**Т. В. Ткач**
***Школа № 2009, г. Москва, Россия***

# МАШИННОЕ ОБУЧЕНИЕ И ОБРАБОТКА БОЛЬШИХ ДАННЫХ В УСЛОВИЯХ СОВРЕМЕННОЙ ШКОЛЫ

***Аннотация***

Современный мир наполнен множеством разнообразных данных, и их объем с каждым годом многократно увеличивается. С одной стороны, Big Data — это данные, которые невозможно обработать на одном компьютере, их сложно структурировать. С другой стороны, это специальные инструменты, подходы и методы обработки данных, которые применяются для того, чтобы человек мог получить конкретные и нужные ему результаты для их дальнейшего эффективного применения. Откуда берутся большие данные? Из интернета — данные соцсетей, блогов, СМИ, форумов, сайтов, интернета вещей. Из архивов и баз данных — корпоративные данные, транзакционная деловая информация. Из разного рода приборов и устройств — показания датчиков, измерителей, регистраторов, например, данные метеорологической службы и операторов сотовой связи. Миллиарды устройств сегодня подключены к интернету. Ожидается, что за несколько лет к глобальной сети будут подключены миллионы новых устройств, которые ежедневно будут генерировать несколько десятков миллионов терабайт данных. Такие цифры должны заинтересовать тех молодых людей, кто ищет свой путь в области информационных технологий. Очевидно, что в ближайшем будущем специалисты, которые умеют работать с большими данными, будут востребованы на рынке труда. «Машинное обучение и большие данные» («Machine learning and Big Data») — это одна из компетенций WorldSkills Junior (Future Skills). В статье представлен опыт подготовки к соревнований такого рода без выхода за рамки школьного предмета «Информатика».

***Ключевые слова:*** машинное обучение, большие данные, реляционная структура данных, визуализация табличных данных, математические модели обработки данных.

***DOI:*** 10.32517/2221-1993-2020-19-7-25-29

Все больше современных организаций осознают тренд использования технологий машинного обучения и больших данных. Так как область эта сравнительно молодая и в вузах еще только-только начинают появляться соответствующие специализации, все это приводит к большой востребованности специалистов на рынке труда. А это, несомненно, делает профессию специалиста по работе с большими данными еще привлекательней. Однако в современной научно-методической литературе внимание больше уделяется этическим и правовым аспектам использования Big Data [11], в то время как педагогические аспекты рассмотрены скупо.

Для ознакомления с профессиями в школах проводятся занятия по программам предпрофессиональной ориентации. Получение первичных профессиональных навыков у школьников обеспечивается богатой палитрой курсов внеурочной деятельности [7] или авторскими компонентами программ по отдельным дисциплинам [9]. Кроме того, для диагностирования сформированности первичных профессиональных навыков проводятся конкурсы профессионального мастерства, в частности, такие как WorldSkills Junior, которые обеспечивают повышение интереса школьников к получению профессии. Одной из компетенций WorldSkills Junior является «Машинное обучение и большие данные» («Machine learning and Big Data»). В течение нескольких лет учащиеся школы № 2009 г. Москвы принимали активное участие в чемпионатах Москвы WorldSkills Russia Junior в данной компетенции, трижды становились призерами, а на чемпионате DigitalSkills 2018 в Казани завоевали полный комплект наград.

***Контактная информация***

**Ткач Татьяна Васильевна**, канд. пед. наук, учитель информатики, школа № 2009, г. Москва, Россия; *адрес:* 117041, Россия, г. Москва, ул. Адмирала Руднева, д. 16, корп. 1; *e-mail:* tkach.tv@sch2009.net

**T. V. Tkach**
School 2009, Moscow, Russia

## MACHINE LEARNING AND BIG DATA PROCESSING IN A MODERN SCHOOL

***Abstract***

The modern world is filled with a variety of data, and their volume is increasing many times every year. On the one hand, Big Data is data that cannot be processed on one computer, it is difficult to structure it. On the other hand, these are special tools, approaches and data processing methods that are used so that a person can get specific and necessary results for their further effective use. Where does big data come from? From the Internet — data from social networks, blogs, media, forums, sites, the Internet of things. From archives and databases — corporate data, transactional business information. From all sorts of instruments and devices — readings from sensors, meters, recorders, for example, data from the meteorological service and cellular operators. Billions of devices are connected to the Internet today. It is expected that within a few years millions of new devices will be connected to the global network, which will generate several tens of millions of terabytes of data every day. Such figures should interest those young people who are looking for their own way in the field of information technology. It is obvious that in the near future, specialists who know how to work with Big Data will be in demand in the labor market. "Machine learning and Big Data" is one of the competencies of WorldSkills Junior (Future Skills). The article presents the experience of preparing for competitions of this kind without going beyond the school subject "Informatics".

***Keywords:*** machine learning, Big Data, relational data structure, tabular data visualization, mathematical data processing models.

Применение технологий машинного обучения и больших данных подразумевает все направления работы с огромным объемом самой разрозненной информации, постоянно обновляемой и разбросанной по разным источникам.

В сфере обработки больших данных и обучения в соответствующей области предполагается решение следующих профессиональных задач:

- проведение анализа больших объемов данных;
- выявление закономерностей в больших объемах данных;
- построение гипотез, теорий, моделей на основе полученных данных и дополнительных вводных;
- создание алгоритмов Big Data;
- формирование бизнес-требований по построению аналитических моделей для рынка;
- обработка больших объемов данных с применением аналитики;
- интеллектуальный анализ данных и процесса машинного обучения.

В последнее время к технологиям машинного обучения и больших данных все чаще прибегают в здравоохранении, банковском секторе, государственном управлении, сельском хозяйстве. Использование знаний и технологий в области машинного обучения и больших данных дает возможность посмотреть на имеющиеся данные под разными, порою очень неожиданными, углами зрения.

Таким образом, можно заключить, что рассматриваемая тема довольно популярна в мире и, в частности, среди молодых людей, желающих стать специалистами данного профиля.

Среди школьных дисциплин именно в рамках курса информатики возможно ознакомление школьников с основами машинного обучения. Однако современные курсы информатики не обеспечивают должного содержания и нуждаются в существенном обновлении [2], в том числе во внедрении вариативных модулей, связанных с предпрофильной ориентацией [10].

Опыт преподавания в профильных информационно-технологических классах ГБОУ «Школа № 2009» показывает, что желание проводить структуризацию и обработку Big Data приходит уже к восьмиклассникам, только-только изучившим основы процедурного программирования. Понятно, что из-за недостатка знаний математического аппарата охватить целиком описанные выше профессиональные задачи невозможно, даже на первоначальном понятийном уровне. Поэтому мотивированным школьникам мы предлагаем ряд упражнений, которые вырабатывают первоначальные навыки, стимулируют любознательность в области вычислительных математических моделей, прививают программистскую культуру в обработке данных.

**На первом этапе введения в тему «Машинное обучение и обработка больших данных» мы знакомим школьников с реляционной структурой данных в MS Access и с форматом CSV.** Файлы CSV являются популярным форматом данных и очень часто используются для обмена информацией между различными приложениями. Рассматриваем особенности настроек при импорте файлов формата CSV, а также возможность автоматизации процесса импорта [3]. Основываясь на нашем опыте, считаем, что понять, как работает реляционная база данных, лучше всего именно в Access. Среда с понятным графическим интерфейсом позволяет создавать схему данных, наглядно отражающую связи между таблицами, дает возможность придумывать и реализовывать разные запросы и отчеты, хорошо готовит школьника к возможному дальнейшему восприятию языка SQL и подобных ему.

**На следующем этапе очень важно показать, как практически те же возможности обработки таблиц мы можем реализовать на языке Python** (в нашей школе изучаем Python с V класса в классах инженерного предпрофиля). Как раз очень полезный материал для VIII—IX классов — подключение внешних библиотек Python. Для работы используем модули csv и pandas, отрабатываем практические навыки решения следующих задач [12]:

- сделать выгрузку csv-таблицы в двумерную структуру данных [5];
- посчитать количество нулевых значений полей;
- вывести записи с определенными номерами, значения полей по признакам;
- удалить записи, поля по признакам;
- разбить поле на два поля;
- создать выборку по запросу [6];
- записать данные из двумерной структуры в формат .csv [4, 5].

**Далее можно заняться визуализацией табличных данных** [12]. Научиться составлять графические представления зависимостей данных совсем несложно. Рекомендуем учащимся библиотеку matplotlib. Она обладает громадным арсеналом инструментов построения различных графиков, диаграмм и к тому же является весьма простой в освоении. На рисунках 1–4 представлено несколько примеров визуализации, выполненных девятиклассниками.

Данный род деятельности очень хорошо сочетается с изучением в IX классе MS Excel. Учащиеся мотивированно, с пониманием дела строят визуализации таблиц с близким им содержанием, например, с какой-нибудь школьной статистикой или спортивной олимпиадной статистикой, делают это в Excel, сохраняют таблицы в формате .csv и учатся строить в Python визуализацию, развивая программистские навыки и способности к аналитике.

**На следующем этапе начинаем погружаться в математику,** знакомимся с понятиями: абсолютная и относительная погрешности, погрешности арифметических действий, численное решение уравнений, аппроксимация экспериментальных данных. Обязательно разбираем такие методы вычислительной математики, как метод дихотомии (половинного деления) для отделения корней уравнения и метод наименьших квадратов для оптимальной аппроксимации [1]. Применение библиотек matplotlib и numpy позволяет хорошо проиллюстрировать все процессы.

Наконец, что еще можно рассматривать уже в IX классе — это **введение в тему «Корреляционный и регрессионный анализ»**. Мы объясняем школьникам, что если составлять таблицы по нашим наблюдениям, то можно обнаружить связи между отдельными признаками, т. е. изменение одних признаков приводит к изменению других. Поэтому выявление таких связей позволяет

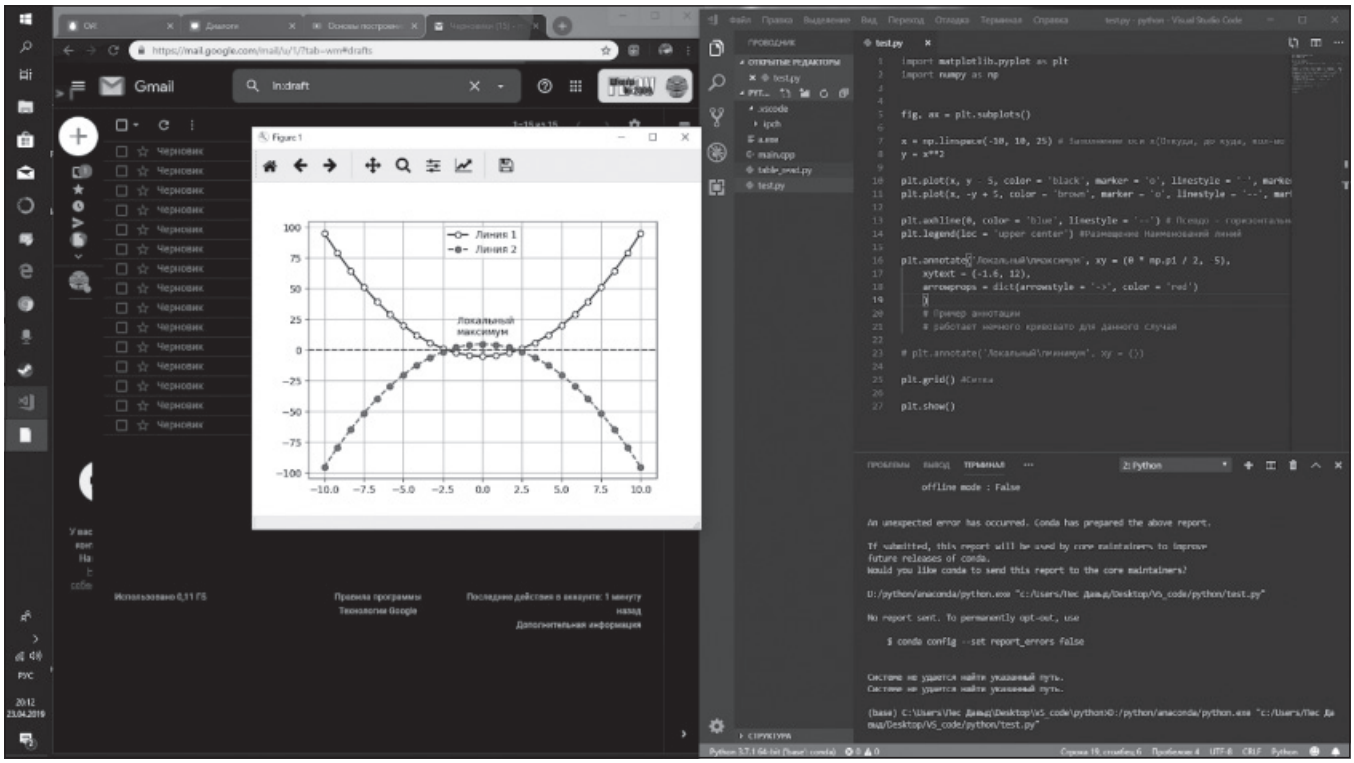


*Рис. 1. Двумерный график*

предвидеть, прогнозировать развитие ситуации при изменении конкретных характеристик исследуемого процесса или явления [8]. Например, все любят смотреть фильмы, и можно собрать из каталогов фильмов реляционную базу данных, содержащую информацию о каждом фильме из какого-либо ограниченного их числа, использовав такие параметры, как жанр, год выпуска, киностудия, директор, режиссер, продюсер, кассовый сбор, рейтинг и т. д. Затем попытка на выборке найти зависимости приводит к хорошим графическим результатам, отражающим количество фильмов в зависимости от страны, бюджет фильма — в зависимости от жанра, кассовый сбор — в зависимости от жанра, рейтинг фильма — в зависимости от жанра. Это позволяет ребятам сформулировать некоторые выводы, имеющие вероятностный характер. В этот момент можно продвинуться

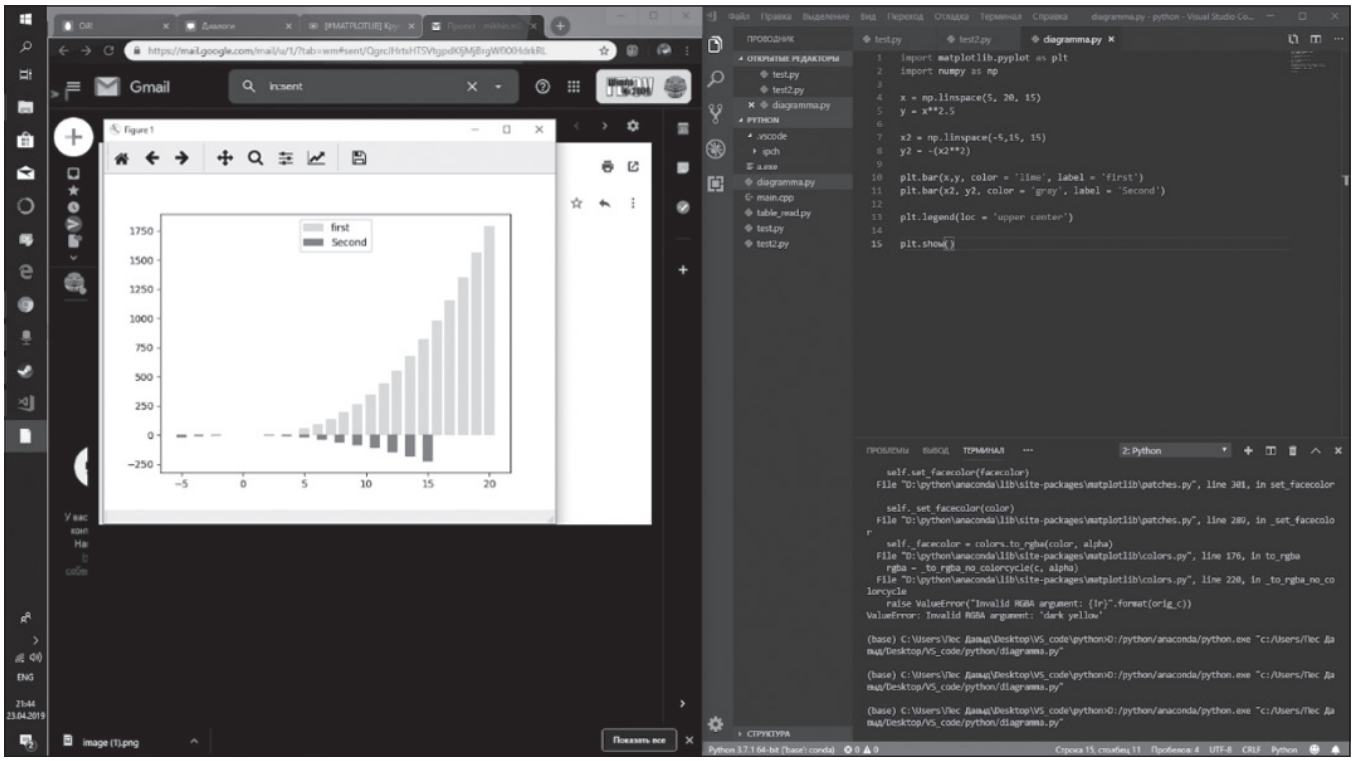


*Рис. 2. Гистограмма*

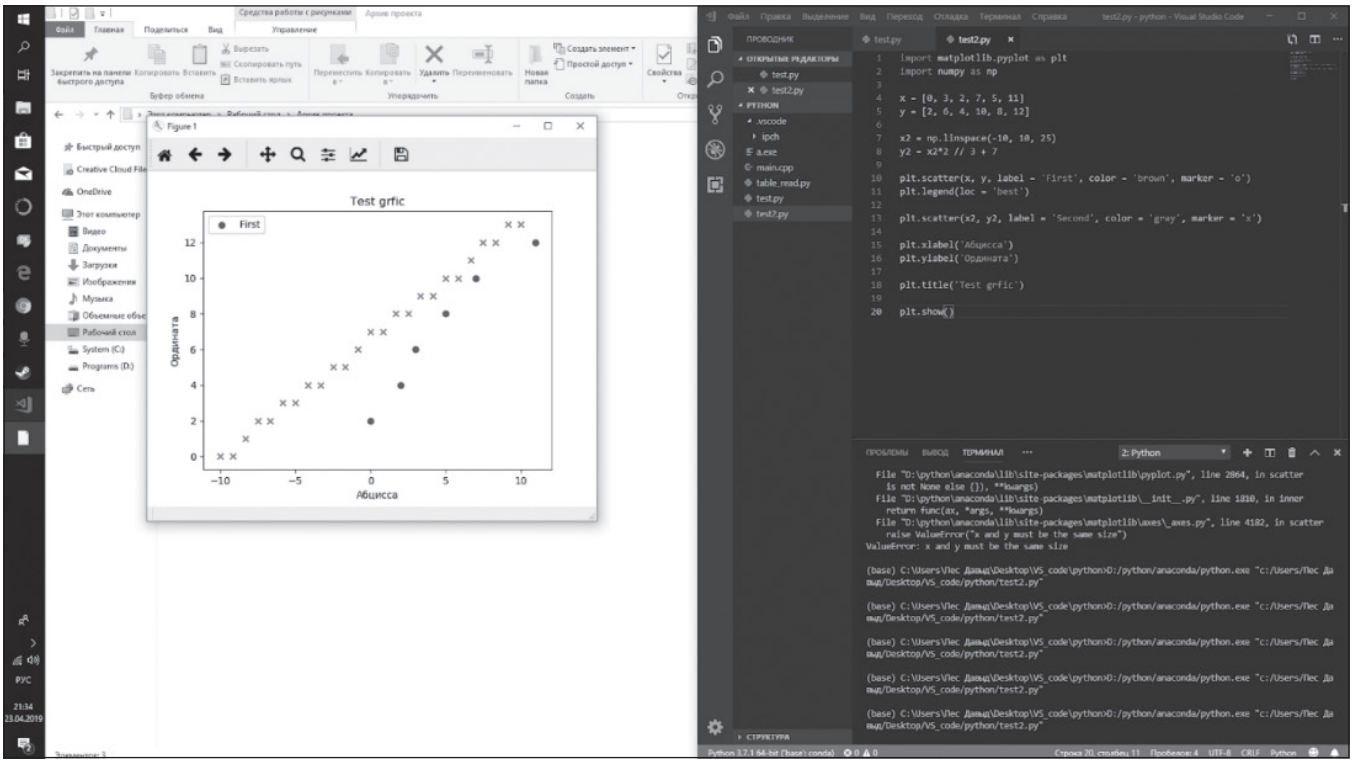


*Рис. 3. Точечная диаграмма*

дальше и объяснить, что в большом датасете зависимости между признаками выявляет специальный функционал Python, который называется метод корреляции. Он приводит к получению таблицы — матрицы корреляции (рис. 5), по значениям которой можно сделать выводы о взаимосвязях атрибутов данных.

Далее мы рассматриваем регрессию как инструмент аппроксимации, она устанавливает форму зависимости в виде непрерывной функции, причем регрессионным анализом мы получаем линию (непрерывный график), наилучшим способом приближающуюся к дискретным наблюдаемым значениям. В качестве примера обычно рассматриваем метод наименьших квадратов.

Если математический уровень школьников позволяет, можно рассказать им, забежав вперед по программе математики, о том, что такое производная и частная про-

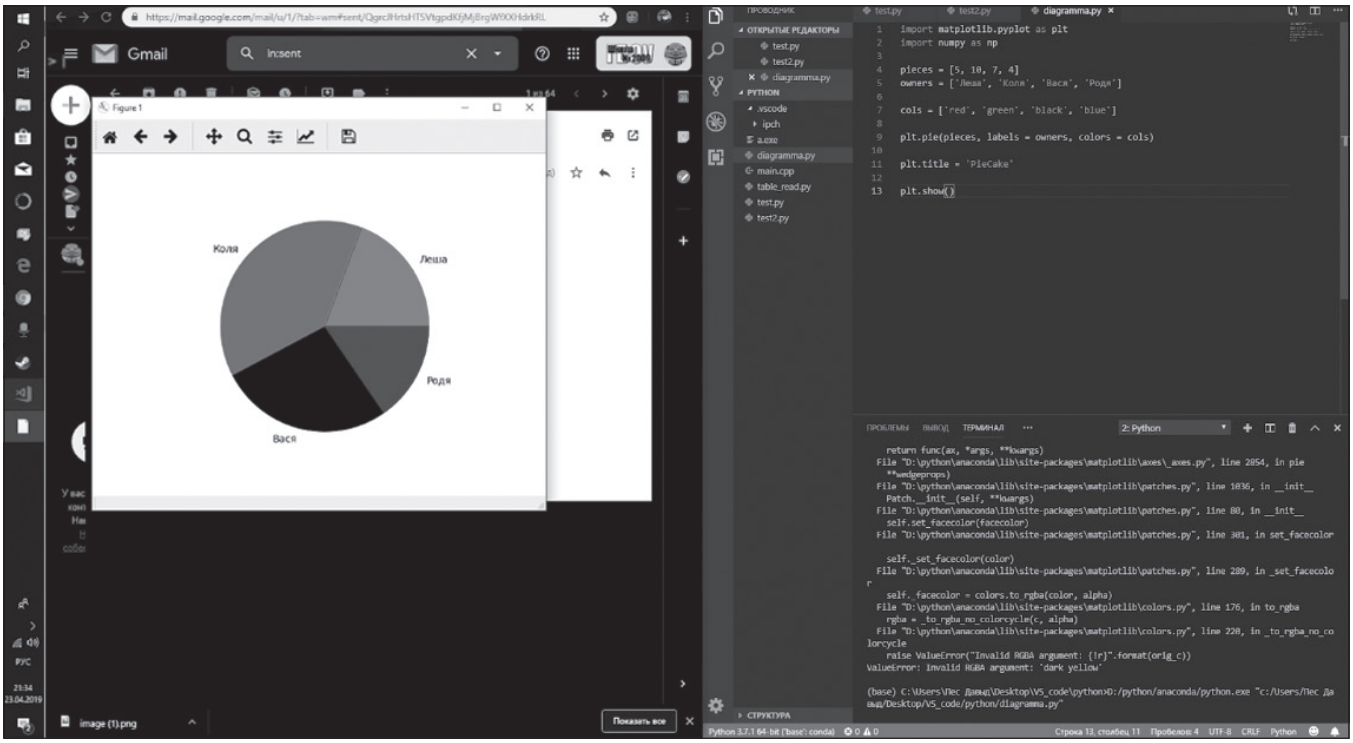


*Рис. 4. Круговая диаграмма*

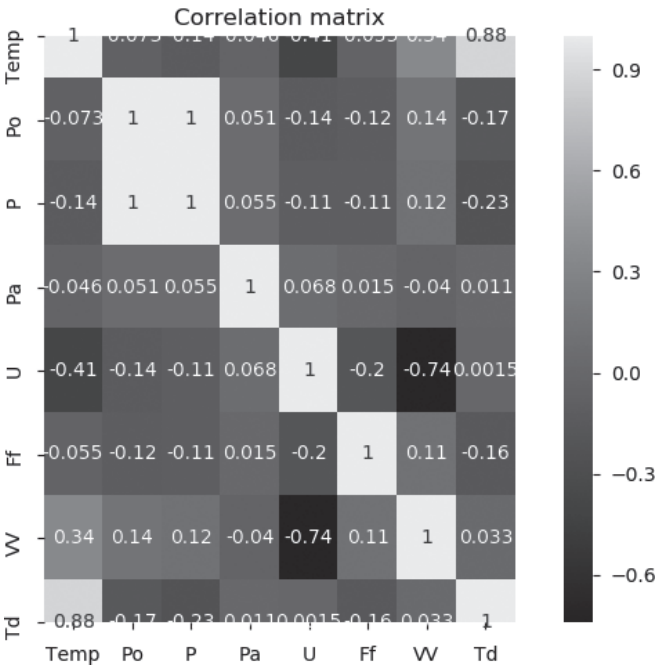


*Рис. 5. Пример матрицы корреляции*

изводная функции. Для наглядного объяснения обычно используем геометрическую интерпретацию (рис. 6) и говорим о скорости изменения функции.

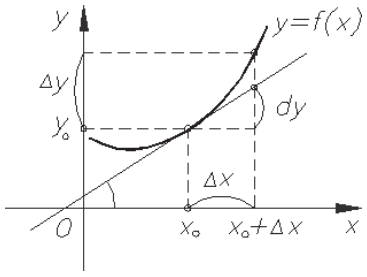


*Рис. 6. Геометрическая интерпретация производной*

Затем можно объяснить метод градиентного спуска на простом примере поиска минимума квадратичной функции.

Мы не можем утверждать, что ученики VIII—IX классов хорошо разбираются с математической составляющей заданий, но они знакомятся со специфической терминологией и привыкают к ней, изучают существующие математические модели обработки данных и привыкают к ним, причем неплохо справляются с выбором того, который предпочтителен в той или иной ситуации.

Уверены, что понимание математических методов придет постепенно, вместе с желанием довести решение проблемы до конца, проанализировать полученные результаты и добиться улучшения их качества. Главная наша задача — увлечь ребят самой возможностью сбора и анализа данных, поиском закономерностей и показать им тот восторг, который приходит в результате новых открытий.

## Список использованных источников

1. *Гельман В. Я.* Решение математических задач средствами Excel. Практикум. СПб.: Питер, 2003. 240 с.

2. *Горячев А. В.* О целесообразности модульной организации курса информатики в основной и старшей школе // Информатизация непрерывного образования — 2018: Материалы Международной научной конференции. В 2 т. / под общ. ред. В. В. Гриншкуна. М.: Российский университет дружбы народов, 2018. С. 450–453.

3. Импорт данных из файлов формата CSV в базу данных MS Access. https://youtu.be/EzGZzfd3wtc

4. Как записать данные в csv-файл модуль csv Python. http://trainingweb.ru/page/write-data-csv-file-csv-module-python

5. Как читать и писать csv-файлы в Python. https://code.tutsplus.com/ru/tutorials/how-to-read-and-write-csv-files-in-python--cms-29907

6. *Каштамов А.* Введение в pandas: анализ данных на Python. https://khashtamov.com/ru/pandas-introduction/

7. *Лалетина О. В., Липатов Д. В.* Роль дополнительного образования в профессиональной ориентации школьников // Отечественная и зарубежная педагогика. 2018. № 2 (49). С. 192–199.

8. Машинное обучение для самых маленьких. https://proglib.io/p/the-simplest-introduction-to-machine-learning/

9. *Павлов Д. И.* О возможном направлении изменений в содержании обучения информатике в основной и старшей школе // Актуальные проблемы обучения математике и информатике в школе и вузе: Материалы IV международной научной конференции. В 2 ч. (г. Москва, 4–5 декабря 2018 года). М.: АКФ «Политоп», 2018. С. 171–175.

10. *Павлов Д. И.* Об изменении методов реализации курса информатики в основной школе // Информатизация непрерывного образования — 2018: Материалы Международной научной конференции. В 2 т. / под общ. ред. В. В. Гриншкуна. М.: Российский университет дружбы народов, 2018. С. 486–491.

11. *Федулов И. Н., Квач И.В.* Правовые аспекты управления big data в странах БРИКС и ШОС // Сервис +. 2019. № 3. С. 85–92.

12. Электронная книга Devpractice Team. Pandas. Работа с данными. 2-е изд. 2020. 170 с. https://devpractice.ru