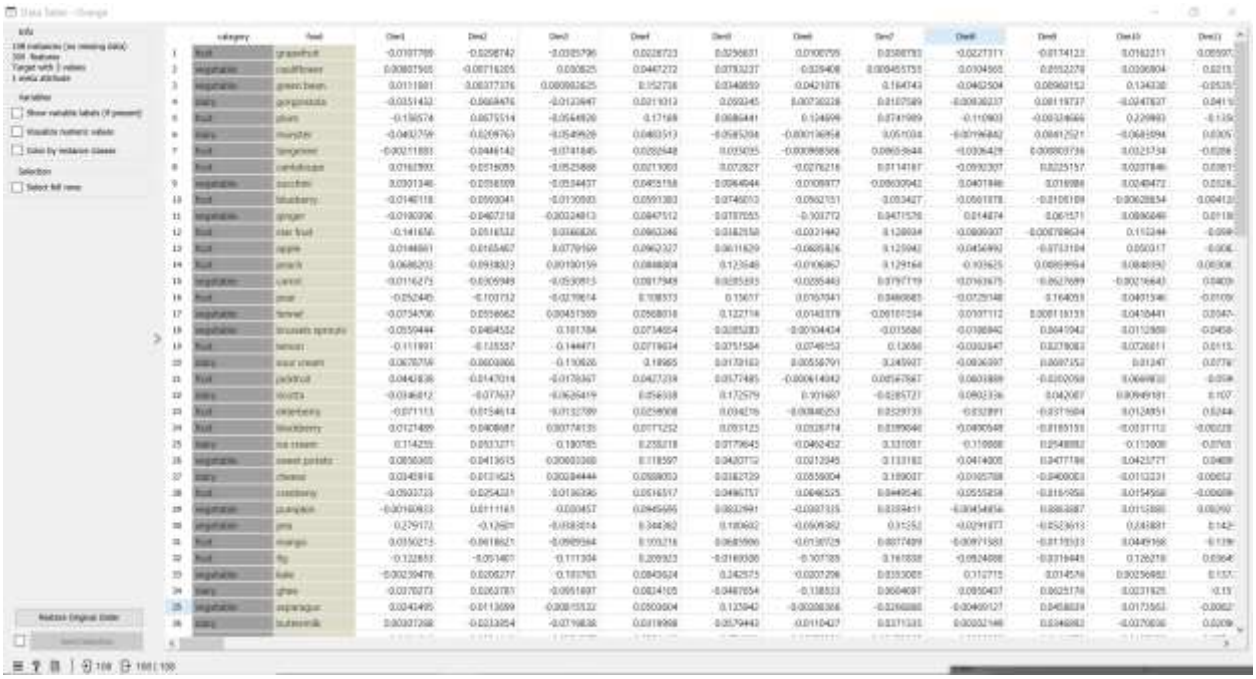

Рисунок-10 Просмотр данных виджета Document Embedding

Далее добавляем виджет Test and Score на холст, и соединяем с виджетом Document Embedding. Виджет Test and Score выводит результаты тестирования алгоритмов классификации (см.рис.11).

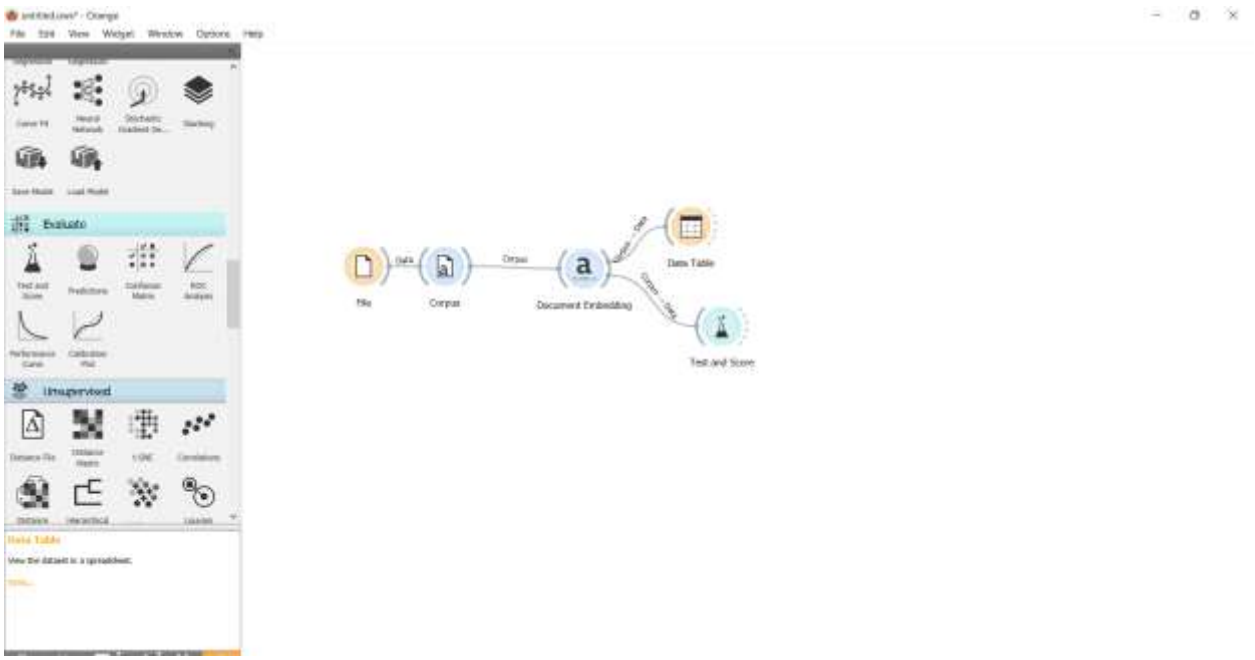

Рисунок-11 Добавление виджета Test and Score на холст

Добавляем виджет Logistic Regression и соединяем с виджетом Test and Score. Логистическая регрессия является статистическим методом, используемый для прогнозирования вероятности возникновения некоторого события путем подгонки данных к логистической кривой. В Orange, логистическая регрессия является одним из инструментов для решения задач классификации. (см.рис.12).

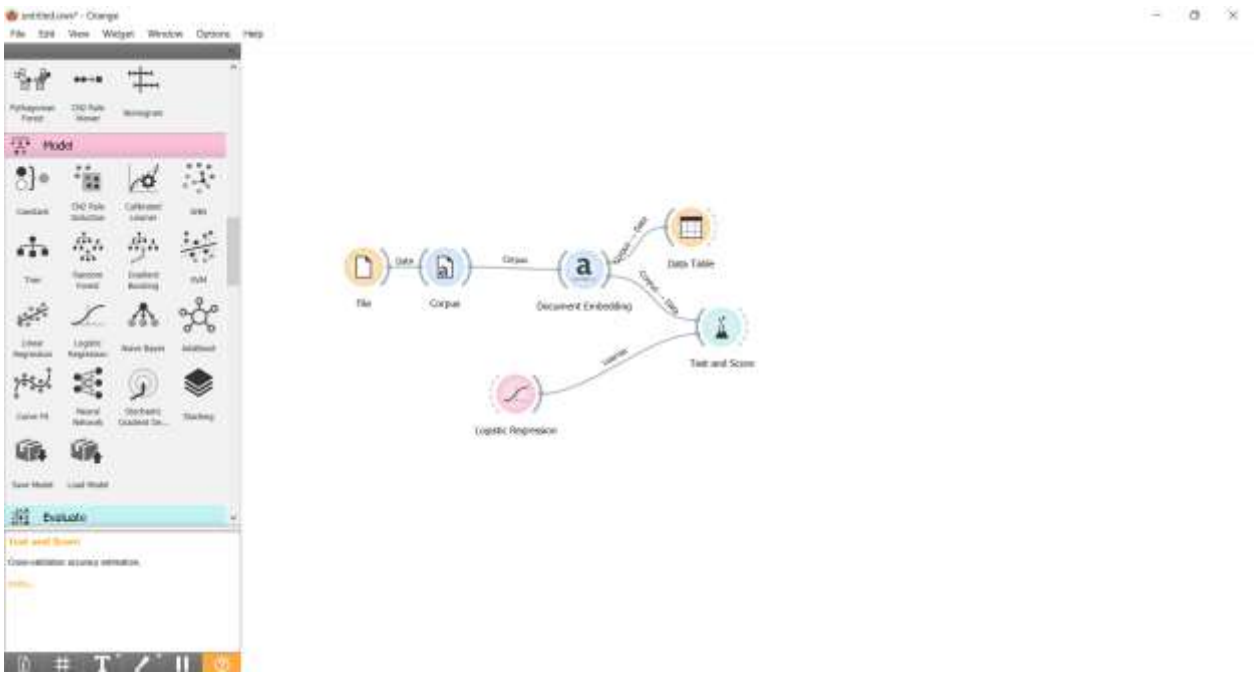

Рисунок - 12 Добавление виджета Logistic Regression на холст

Открываем виджет Test and Score, и можем увидеть, что появились результаты логической регрессии. Точность классификации (CA) показывает

результат 0.981, данный результат показывает, что существуют несколько неправильных классификаций (см.рис.13).

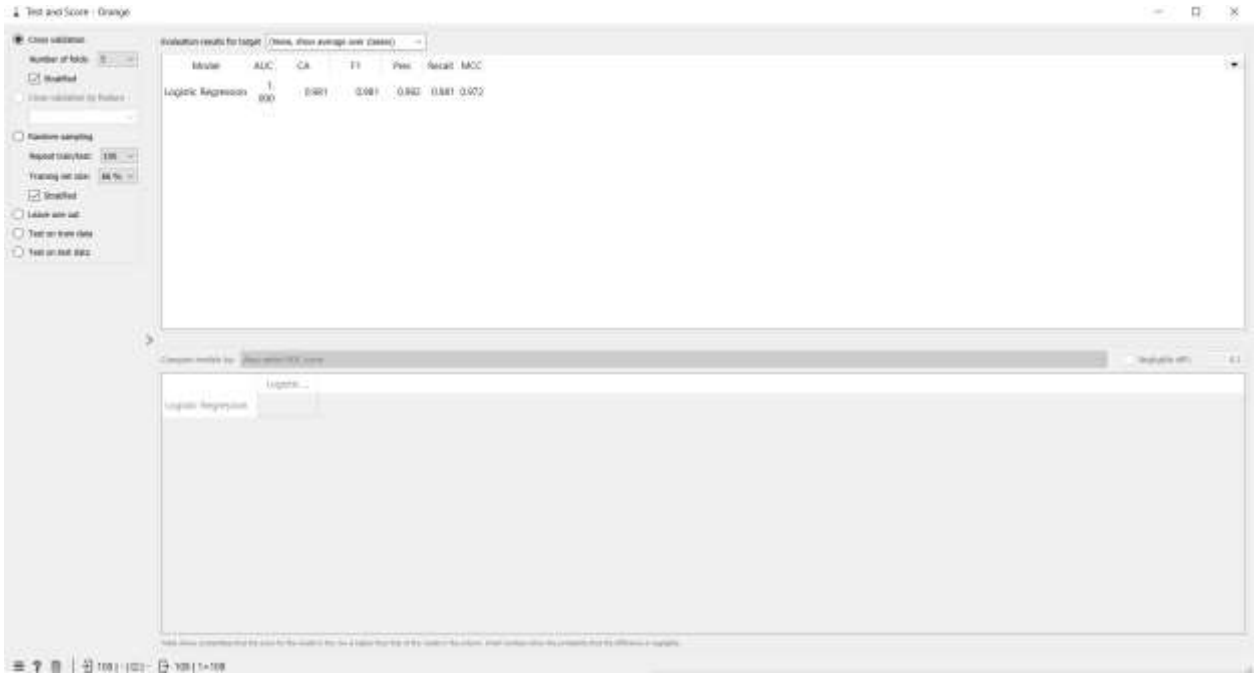

Рисунок - 13 Просмотр данных Test and Score

Для того, чтобы посмотреть количество неправильных классификаций, добавим виджет Confusion Matrix и соединим с виджетов Test and Score (см.рис.14).

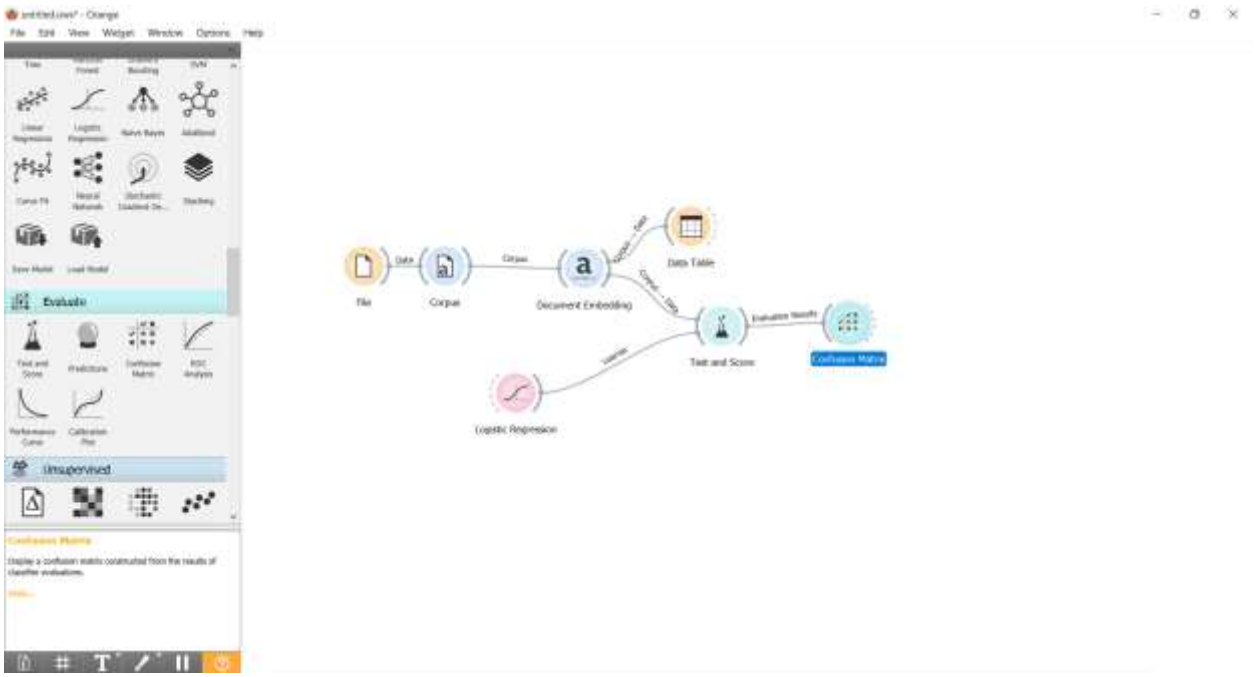

Рисунок - 14 Добавление виджета Confusion Matrix

С помощью матрицы путаницы можем увидеть, что неправильно квалифицировало два овоща (см.рис.15).

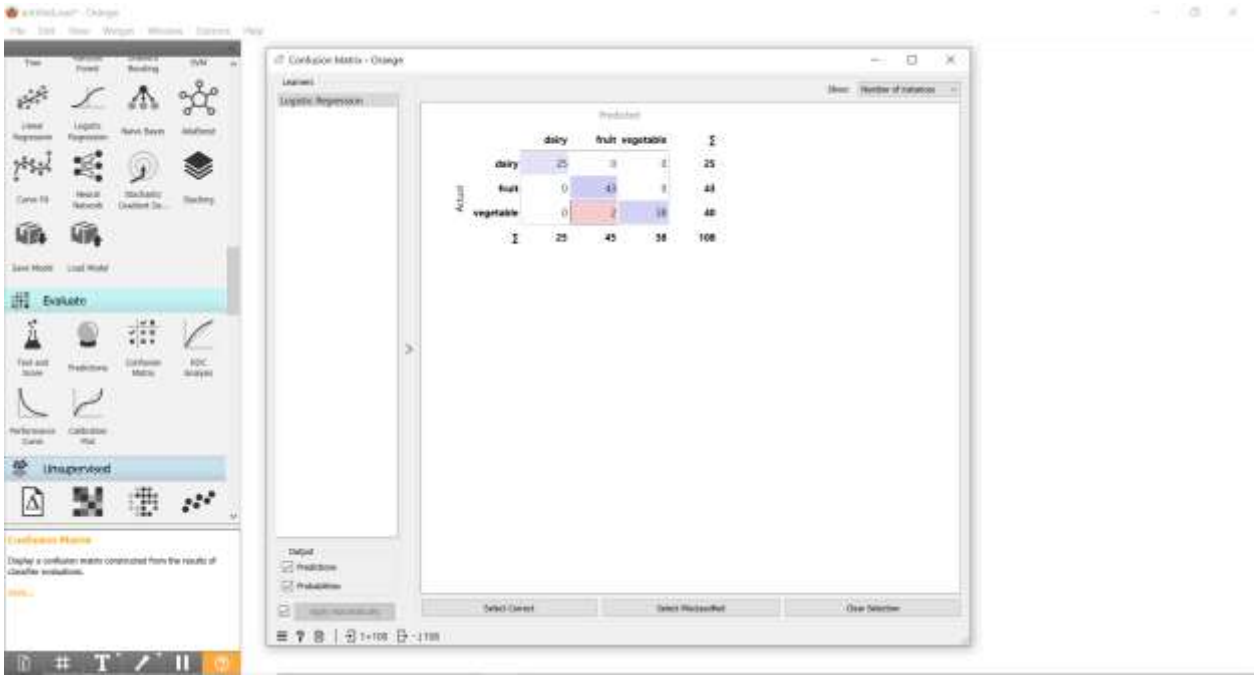

Рисунок - 15 Просмотр данных виджета Confusion Matrix

Для того, чтобы посмотреть название продуктов ошибочной классификации, добавляем виджет Data Table, и соединяем с виджетом Confusion Matrix (см.рис.16).

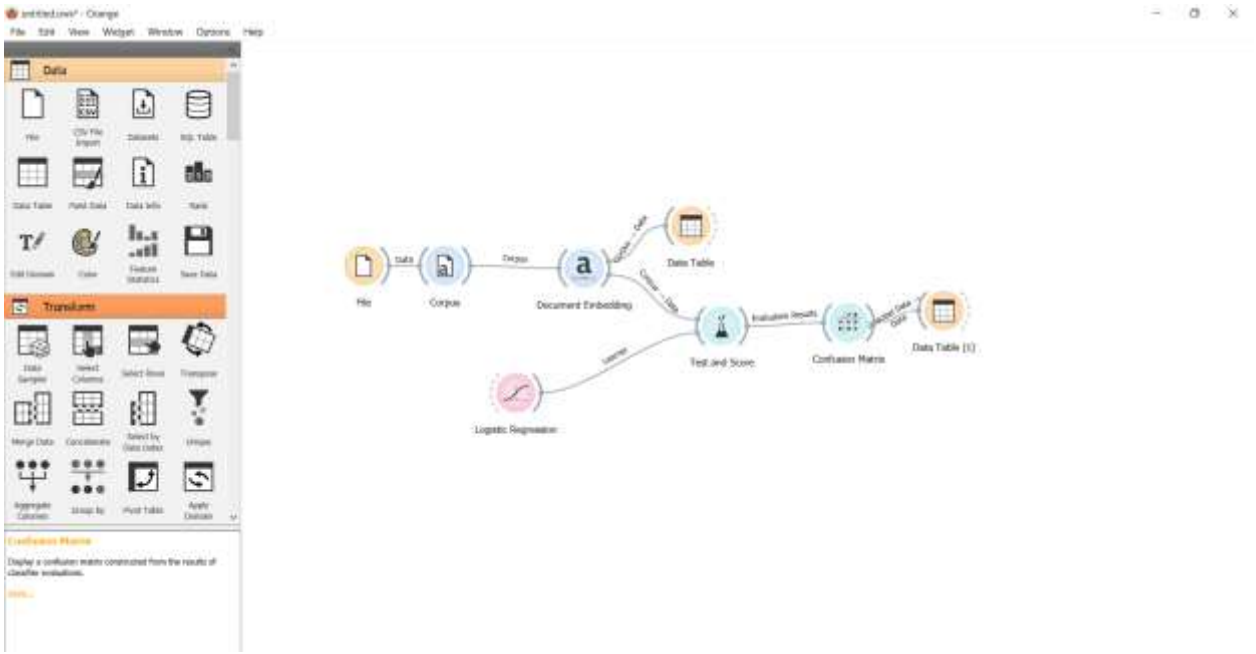

Рисунок - 16 Добавление виджета Data Table на холст

Открываем виджет Data Table и виджет Confusion Matrix. В виджете Confusion Matrix выбираем ошибочные классификации, и в виджете Data Table можно увидеть их название. В таблице можем увидеть, что продукты тыква и инжир неправильно классифицировались (см.рис.17).

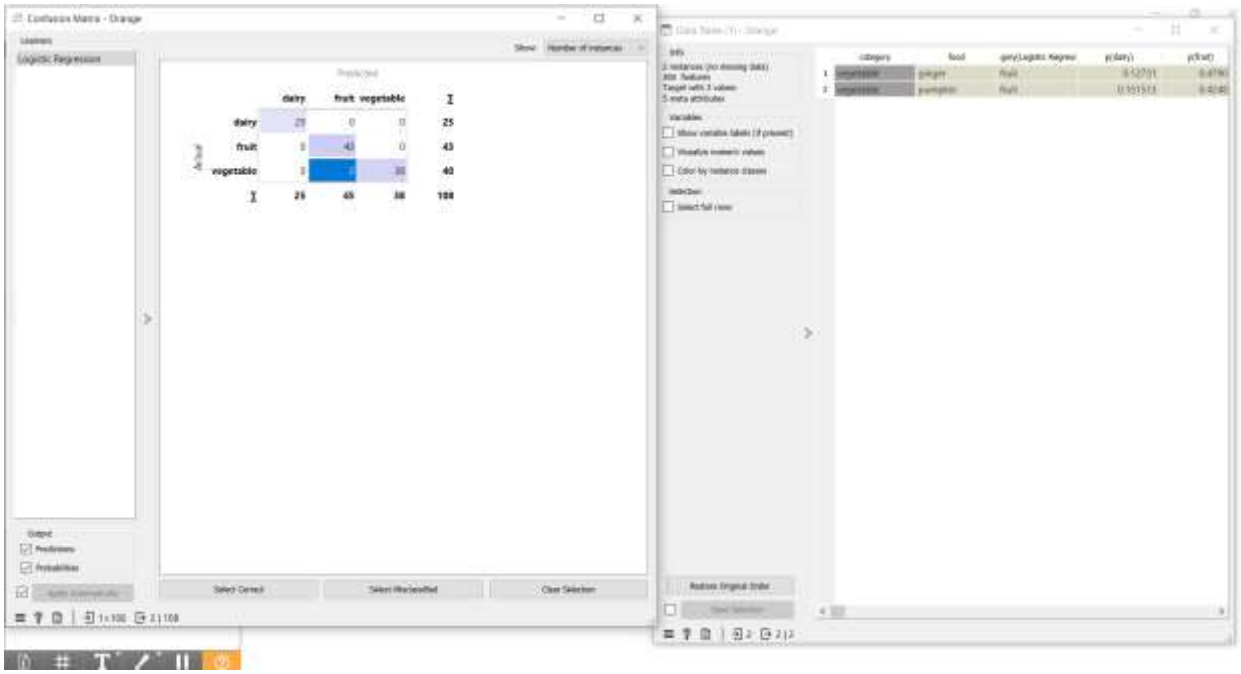

Рисунок - 17 Просмотр данных виджетов Confusion Matrix и Data Table

Также можно попробовать классифицировать продукты которых нет в наборе данных. Для этого добавим виджет Word List (см.рис.18).

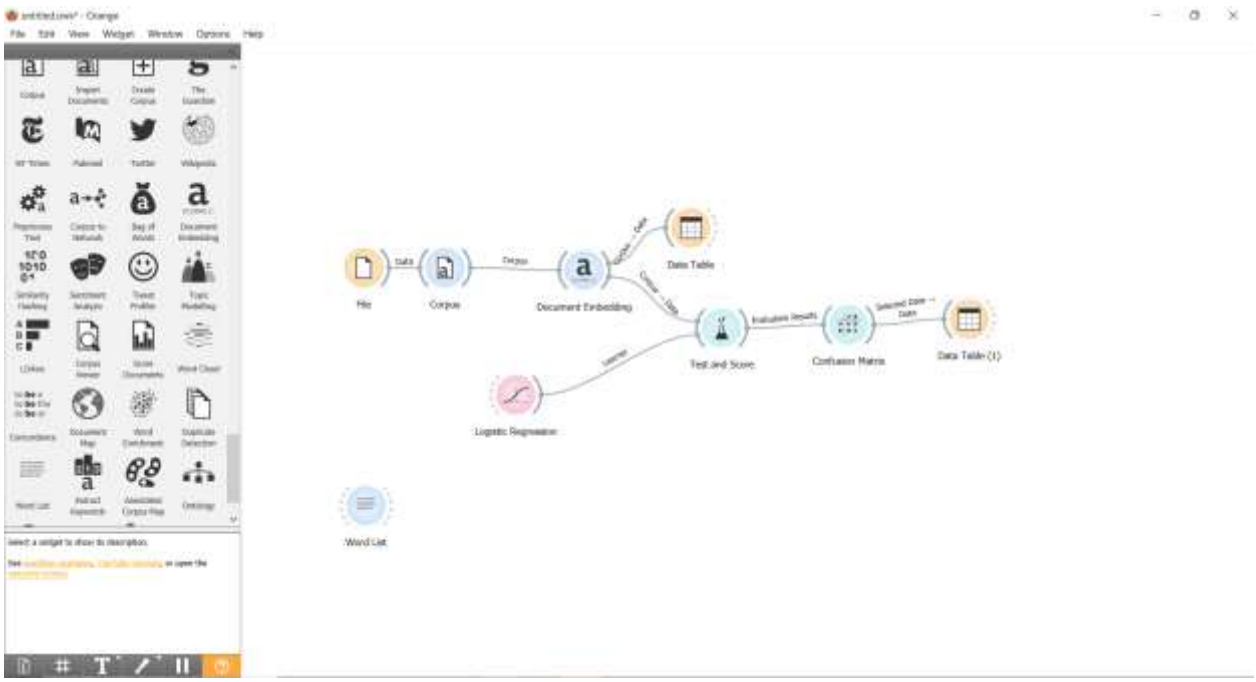

Рисунок - 18 Добавления виджета Word list на холст

Открываем виджет Word List. Для того, чтобы добавить слова нажимаем на кнопку +, и добавляем слова финики, клементина, камамбер и гауда на английском языке (см.рис.19).

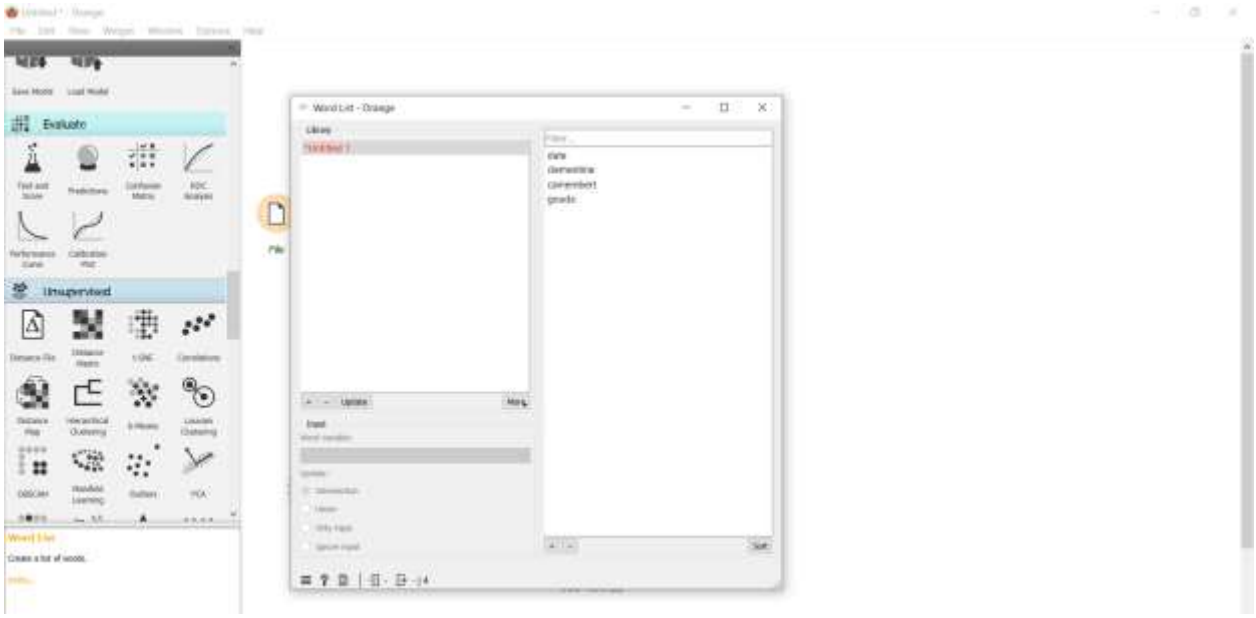

Рисунок - 19 Добавление слов в виджет Word List

Добавляем виджет Corpus и соединяем с виджетом Word List. После соединения в паявшимся окне выбираем соединение Words к Data (см.рис.20).

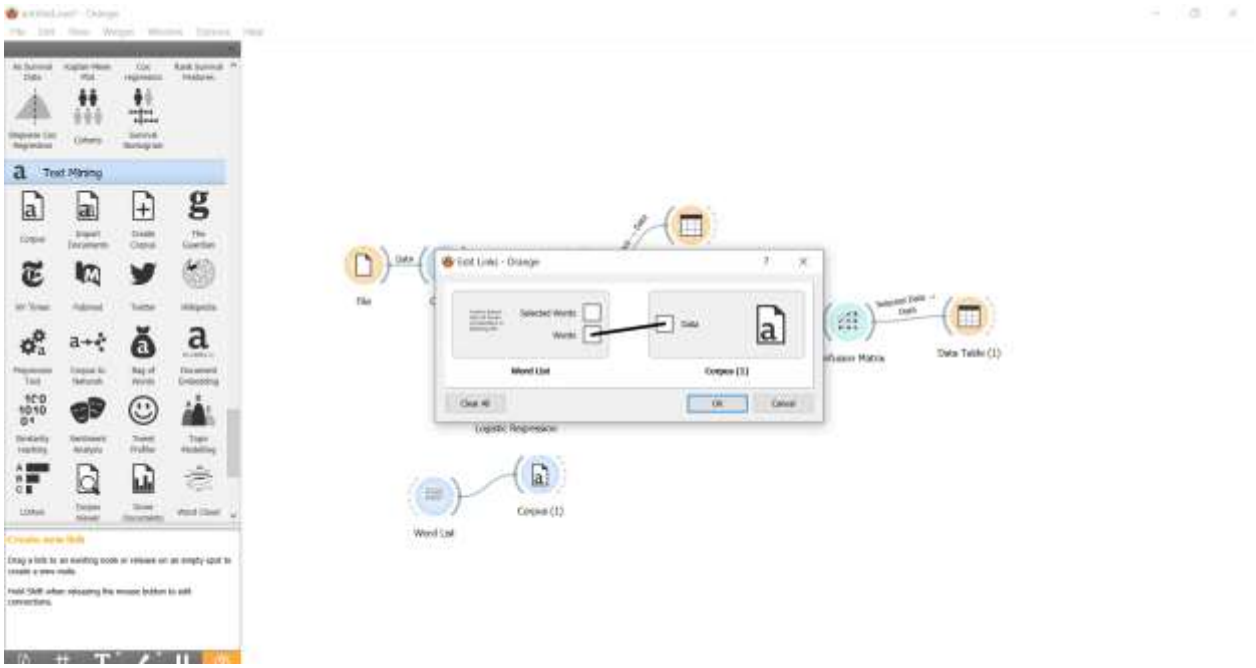

Рисунок - 20 Добавление виджета Corpus на холст

Необходимо проверить, чтобы в виджете Corpus (1) был указан английский язык (см.рис.21).

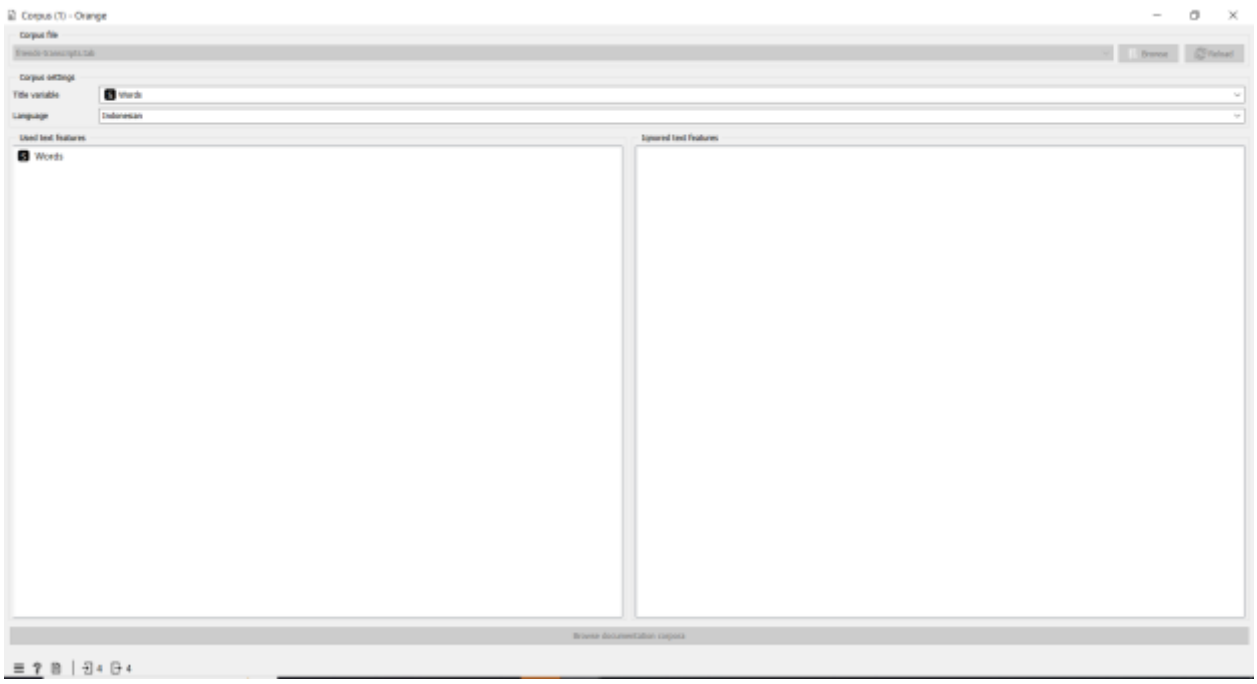

Рисунок - 21 Просмотр виджета Corpus

Далее добавляем виджет Document Embedding на холст, и соединяем с виджетом Corpus (1) (см.рис.22).

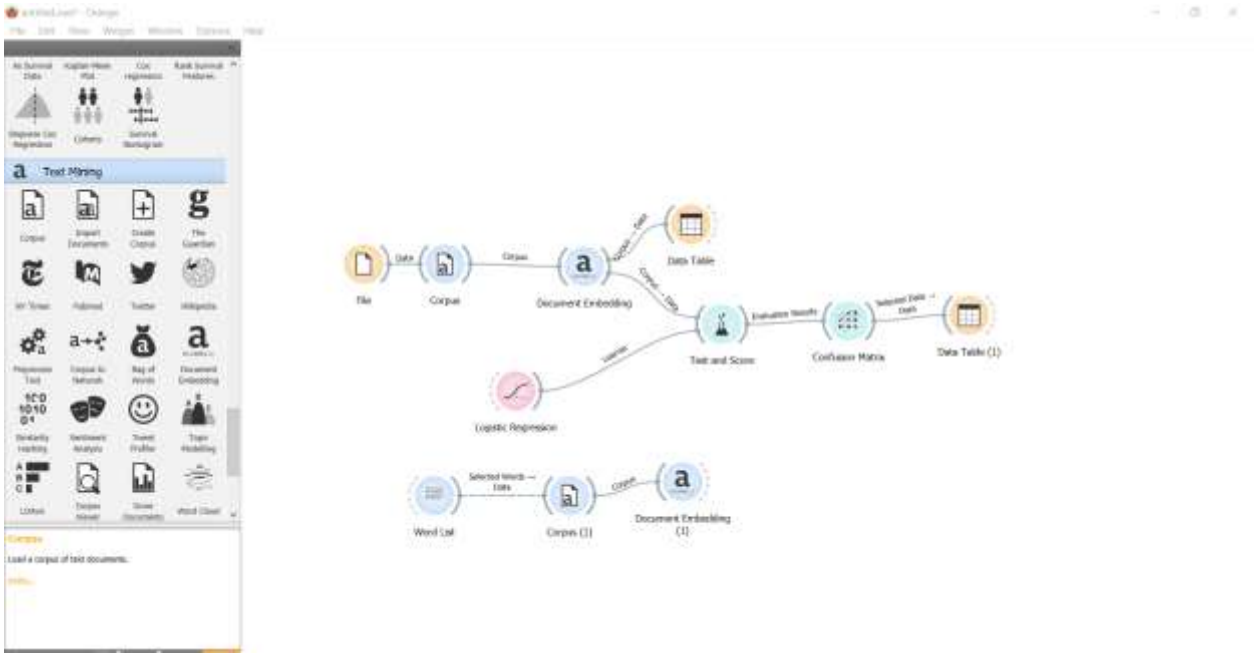

Рисунок - 22 Добавление виджета Document Embedding на холст

Откроем виджет Document Embedding (1), для того, чтобы убедиться, что выбран английский язык (см.рис.23).

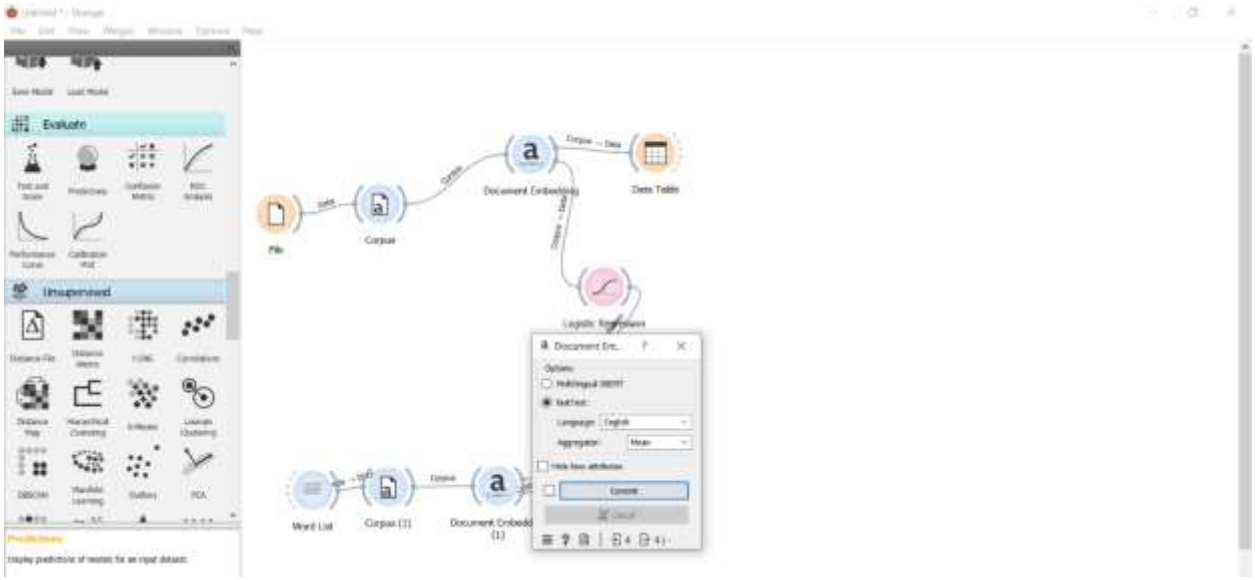

Рисунок - 23 Просмотр виджета Document Embedding (1)

Добавляем виджет Data Table и соединяем с виджетом Document Embedding (1) (см.рис.24).

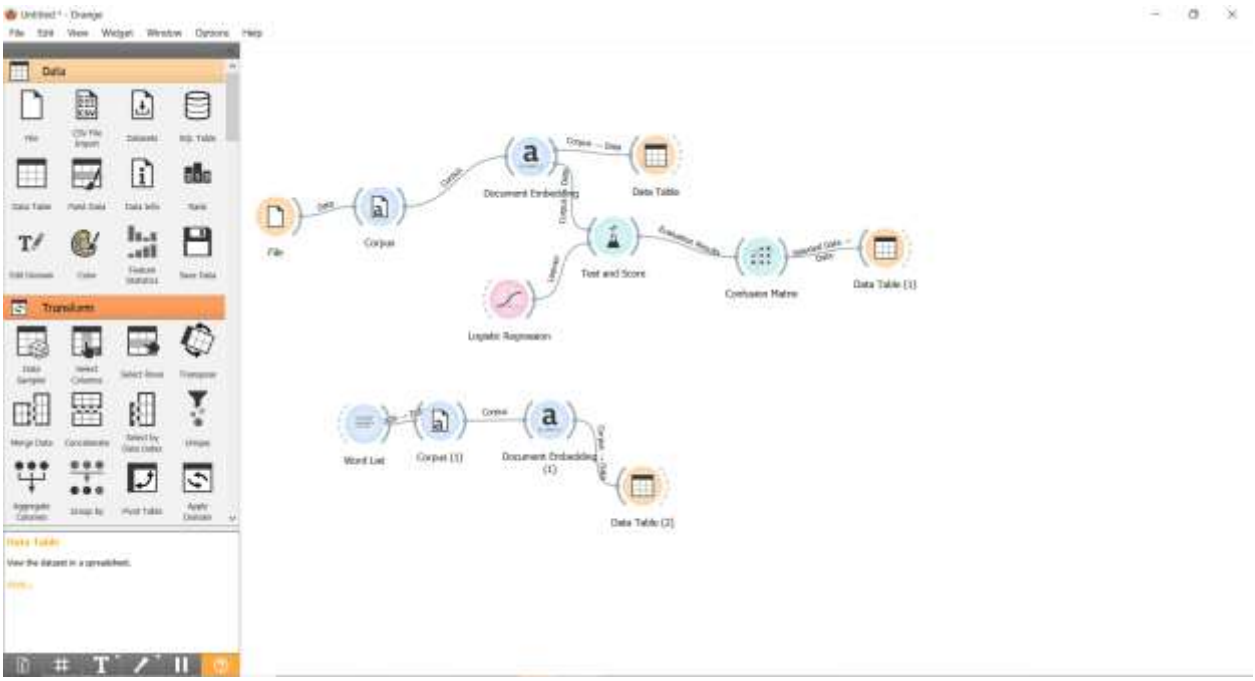

Рисунок - 24 Добавление виджета Data Table на холст

Открываем виджет Data Table (2), и можем увидеть, что все слова, которые добавляли в виджете Word List отображаются (см.рис.25).

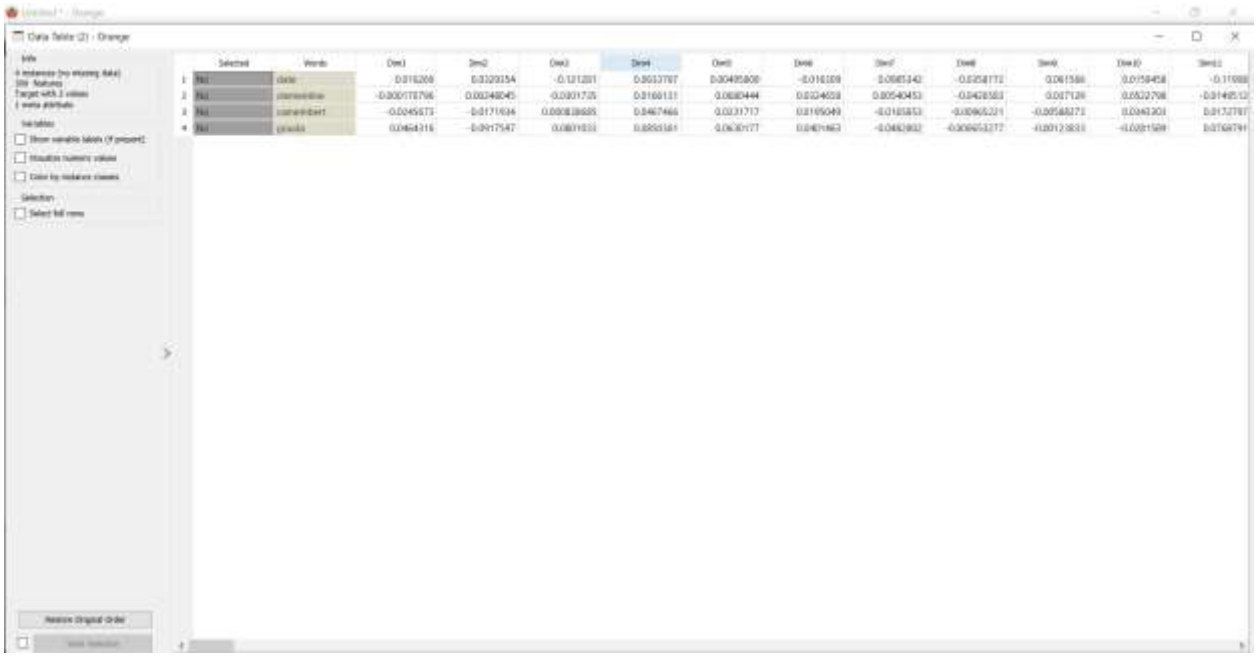

Рисунок - 25 Просмотр данных виджета Data Table

Далее необходимо удалить виджеты Test and Score, Confusion Matrix и Data Table (1) с холста (см.рис.26).

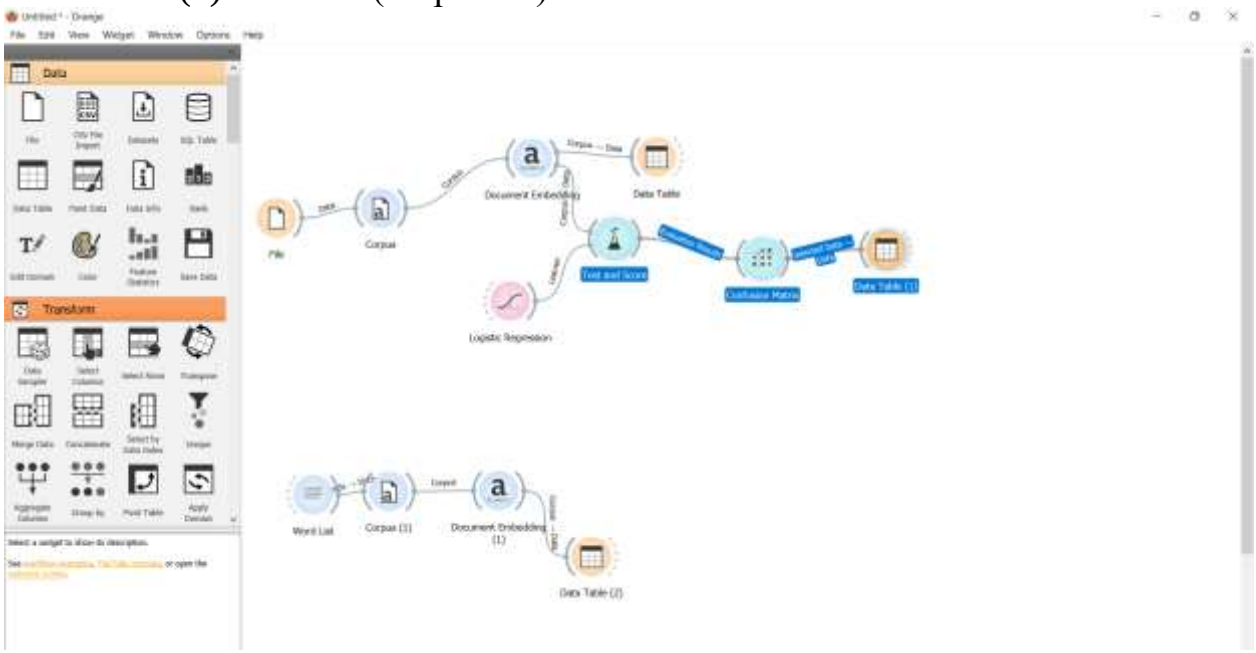

Рисунок - 26 Удаление виджетов Test and Score, Confusion Matrix и Data Table (1) с холста

Виджет Logistic Regression соединяем с виджетом Document Embedding (см.рис.27).

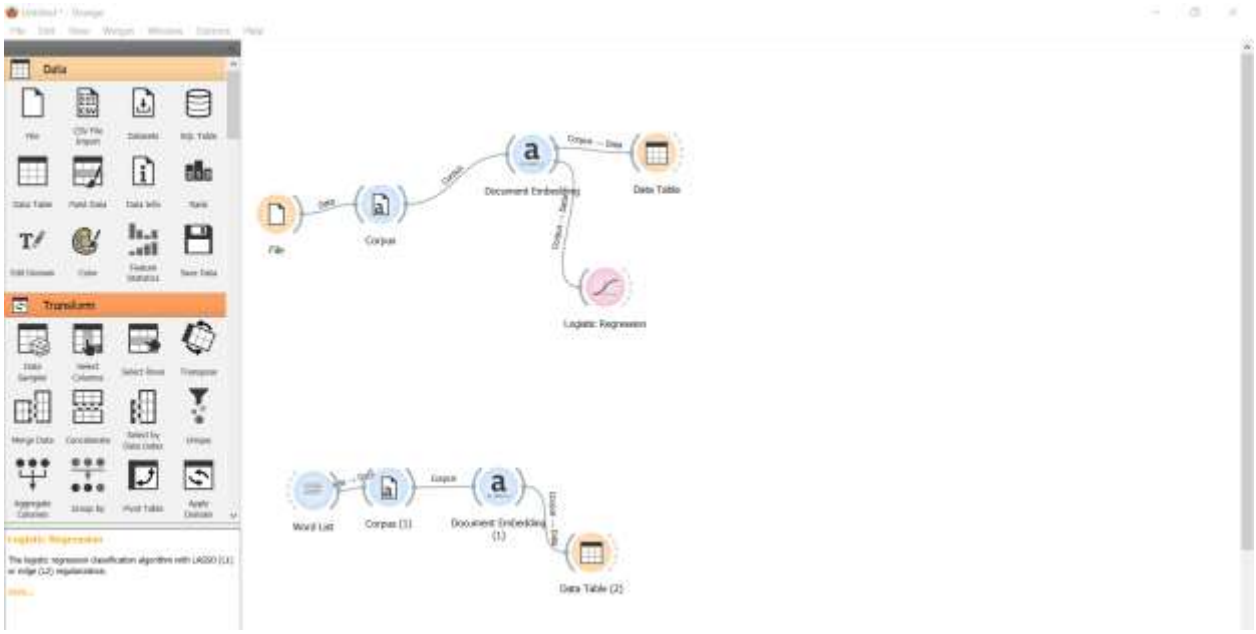

Рисунок - 27 Соединение виджета Logistic Regression с виджетом Document Embedding

Далее добавляем виджет Predictions на холст, и соединяем с виджетами Logistic Regression и Document Embedding (1). Виджет Predictions показывает прогнозы моделей на основе данных (см.рис.28).

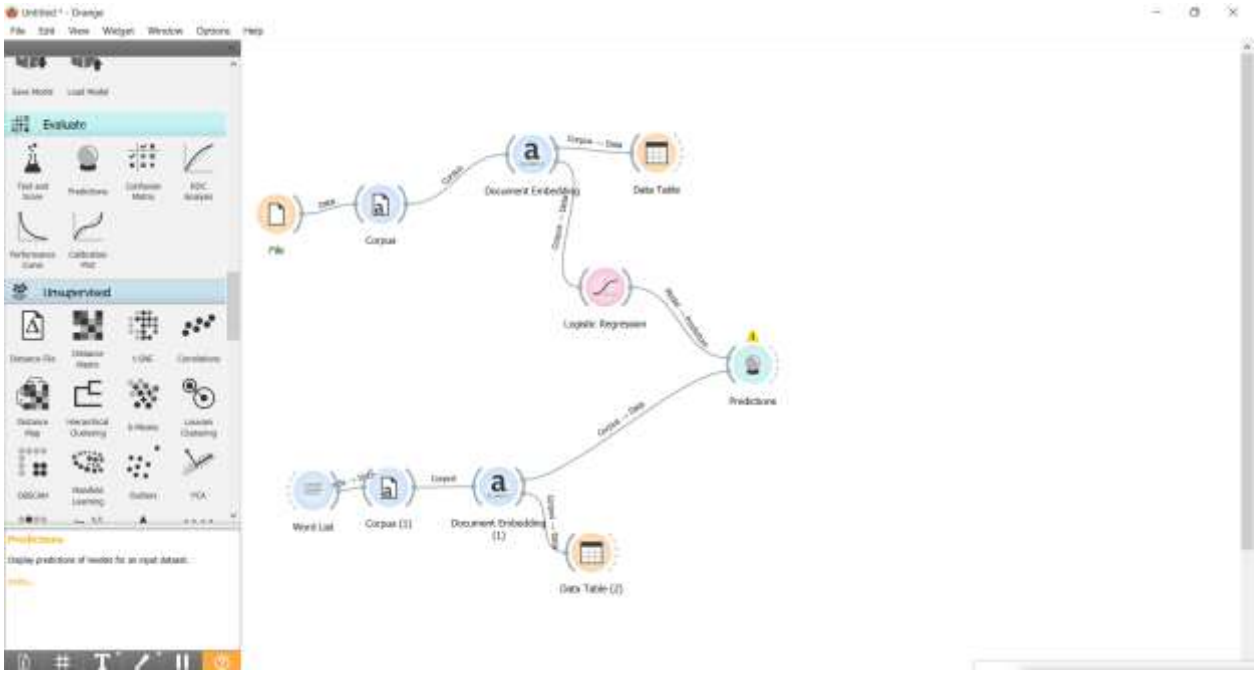

Рисунок - 28 Добавление виджета Predictions на холст

Открываем виджет Predictions, и можем увидеть, что финики и клементина отнеслись к категории фрукты, а виды сыра камамбер и гауда отнеслись к молочным продуктам (см.рис.29).

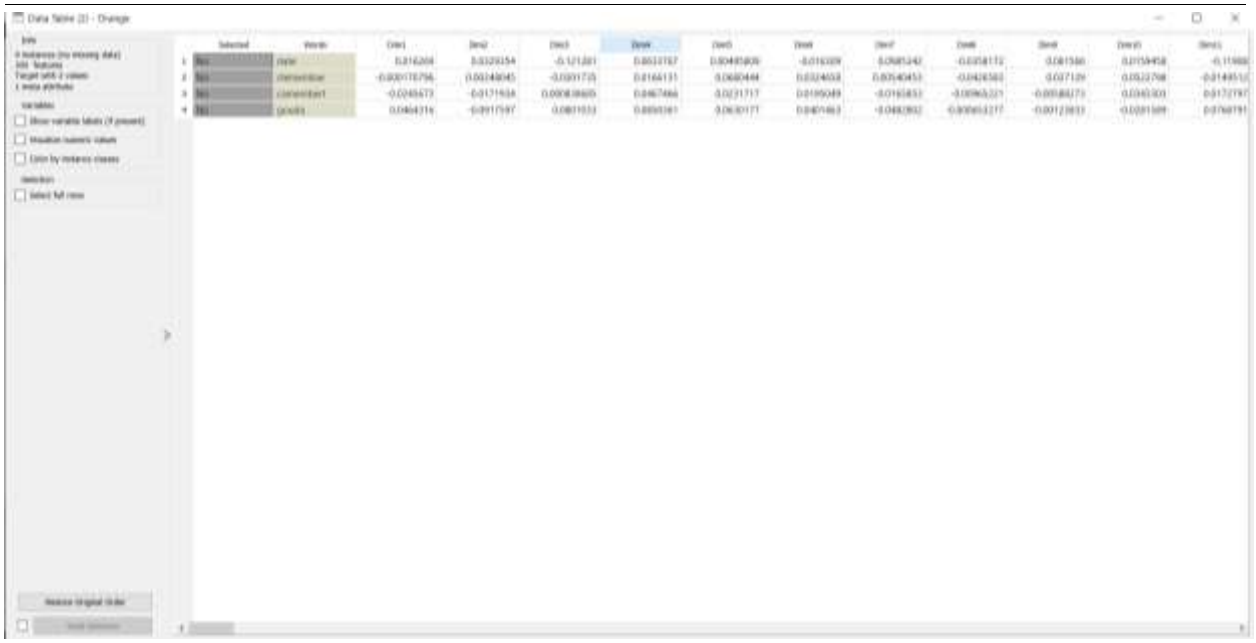

Рисунок - 29 Просмотр данных виджета Predictions

В ходе решения задач классификации получились две итоговые схемы.

Первая итоговая схема решения задачи классификации с готовым набором данных состоящий из слов названий различных продуктов и их категорий (см.рис.30).

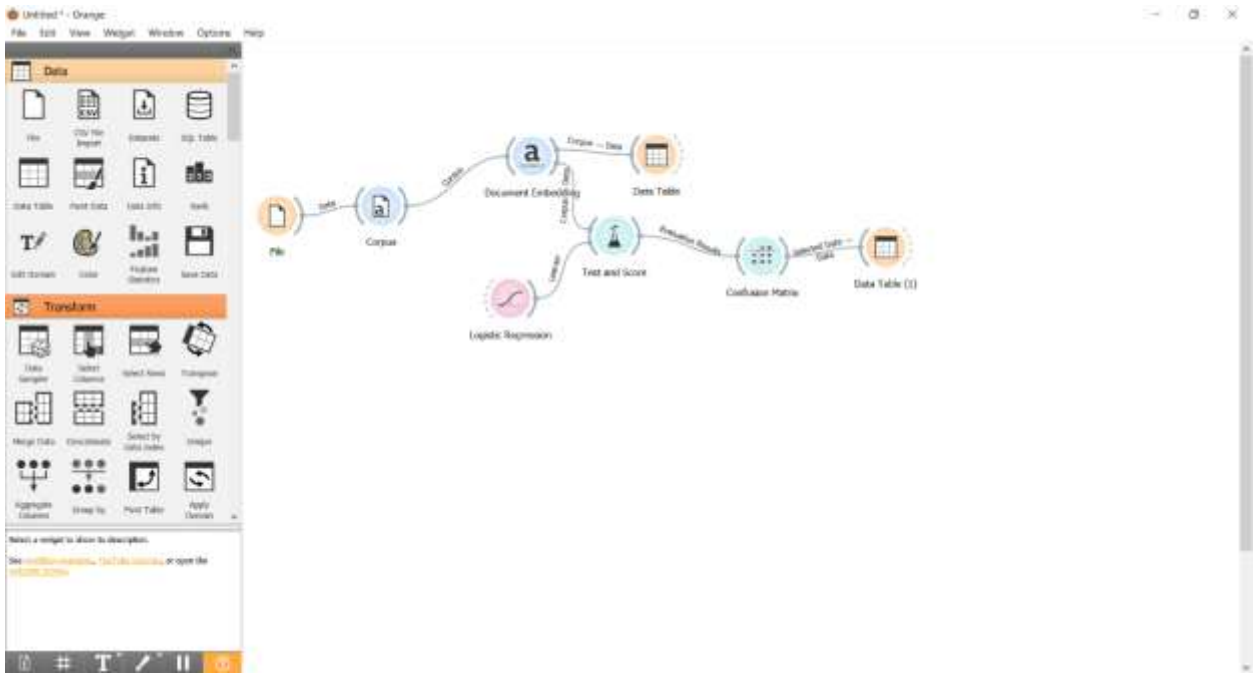

Рисунок - 30 Итоговая схема решения задачи классификации с готовым набором данных продуктов и их категорий

Вторая итоговая схема решения задачи классификации с готовым набором данных состоящий из слов названий различных продуктов и их категорий и с добавлением слов, которых нет в наборе данных (см.рис.31).

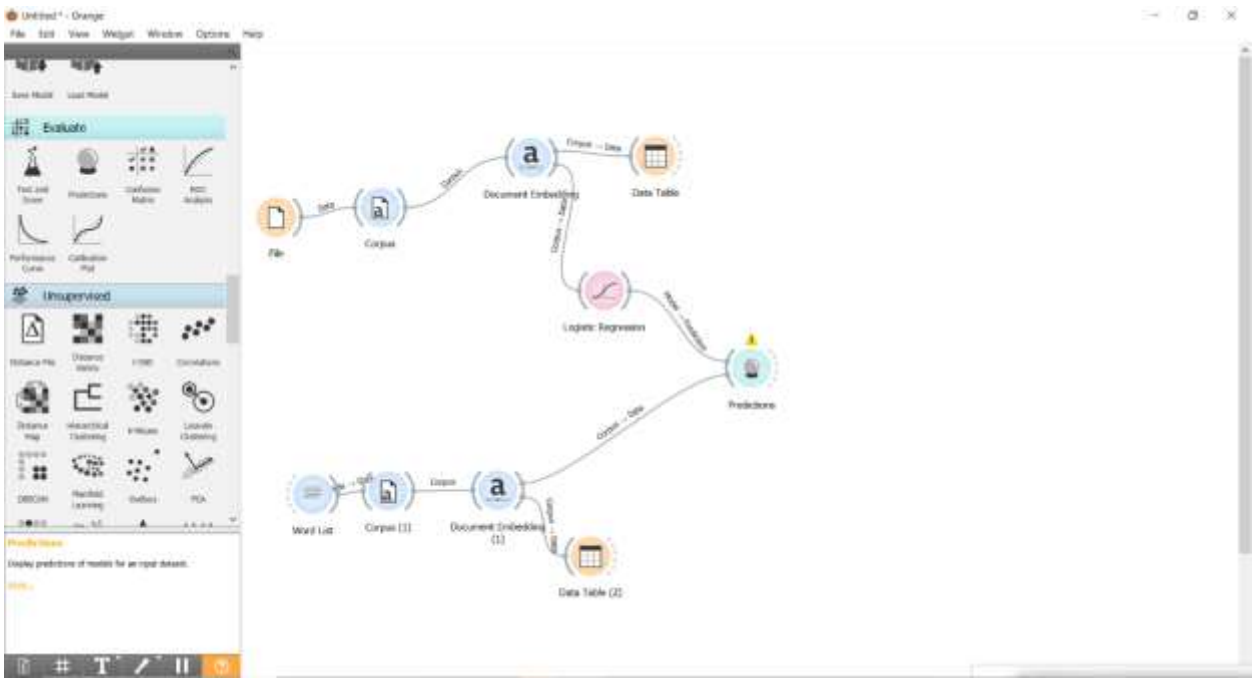

Рисунок - 31 Итоговая схема решения задачи классификации с готовым набором данных состоящий из слов названий различных продуктов и их категорий и с добавлением слов, которых нет в наборе данных

4 Выводы

В данной работе была выполнена задача классификации набора данных состоящий из слов названий различных продуктов и их категорий с помощью программного пакета визуального программирования на основе компонентов для визуализации данных Orange. С помощью виджетов Corpus, Corpus Viewer, Document Embedding, Data Table, Logistic Regression, Confusion Matrix, Test and Score, Word List, Predictions выполнили классификацию набора данных состоящий из слов названий различных продуктов и их категорий и получили две итоговых схемы.

Библиографический список

1. Мастевной С. С., Петрова А. Н. Data mining: обзор методов и области их применения // Наука, инновации и технологии: от идей к внедрению. 2022. С. 38-40.
2. Костырева С. А. и др. Решение задачи классификации с помощью визуального программирования в Orange // Точная наука. 2023. № 140. С. 18-21.
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.
4. Мифтахова А. А. Применение метода дерева решений для решения задач классификации и прогнозирования //Инфокоммуникационные технологии. 2016. Т. 14. №. 1. С. 64-70.

5. Малышенко К. А., Малышенко В. А., Анашкина М. В. Определение сорта вина на основе статистического анализа химических показателей // Дистанционные образовательные технологии. 2019. С. 297-307.