Отчёт по проекту

**“Регрессионный анализ возрастных изменений моллюсков на “Abalone Age Prediction” наборе данных”**

**1. Введение**

В кухнях всего мира морские ушки (Abalone) считаются деликатесом, а из их раковин делают украшения из-за их переливающегося блеска. Размер морских улиток варьируется от маленьких до очень крупных. Из-за экономической ценности и спроса на морское ушко его часто собирают на фермах, поэтому для прогнозирования его возраста необходимы физические измерения. При традиционном методе раковину прорезают по конусу, окрашивают, а затем пересчитывают под микроскопом их кольца, чтобы определить ее возраст. Для определения возраста морского ушка используется трудоемкий процесс прорезания раковины через конус, ее окрашивания и подсчета колец под микроскопом.

**2. Постановка задачи**

Данных проект предназначен для решения проблемы необходимости трудоёмкого процесса традиционного измерения возраста моллюсков Abalone.

Задача состоит в поиске оптимальной модели для решения задачи при помощи инструментов Apache PySpark и различных вспомогательных библиотек Python.

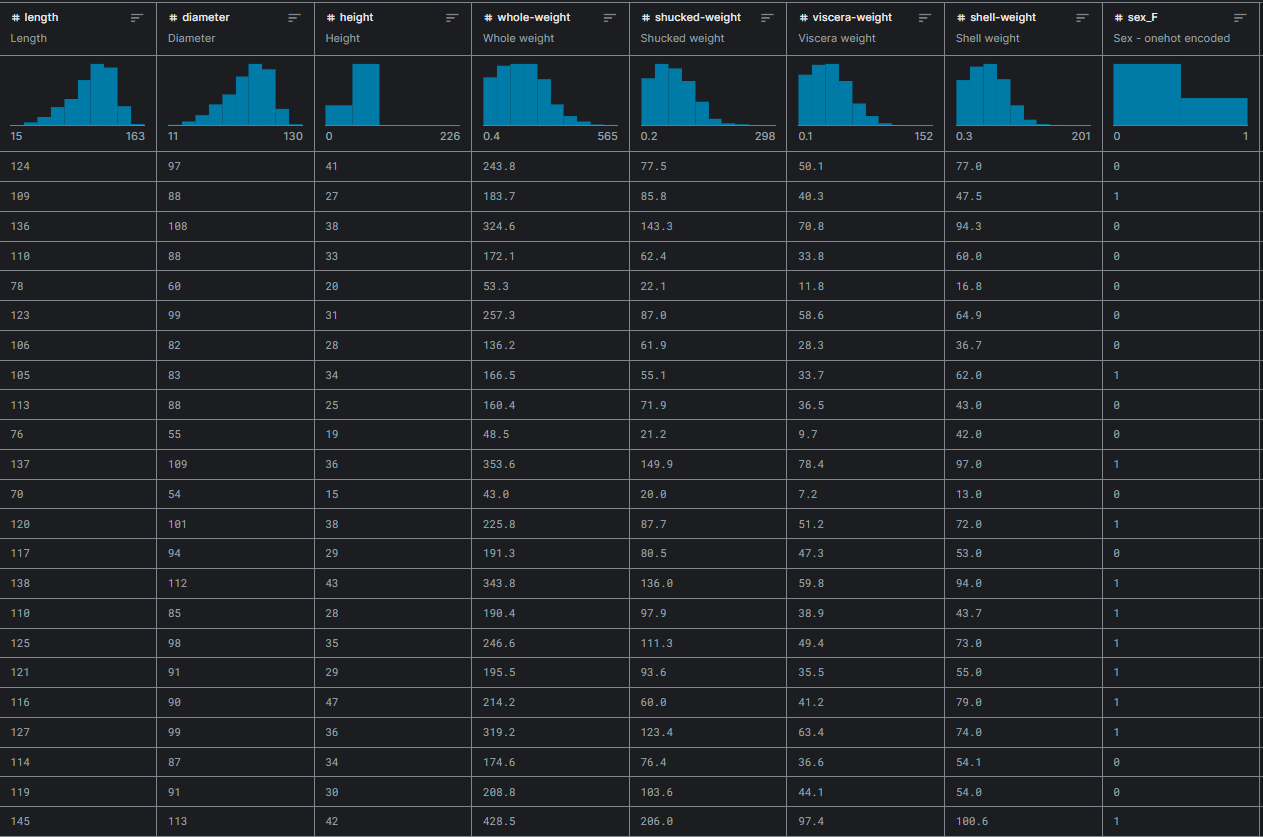
Был использован регрессионный анализ для определения возраста моллюсков по их размерам, полу и весу без использования данных полученных при прорезании раковин, что поможет облегчить процесс определения возраста моллюсков.

3. Описание набора данных

Набор данных Abalone Age Prediction представляет собой таблицу со следующими атрибутами:

* длина (наибольшее измерение ракушки) в мм (15-163)
* диаметр (перпендикуляр длине) в мм (11-130)
* высота в мм (4-50)
* общий вес в граммах (0.4-570)
* вес очищенного (граммы) (0.2-297)
* вес ракушки (граммы) (0.3-201)
* вес внутренности (граммы) (0.1-152)
* пол (мужской, женский, младенец)
* количество колец/возраст - (1-29 лет)

На следующем скриншоте можно видеть общую схему набора данных:



**3. Ход работы**

Гипотеза:

Возможно создать модель, достаточно точную для замены традиционного метода измерения возраста моллюсков Abalone при использовании регрессионных моделей Apache PySpark.

Для решения задачи были выбраны следующие модели:

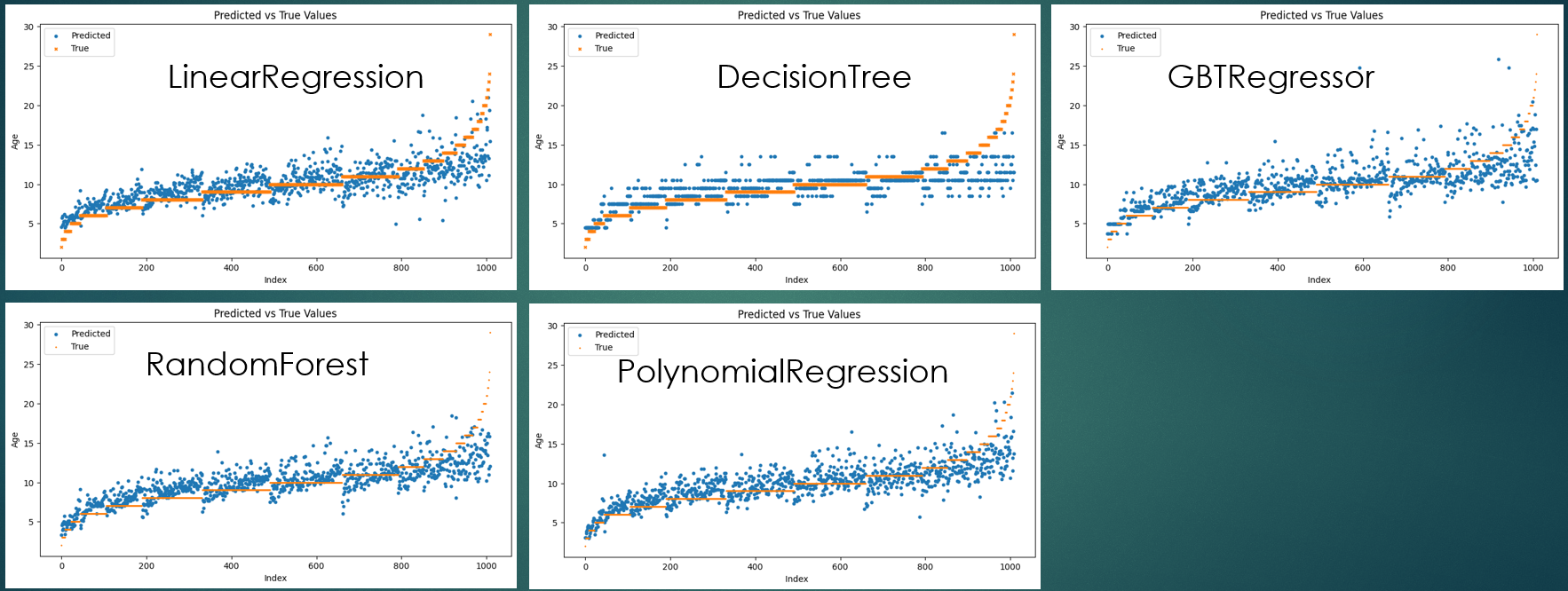
* Линейная регрессия
* Решающие деревья
* Случайный лес
* Полиномиальная регрессия
* Деревья на повышении градиента

Каждый метод может оказаться подходящим для данного датасета из-за своей структуры, поэтому был проведён анализ, подобраны параметры и было проведено сравнение результатов. Для более гибких моделей был использован GridSearch для поиска оптимальных гиперпараметров.

Для измерения ошибки была использована мера различия RMSE, в итоге были получены следующие результаты:

* Линейная регрессия (RMSE = 2.4818)
* Решающие деревья (RMSE = 2.4818)
* Случайный лес (RMSE = 2.10413 с GridSearch и ~2.25 без использования GridSearch)
* Полиномиальная регрессия (RMSE = 2.0575)
* Деревья на повышении градиента (Gradient-Boosted Trees) (RMSE = 2.3222)

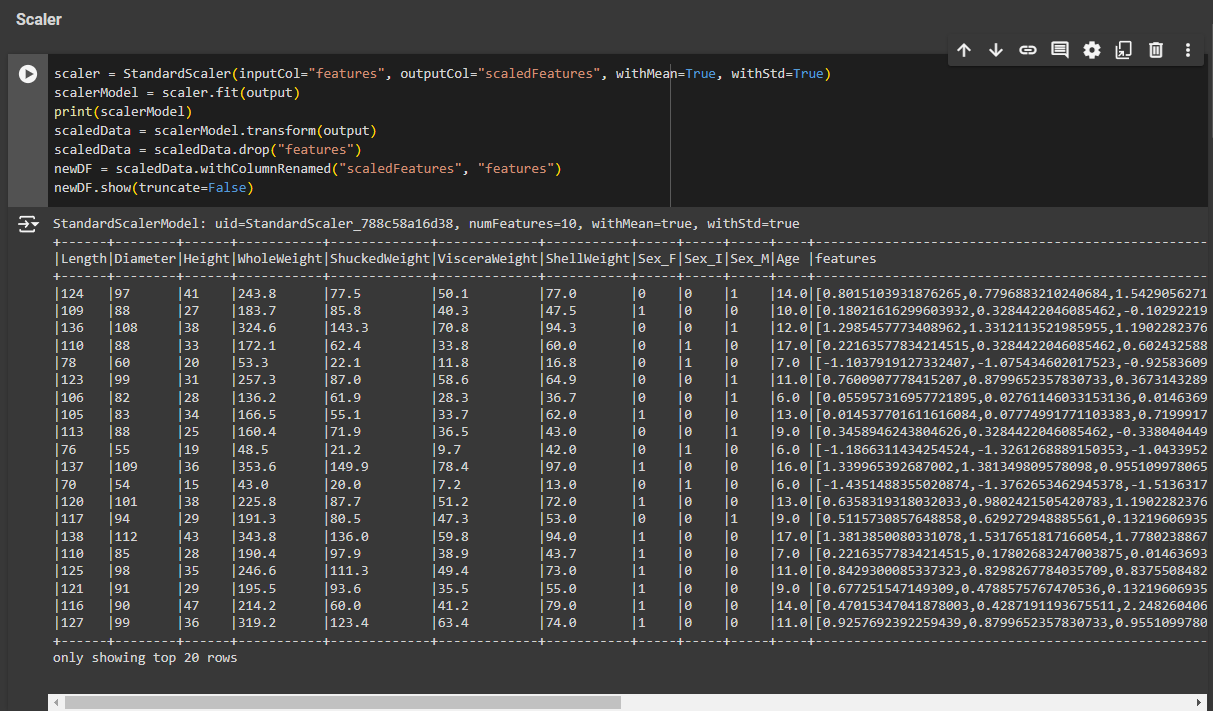
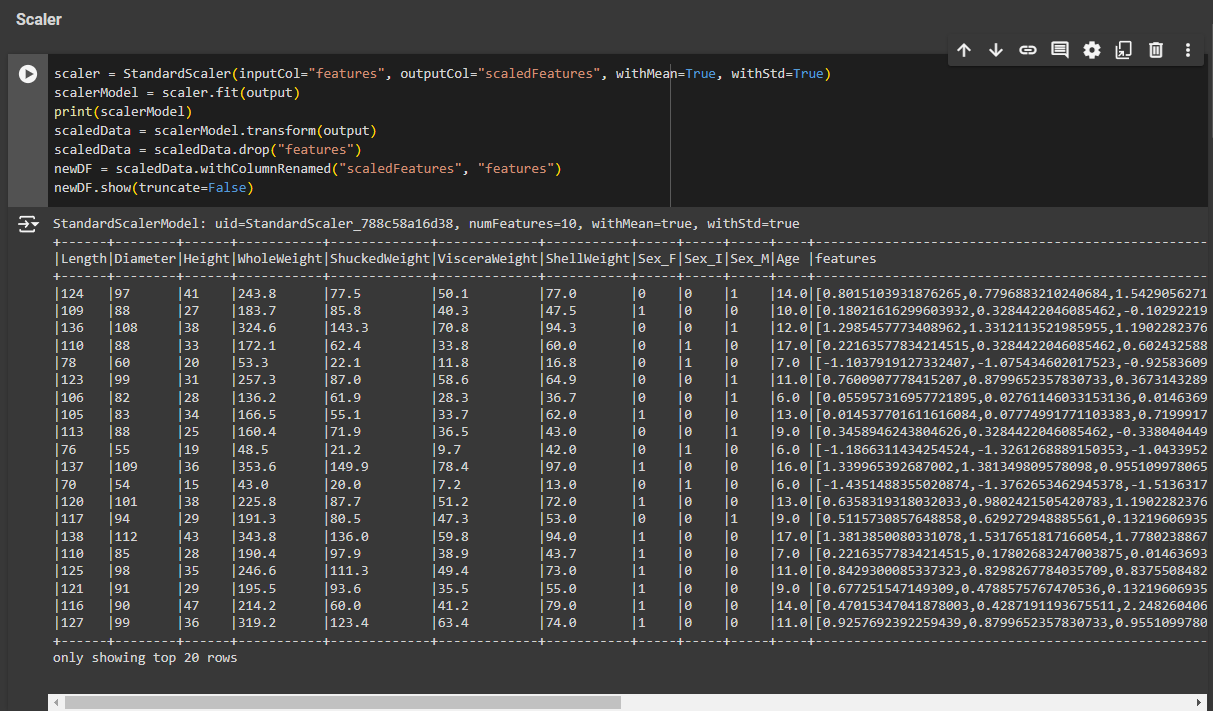
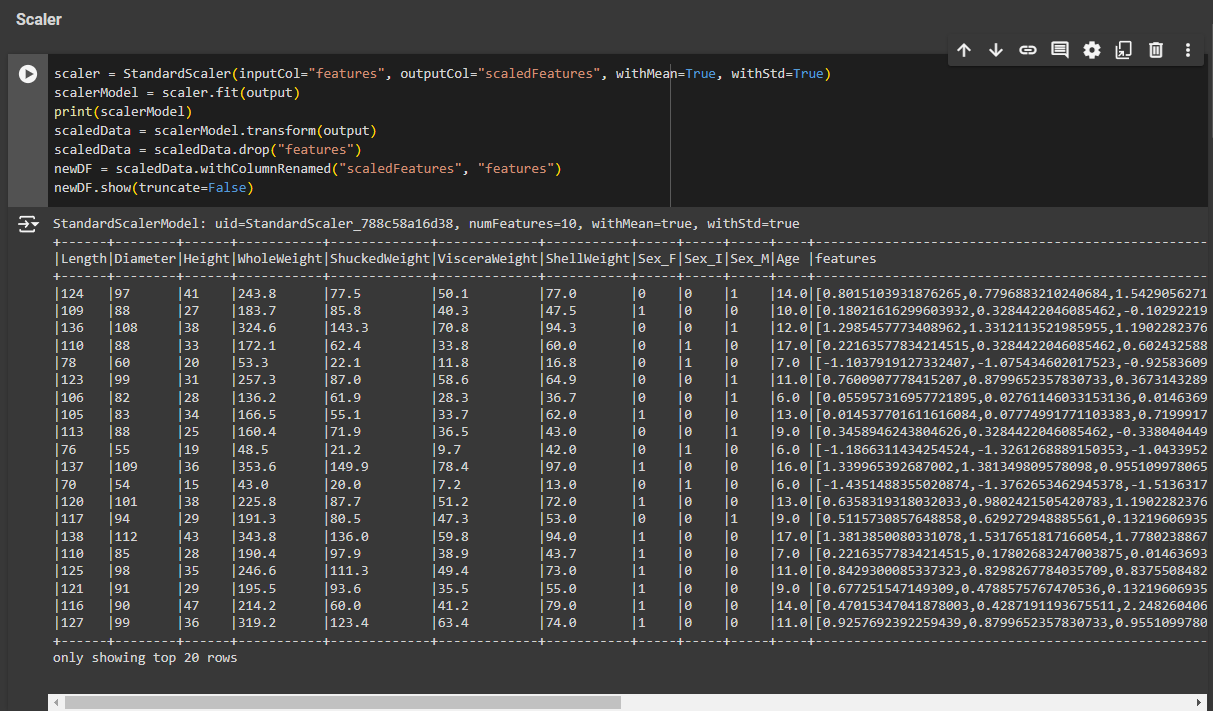
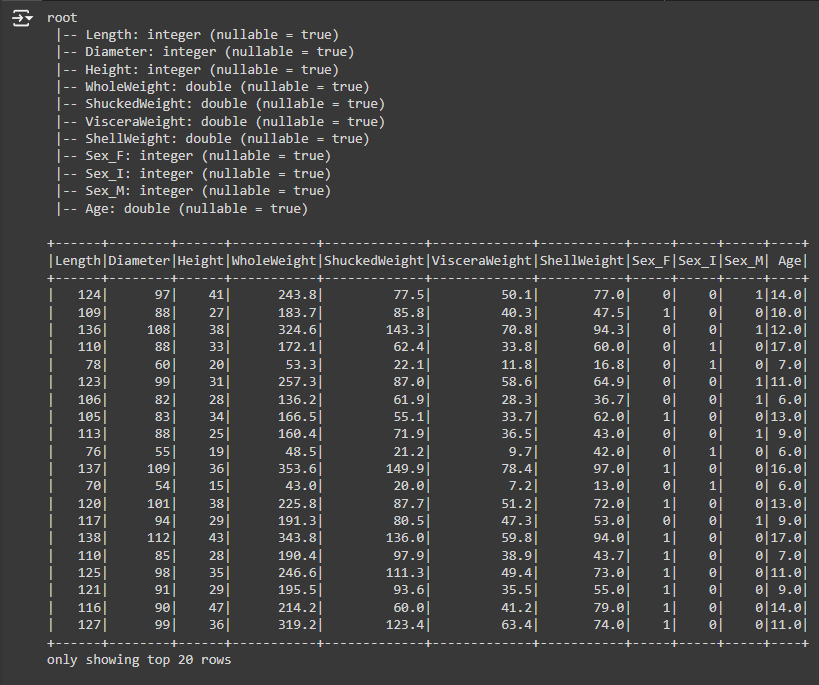
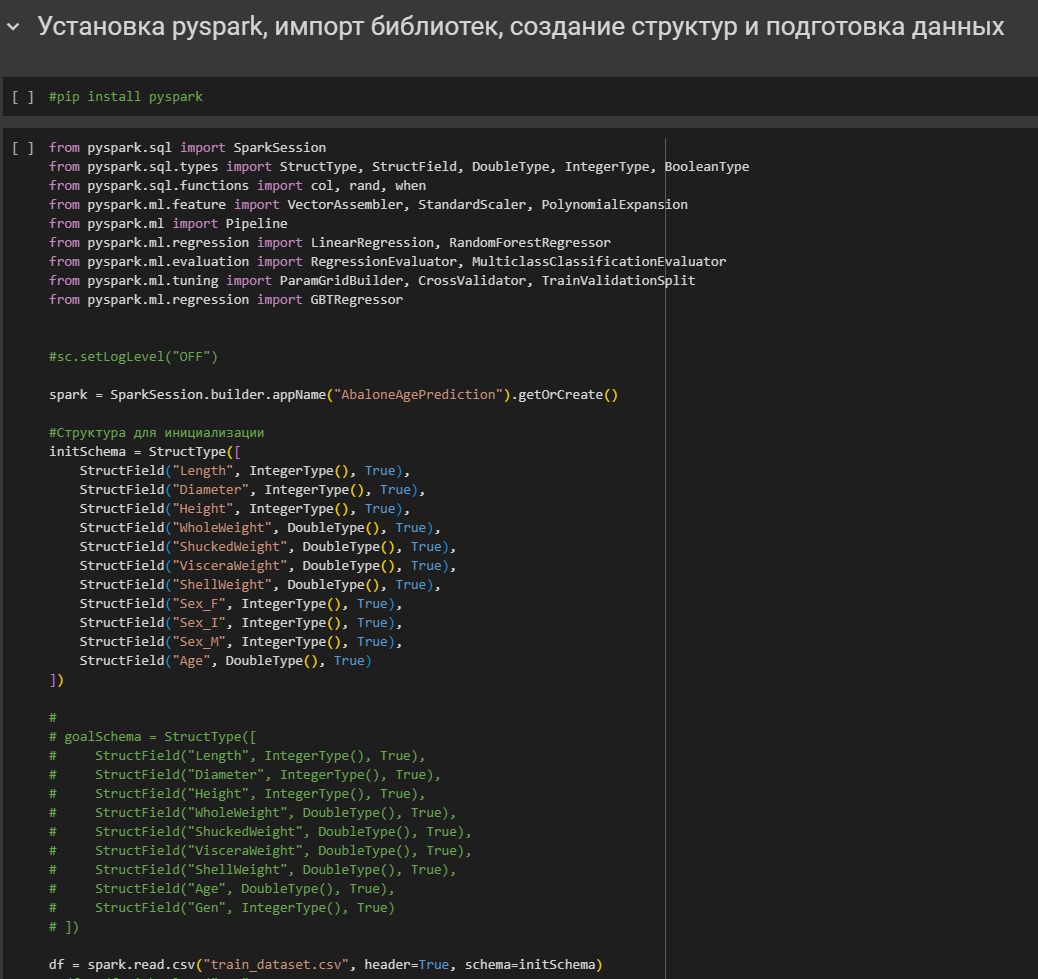
Визуализация истинных и предсказанных значений при помощи Matplotlib на следующем скриншоте:

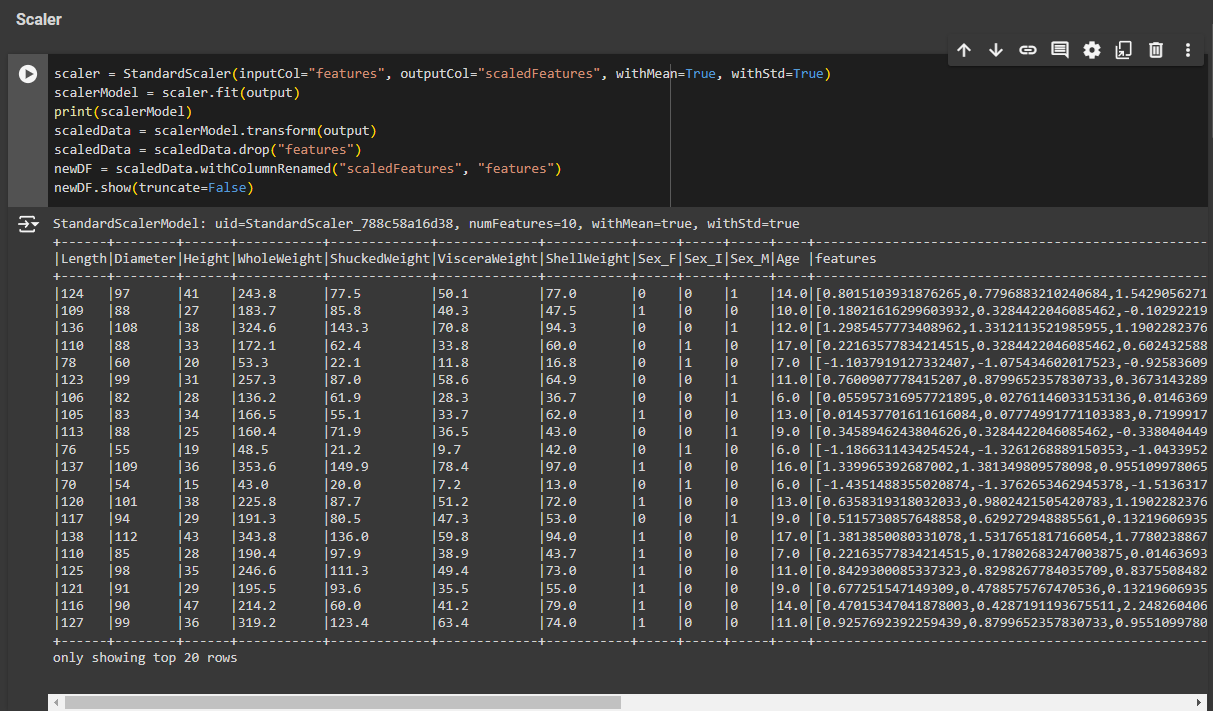


Оранжевым показаны истинные значения, а голубые точки – предсказанное значение возраста.

На этом скриншоте можно видеть визуализацию полученных результатов, где можно видеть хоть и небольшое, но превосходство модели PolynomialRegression.

Фрагмент кода с предобработкой датасета и наиболее точной моделью на следующий скриншотах:





Полный код можно найти в GitHub репозитории: [GitHub репозитории](https://github.com/CherryYu-beep/BidDataProject).

Также был проведён анализ различных решений других пользователей на сайте, где был найден набор данных Kaggle.com. В основном они использовали библиотеки pytorch или scykitlearn, использовались различные модели регрессии, KNN, кластеризация, случайный лес. Наиболее точным из них оказалось “Easy Regression to Determine Abalone Age”: [решение “Easy Regression to Determine Abalone Age”](https://www.kaggle.com/code/greegtitan/easy-regression-to-determine-abalone-age). Рейтинг RMSE = 2.1298 при использовании модели линейной регрессии и случайного леса.

**4. Вывод**

* Была найдена модель, позволяющая оценивать возраст моллюсков Abalone с точностью предсказания RMSE=~2.05, что позволяет использовать модель как замену традиционному методу измерения при отсутствии требования очень точного результата.
* Данная модель оказалась на ~0.06 RMSE точнее, чем аналогичные модели от других пользователей