УДК 004.853

ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗА УСПЕВАЕМОСТИ СТУДЕНТОВ

Балахнин Е.Е., студент гр. ПИб-192, II курс
Жмуровский К.В., студент гр. ПИб-192, II курс
Тютиков А.Н., студент гр. ПИб-192, II курс
Научный руководитель: Киреева К.А., ассистент
Кузбасский государственный технический университет
имени Т.Ф. Горбачева, г. Кемерово

Машинное обучение – класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для построения каких-либо методов машинного обучения используются такие средства как: математическая статистика, численные методы, математический анализ, метод оптимизации, теория вероятностей, теория графов, а также различные техники работы с данными в цифровой форме [1].

Методы машинного обучения все еще находятся на стадии развития, но некоторые уже максимально изучены и используются для решения некоторых задач любой сложности и любой направленности. Идея развития методов машинного обучения заключается в том, что для каждой задачи требуется свой подход для ее решения, а также на разных компьютерах используется свой метод решения.

Для решения задач компьютером с применением искусственного интеллекта, требуется практика и постоянная автоматическая поднастройка. Любая модель машинного обучения нуждается в постоянной тренировке с использованием базы данных и, во многих ситуациях, в подсказке самого человека.

Искусственному интеллекту нужно постоянно предоставлять опыт, то есть ему необходимы данные, которые постоянно обновляются. Чем больше данных получает искусственный интеллект, тем точнее компьютер сможет взаимодействовать с ними, а также с данными, которые получает в будущем. Чем выше точность взаимодействия с данными, тем выше успешность выполнения поставленной задачи, а также выше степень прогностической точности.

Машинное обучение работает следующим образом – сбор данных, очистка ненужных значений, выборка значений или разделение на категории, выполнение поставленной задачи или обучение, оценка данных и их корректировка, оптимизация модели.

Подробнее остановимся на этапах работы машинного обучения.

Сбор данных. Машинное обучение основывается на данных. Первый шаг для решения поставленной задачи – убедиться в том, что данные верны и

относятся именно к той задаче, которую необходимо решить. Также следует оценить возможность для сбора дополнительных данных для сравнения и оценки. Нужно обдумать их источник и необходимый формат.

Очистка значений. Большинство данных формируются из разных источников, отображаются в различных форматах и языках. Среди всех поступающих данных, могут оказаться и те, которые не имеют никакого отношения к поставленной задачи, и которые требуют удаления. А может произойти нехватка данных и их нужно добавить. От правильной подготовки базы данных прямым образом зависит их пригодность к использованию, и достоверность результатов.

Разделение. Если размер данных слишком велик, то может потребоваться только часть. Обычно это называют выборкой данных. Часть, которая была выбрана в качестве решения поставленной задачи, делится на две группы. Одна группа - для использования алгоритмом, а вторая группа – для оценки его действий.

Обучение. Данный этап фактически направлен на поиск математической функции, которая выполнит указанную задачу со 100 % результатом. Обучение происходит в зависимости от типа используемой модели.

Оценка. После того, как алгоритм закончил выполнение задачи, его эффективность оценивается на данных, с которыми он не сталкивался. Дополнительная корректировка данных, после выполнения задачи, корректируются при необходимости. Данный процесс позволяет предотвратить переобучение – явление, при котором алгоритм хорошо работает только на учебных данных.

Оптимизация. Модель оптимизируется, чтобы при интеграции в приложение весить как можно меньше и как можно быстрее работать.

Суть исследовательской работы заключается в том, чтобы выяснить, существует ли взаимосвязь между вторичными факторами студента и его успеваемостью. Для этой работы было использовано машинное обучение, которое может найти и предсказать результаты студента. Был использован готовый алгоритм обучения нейронной сети – логическая регрессия. В нее подаются подготовленные данные об успеваемости студентов, по которым определяется успеваемость студента. Для определения успешности данной модели, было проведено несколько экспериментов.

Для предсказания успеваемости обучающего, использовались входные параметры такие как:

1. Schools – школа, в которой обучался студент (двоичный).
2. Sex – пол студента (двоичный: «F» - женский, «M» - мужской).
3. Age – возраст студента (числовой: от 15 до 22 лет).
4. Address – тип домашнего адреса студента (двоичный: «U» - городской или «R» - сельский).
5. Famsize – размер семейства (двоичный: «LE3» - меньше или равно 3 или «GT3» - больше 3).
6. Pstatus – статус совместного проживания родителей (двоичный: «T» - живут вместе или «A» - порознь).

7. Medu – образование матери (числовое: 0 – нет, 1 – начальное образование, 2 – неполное среднее общее образование, 3 – среднее общее образование, 4 – высшее образование).
8. Fedu – образование отца (числовое: 0 – нет, 1 – начальное образование, 2 – неполное среднее общее образование, 3 – среднее общее образование, 4 – высшее образование).
9. Mjob – работа матери (номинальное: «учитель», «услуги», «другое»).
10. Fjob – работа отца (номинальное: «учитель», «услуги», «другое»).
11. Reason – причина выбора школы (номинальное: «близость к дому», «репутация школы» или «другое»).
12. Guardian – опекун ученика (номинальное: «мать», «отец» или «другое»).
13. Traveltime – домой и в школу, время пути (числовое: 1 – 1 час).
14. Studytime – еженедельное учебное время (числовое: 1 – 10 часов).
15. Failures – количество прошлых классных неудач (числовое: n, если 1 <= n < 3, иначе 4).
16. Schoolsup – дополнительная образовательная программа (двоичное: «Да» или «Нет»).
17. Famsup – семейная дополнительная образовательная программа (двоичное: «Да» или «Нет»).
18. Paid – дополнительные платные занятия по предмету курса (двоичное: «Да» или «Нет»).
19. Activities – участвовал ли в внеклассных мероприятиях (двоичное: «Да» или «Нет»).
20. Nursery – посещал детский сад (двоичное: «Да» или «Нет»).
21. Higher – хочет получить высшее образование (двоичное: «Да» или «Нет»).
22. Internet – доступ в интернет дома (двоичный: «Да» или «Нет»).
23. Romantic – с романтическими отношениями (бинарное: «Да» или «Нет).
24. Famrel – качество семейных отношений (числовое: от 1 – очень низко до 5 – очень высоко).
25. Freetime – свободное время после школы (числовое: от 1 – очень мало до 5 – очень много).
26. Goout – встреча с друзьями (числовое: от 1 – очень редко до 5 – очень часто).
27. Dalc – потребление алкоголя в течение рабочего дня (числовое: от 1 – очень редко до 5 – очень часто).
28. Walc – потребление алкоголя в выходные дни (числовое: от 1 – очень редко до 5 – очень часто).
29. Health – текущее состояние здоровья (числовое: от 1 – очень плохое до 5 – отличное).
30. Adsences – количество пропусков в школе (числовое: от 0 до 93).

После заполнения данных, машинное обучение прогнозировало успеваемость студента по 20-ти бальной шкале (американской).

Приложение состоит из трех блоков (рис. 1).

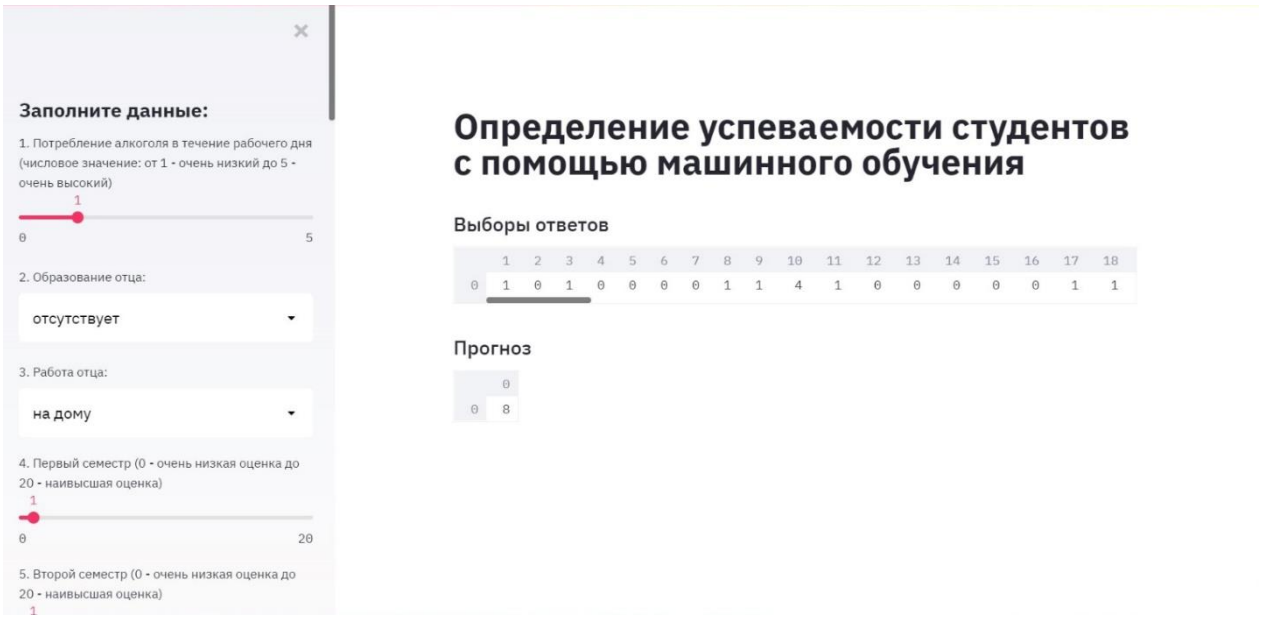

Рисунок 1 – Готовый сайт

Первый блок – это входные параметры, получаемые машиной. В нем студент заполняет данные, которые потребуются для прогнозирования его успеваемости (рис. 2).

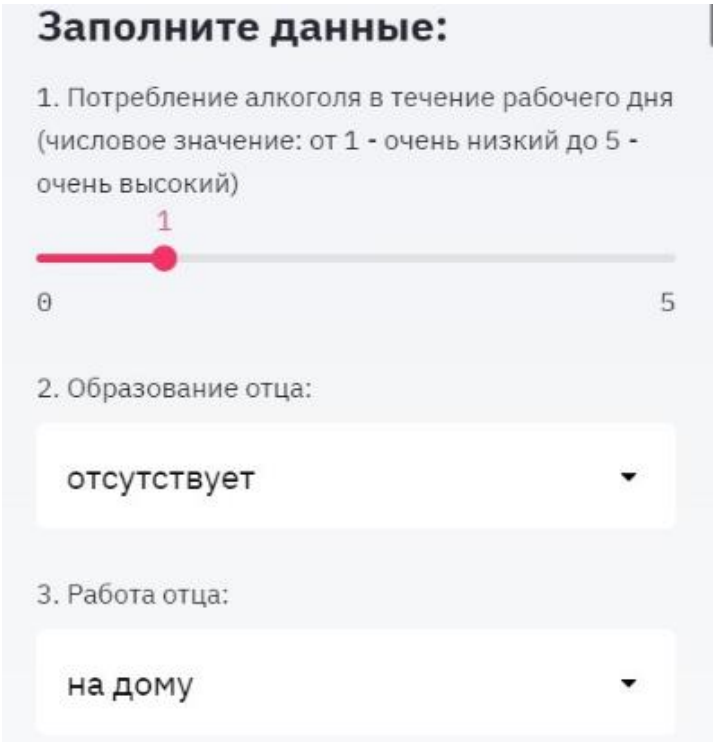

Рисунок 2 – Блок входных параметров

Второй блок – те параметры, которые студент уже ввел. Данные параметры можно корректировать и видеть их в блоке (рис. 3).

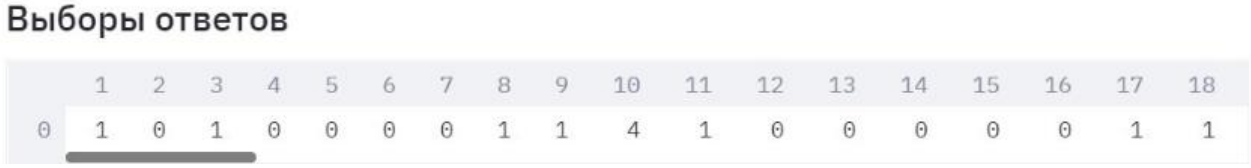

Рисунок 3 – Блок введенных параметров

Третий блок – результат предсказания машины по данным из второго блока. Весь процесс вычисления можно наблюдать, после ввода данных. Конечное число – конечная оценка за семестр от 1 до 20, по американской системе оценивания (рис. 4).

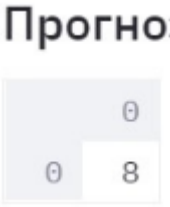

Рисунок 4 – Блок результата прогноза

Приложение разработано на языке программирования Python с использованием PyCharm.

Были импортированы библиотеки, которые потребовались для создания приложения: библиотека «Streamlit» отвечает за создание сайта и его визуальную составляющую, библиотека «Pandas» используется для работы с данными и таблицами, библиотека «Joblib» нужна для подключения нашего машинного обучения, библиотека «pymysql» необходима для подключения базы данных (рис. 5).

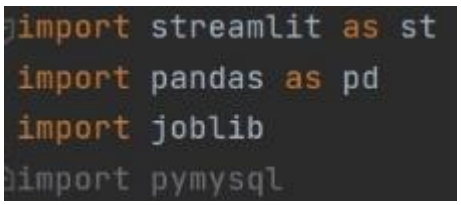

Рисунок 5 – Подключенные библиотеки

Чтобы машина обучения работала правильно, к ней необходимо прописать путь. Для удобной работы студентам необходимо разработать визуальную часть приложения (рис. 6).

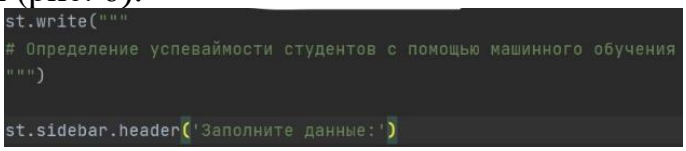

Рисунок 6 – Метод, отвечающий за визуальную часть сайта

Также потребуется метод, который будет отвечать за конвертацию данных, которые ввел студент, в данные, которые будут понятны для машины (рис. 7).

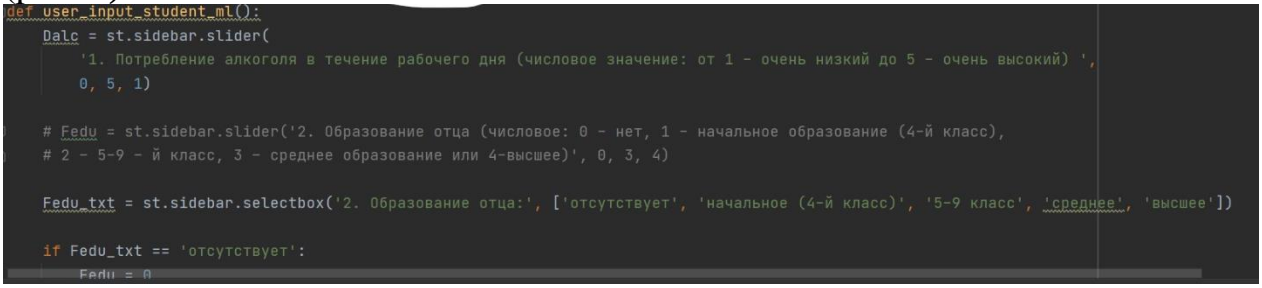

Рисунок 7 – Метод конвертирования данных

Для прогнозирования необходимо использовать метод, в котором будет находиться массив из подготовленных данных (рис. 8).

```
studytime = st.sidebar.slider('27. Еженедельное учебное время (1-10 часов)', 1, 10, 1)

traveltime = st.sidebar.slider('28. От дома в школу время в пути (1 - 4 часа)', 1, 4, 1)

data = {'1': Dalc,
        '2': Fedu,
        '3': Fjob_at_home,
        '4': Fjob_health,
        '5': Fjob_other,
        '6': Fjob_services,
        '7': Fjob_teacher,
        '8': G1,
        '9': G2,
        '10': Medu,
        '11': Mjob_at_home,
        '12': Mjob_health,
        '13': Mjob_other,
        '14': Mjob_services,
        '15': Mjob_teacher,
        '16': Pstatus_A,
        '17': Pstatus_T,
        '18': Walc,
        '19': absences,
        '20': activities_no,
        '21': activities_yes,
        '22': address_R,
```

Рисунок 8 – Метод для массива с данными

Таким образом, было создано приложение, которое способно спрогнозировать успеваемость студента за счет влияния внешних факторов. Были настроены и запущены нейронные сети таким образом, чтобы они смогли найти зависимость между выборкой данных и их результатами. На основе этого можно сделать следующее утверждение. Учеба не происходит в «вакууме» и два умных студента могут иметь совершенно противоположные результаты обучения между собой.

Список литературы

1. Бринк, Х. Машинное обучение/ Х. Бринк, Д. Ричардс, М. Феверолф. – 2018. – 336 с.