**Министерство образования и науки Российской Федерации**

федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский Томский политехнический университет»

Инженерная школа информационных технологий и робототехники

Отделение информационных технологий

Направление Программная инженерия

**ОТЧЁТ О ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ**

г. Томск, ФГАОУ ВО НИ ТПУ

Выполнили обучающиеся гр. 8К71:   
Галлингер Владислав Андреевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (подпись)

Семенюта Антон Вадимович \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_г.

Проверили:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
(должность руководителя от отделения) (ФИО)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (оценка) (подпись)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_г.

Томск 2020

**Оглавление**

[Список терминов 3](#_Toc52730809)

[Введение 6](#_Toc52730810)

[1 Предобработка «сырых» данных 8](#_Toc52730811)

[2 Разведочный анализ (Exploratory Data Analysis) 20](#_Toc52730812)

[3 Формулирование и проверка статистических гипотез 32](#_Toc52730813)

[4 Поиск структуры в данных с помощью алгоритмов машинного обучения без учителя 58](#_Toc52730814)

[5 Построение предсказательной модели 59](#_Toc52730815)

[Вывод 68](#_Toc52730816)

[Список литературы 69](#_Toc52730817)

# Список терминов

**Python** – скриптовый язык программирования.

**Google Collab** – бесплатный облачный сервис на основе Jupyter Notebook.

**Jupyter Lab** – среда для интерактивных повторно воспроизводимых вычислений, основанная на Jupyter Notebook.

**Датасет** (англ. Dataset – набор данных) – термин, обозначающий набор данных.

**Доверительный интервал** – термин, используемый в математической статистике при интервальной оценке статистических параметров. Доверительным называют интервал, который покрывает неизвестный параметр с заданной надежностью.

**P-уровень значимости** – вероятность получить для данной вероятностной модели распределения значений случайной величины такое же или более экстремальное значение статистики (среднего арифметического, медианы и др.), по сравнению с ранее наблюдаемыми, при условии, что нулевая гипотеза верна.

**Достигаемый уровень значимости** – это наименьшая величина уровня значимости, при которой нулевая гипотеза отвергается для данного значения статистики некоторого критерия T.

**Статистическая гипотеза** – это предположение о свойствах случайных величин или событий, которое мы хотим проверить по имеющимся данным.

**Корреляция** – статистическая взаимосвязь двух или более случайных величин.

**Агломеративная кластеризация** – один из методов иерархической кластеризации, алгоритм упорядочивания данных, направленный на создание иерархии вложенных классов. С помощью данного метода новые кластеры создаются путем объединения более мелких кластеров, и таким образом, дерево создается от листьев к стволу.

**Кластер** – объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определенными свойствами.

**Машинное обучение** – класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решения множества сходных задач.

**Искусственная нейронная сеть** – математическая модель, воплощенная с помощью кода, построенная по принципу организации и функционирования сетей нервных клеток живого организма.

**Логистическая регрессия** – статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путем его сравнения с логистической кривой.

**Линейная регрессия** – статистическая модель, используемая для определения зависимости одной переменной y от другой или нескольких других переменных x с линейной функцией зависимости.

**Функция активации** – функция, применяемая в искусственных нейронных сетях, определяющая выходной сигнал, который определяется входным сигналом или набором входным сигналов.

**Случайный лес** – алгоритм машинного обучения, заключающийся в использовании комитета (ансамбля) решающих деревьев. Алгоритм применяется для задач классификации, регрессии и кластеризации.

**Метод опорных векторов** – алгоритм обучения с учителем, использующийся для задач классификации и регрессионного анализа.

**Метод k-ближайших соседей** – метрический алгоритм для автоматической классификации объектов или регрессии. В случае метода для классификации объект присваивается тому классу, который является наиболее распространенным среди k соседей данного элемента, классы которых уже известны.

# Введение

«Современные проблемы требуют современных решений!», – такой девиз звучит в настоящее время. Технологии стремительно развиваются, на замену человеческому труду приходит автоматика, высвобождая время от рутинных задач. Благодаря этому человек может заниматься более высокоинтеллектуальным трудом, выстраивая тактические или же стратегические планы по развитию своего дела. Собирая данные и проводя анализ, становится ясно, в каком направлении идти дальше. Но иногда данных бывает настолько много, а задачи бывают настолько сложными и многопараметрическими, что в ручную анализ провести стоит очень много времени и сил. В таких случаях на помощь человеку приходит молодая наука, называемая Data Science, или наука о данных. Она содержит множество различных методик и инструментов по работе с данными различного рода. Как раз об этом и пойдет речь в этой работе.

Одним из важнейших направлений стратегии развития ТПУ является повышение качества образования. Качество образования оказывает большое влияние на компетентность будущих специалистов в своей области, спрос на программы обучения ВУЗа и, соответсвенно, на рейтинги ВУЗа. Для повышения качества образования применяются различные подходы: анкетирование, опросы, осмотры, проверки, экспертизы и интеллектуальный анализ накопленных данных и сведений. Последний метод становится все более популярным, так как позволяет более углубленно изучить имеющиеся данные и получить порой неочевидные выводы. Благодаря этому удается решать поставленные задачи и получать удовлетворительные результаты.

Задачами данной производственной практики являются:

1. Проведение разведочного анализа на поиск зависимостей в данных;
2. Выявление признаков студентов ТПУ, наибольшим образом оказывающие влияние на обучаемость студента (т.е. необходимо выявить такие признаки, степень воздействия которых на результативность обучения наивысшая);
3. Построение предиктивной модели, которая наилучшим образом будет определять вероятность того, будет ли студент учиться хорошо или плохо.

Помимо указанных целей, не менее важным является предобработка данных. Качество входных данных напрямую отражается на качестве исследований.

Также, были применены алгоритмы машинного обучения для автоматического поиска зависимостей в данных. Результаты такого поиска можно найти ниже.

Для предобработки и анализа данных мы будем использовать скриптовый язык программирования Python и среды разработки Google Collab и Jupyter Lab.

Python прост в изучении для его использования в качестве инструмента для анализа данных, так как имеет в наличие большое число библиотек, такими, как NumPy, Pandas, Matplotlib, Scipy, Sklearn, Keras.

Google Collab позволяет удаленно взаимодействовать с файлом, хранящемся в облаке. Также он всегда содержит актуальные версии всех используемых модулей.

Jupyter Lab используется в качестве среды разработки без онлайн-соединения. В отличие от Google Collab, эта среда использует вычислительные ресурсы локальной машины.

# 1 Предобработка «сырых» данных

На вход мы получили датасет, состоящий из 8551 студента и 26 колонок-характеристик обучающихся (рисунок 1).

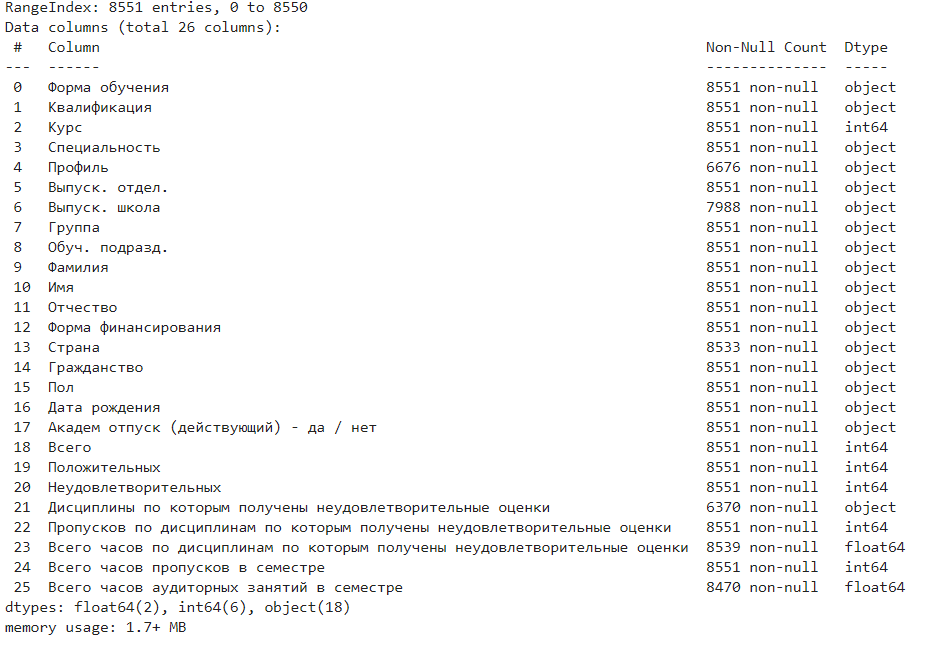


Рисунок 1 — Краткая информационная сводка о взятом датасете

Имеем следующие характеристики:

1. «Форма обучения» — Форма обучения студента (Очная, заочная, очно-заочная);
2. «Квалификация» — Степень, которая будет присвоена студенту (бакалавр, магистр, специалист);
3. «Курс» — Курс, на котором студент обучается (1-5);
4. «Специальность» — Номер и название специальности студента;
5. «Профиль» — Название профиля студента;
6. «Выпуск. отдел.» — Выпускающее студента отделение;
7. «Выпуск. школа» — Выпускающая студента школа;
8. «Группа» — Группа, в которой обучается студент;
9. «Обуч. подразд.» — Обучающее подразделение студента;
10. «Фамилия» — Фамилия студента;
11. «Имя» — Имя студента;
12. «Отчество» — Отчество студента;
13. «Форма финансирования» — Форма финансирования студента (на основе бюджетного финансирования, на договорной основе, по целевому приёму);
14. «Страна» — Страна постоянной прописки студента;
15. «Гражданство» — Гражданство студента;
16. «Пол» — Пол студента (мужской, женский);
17. «Дата рождения» — Дата рождения студента (формат DD/MM/YYYY);
18. «Академ отпуск (действующий) - да / нет» — Находится ли студент в академическом отпуске (Да, Нет);
19. «Всего» — Число дисциплин студента в прошлом семестре;
20. «Положительных» — Число сданных дисциплин у студента в прошлом семестре;
21. «Неудовлетворительных» — Число несданных дисциплин у студента в прошлом семестре;
22. «Дисциплины по которым получены неудовлетворительные оценки» — Перечень несданных дисциплин;
23. «Пропусков по дисциплинам по которым получены неудовлетворительные оценки» — Количество пропусков студента по несданным дисциплинам;
24. «Всего часов по дисциплинам по которым получены неудовлетворительные оценки» — Количество часов по несданным дисциплина в учебном плане студента;
25. «Всего часов пропусков в семестре» — Количество часов пропусков студента в прошлом семестре;
26. «Всего часов аудиторных занятий в семестре» — Количество часов аудиторных занятий студента в прошлом семестре.

Данные взяты на конец сессии весеннего семестра 2019 года.

На рисунке 2 представлены основные статистики числовых переменных датасета.



Рисунок 2 — Основные статистики числовых переменных датасета

На рисунках 3-4 представлены первые 5 студентов набора данных. По ним можно составить представление, как датасет выглядит в целом. Представленные ниже данные – результат функции PandasDataframe.head().

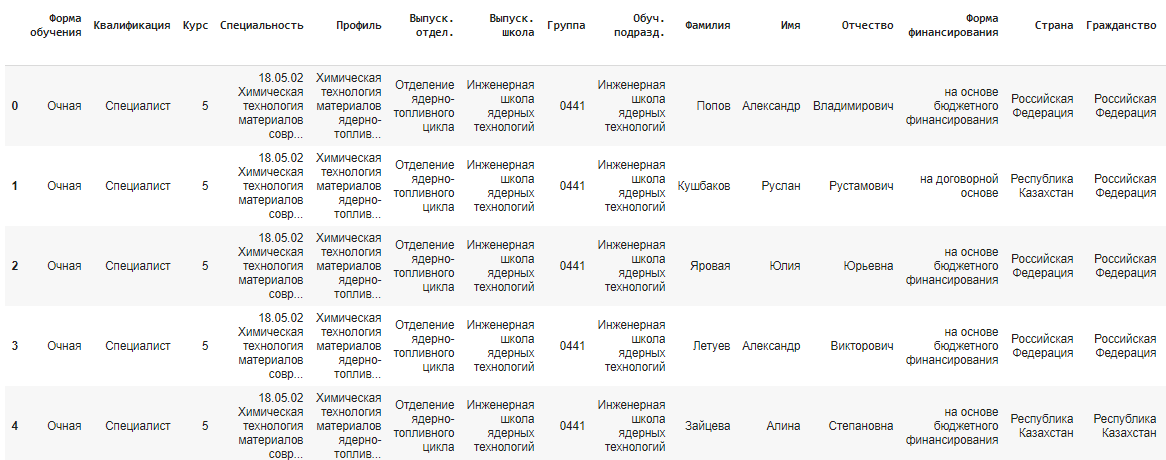


Рисунок 3 — Первые пять студентов набора данных (часть 1)

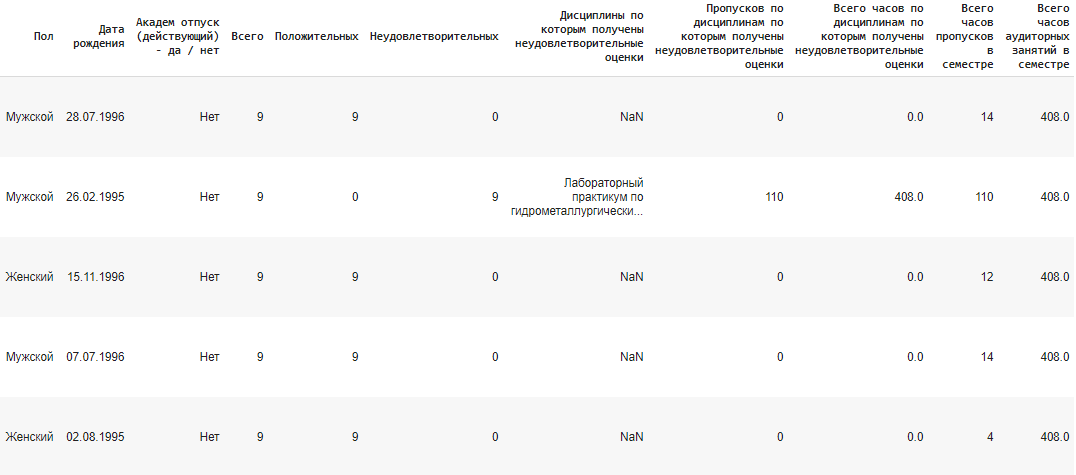


Рисунок 4 — Первые пять студентов набора данных (часть 2)

Проведем обработку персональных данных. Прежде всего объединим столбцы «Фамилия», «Имя» и «Отчество» и сформируем новый столбец «Полное имя». Также сформируем столбец «Индекс студента» путём их последовательной нумерации от 0 до 8550.

Вынесем новые столбцы в отдельный датафрейм students\_names. Также удалим из исходного датасета столбцы «Фамилия», «Имя», «Отчество». Результат представлен на рисунке 5.



Рисунок 5 — Новый датафрейм с персональными данными

Посмотрим на распределение студентов по количеству неудовлетворительных оценок (рисунок 6). Видим, что данное распределение ненормально. Студентов, не имеющих неудовлетворительных оценок, большинство. С возрастанием количества несданных дисциплин количество студентов убывает экспоненциально.

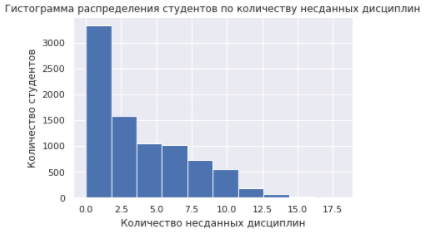


Рисунок 6 — Новый датафрейм с персональными данными

Поработаем со столбцом «Дисциплины по которым получены неудовлетворительные оценки». При изучении датасета стало понятно, что некоторые из несданных дисциплин — факультативные предметы, которые отображаются у каждого члена группы, если хотя бы один из студентов на них записался (рисунок 7). При этом у оставшихся студентов такие предметы идут в данный столбец. Было решено, что если удалить все факультативные дисциплины из датасета, данные станут нагляднее и проще для анализа.

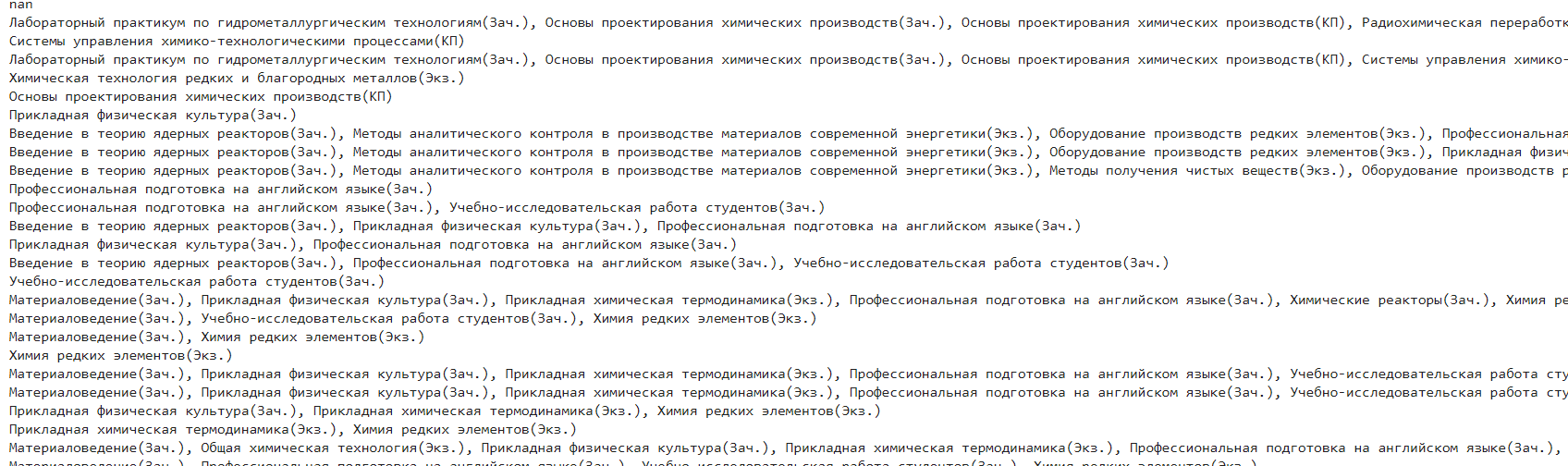


Рисунок 7 — Начало списка уникальных наборов несданных дисциплин

На рисунке 8 представлен список регулярных выражений, с помощью которых мы удаляем все факультативы из рассмотрения. Для поиска и замены факультативных дисциплин в списке используем библиотеку re (библиотека для взаимодействия с регулярными выражениями в Python).

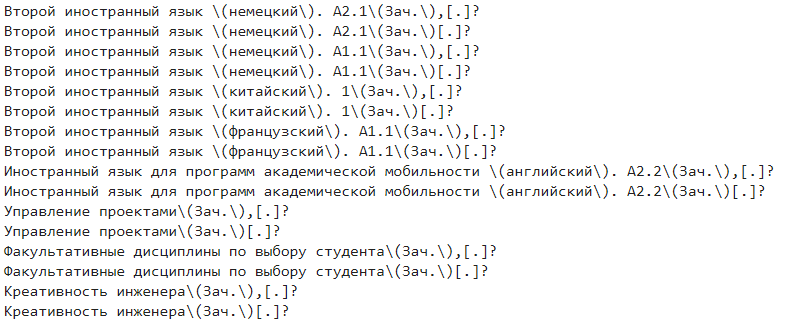


Рисунок 8 — Список регулярных выражений для удаления факультативов

Результат применения регулярных выражений представлен на рисунке 9. Кроме того, была произведена замена значения nan на категориальную строку «Нет» и убраны излишние пробелы в начале и в конце каждого строки и дублирование пробелов между словами.

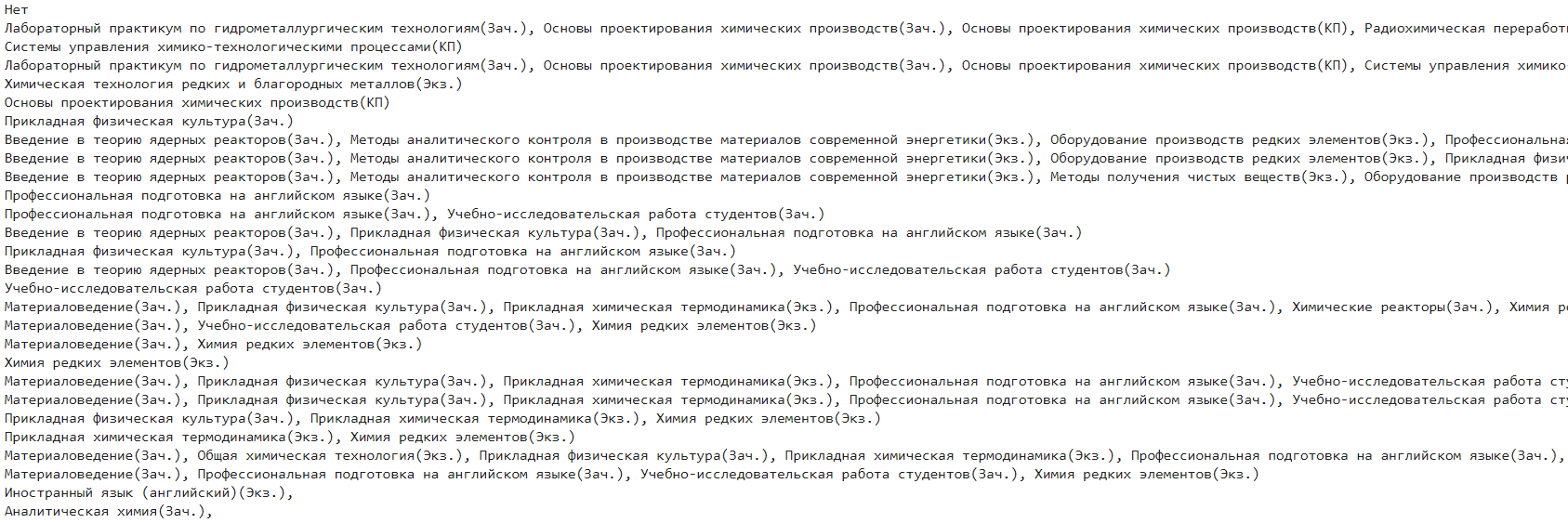


Рисунок 9 — Начало списка уникальных наборов несданных дисциплин (обработанного)

Также во всех категориальных переменных заменим nan на «Нет», а в числовых — на медиану всех значений признака.

Выделим целевую переменную. Было решено создать числовую переменную «Успешность», зависящую от столбцов «Всего» и «Положительных. Для расчета успешности каждого студента будем использовать следующую формулу:

Успешность =

На рисунке 10 представлена вычисленная успешность для первых пяти студентов в списке.

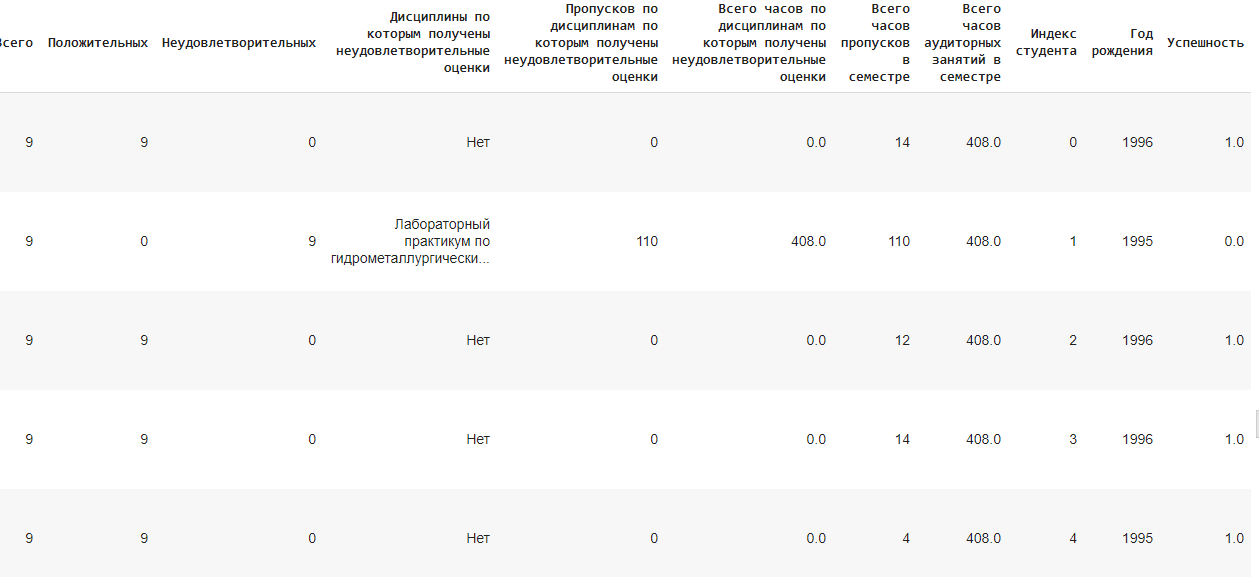


Рисунок 10 — Вычисленный столбец «Успешность»

На рисунке 11 представлена гистограмма распределения студентов по их успешности. Мы видим, что распределение ненормально, имеется повышение плотности по краям и резкое снижение в центре распределения.

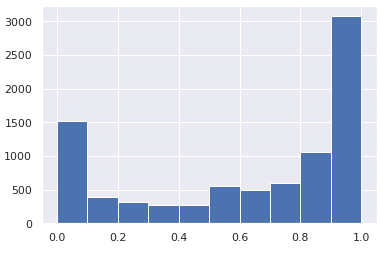


Рисунок 11 — Гистограмма распределения студентов по их успешности

Далее разобьем всех студентов на три класса. В класс «0» мы запишем студентов, чья успеваемость меньше или равна 0,25. В класс «1» — от 0,25 до 0,75, не включая граничные значения. Оставшиеся значения — больше или равно 0,75 — запишем в класс «2». Функция, классифицирующая студентов по их успеваемости, представлена на рисунке 12.

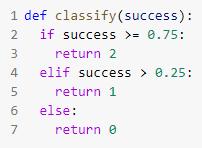


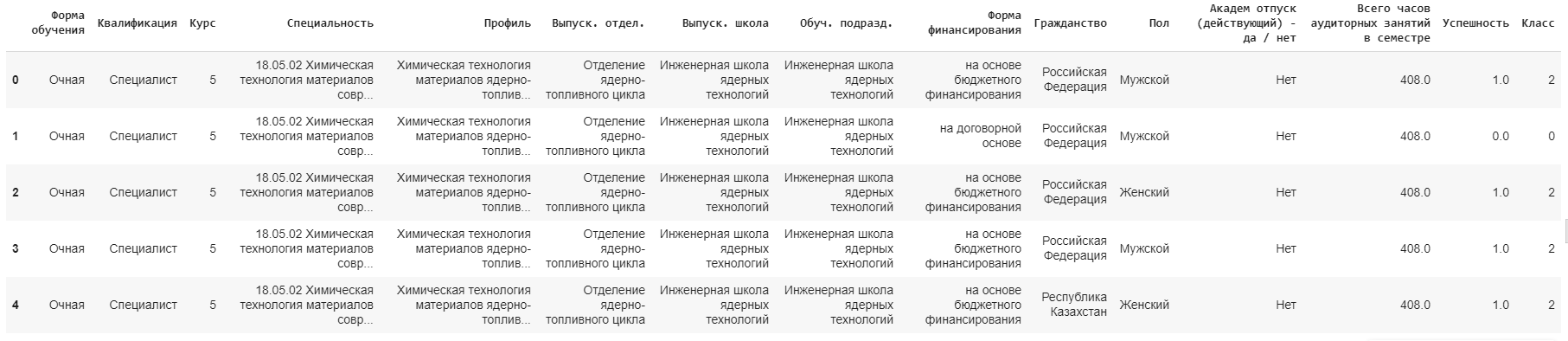
Рисунок 12 — Вычисленный столбец «Успешность»

Следующим шагом мы «очистим» датасет от признаком, не представляющих интереса для дальнейшего его анализа, создания и обучения модели.

Удалим из датасета следующие столбцы:

* «Всего» — Избыточный столбец, использовавшийся для построения признака «Успешность»;
* «Положительных» — Избыточный столбец, использовавшийся для построения признака «Успешность»;
* «Неудовлетворительных» — Избыточный столбец, зависящий от столбцов «Всего» и «Положительных»;
* «Группа» — Малоинформативный признак, большой разброс успешности у студентов каждой группы, появляются новые категории признака каждый год;
* «Страна» — Признак, практически полностью дублирующий столбец «Гражданство»;
* «Дисциплина по которым получены неудовлетворительные оценки» — Признак, который мы не сможем использовать для предсказания, так как на начало семестра мы не знаем его значений;
* «Индекс студента» — Малоинформативный признак;
* «Год рождения» — Признак, сильно коррелирующий со столбцом «Курс»;
* «Выпуск. школа» — Признак, сильно коррелирующий со столбцом «Выпуск. отдел.»;
* «Всего часов по дисциплинам по которым получены неудовлетворительные оценки» — Признак, который мы не сможем использовать для предсказания, так как на начало семестра мы не знаем его значений;
* «Пропусков по дисциплинам по которым получены неудовлетворительные оценки» — Признак, который мы не сможем использовать для предсказания, так как на начало семестра мы не знаем его значений;
* «Всего часов пропусков в семестре» — Признак, который мы не сможем использовать для предсказания, так как на начало семестра мы не знаем его значений;
* «Всего аудиторных занятий в семестре» — Линейно коррелирует с одним из признаков. В дальнейшем рассмотрим детальнее.

Итоговый набор данных представлен на рисунке 13.



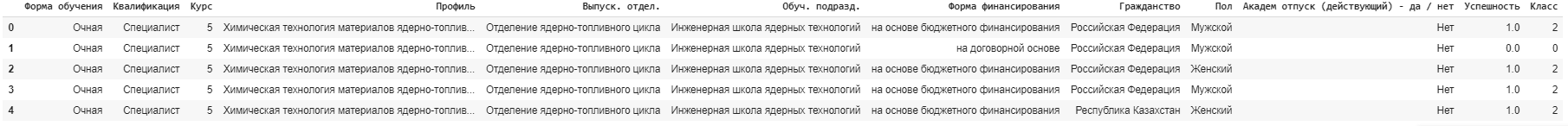


Рисунок 13 — Новый набор данных

Исследуем гистограмму распределения студентов по их классам (рисунок 14). Мы видим, что большинство студентов относятся ко второму классу, а к первому относится чуть меньше, чем к нулевому.

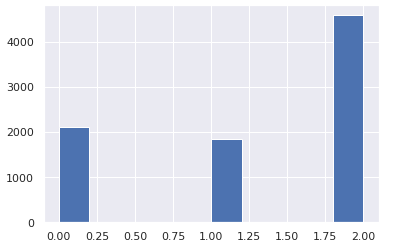


Рисунок 14 — Гистограмма распределения студентов по классам

Последнее, что мы сделаем в этом разделе — рассмотрим таблицу корреляции столбцов данного датасета (рисунок 15). Будем считать, что признаки линейно зависимы, если коэффициент корреляции Пирсона больше или равен 0,75 по модулю. Для категориальных переменных мы использовали функцию get\_dummies, которая строит на их основе новые бинарные признаки.

Прежде всего, имеются линейные зависимости между целевыми признаками «Класс» и «Успешность», что подчеркивает построение нами одного признака на основе другого. Следующая зависимость — между признаками «Форма Обучения\_Очная» и «Всего аудиторных занятий в семестре». Следовательно, придется удалить признак «Всего аудиторных занятий в семестре». Как видно по таблице, других значений больших или равных 0,75 по модулю нет, следовательно, линейных зависимостей в нашем датасете не осталось.



Рисунок 15 — Таблица корреляции столбцов датасета

На этом мы заканчиваем предобработку данных. Мы избавились от факультативных дисциплин в списке, вырезали неинформативные признаки из датасета, избавились от линейных зависимостей и сформировали два новых целевых признака «Успешность» и «Класс».

# 2 Разведочный анализ (Exploratory Data Analysis)

На рисунке 16 представлен словарь частот несданных дисциплин. Заметно, что студентов, не имеющих долгов — большинство.

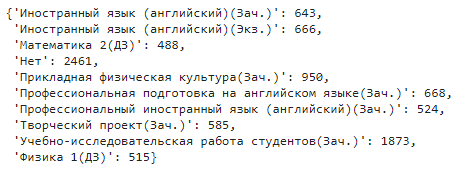


Рисунок 16 — Словарь частот несданных дисциплин

Самая часто встречающаяся несданная дисциплина — «Учебно-исследовательская работа студентов». У данного предмета немного строгих сроков сдачи, возможно, поэтому студенты недооценивают этот предмет.

Также можно допустить, что студенты несерьезно относятся к предметам «Прикладная физическая культура» и «Иностранный язык (английский)», считая, что следует уделить внимание более сложным дисциплинам.

Также в десятку самых частых несданных дисциплин входят «Физика» и «Математика». Что неудивительно, они являются одними из самых сложных общих дисциплин в любом учебном плане.

На рисунке 17 представлена гистограмма распределения студентов по несданным дисциплинам.

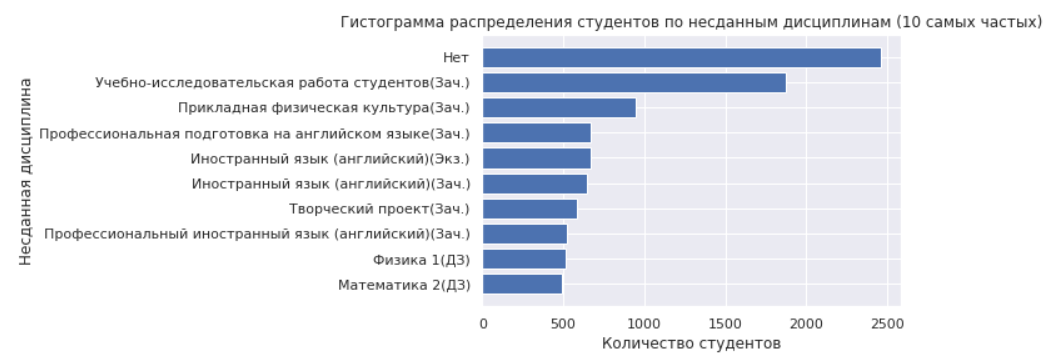


Рисунок 17 — Гистограмма распределения студентов по несданным дисциплинам

На рисунке 18 представлен словарь частот гражданств. Мы видим, что студентов из Российской Федерации большинство, но также присутствует много студентов из Республики Казахстан и Республики Узбекистан.

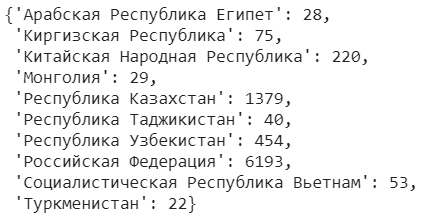


Рисунок 18 — Словарь частот гражданств

На рисунке 19 представлена гистограмма распределения студентов по гражданству.

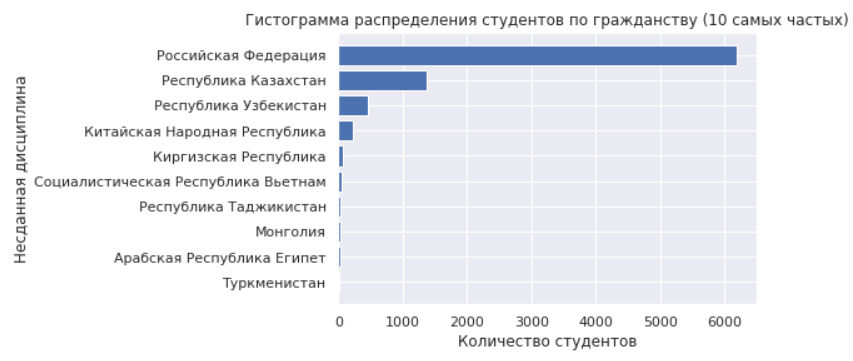


Рисунок 19 — Гистограмма распределения студентов по несданным дисциплинам

На рисунке 20 мы видим словарь частот значений признака «Год рождения». На конец весны 2019 года среди студентов большинство родилось с 1997-го по 2000-ый год

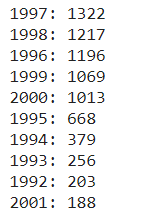


Рисунок 20 — Словарь частот значений признака «Год рождения»

На рисунке 21 представлена гистограмма распределения студентов по годам рождения.

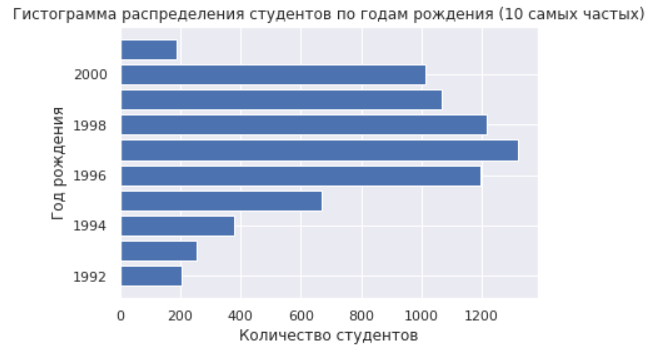


Рисунок 21 — Гистограмма распределения студентов по годам рождения

На рисунке 22 мы видим словарь частот различных форм обучения. Заметно, что большинство студентов обучаются очно. Очно-заочных студентов меньшинство.

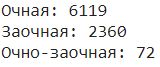


Рисунок 22 — Словарь частот форм обучения

На рисунке 23 представлена гистограмма распределения студентов по формам обучения.

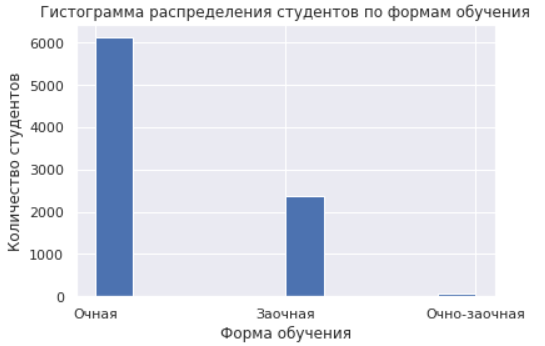


Рисунок 23 — Гистограмма распределения студентов по формам обучения

На рисунке 24 представлен словарь частот различных квалификаций, присваиваемых студентам после окончания обучения. Обучающихся в бакалавриате большинство, обучающихся в магистратуре и специалитете — примерно равное количество.

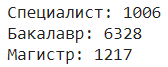


Рисунок 24 — Словарь частот форм обучения

На рисунке 25 представлена гистограмма распределения студентов по квалификациям.

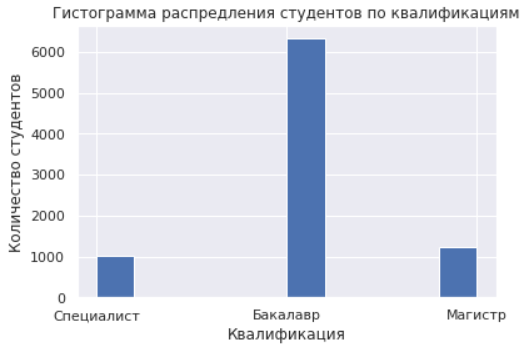


Рисунок 25 — Гистограмма распределения студентов по квалификациям

Рассмотрим бакалавриат подробнее. На рисунке 26 представлен словарь частот значений поля «Курс» для студентов бакалаврита. Подавляющее меньшинство студентов обучаются на пятом курсе, на остальных курсах обучающихся примерно равное количество.

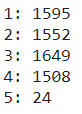


Рисунок 26 — Словарь частот форм обучения

На рисунке 27 представлена гистограмма распределения студентов бакалавриата по курсам обучения.

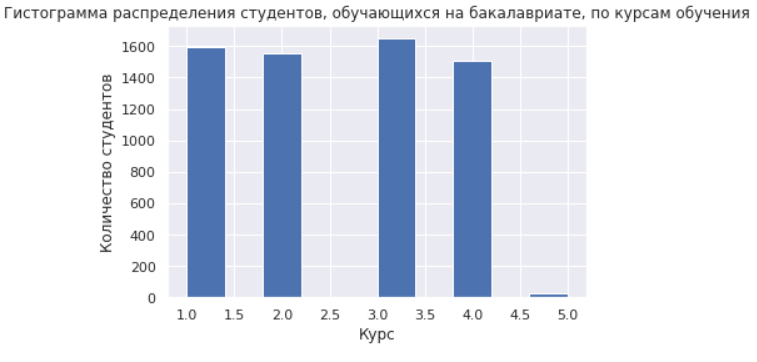


Рисунок 27 — Гистограмма распределения студентов бакалавриата по курсам обучения

На рисунке 28 представлен словарь частот форм финансирования студентов. Заметно, что большинство обучается на основе бюджетного финансирования. Меньшинство — по целевому приему.

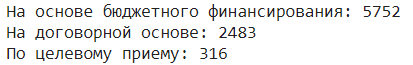


Рисунок 28 — Словарь частот форм обучения

На рисунке 29 представлена гистограмма распределения студентов по формам финансирования.

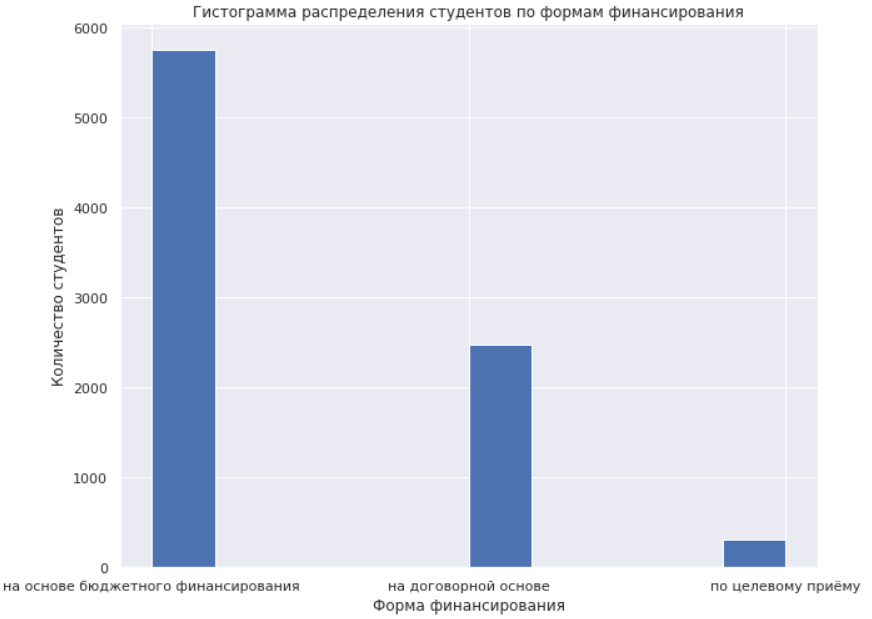


Рисунок 29 — Гистограмма распределения студентов по формам финансирования

На рисунке 30 представлен словарь частот пола студентов. Большинство студентов являются мужчинам (студентов-мужчин примерно в 2,4 раза больше, чем студентов женщин).



Рисунок 30 — Словарь частот пола студентов

На рисунке 31 представлена гистограмма распределения студентов по полу.

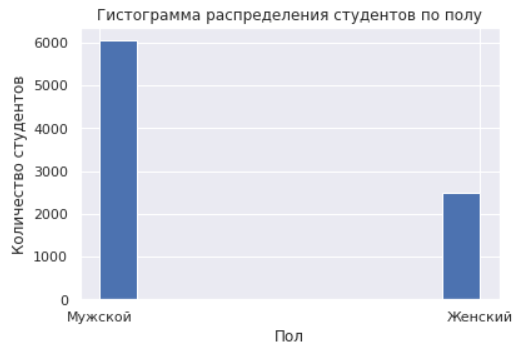


Рисунок 31 — Гистограмма распределения студентов по формам финансирования

На рисунке 32 представлен словарь частот инженерных школ, в которых обучаются студенты. Больше всего студентов обучаются в Инженерной школе природных ресурсов и Инженерной школе энергетики.

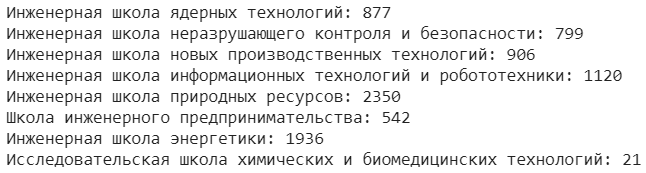


Рисунок 31 — Словарь частот инженерных школ

На рисунке 33 представлена гистограмма распределения студентов по инженерным школам.



Рисунок 33 — Гистограмма распределения студентов по инженерным школам

На рисунке 34 представлен словарь частот значений признака «Академ отпуск (действующий) - да / нет». Большинство студентов на конец весны 2019 года в академическом отпуске не находились.



Рисунок 34 — Словарь частот значений признака «Академ отпуск (действующий) – да / нет»

На рисунке 35 представлена гистограмма распределения студентов по значениям признака «Академ отпуск (действующий) - да / нет».



Рисунок 35 — Гистограмма распределения студентов по значениям признака «Академ отпуск (действующий) – да / нет»

На рисунке 36 представлена диаграмма размаха неудовлетворительных оценок по дисциплинам. Мы видим, что среднее арифметическое количество несданных дисциплин находится в районе двух, а медиана равна 3,28 (рисунок 37). Среднеквадратичное отклонение же равно 3,36 несданных дисциплин.

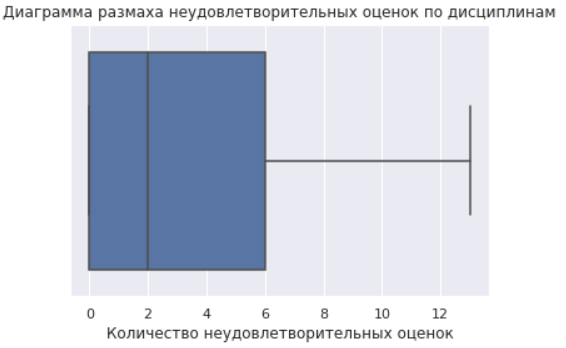


Рисунок 36 — Диаграмма размаха неудовлетворительных оценок по дисциплинам

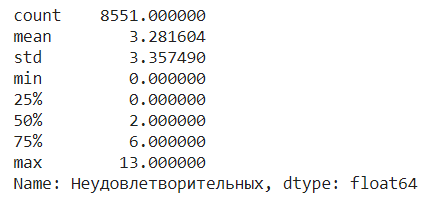


Рисунок 37 — Статистическая сводка о несданных дисциплинах

# 3 Формулирование и проверка статистических гипотез

Продолжим исследование датасета и попробуем сформулировать статистические гипотезы.

Для начала проведем изучим влияние пола на класс, к которому студент относится. На рисунке 38 представлена таблица сопряженности класса студента и его пола. Заметно, что в каждом из классов мужчин больше, чем женщин, однако во втором классе процентное отношение женщин резко возрастает. Возможно это повлияет на прогнозирование успешности студента.

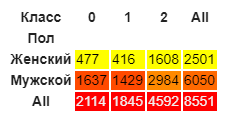


Рисунок 38 — Таблица сопряженности класса студента и его пола

На рисунке 39 представлены доли представителей второго класса среди мужчин и женщин соответственно. Видим, что доля представителей второго класса среди женщин больше.

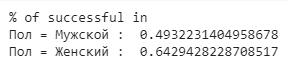


Рисунок 39 — Доли представителей второго класса среди мужчин и женщин

На рисунках 40-41 приведены доверительные интервалы для показателей успешности среди мужчин и женщин соответственно. Интервалы не пересекаются, что дает нам основание сформулировать первую статистическую гипотезу: «Средняя успешность студентов-женщин значимо выше, чем успешность студентов-мужчин».



Рисунок 40 — Доверительный интервал для доли представителей второго класса среди мужчин



Рисунок 41 — Доверительный интервал для доли представителей второго класса среди мужчин

Проверим, являются ли распределения успешности мужчин и женщин нормальными. На рисунке 42 приведены соответствующие Q-Q графики. На графиках видно, что распределения далеки от нормальных.

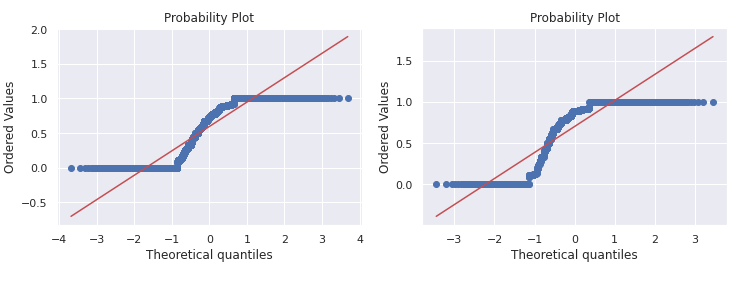


Рисунок 42 — Q-Q графики для распределения успешности мужчин и женщин нормальными

Для большей точности проверим нормальность распределений критерием Шапиро-Уилка. При взятом уровне значимости α = 0,05 p-значение и в том и в другом случаях равны нулю, следовательно, отвергаем нулевые гипотезы, что распределения нормальны (рисунки 43-44).



Рисунок 43 — Достигаемый уровень значимости по критерию Шапиро-Уилка для студентов-мужчин



Рисунок 44 — Достигаемый уровень значимости по критерию Шапиро-Уилка для студентов-женщин

Мы не можем определить вид распределения, поэтому придется использовать непараметрические критерии. Мы сравниваем две независимые выборки, поэтому будем использовать критерий Манна-Уитни и перестановочный критерий.

На рисунке 45 представлен достигаемый уровень значимости по критерию Манна Уитни. Он меньше α = 0,05, следовательно, отвергаем нулевую гипотезу в пользу двухсторонней альтернативной о том, что распределение одной выборки имеет сдвиг относительно другой.



Рисунок 45 — Достигаемый уровень значимости по критерию Манна-Уитни

Создадим аппарат для подсчета p-значения для перестановочного критерия. Для этого реализуем функцию permutation\_test (рисунки 46-47).

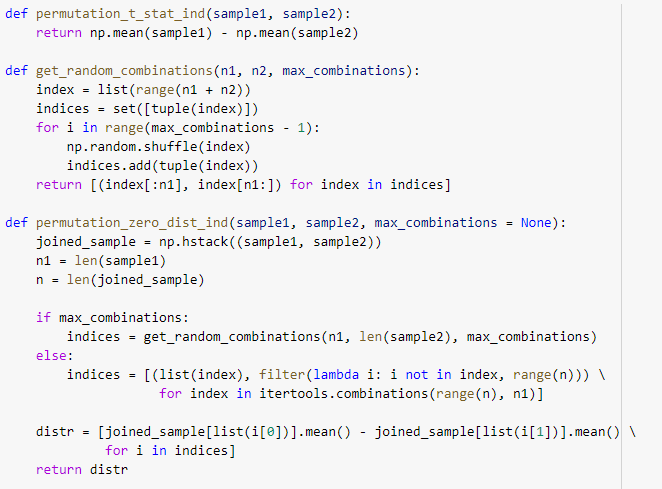


Рисунок 46 — Создание функции для подсчета p-значения перестановочного критерия

(часть 1)

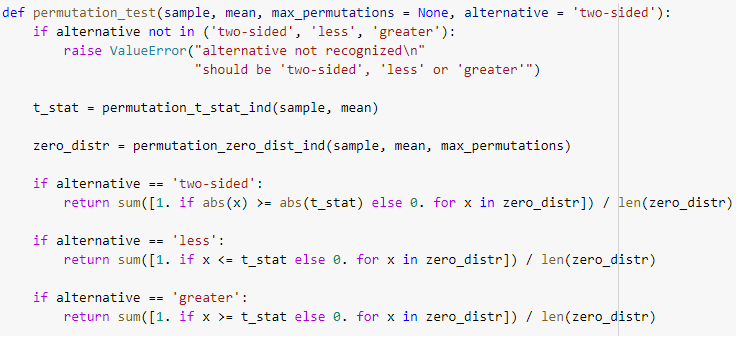


Рисунок 47 — Создание функции для подсчета p-значения перестановочного критерия

(часть 2)

Теперь вычислим p-значение перестановочного критерия. Как мы видим, для 10000 перестановок оно равно нулю, поэтому мы снова отвергаем нулевую гипотезу (рисунок 48).

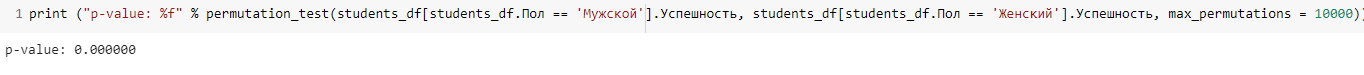


Рисунок 48 — Подсчет p-значения перестановочного критерия

Попробуем увеличить число перестановок. При использовании 50000 перестановок p-значение также равно нулю (рисунок 49).

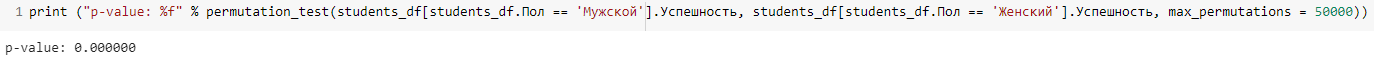


Рисунок 49 — Подсчет p-значения перестановочного критерия

Также взглянем на категориальный график успешности студентов мужского и женского пола (рисунок 50). По нему также ясно видно, что среднее значение успешности студентов женского пола выше, чем мужского.

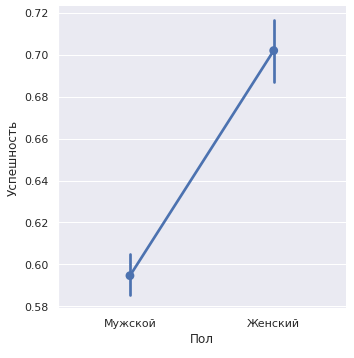


Рисунок 50 — Категориальный график успешности студентов мужского и женского пола

Таким образом, с определенной долей уверенности можно сделать вывод, что студенты женского пола в среднем действительно имеют меньше несданных дисциплин.

На рисунке 51 представлена таблица сопряженности класса студента, его пола и обучающего подразделения. В каждой инженерной школе студентов-мужчин больше, чем женщин. Также заметим, что больше всего студентов обучается в Инженерной школе природных ресурсов и Инженерной школе энергетики.



Рисунок 51 — Таблица сопряженности класса студента, его пола и обучающего подразделения

На категориальном графике успешности студентов разных полов и инженерных школ мы видим, что наиболее высокую успешность имеют студенты Инженерной школы ядерных технологий и Инженерной школы неразрушающего контроля и безопасности (рисунок 52). Сформулируем две гипотезы: «Средняя успешность студентов ИШНКБ выше, чем остальных (без учета студентов ИШЯТ)» и «Средняя успешность студентов ИШЯТ выше, чем остальных (без учета студентов ИШНКБ)».

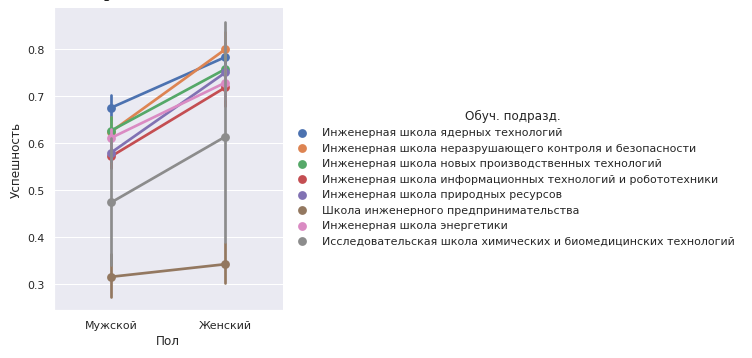


Рисунок 52 — Категориальный график успешности студентов разных полов и инженерных школ

На рисунках 53-54 мы видим результаты применения критерия Манна-Уитни к обеим выборкам. P-значения в обоих случаях во много раз меньше α, поэтому можем отвергнуть нулевые гипотезы в пользу альтернативных (распределение одной выборки имеет сдвиг относительно другой).



Рисунок 53 — P-значения для критерия Манна-Уитни для выборок студентов ИШЯТ и остальных (без студентов ИШНКБ)



Рисунок 54 — P-значения для критерия Манна-Уитни для выборок студентов ИШНКБ и остальных (без студентов ИШЯТ)

Перестановочный критерий для обоих случаев показывает одинаковое значение p-значения, равное 0,0001, что гораздо меньше принятого уровня значимости (рисунки 55-56).

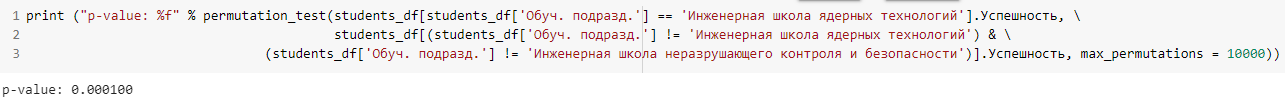


Рисунок 55 — P-значения для перестановочного критерия для выборок студентов ИШНКБ и остальных (без студентов ИШЯТ)

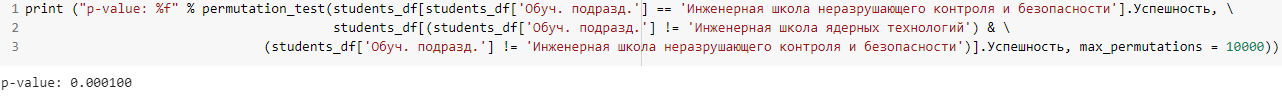


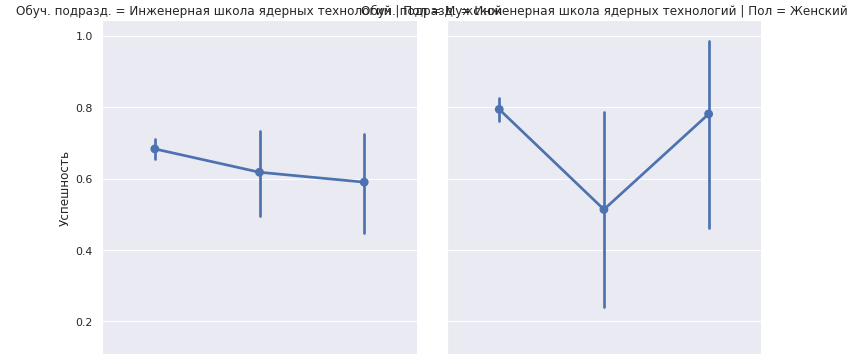
Рисунок 56 — P-значения для перестановочного критерия для выборок студентов ИШНКБ и остальных (без студентов ИШЯТ)

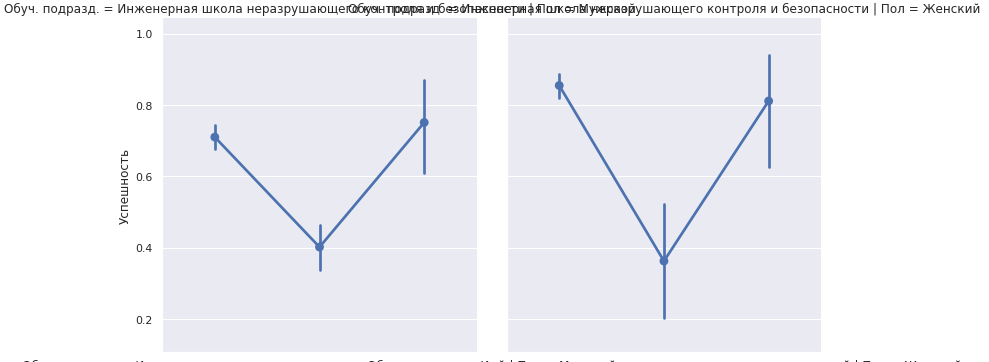
Следовательно, можем принять обе наших гипотезы.

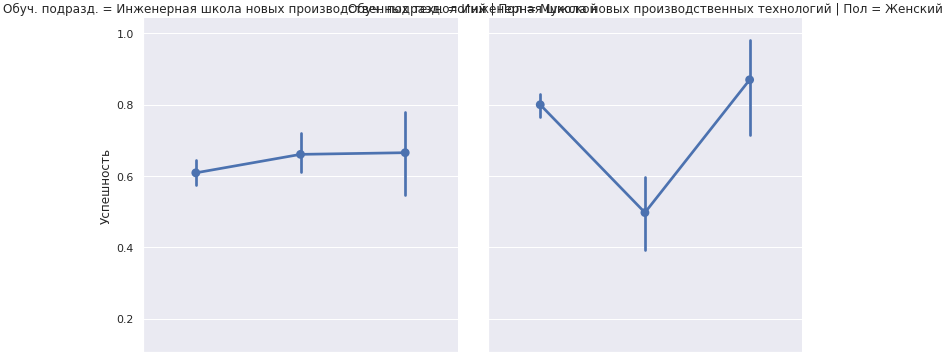
Теперь сравним студентов по форме финансирования. На рисунке 57 представлены диаграммы классов студентов для каждой формы финансирования. Сразу заметим, что для студентов, обучающихся на основе бюджетного финансирования и по целевому приему, преобладает второй класс, а для студентов, обучающихся на договорной основе, — нулевой.

 Рисунок 57 — Диаграммы классов студентов для каждой формы финансирования

На категориальных графиках успешности студентов разных полов, инженерных школ и форм обучения мы видим, что наиболее высокую успешность имеют студенты, обучающиеся по целевому приему, хотя бывает и лидируют обучающиеся на бюджетной основе (рисунки 58-66). Сформулируем гипотезу: «Средняя успешность студентов, обучающихся по целевому приему выше, чем остальных».

 Рисунок 59 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 1)

 Рисунок 60 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 2)

 Рисунок 61 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 3)

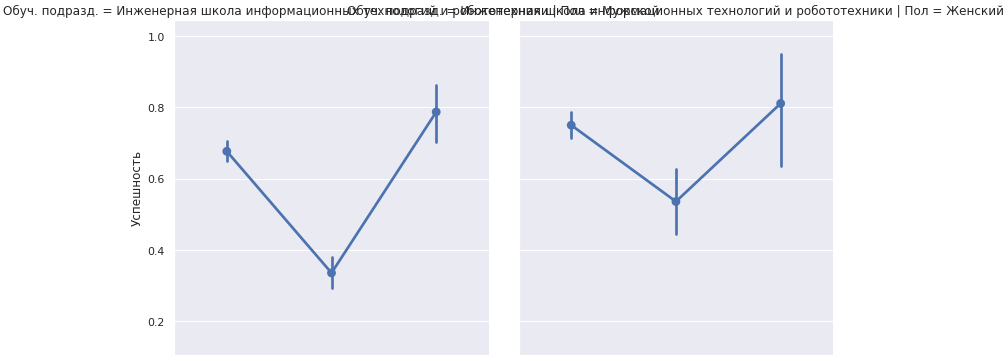
 Рисунок 62 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 4)

 Рисунок 63 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 5)

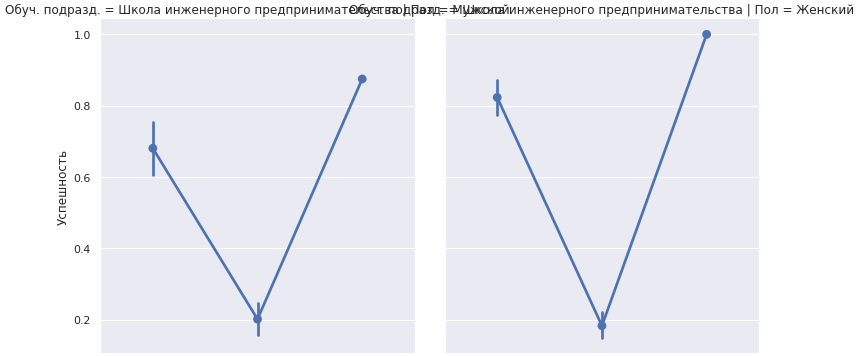
 Рисунок 64 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 6)

 Рисунок 65 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 7)

 Рисунок 66 — Категориальные графики успешности студентов разных полов, инженерных школ и форм обучения (часть 8)

Проверим это статистическими критериями. Средние значения успешности для студентов, обучающихся по целевому приему, и остальных равны, соответственно, 0,79 и 0,62 (рисунки 67-68).

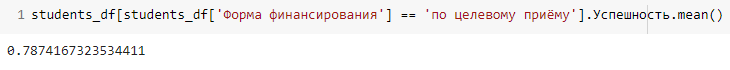
 Рисунок 67 — Среднее значение успешности для студентов, обучающихся по целевому приему



Рисунок 68 — Среднее значение успешности для студентов, не обучающихся по целевому приему

Используем критерий Манна-Уитни. На рисунке 69 видно, что достигаемый уровень значимости значительно меньше α=0,05.



Рисунок 69 — P-значение для критерия Манна-Уитни

Используем перестановочный критерий. На рисунке 70 видно, что достигаемый уровень значимости значительно меньше α=0,05.

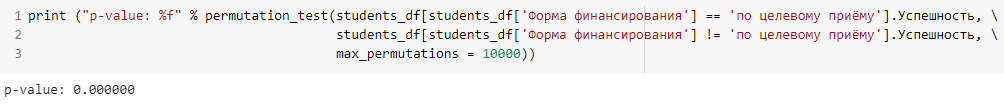


Рисунок 70 — P-значение для перестановочного критерия

Скорее сего, студенты, обучающиеся по целевому приему, действительно имеют меньше долгов в процентном соотношении, чем остальные студенты.

На категориальном графике успешности студентов разных полов, инженерных школ и квалификаций, получаемых после обучения, мы не видим четкой картины и преобладания успешности одних студентов над другими (рисунки 71-74).

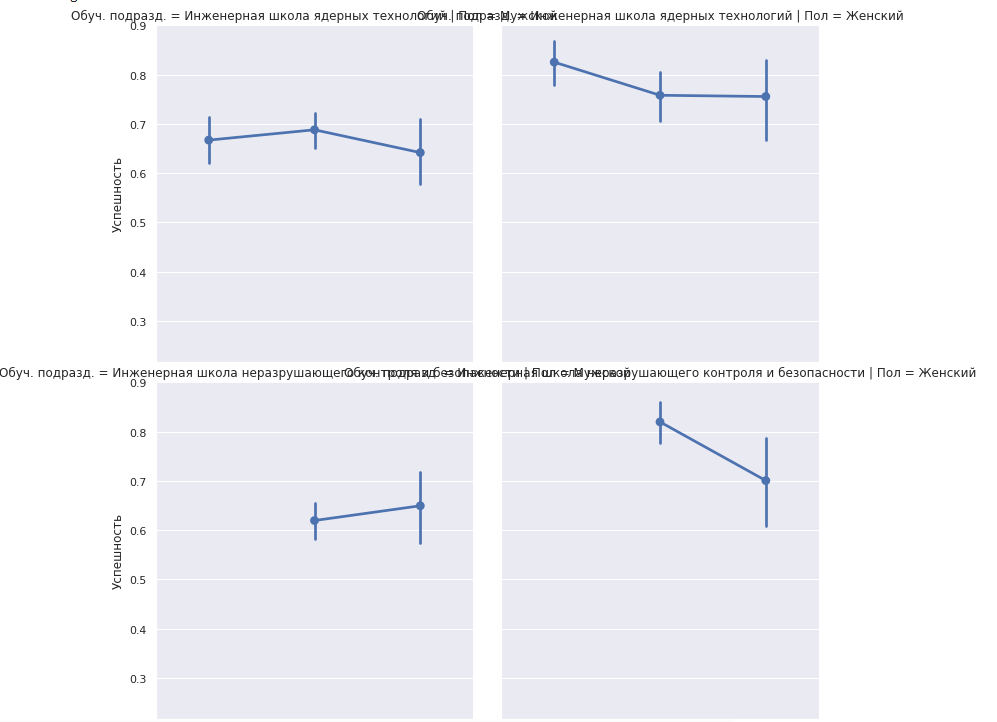
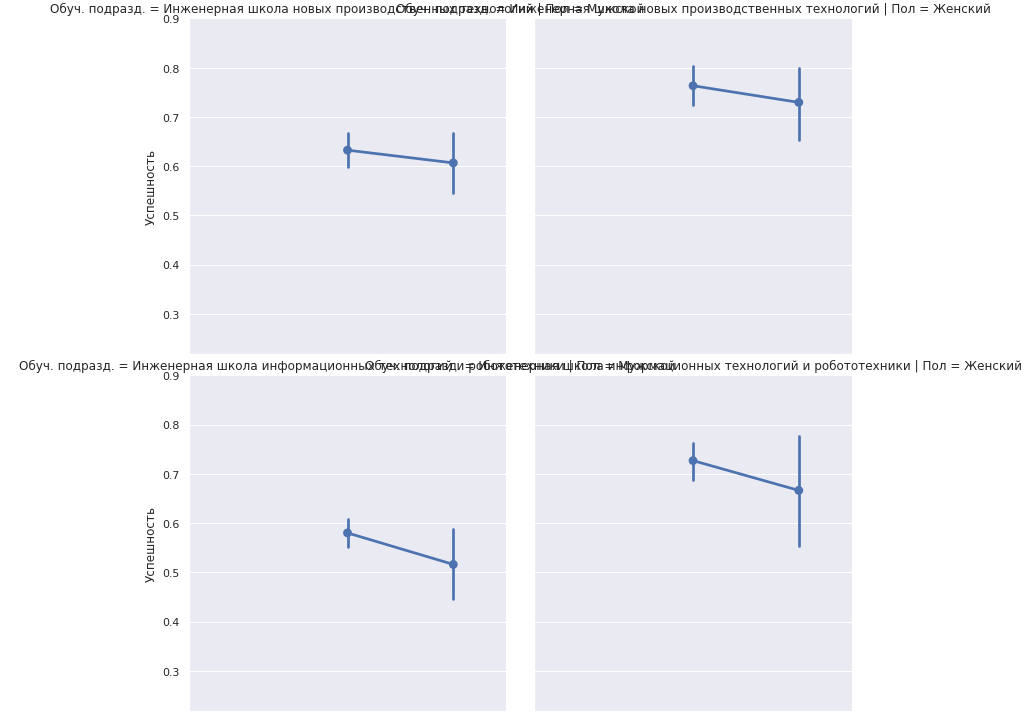


Рисунок 71 — Категориальные графики успешности студентов разных полов, инженерных школ и квалификаций (часть 1)

 Рисунок 72 — Категориальные графики успешности студентов разных полов, инженерных школ и квалификаций (часть 2)

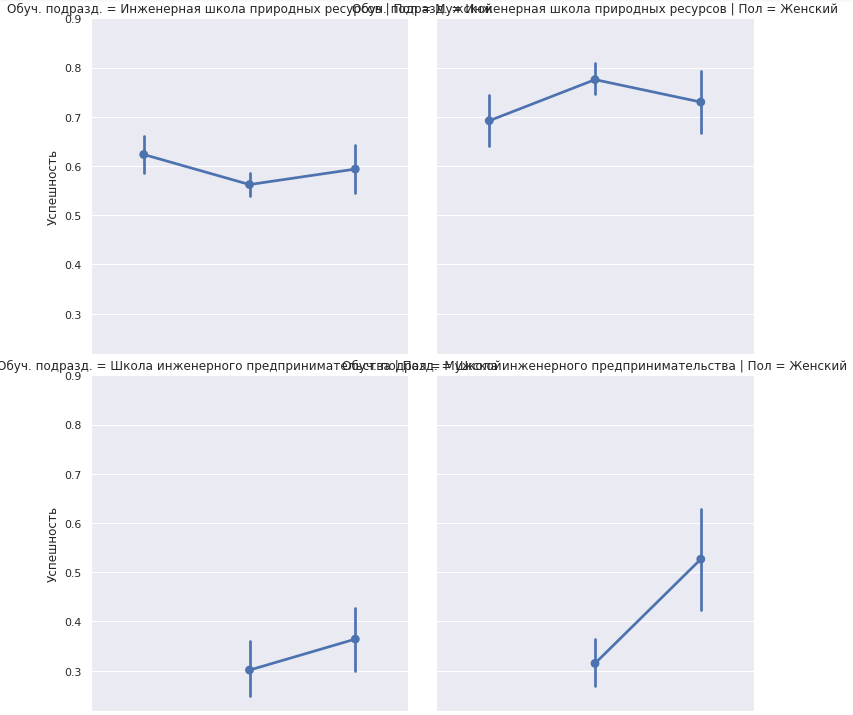
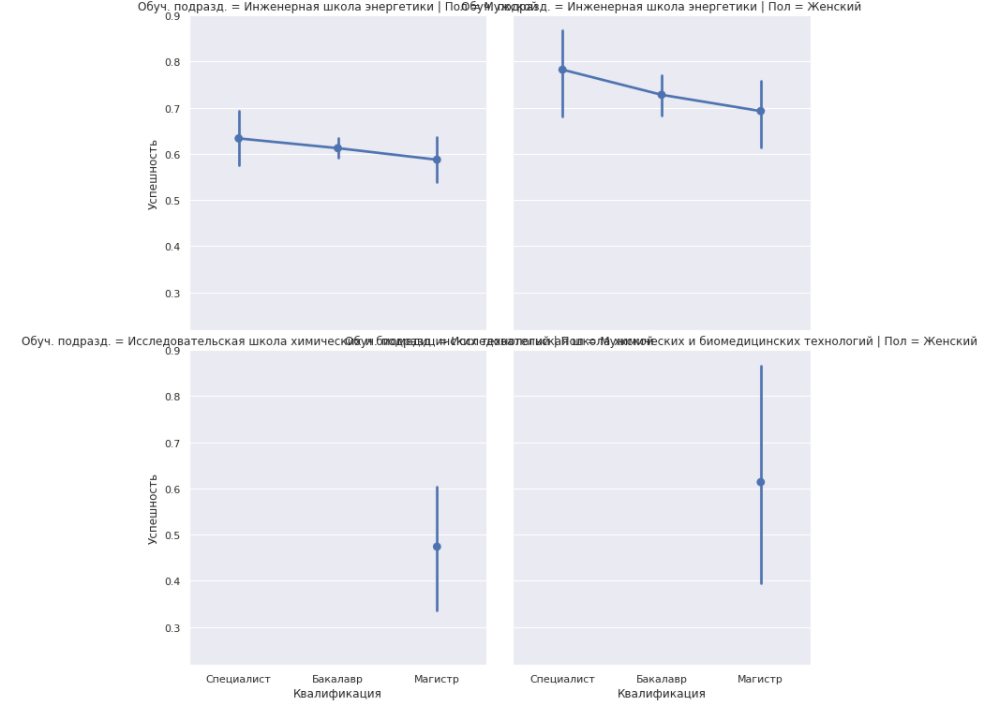
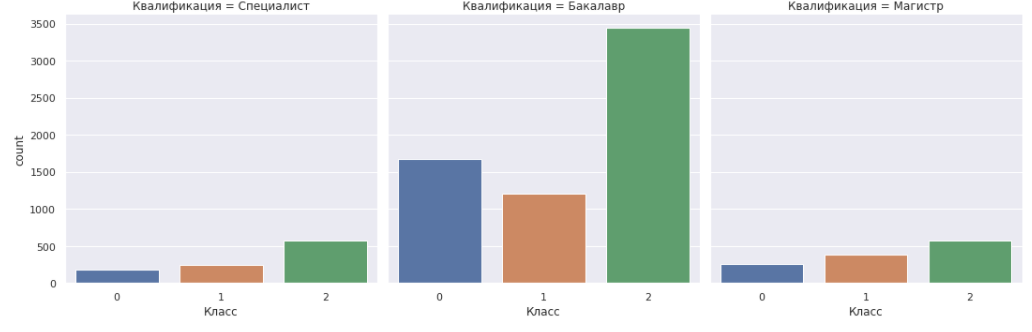


Рисунок 73 — Категориальные графики успешности студентов разных полов, инженерных школ и квалификаций (часть 3)

 Рисунок 74 — Категориальные графики успешности студентов разных полов, инженерных школ и квалификаций (часть 4)

Рассмотрим диаграммы классов студентов различных квалификаций (рисунок 75). Во всех трех графиках большинство студентов относится ко второму классу, однако в графике, относящемся к баклавриату, достаточно весомую часть всех студентов забирает нулевой класс.

 Рисунок 75 — Категориальные графики успешности студентов разных полов, инженерных школ и квалификаций (часть 4)

На рисунке 76 представлена таблица сопряженности класса студента и квалификации, которую он получит после обучения. Здесь уже становится видно, что специалитет обладает наибольшей долей успешных студентов.

Сформулируем гипотезу: «Средняя успешность студентов, обучающихся на специалитете выше, чем остальных».

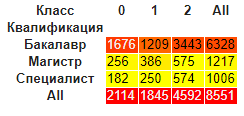


Рисунок 76 — Таблица сопряженности класса студента и его будущей квалификации

На рисунке 77 представлены средние значения успешности студентов, обучающихся на специалитете, и остальных.

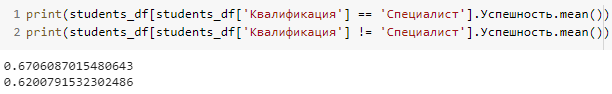


Рисунок 77 — Средние значения успешности студентов, обучающихся на специалитете, и остальных

Используем критерий Манна-Уитни. На рисунке 78 видно, что достигаемый уровень значимости значительно меньше α=0,05.



Рисунок 78 — P-значение критерия Манна-Уитни

Используем перестановочный критерий. На рисунке 79 видно, что достигаемый уровень значимости значительно меньше α=0,05.

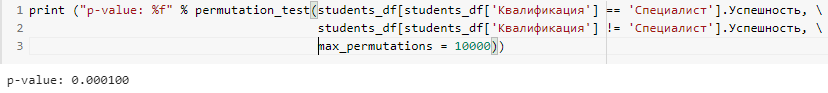


Рисунок 79 — P-значение перестановочного критерия

Следовательно, мы принимаем нашу гипотезу.

Далее рассмотрим гражданство. Для начала посмотрим, как соотносятся студенты из Российской Федерации и остальные (рисунок 80).



Рисунок 80 — Диаграмма распределения студентов (из РФ / не из РФ)

На категориальном графике успешности студентов из РФ и остальных мы видим явное преобладание успешности студентов из Российской Федерации. (рисунок 81). Можем сформулировать следующую гипотезу: «Среднее значение успешности студентов из Российской Федерации выше, чем остальных».

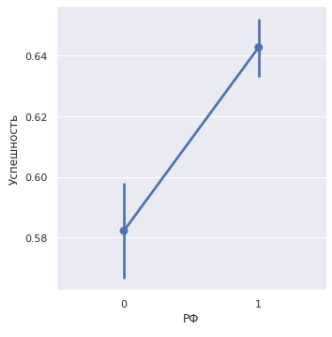


Рисунок 81 — Категориальный график успешности студентов из РФ и остальных

На рисунке 82 представлены средние значения успешности студентов из Российской Федерации и остальных.

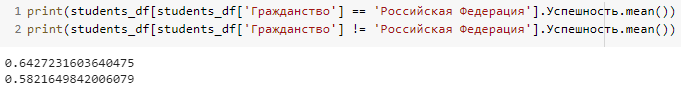


Рисунок 82 — Средние значения успешности студентов из РФ и остальных

На рисунке 83 видно, что достигаемый уровень по критерию Манна-Уитни меньше 0,05.



Рисунок 83 — Достигаемый уровень значимости по критерию Манна-Уитни

На рисунке 84 видно, что достигаемый уровень значимости по перестановочному критерию меньше 0,05.

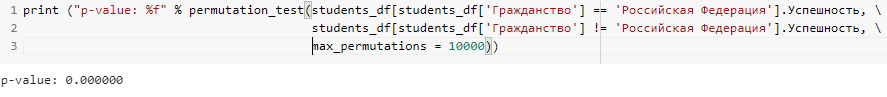


Рисунок 84 — Достигаемый уровень значимости по перестановочному критерию

Следовательно, принимаем гипотезу о тому, что среднее значение успешности студентов из Российской Федерации выше, чем остальных.

На категориальном графике успешности и курса студентов мы видим, что пик успешности происходит на четвертом курсе. (рисунок 85). Можем сформулировать следующую гипотезу: «Среднее значение успешности студентов четвертого курса выше, чем остальных».

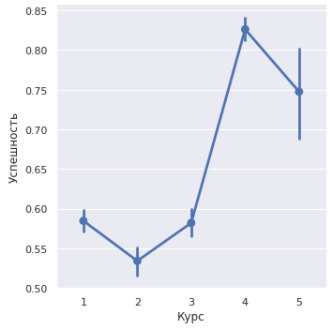


Рисунок 85 — Категориальный график успешности и курса студентов

Рассмотрим диаграммы классов студентов различных курсов (рисунок 86). Видно, что большая часть четверокурсников принадлежит второму классу. При этом на всех курсах доли студентов, относящихся к нулевому и первому классу, остается одинаковой.

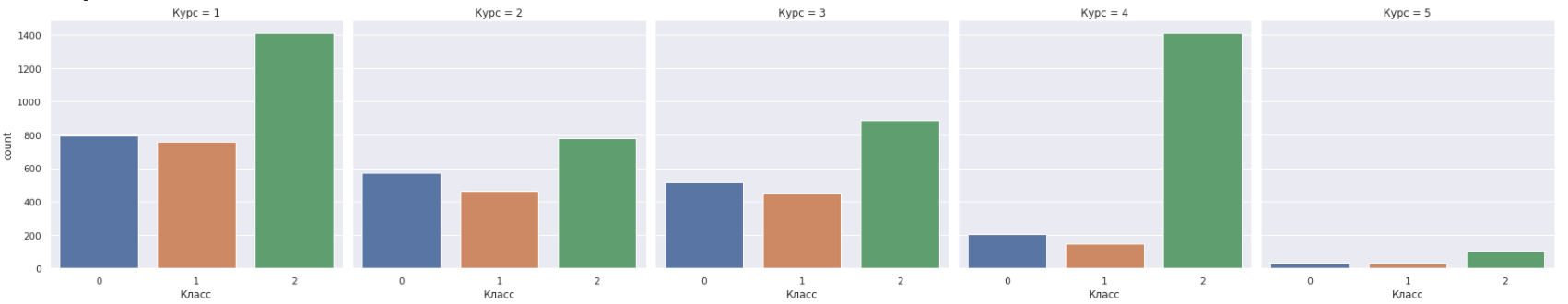


Рисунок 86 — Диаграммы классов студентов различных курсов

На рисунке 87 представлены средние значения успешности студентов четвертого курса и остальных.

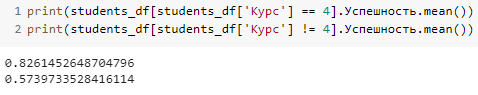


Рисунок 87 — Средние значения успешности студентов четвертого курса и остальных

На рисунке 88 видно, что достигаемый уровень по критерию Манна-Уитни меньше 0,05. Следовательно, четверокурсники действительно имеют меньше долгов.

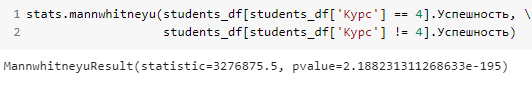


Рисунок 88 — Достигаемый уровень значимости по критерию Манна-Уитни

На рисунке 89 видно, что достигаемый уровень по перестановочному критерию также меньше 0,05.

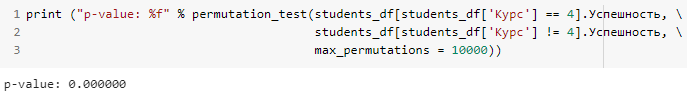


Рисунок 89 — Достигаемый уровень значимости по перестановочному критерию

Следовательно, принимаем нашу гипотезу. Среднее значение успешности студентов четвертого курса действительно значимо больше, чем у остальных студентов.

Далее рассмотрим диаграммы классов студентов, находящихся и не находящихся в академическом отпуске (рисунок 90). Видно, что на каждом графике студентов, относящихся ко второму классу, большинство.

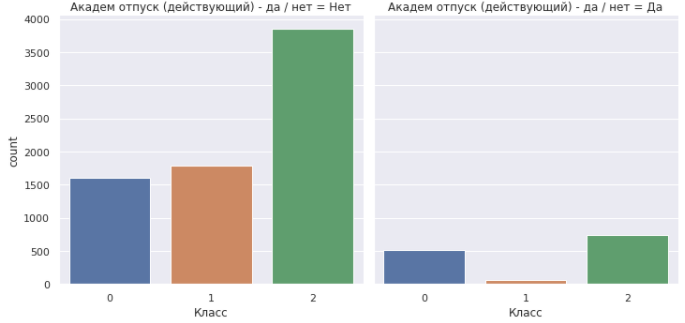


Рисунок 90 — Диаграммы классов студентов, находящихся и не находящихся в академическом отпуске

На рисунке 91 представлены средние значения успешности студентов, находящихся и не находящихся в академическом отпуске. Заметим, что у вторых оно выше.

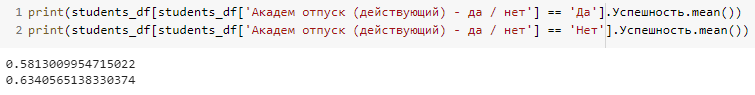


Рисунок 91 — Средние значения успешности студентов, находящихся и не находящихся в академическом отпуске

На рисунке 92 видно, что достигаемый уровень значимости по критерию Манна-Уитни меньше 0,05, хотя и относительно близок к нему. Тем не менее, можно считать, что студенты, не находящиеся в академическом отпуске, имеют меньше долгов.



Рисунок 92 — Достигаемый уровень значимости по критерию Манна-Уитни

На рисунке 93 видно, что достигаемый уровень по перестановочному критерию также меньше 0,05.

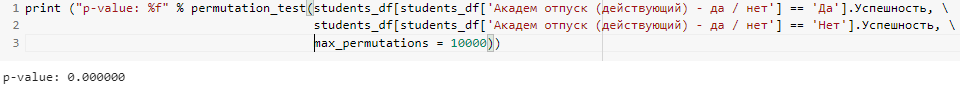


Рисунок 93 — Достигаемый уровень значимости по перестановочному критерию

Гипотезу мы снова принимаем.

На этом мы завершаем формулирование и проверку статистических гипотез. Всего выводов несколько:

* Студенты женского пола имеют меньше несданных дисциплин, чем студенты мужского;
* Студенты ИШЯТ и ИШНКБ имеют меньше несданных дисциплин по сравнению с остальными;
* Студенты, обучающиеся по целевому приему, имеют меньше несданных дисциплин по сравнению со студентами с другими формами финансирования;
* Студенты, обучающиеся на специалитете, имеют меньше несданных дисциплин по сравнению с остальными;
* Студенты, имеющие гражданство Российской Федерации, имеют меньше несданных дисциплин по сравнению с остальными;
* Студенты-четверокурсники обучаются успешнее студентов других курсов;
* Студенты, не состоящие в академическом отпуске, обучаются успешнее.

## 4 Поиск структуры в данных с помощью алгоритмов машинного обучения без учителя

Агломеративная кластеризация

lin = linkage(X, method='average', metric='cosine')

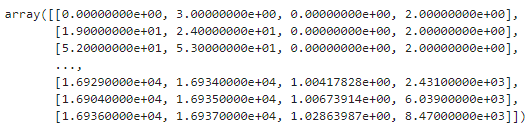


Рисунок 94 – массив значений агломеративного алгоритма

dend = dendrogram(lin)

На рисунке 95 представлен график кластеров студентов. Можно заметить, что алгоритм выявил несколько кластеров, помеченных цветами. Так, кластер голубого цвета, красного, зеленого, фиолетового выделяют определенные группы студентов, схожих по определенным признакам между собой. Синий же кластер, вероятно, включает в себя студентов, которые являются средними по всем параметрам.

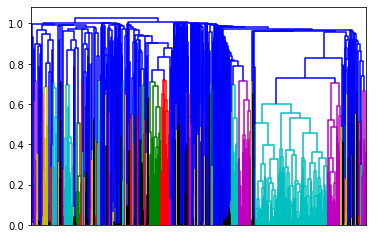


Рисунок 95 – график как результат работы агломеративной кластеризации

# 5 Построение предсказательной модели

Ниже представлен список признаков, которые не нужны при обучении модели

* Всего
* Положительных
* Неудовлетворительных
* Успешность
* Специальность
* Выпуск. школа
* Группа
* Страна
* Дисциплины по которым получены неудовлетворительные оценки
* Индекс студента
* Год рождения
* Всего часов по дисциплинам по которым получены неудовлетворительные оценки

Далее, выполним преобразование данных к бинарному виду для категориальных данных (кроме первых двух признаков), используя метод get\_dummies библиотеки pandas.

data\_dummy = pd.get\_dummies(data)



Рисунок 96 – таблица после преобразования

Нормализуем данные

for feature in (data\_dummy.columns[:2]):

mean = np.mean(data\_dummy[feature])

std = np.std(data\_dummy[feature])

data\_dummy[feature] = (data\_dummy[feature]-mean)/std



Рисунок 97 – данные для предиктивных моделей после нормализации

Преобразуем целевой вектор в матрицу соответственно числу классов (в нашем случае классов три: это 0, 1 и 2).

target\_onehot = tf.keras.utils.to\_categorical(target)

values = data\_dummy.values

Создаем наборы для тренировочной и тестовой выборки (разбиение 2 к 1). Разбиение производится для целевого вектора и для целевой матрицы.

X\_train, X\_test, y\_train\_onehot, y\_test\_onehot = train\_test\_split(values, target\_onehot, test\_size=0.33, random\_state=42)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(values, target, test\_size=0.33, random\_state=42)

**Логистическая и линейная регрессия**

Здесь применяется модель логистической и линейной регрессии с помощью таких библиотек как keras и sklearn. Размер входного вектора: 236, функция активации на входном слое: ReLU, функция активации на выходном слое: softmax (softmax является обобщением логистической функции для многомерного случая), функция оптимизации: Adadelta, функция потерь: категориальная кросс-энтропия, метрика качества: accuracy (точность). Всего используется 100 эпох (итераций обучения).

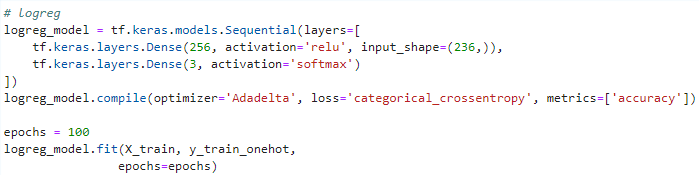


Рисунок 98 – код модели логистической регрессии

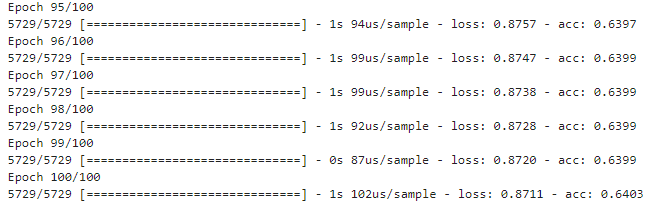


Рисунок 99 – процесс обучения на выборке данных



Рисунок 100 – код для вычисления величины потерь и точности



Рисунок 101 – результаты тестирования модели на тестовой выборке

Теперь опробуем модель линейной регрессии на подготовленных данных.



Рисунок 102 – код создания модели и получения массива предсказанных значений на основе тестовой выборки и весов самой модели



Рисунок 103 – массив предсказанных значений

Теперь же произведем округление по математическому правилу.



Рисунок 104 – функция округления



Рисунок 104 – массив преобразованных к целому виду чисел

Напишем функцию подсчета точности, которая будет использоваться в дальнейшем.

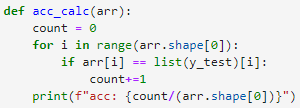


Рисунок 105 – функция вычисления точности

Вызываем функцию acc\_calc(ans2) и получаем ответ 

**Полносвязная нейронная сеть**

В данном случае полносвязная нейронная сеть – это сеть с одним скрытым слоем. Так же, как и в предыдущем случае используются функции активации ReLU на входном и скрытом слое и softmax для выходного слоя, а для компиляции применены оптимизатор Adadelta, функция потерь категориальная кросс-энтропия и в качестве метрики выбрана метрика accuracy.

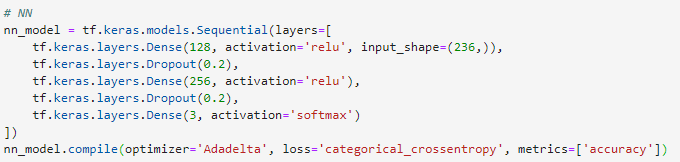


Рисунок 106 – код построения модели нейронной сети с одним скрытым слоем и ее сборку

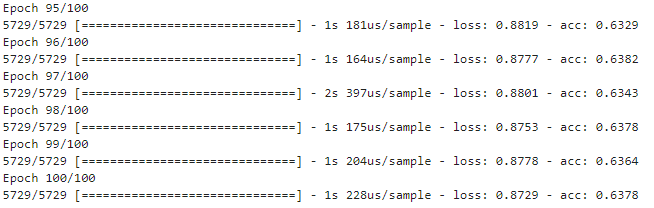


Рисунок 107 – процесс обучения нейронной сети

Ровно так же, как и для логистической регрессии, выполняется расчет показателей величины потерь и точности.



Рисунок 108 – код, выполняющий вычисление оценок качества модели



Рисунок 109 – результаты тестирования модели на тестовой выборке

**Случайный лес**

Применим еще одну модель классификации. На этот раз опробуем алгоритм случайного леса с параметрами максимальной глубины дерева = 2 и обучим полученную модель на данных тренировочной выборки.

clf = RandomForestClassifier(max\_depth=2, random\_state=42)

clf.fit(X\_train, y\_train)

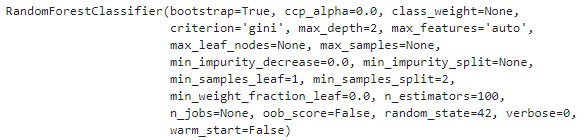


Рисунок 110 – полный список параметров модели

ans3 = clf.predict(X\_test)



Рисунок 111 – массив предсказанных классов (принадлежность студента к тому или иному классу)

В итоге, применив функцию acc\_calc(ans3) получаем ответ:



**Метод опорных векторов**

Далее используем достаточно популярный метод под названием Метод Опорных Векторов. В нашем случае данный алгоритм применяется как классификатор. Суть данного метода заключается в следующем: алгоритм переводит исходный вектор в пространство более высокой размерности и пытается разделить классы между собой гиперплоскостью с максимальным зазором. Используя по умолчанию заданный параметр регуляризации C = 1 и ядро в качестве РБФ (радиально-базисной функции), обучим данный метод машинного обучения и сделаем предсказание.

clf2 = svm.SVC()

clf2.fit(X\_train, y\_train)

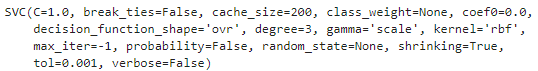


Рисунок 112 – полный список параметров модели

ans4 = clf2.predict(X\_test)

acc\_calc(ans4) Ответ:

Можно заметить, что данный алгоритм показал пока что самую лучшую точность среди всех использованных ранее алгоритмов.

**Алгоритм k-ближайших соседей**

Последним алгоритмом будет алгоритм ближайших соседей.

neigh = KNeighborsClassifier(n\_neighbors=3)

neigh.fit(X\_train, y\_train)



Рисунок 113 – полный список параметров модели

ans5 = neigh.predict(X\_test)

acc\_calc(ans5)

Ответ: 

Попытка совмещения ответов различных алгоритмов результат не улучшила: точность упала с 75.7% до 74.2%.

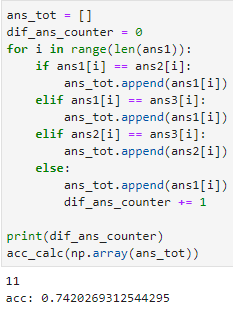


Рисунок 114 – код с выведенным результатом. Всего имеется 11 ответов, которые не совпали у трех алгоритмов

# Вывод

В ходе производственной практики был исследован набор данных о характеристиках студентов Томского Политехнического Университета и их успеваемости, а также составлена и обучена прогнозирующая модель.

Данные были предобработаны: персональные данные были вынесены в отдельный файл, пропущенные значения были изменены соответствующим образом, были удалены факультативы. Также удалены неинформативные и линейно-зависимые признаки.

Был проведен разведочный анализ. Получены следующие выводы:

* Студенты женского пола имеют меньше несданных дисциплин, чем студенты мужского;
* Студенты ИШЯТ и ИШНКБ имеют меньше несданных дисциплин по сравнению с остальными;
* Студенты, обучающиеся по целевому приему, имеют меньше несданных дисциплин по сравнению со студентами с другими формами финансирования;
* Студенты, обучающиеся на специалитете, имеют меньше несданных дисциплин по сравнению с остальными;
* Студенты, имеющие гражданство Российской Федерации, имеют меньше несданных дисциплин по сравнению с остальными;
* Студенты-четверокурсники обучаются успешнее студентов других курсов;
* Студенты, не состоящие в академическом отпуске, обучаются успешнее.

Было составлено и обучено несколько прогнозирующих моделей. Самая эффективная из них по показателю точности оказалось модель на основе опорных векторов. Точность модели – 0,7576.

В дальнейшем планируется развивать работу. Будет добавлено и изучено еще несколько датасетов, составлено больше моделей с другими параметрами.

# Список литературы

1. Спицын В.Г., Цой Ю.Р. Интеллектуальные системы: Учебное пособие – Томск: Издательство ТПУ, 2012. – 72 с.
2. С. Хайкин, «Нейронные сети: полный курс, 2-е издание». : Пер. с англ. – М. : Издательский дом «Вильямс», 2006. – 1104 с.
3. Python для анализа данных. [Электронный ресурс]. – Режим доступа: https://www.coursera.org/specializations/machine-learning-data-analysis, свободный. – Загл. с экрана. (Дата обращения 23.09.2020
4. Python для анализа данных. [Электронный ресурс]. – Режим доступа: https://www.coursera.org/learn/python-for-data-science/home/welcome, свободный. – Загл. с экрана. (Дата обращения 23.09.2020)
5. Разведочный анализ данных. [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Разведочный\_анализ\_данных, свободный. – Загл. с экрана. (Дата обращения 23.09.2020)