МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Слушатель Княжевский Ярослав Владимирович

Москва, 2022

Оглавление

1.Аналитическая часть

* 1. Постановка задачи
  2. Описание используемых методов
  3. Разведочный анализ данных

2. Практическая часть

* 1. Предобработка данных
  2. Разработка и обучение модели
  3. Тестирование модели
  4. Написать нейронную сеть, которая будет рекомендовать соотношение матрица

3. Вывод

**1.Аналитическая часть**

* 1. Постановка задачи

**Тема:** Прогнозирование конечных свойств новых материалов (композиционных материалов).

**Описание:**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

**На входе** имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

**Актуальность:** Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

**Датасет** со свойствами композитов. Объединение делать по индексу тип объединения INNER

<https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw_SOi-vyNf2/view?usp=sharing>

Требуется:

1. Изучить теоретические основы и методы решения поставленной задачи.
2. Провести разведочный анализ предложенных данных. Необходимо нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек. Необходимо также для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков.
3. Провести предобработку данных (удаление шумов, нормализация и т.д.).
4. Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении модели необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.
5. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.
6. Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании 4 или 5 (один или два прогноза, на выбор учащегося).
7. Оценить точность модели на тренировочном и тестовом датасете.
8. Создать репозиторий в GitHub / GitLab и разместить там код исследования. Оформить файл README.

В целях формирования понимания по проблематике вопроса были изучены требования к ВКР, где были изложены общие сведения по рассматриваемой задаче, а также внешние источники раскрывающие данный вопрос с иными подходами. В соответствии с изданием "Проектирование и прогнозирование механических свойств однонаправленного слоя из композиционного материала", композиционными материалами (КМ) называют материалы, состоящие из двух и более нерастворимых друг в друге компонентов с четко обозначенной границей раздела и прочным взаимодействием по всей площади контакта. Один из компонентов КМ является непрерывной фазой и называется матрицей, в которой размещены нерастворимые в ней материалы другой природы, называемые армирующими или наполнителями.

Композиционные материалы обладают комплексом свойств, которыми обладают компоненты, и свойствами, которыми отдельные компоненты не обладают.

Для обеспечения прочного взаимодействия между компонентами КМ матрица должна обеспечить хорошую смачиваемость всей поверхности распределенного в ней наполнителя, не вызывая химических реакций в них, должна обеспечить формуемость состава в монолитное изделие без нарушения формы и размера частиц наполнителя и их взаимного расположения в течение всего периода эксплуатации изделия.

Распределение наполнителя в матрице может быть хаотичным или подчиненным определенной схеме. В первом случае материал будет изотропным, во втором - анизотропным. К изотропным принадлежат КМ, содержащие в качестве наполнителя частицы сферической формы, чешуйки, рубленные волокна, равномерно распределенные в полимерной или металлической матрице. Заданную схему взаимного распределения арматуры в анизотропном материале легче достигнуть с помощью непрерывных волокон. По определенной схеме можно подготовить волокна в виде лент, тканей заданного плетения, трикотажа.

По типу матрицы композиционные материалы подразделяются на отверждающие (реактопласты), термопластичные композиционные материалы (ТКМ), композиционные материалы на металлической матрице.

По природе наполнителей композиционные материалы подразделяются на минеральнонаполненные (стекло, базальт), углероднонаполненные (углеродные волокна), органонаполненные (органические волокна), металлонаполненные. По форме частиц наполнителя различают композиционные материалы с дисперсными наполнителями (порошки, микросферы, чешуйки), композиционные материалы, наполненные короткими волокнами (3÷12 мм), и КМ с непрерывными волокнами.

По распределению наполнителей в матрице различают изотропные и анизотропные композиционные материалы. К первым относятся КМ с дисперсными наполнителями и наполненные короткими волокнами. Ко вторым - КМ с непрерывными волокнами.

Отдельную классификационную группу образуют гибридные КМ, армированные системами волокон различной природы (стеклянные, органические, углеродные и т.д.).

Армирующие компоненты композиционных материалов в виде волокон, нитей, тканей и др. сами по себе не являются конструкционными материалами. Для того чтобы получить армированный пластик, необходимо соединить волокна друг с другом, обеспечить передачу усилий между ними, зафиксировать форму изделий и взаимное расположение армирующей компоненты, т. е. создать монолитную структуру композита. Для этого используются различные связующие. В качестве связующих применяются полимерные, металлические и керамические материалы.

Физико-механические и технологические свойства связующих должны позволять получать конструкции прочные, легкие, малоэнергоемкие, технологичные. При выборе связующего, помимо удовлетворения прочностным, технологическим, адгезионным свойствам, необходимо обратить внимание на следующее: относительное удлинение матрицы при растяжении, сжатии, сдвиге должно быть не меньше относительного удлинения волокон. Иначе при действии внешней нагрузки разрушение в первую очередь начнется в связующем, что существенно снизит несущую способность конструкции. Нужно стремиться к тому, чтобы коэффициенты теплового расширения связующего и волокна были близкими. Как правило, коэффициенты теплового расширения смол значительно больше, чем волокон, что приводит к большим внутренним напряжениям, снижающим несущую способность конструкции. Связующее должно обладать малой вязкостью и хорошей адгезионной способностью к армирующему материалу. Малая вязкость необходима для того, чтобы связующее могло проникнуть и склеить между собой волокна и после отверждения получилась монолитная система. Выбранный тип связующего должен обеспечить легкость и простоту технологического формования и заполнения формы при возможно более низких температурах и давлениях. После полимеризации (отвердевания) оно должно иметь монолитную структуру и обеспечить передачу внешних усилий к волокнам. Связующее должно обладать высокой адгезионной прочностью к волокнам. Для повышения адгезионной прочности волокна покрывают очень тонким слоем специальных веществ, называемых аппретами. Аппреты не только обеспечивают высокую адгезионную прочность между связующим и волокнами, но и снижают внутренние напряжения на границе волокно-матрица, возникшие из-за разности коэффициентов линейного расширения. От связующего в значительной степени зависят такие свойства композиционных материалов, как теплостойкость, влагостойкость, длительная прочность, стойкость под действием агрессивных сред. Теплостойкость должна быть не меньше температуры эксплуатации изделий из КМ.

После формования изделия из композиционных материалов отвержденное связующее называется матрицей. Часто эти понятия являются синонимами.

Пропитанный связующим армирующий каркас при формировании обычно подвергается действию значительного давления. Важно, чтобы пузырьки воздуха или технологического газа - носителя связующего - не задерживались в композите. Поскольку для достижения высоких механических характеристик объемное соотношение армирующего материала и полимерной матрицы в текстильных композитах должно составлять около 70/30, важно иметь способ удаления избытка связующего, введенного при пропитке. После получения требуемой степени пропитки полимерное связующее отверждается. Отверждение происходит при нормальной или повышенной температуре в одну или несколько стадий в зависимости от вида армирующего материала, типа связующего, способа изготовления композита, необходимой скорости отверждения, степени отверждения и требований к свойствам конечного продукта. Отверждение при высокой температуре увеличивает термостойкость матрицы и композита, однако при этом возрастает их хрупкость.

Необходимо указать, что используемые в рамках рассматриваемого данной работой Dataset "нашивки" имеют отношение к "наполнителям", используемым в композитных материалах. Так, например, при разработке технологии создания элементов газотурбинного двигателя из полимерных композиционных материалов применяется автоматизированная нашивка ровингом, представленная волокнистыми наполнителями: углеродными волокнами HTS 45 12R и IMS 65 24K (TohoTenax, Япония), и армидным волокном Русар-С (ООО НПП "Термотекс", г. Мытищи, Россия). В связи с этим необхдоимо кратко сказать о том, что же представляет из себя "наполнитель", он же армирующая структура.

Непропитанные волокнистые материалы (волокна, нити, жгуты, ткани, трикотаж и др.), образующие определенную плоскую или пространственную структуру, называют армирующими каркасами. Армирующие каркасы подразделяются на четыре категории: дискретную, непрерывную, одномерную, плоского переплетения (двухмерную) и пространственную интегрированную.

В дискретной волокнистой системе ориентация волокон хаотична и не поддается точному контролю.

Для структур второй категории характерна наивысшая степень непрерывности и линейности волокон. В этих структурах реализуется наибольшая эффективность свойств волокон. Подобные структуры характерны для намоточных волокнистых систем. Недостаток армирующих каркасов намоточных волокнистых систем заключается в их низкой внутри- и межслоевой прочности.

Третью категорию волокнистого армирования составляют системы плоского переплетения, с помощью которых можно получить различные геометрии структур армирующих каркасов. В случае тканых материалов путем изменения частоты переплетения нитей основы и утка можно образовывать полотняное (наибольшая частота переплетения), саржевое или сатиновое переплетения (наименьшая частота переплетения). Чем меньше извитость волокон в армирующем каркасе, тем больше эффективность использования его прочностных и жесткостных свойств и меньшая структурная целостность. Полотняное переплетение характеризуется большой структурной целостностью за счет снижения эффективности использования прочности нити. И, наоборот, большая эффективность использования прочности нитей у ткани сатинового переплетения характеризуется меньшей структурной целостностью.

Четвертую категорию волокнистых систем образуют интегрированные пространственные системы, в которых армирующие волокна ориентированы в разных направлениях в плоскости и пространстве. Основная положительная черта таких структур – отсутствие выраженных направлений слабого сопротивления сдвигу, отрыву и расслоению благодаря многонаправленному армированию. В основе всех типов армирующих каркасов лежит нить. Свойства нити зависят в свою очередь от физических свойств и структуры составляющих ее волокон, о чем в данной работе мы говорить не будем.

В условиях расширения разнообразия материалов, используемых при проектировании нового композита возрастает потребность в определении характеристики нового композита с наименьшими финансовыми затратами. Для решения данной задачи как правио использую два пути: физические испытания образцов материалов, или прогнозирование характеристик, в том числе на основе физико-математических моделей. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента). В рамках текущей работы мы попробуем спроектировать модель, способную с высокой долей вероятности прогнозировать модули упругости при растяжении и прочности при растяжении, а также напишем нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.

Для разрешения указанных задач необхдоимо проанализировать представленную для обучения моделей базу данных.

* 1. Описание используемых методов

- Визуализация данных — это представление данных в виде, который обеспечивает наиболее эффективную работу человека по их изучению. Визуализация данных находит широкое применение в научных и [статистических](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) исследованиях (в частности, в прогнозировании, [интеллектуальном анализе данных](https://ru.wikipedia.org/wiki/Data_mining), [бизнес-анализе](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D0%B7%D0%BD%D0%B5%D1%81-%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7)), в [педагогическом дизайне](https://ru.wikipedia.org/w/index.php?title=%D0%9F%D0%B5%D0%B4%D0%B0%D0%B3%D0%BE%D0%B3%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B4%D0%B8%D0%B7%D0%B0%D0%B9%D0%BD&action=edit&redlink=1) для обучения и тестирования, в новостных сводках и аналитических обзорах. Визуализация данных связана с [визуализацией информации](https://ru.wikipedia.org/wiki/%D0%92%D0%B8%D0%B7%D1%83%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_%D0%B8%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D0%B8), [инфографикой](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%84%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0" \o "Инфографика), [визуализацией научных данных](https://ru.wikipedia.org/w/index.php?title=%D0%92%D0%B8%D0%B7%D1%83%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_%D0%BD%D0%B0%D1%83%D1%87%D0%BD%D1%8B%D1%85_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85&action=edit&redlink=1), [разведочным анализом данных](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D0%B7%D0%B2%D0%B5%D0%B4%D0%BE%D1%87%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85) и [статистической графикой](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D0%BA%D0%B0).

В данной работе использовались такие методы визуализации данных, как диаграмма « ящик с усами» (рис.1) , гистограммы (рис.2) и парные графики (рис. 3).

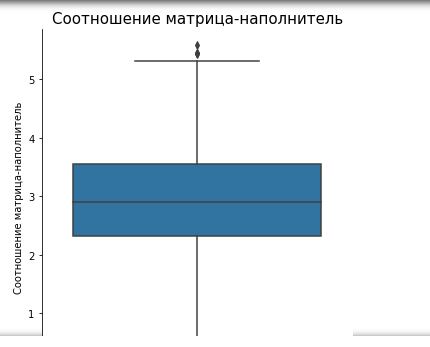


рис.1

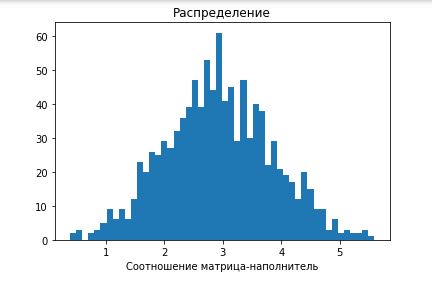


рис. 2

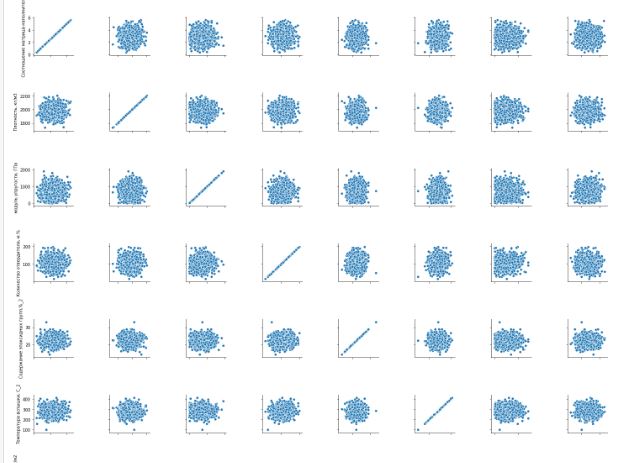


рис.3

- Корреляция (от лат. correlatio), корреляционная зависимость — взаимозависимость  двух или нескольких случайных величин. Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшению или увеличению) другой(-их) переменной(-ых).

При расчете корреляций пытаются определить, существует ли статистически достоверная связь между двумя или несколькими переменными в одной или нескольких выборках. Например, **взаимосвязь** между ростом и весом детей, взаимосвязь между успеваемостью и результатами выполнения теста IQ, между стажем работы и производительностью труда.

Важно понимать, что корреляционная зависимость отражает только **взаимосвязь между переменными** и не говорит о причинно-следственных связях.

Пример корреляции из работы, визуализация (рис.4)



рис.4

- Предобработка данных - замена, модификация или удаление частей набора данных с целью повышения непротиворечивости, полноты и корректности. набора данных, а также уменьшения избыточности.

## - Логистическая регрессия.

Логистическая регрессия моделирует функцию среднего распределения Бернулли как линейное уравнение (среднее значение равно вероятности p события Бернулли). Используя ссылку logit как функцию от среднего значения (p), логарифм шансов (log-odds) можно получить аналитически и использовать в качестве ответа так называемой обобщенной линейной модели. Вдобавок к предсказанию, это позволяет интерпретировать модель в причинно-следственной связи. Этого не получится достичь с помощью линейного персептрона.

В отличие от обычной регрессии, в методе логистической регрессии не происходит предсказание значения числовой переменной на основе выборки исходных значений. Вместо этого, значением функции является вероятность того, что данное исходное значение принадлежит к определенному классу. Для простоты, давайте предположим, что у нас есть только два класса и вероятность, которую мы будем определять,   вероятности того, что некоторое значение принадлежит классу "+". И конечно  . Таким образом, результат логистической регрессии всегда будет находиться в интервале [0, 1].

Основная идея логистической регрессии заключается в том, что пространство исходных значений возможно разделить линейной границей (т.е. прямой) на два различных класса (области). В случае двух измерений  под линейной границей подразумеватся просто прямая линия без изгибов. В случае трех — плоскость, и так далее. Эта граница задается в зависимости от имеющихся исходных данных и обучающего алгоритма. Чтобы все работало, точки исходных данных должны разделяться линейной границей на две вышеупомянутых области. Если точки исходных данных удовлетворяют этому требованию, то их можно назвать линейно разделяемыми.

Для понимания геометрического подтекста «разделения» исходного пространства на две области возьмем две исходные переменные -  и , тогда функция, соответствующая границе, примет вид:

Рассмотрим точку *(a,b).* Подставляя значения  и  в граничную функцию, получим результат . Теперь, в зависимости от положения *(a,b)* следует рассмотреть три варианта:

* *(a,b)*  лежит в области, ограниченной точками класса "+". Тогда , будет положительной, находясь где-то в пределах (0,image). С математической точки зрения, чем больше величина этого значения, тем больше расстояние между точкой и границей. А это означает большую вероятность того, что *(a,b)*  принадлежит классу "+". Следовательно,   будет находиться в пределах (0,5, 1].
* *(a,b)*  лежит в области, ограниченной точками класса "-". Теперь,  будет отрицательной, находясь в пределах (-∞, 0). Но, как и в случае с положительным значением, чем больше величина выходного значения по модулю, тем больше вероятность, что *(a,b)*  принадлежит классу "-", и    находится в интервале [0, 0.5).
* *(a,b)*  лежит на самой границе. В этом случае, . Это означает, что модель действительно не может определить, принадлежит ли *(a,b)*  к классу "+" или к классу "-". И в результате,    будет равняться 0,5.

Итак, имеется функция, с помощью которой возможно получить значение в пределах (-∞,∞) имея точку исходных данных. Но далее требуется преобразовать полученное значение в вероятность , пределы которой [0, 1]. Это выполняется с помощью функции **отношения шансов (OR)**.

Обозначим *P(X)* вероятностью происходящего события *X*. Тогда, отношение шансов (*OR(X))* определяется из  , а это — отношение вероятностей того, произойдет ли событие или не произойдет. Очевидно, что вероятность и отношение шансов содержат одинаковую информацию. Но, в то время как *P(X)*  находится в пределах от 0 до 1, *OR(X)* находится в пределах от 0 до ∞.

Это значит, что необходимо выполнить еще один шаг, так как используемая нами граничная функция выдает значения от -∞ до ∞. Далее необходимо подсчитать логарифм *OR(X)*, что называется логарифмом отношения шансов. В математическом смысле, *OR(X)* имеет пределы от 0 до ∞, а   — от -∞image до ∞.

В итоге, разработан способ интерпретации результатов, подставленных в граничную функцию исходных значений. В используемой модели граничная функция определяет логарифм отношения шансов класса "+". В сущности, в нашем двухмерном примере, при наличии точки *(a,b)*  , алгоритм логистической регрессии будет выглядеть следующим образом:

Шаг 1. Вычислить значение  граничной функции (или, как вариант, функцию отношения шансов). Для простоты обозначим эту величину *t*.

Шаг 2. Вычислить отношение шансов: . (так как *t* является логарифмом ).

Шаг 3. Имея значение , вычислить  с помощью простой зависимости.

Получив значение *t* в шаге 1, можно объединить шаги 2 и 3:

Правая часть уравнения, указанного выше, называется логистической функцией. Отсюда и название, данное этой модели обучения.

## - Случайный лес.

Случайный лес является очень популярным и эффективным алгоритмом машинного обучения. Это разновидность ансамблевого алгоритма, называемого бэггингом.

Бутстрэп является эффективным статистическим методом для оценки какой-либо величины вроде среднего значения. Берется множество подвыборок из ваших данных, считается среднее значение для каждой, а затем усредняются результаты для получения лучшей оценки действительного среднего значения.

В бэггинге используется тот же подход, но для оценки всех статистических моделей чаще всего используют деревья решений. Тренировочные данные разбивают на множество выборок, для каждой из которой создают модель. Когда необходимо получить предсказание, то это делает каждая модель, а затем прогнозы усредняют для более сглаженного выходного значения.

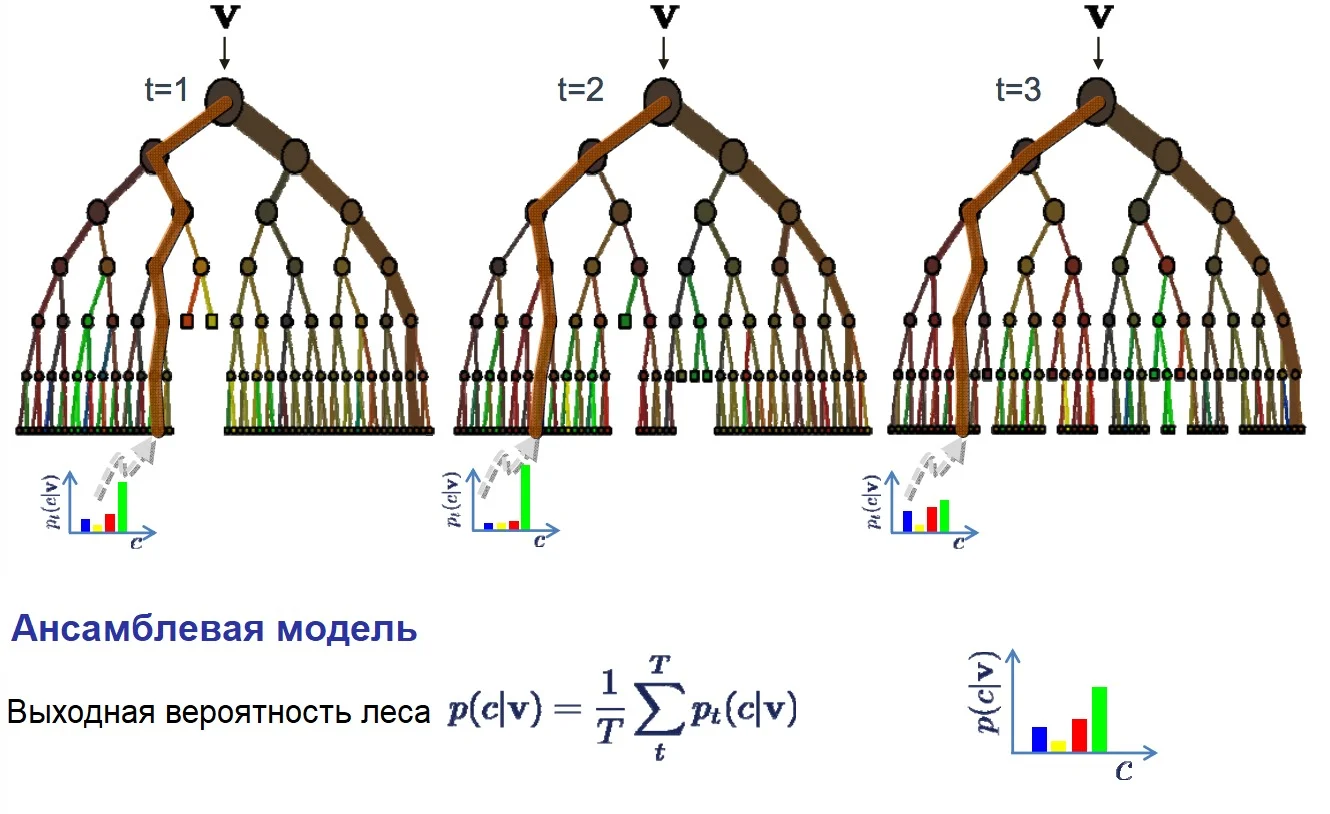


рис. 5

В алгоритме случайного леса для всех выборок из тренировочных данных строятся деревья решений. При построении деревьев для создания каждого узла выбираются случайные атрибуты. В отдельности полученные модели не отличаются высокой точностью, но при их объединении качество предсказания значительно улучшается.

Деревья решений являются одним из эффективно работающих инструментов интеллектуального анализа данных и предсказательной аналитики. Этот метод позволяет решать задачи классификации и регрессии.

Деревья решения по сути являются иерархическими древовидными структурами, состоящими из решающих правил вида «Если ..., то ...». Правила автоматически создаются в ходе обучения на  выборке  и, поскольку они генерируются практически на естественном языке (например, «Если **объём продаж** более 1000 шт., то товар **перспективный**»), то деревья решений как аналитические модели более вербализуемы и интерпретируемы, чем, скажем, нейронные сети.

Поскольку правила при таком алгоритме получаются путём обобщенного анализа входных данных (обучающих примеров), описывающих предметную область, то по аналогии с соответствующим методом логики их называют индуктивными правилами, а сам процесс обучения — индукцией деревьев решений.

В обучающей выборке для примеров требуется задать целевое значение, т.к. деревья решений - это модели, которые строятся на основе обучения с учителем. При этом, если целевая переменная дискретна (метка класса), то модель называют деревом классификации, а если непрерывна , то деревом регрессии.

Собственно, само дерево решений — это метод представления решающих правил в иерархической структуре, которая состоит из элементов двух типов — узлов (node) и листьев (leaf). В узлах находятся решающие правила и выполняется проверка на соответствие примеров этому правилу по какой-либо характеристике обучающего множества.

В самом простом случае, в результате проверки, множество примеров, попавших в узел, разбивается на два подмножества, в одно из которых попадают примеры, удовлетворяющие правилу, а в другое — не удовлетворяющие.

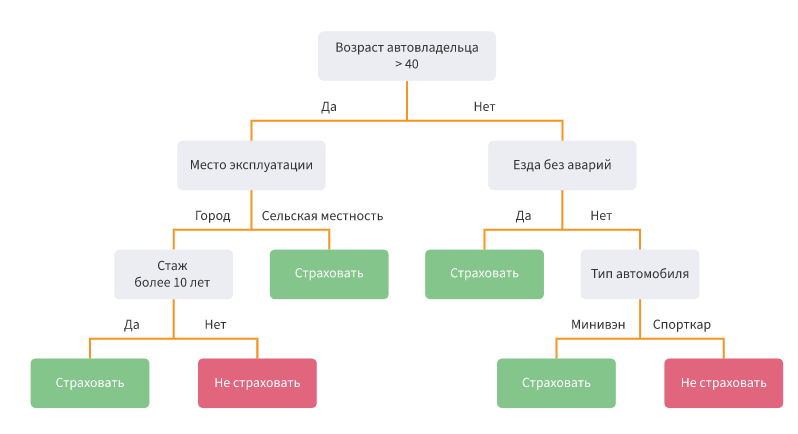


рис. 6

Затем к каждому подмножеству вновь применяется правило и процедура рекурсивно повторяется, пока не будет достигнуто некоторое условие остановки алгоритма. В итоге на последнем узле проверка и разбиение не производятся и он объявляется листом. Лист определяет конечное решение для каждого попавшего в него примера. Для дерева классификации — это класс, ассоциируемый с узлом, а для дерева регрессии — соответствующий листу модальный интервал целевой переменной.

Основные этапы построения дерева решений

Построение осуществляется в 4 этапа:

* Выбираем атрибут для осуществления разбиения в данном узле.
* Определяем критерий остановки обучения.
* Выбираем метод отсечения ветвей.

Оцениваем точность построенного дерева.

Если алгоритм с высокой дисперсией, например, деревья решений, показывает хороший результат на ваших данных, то этот результат зачастую можно улучшить, применив бэггинг.

Рассмотрим задачу регрессии с базовыми алгоритмами . Предположим, что существует истинная функция ответа для всех объектов *y(x)*, а также задано распределение на объектах *p(x)*. В этом случае мы можем записать ошибку каждой функции регрессии

и записать матожидание среднеквадратичной ошибки

Средняя ошибка построенных функций регрессии имеет вид

Предположим, что ошибки несмещены и некоррелированы:

Построим теперь новую функцию регрессии, которая будет усреднять ответы построенных нами функций:

При помощи усреднения ответов мы значительно уменьшаем ошибку.

Бэггинг позволяет снизить дисперсию (variance) обучаемого классификатора, уменьшая величину, на сколько ошибка будет отличаться, если обучать модель на разных наборах данных, или другими словами, предотвращает переобучение. Эффективность такого подхода достигается по причине того, что базовые алгоритмы, обученные по различным подвыборкам, получаются достаточно различными, и их ошибки взаимно компенсируются при обобщении, а также за счёт того, что объекты-выбросы могут не попадать в некоторые обучающие подвыборки.

Среди прочих методов Data Mining, метод дерева принятия решений имеет несколько достоинств:

1. Простой в понимании и интерпретации. Люди способны интерпретировать результаты модели дерева принятия решений после краткого объяснения

2. Не требует подготовки данных. Прочие техники требуют нормализации данных, добавления фиктивных переменных, а также удаления пропущенных данных.

3. Способен работать как с категориальными, так и с интервальными переменными. Прочие методы работают лишь с теми данными, где присутствует лишь один тип переменных. Например, метод отношений может быть применѐн только на номинальных переменных, а метод нейронных сетей только на переменных, измеренных по интервальной шкале.

4. Использует модель «белого ящика». Если определѐнная ситуация наблюдается в модели, то еѐ можно объяснить при помощи булевой логики. Примером «черного ящика» может быть искусственная нейронная сеть, так как результаты данной модели поддаются объяснению с трудом.

5. Позволяет оценить модель при помощи статистических тестов. Это даѐт возможность оценить надѐжность модели.

6. Является надѐжным методом. Метод хорошо работает даже в том случае, если были нарушены первоначальные предположения, включѐнные в модель.

7. Позволяет работать с большим объѐмом информации без специальных подготовительных процедур. Данный метод не требует специального оборудования для работы с большими базами данных.

При этом данный метод имеет и ряд недостатков:

1. Проблема получения оптимального дерева решений является NPполной с точки зрения некоторых аспектов оптимальности даже для простых задач. Таким образом, практическое применение алгоритма деревьев решений основано на эвристических алгоритмах, таких как алгоритм «жадности», где единственно оптимальное решение выбирается локально в каждом узле. Такие алгоритмы не могут обеспечить оптимальность всего дерева в целом.

2. В процессе построения дерева решений могут создаваться слишком сложные конструкции, которые недостаточно полно представляют данные. Данная проблема называется переобучением. Для того, чтобы еѐ избежать, необходимо использовать метод «регулирования глубины дерева».

3. Существуют концепты, которые сложно понять из модели, так как модель описывает их сложным путѐм. Данное явление может быть вызвано проблемами XOR, чѐтности или мультиплексарности. В этом случае мы имеем дело с непомерно большими деревьями. Существует несколько подходов решения данной проблемы, например, попытка изменить репрезентацию концепта в модели (составление новых суждений), или использование алгоритмов, которые более полно описывают и репрезентируют концепт (например, метод статистических отношений, индуктивная логика программирования).

4. Для данных, которые включают категориальные переменные с большим набором уровней (закрытий), больший информационный вес присваивается тем атрибутам, которые имеют большее количество уровней.

## 

## -Многослойный перцептрон (нейронные сети)

Перцептрон принимает обратную логит (логистическую) функцию от wx и не использует вероятностные предположения ни для модели, ни для ее параметра. Онлайновое обучение даст вам точно такие же оценки для весов / параметров модели, но вы не сможете интерпретировать их в причинно-следственной связи из-за отсутствия p-значений, доверительных интервалов и, следовательно, базовой вероятностной модели.

**Однослойный персептрон** (англ. *Single-layer perceptron*) — перцептрон, каждый S-элемент которого однозначно соответствует одному А-элементу, S-A связи всегда имеют вес *1*, а порог любого А-элемента равен *1*. Часть однослойного персептрона соответствует модели искусственного нейрона.

Его уникальность состоит в том, что каждый S-элемент однозначно соответствует одному A-элементу, все S-A связи имеют вес, равный *+1*, а порог A элементов равен *1*. Часть однослойного перцептрона, не содержащая входы, соответствует искусственному нейрону, как показано на картинке. Таким образом, однослойный перцептрон — это искусственный нейрон, который на вход принимает только *0* и *1*.

Однослойный персептрон также может быть и элементарным персептроном, у которого только по одному слою S,A,R-элементов.

Задача обучения перцептрона — подобрать такие , чтобы  как можно чаще совпадал с *y(x)*— значением в обучающей выборке (здесь σ — функция активации). Для удобства, чтобы не тащить за собой свободный член , добавим в вектор *x*лишнюю «виртуальную размерность» и будем считать, что . Тогда   можно заменить на  .

Чтобы обучить эту функцию, сначала выбирается функция ошибки, которую затем оптимизируется [градиентным спуском](https://neerc.ifmo.ru/wiki/index.php?title=%D0%A1%D1%82%D0%BE%D1%85%D0%B0%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B3%D1%80%D0%B0%D0%B4%D0%B8%D0%B5%D0%BD%D1%82%D0%BD%D1%8B%D0%B9_%D1%81%D0%BF%D1%83%D1%81%D0%BA). Число неверно классифицированных примеров здесь использовать неуместно, потому что эта функция кусочно-гладкая, с массой разрывов: она будет принимать только целые значения и резко меняться при переходе от одного числа неверно классифицированных примеров к другому. Поэтому следует использовать другую функцию, так называемый *критерий перцептрона*:

где *M* — множество примеров, которые перцептрон с весами *w* классифицирует неправильно.

Иначе говоря, надо минимизировать суммарное отклонение наших ответов от правильных, но только в неправильную сторону; верный ответ ничего не вносит в функцию ошибки. Умножение на *y(x)*здесь нужно для того, чтобы знак произведения всегда получался отрицательным: если правильный ответ *−1*, значит, перцептрон выдал положительное число (иначе бы ответ был верным), и наоборот. В результате получается кусочно-линейная функция, дифференцируемая почти везде, а этого вполне достаточно.

Теперь  можно оптимизировать градиентным спуском. На очередном шаге получаем:

Алгоритм такой — мы последовательно проходим примеры *,,…* из обучающего множества, и для каждого :

* если он классифицирован правильно, не меняем ничего;
* а если неправильно, прибавляем .

Ошибка на примере  при этом, очевидно, уменьшается, но, конечно, совершенно никто не гарантирует, что вместе с тем не увеличится ошибка от других примеров. Это правило обновления весов так и называется — правило обучения перцептрона, и это было основной математической идеей работы Розенблатта.

Мы применили MLPClassifier из библиотеки sklearn, задали только два параметра (линейная функция активации, порог на остановку оптимизации весов), остальные взяты по умолчанию. По умолчанию здесь 1 скрытый слой и 100 нейронов, метод оптимизации весов - Adam, то есть стохастический градиент.

* 1. Разведочный анализ данных

 анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов [визуализации](https://ru.wikipedia.org/wiki/%D0%92%D0%B8%D0%B7%D1%83%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85).

Понятие введено математиком [Джоном Тьюки](https://ru.wikipedia.org/wiki/%D0%A2%D1%8C%D1%8E%D0%BA%D0%B8,_%D0%94%D0%B6%D0%BE%D0%BD), который сформулировал цели такого анализа следующим образом:

* максимальное «проникновение» в данные,
* выявление основных структур,
* выбор наиболее важных переменных,
* обнаружение отклонений и аномалий,
* проверка основных гипотез,
* разработка начальных моделей.

Основные средства разведочного анализа — изучение [вероятностных распределений](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9) переменных, построение и анализ [корреляционных матриц](https://ru.wikipedia.org/w/index.php?title=%D0%9A%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BC%D0%B0%D1%82%D1%80%D0%B8%D1%86%D0%B0&action=edit&redlink=1), [факторный анализ](https://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%BA%D1%82%D0%BE%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7), [дискриминантный анализ](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BA%D1%80%D0%B8%D0%BC%D0%B8%D0%BD%D0%B0%D0%BD%D1%82%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7), [многомерное шкалирование](https://ru.wikipedia.org/wiki/%D0%9C%D0%BD%D0%BE%D0%B3%D0%BE%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D0%B5_%D1%88%D0%BA%D0%B0%D0%BB%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5).

В данной работе на этапе построения «ящиков с усами» видны выбросы, гистограммы свидетельствуют, что данные распределены в основном нормально, только угол нашивки исключение. Парные графики не дают информации. Корреляция показывает, что видны пары, количество отвердителя - температура вспышки, плотность нашивки - угол нашивки.

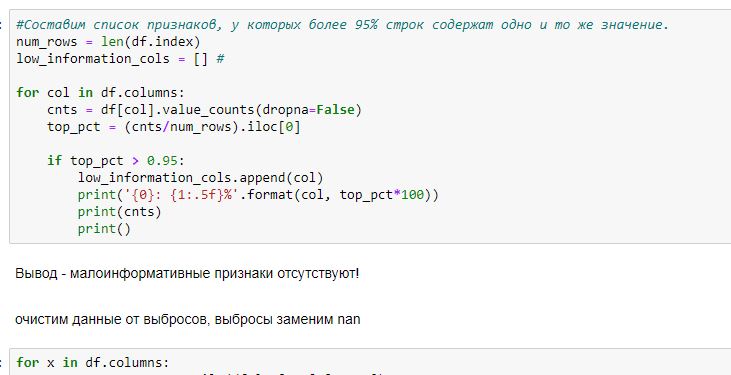
Преимущественно данные стремятся к нормальному распредедению.Угол нашивки как и отражено в датасете, имеет только два значения 90 градусов и 0 градусов, что отражает общий подход к проведению нашивки материалов.

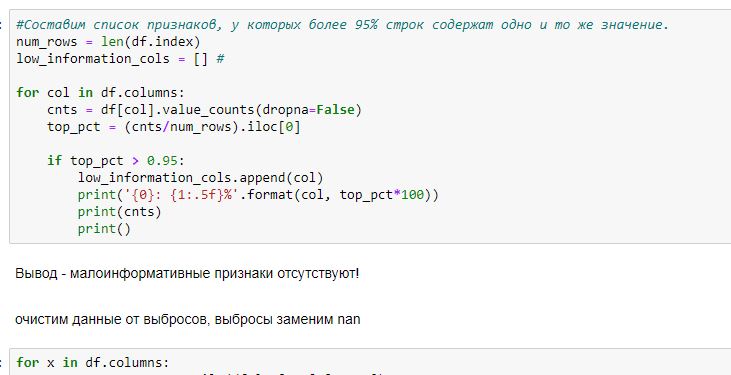
Теория составления композитных материалов в том числе указвает на закономерности, заметные в данном графике. Так на качество материла влияет температура вспышки и количество отвердителя в т.ч. из-за взаимодействия отвердителя с матрицей и наполнителем под влиянием температуры. Также имеют значение угол нашивки и плотность нашивки, что влияет на свойства материала. Менее яркие зависимости выявлены в потреблении смолы и соотношении матрицы-наполнителя, плотности и плотности нашивки, модуля упругости и плотности нашивки.

Проведенное ислледование попарных показателей демонстрирует отсутствие четкой зависимости показателей. Т.е. зависимость меж показателями не описывается линейно, нет какой-то прямой взаимосвязи, требуется участие нескольких показателей. Отсутствие линейной зависимости будет подтверждено далее при построении регрессии.

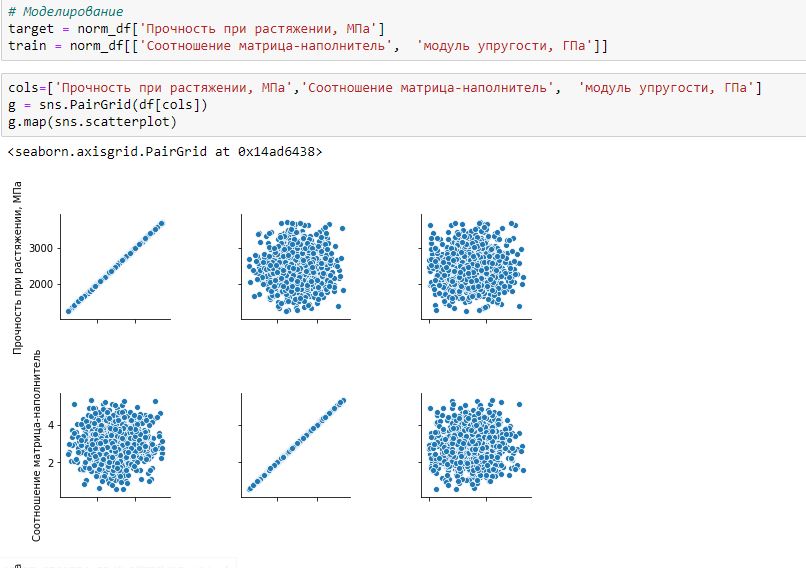
**2. Практическая часть**

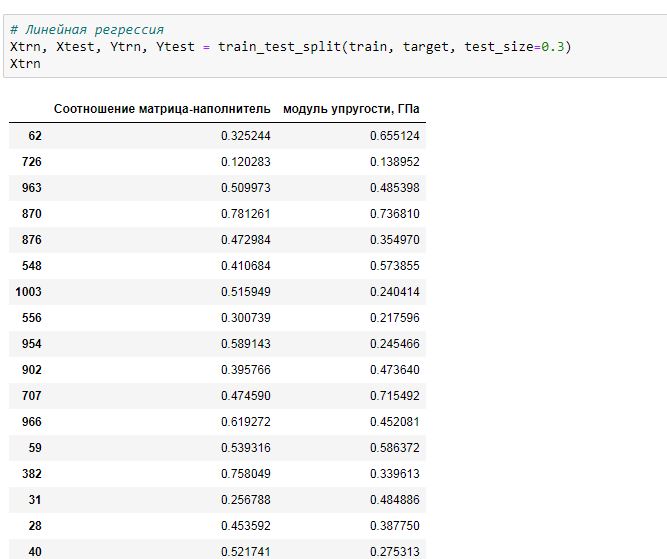
* 1. Предобработка данных





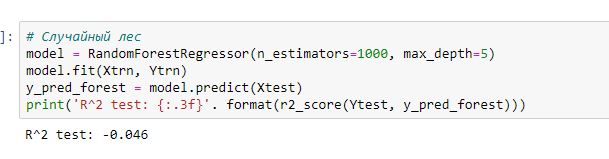
* 1. Разработка и обучение модели





* 1. Тестирование модели



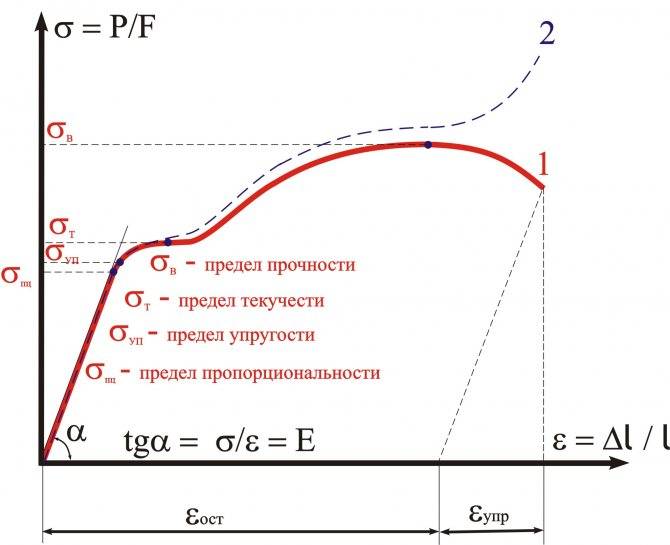


В рамках задачи предполагается прогнозировать 2 модели.

1. 'Модуль упругости при растяжении, ГПа', матрица корреляции показала зависимость с показателями: 'модуль упругости, ГПа', 'Содержание эпоксидных групп,%\_2','Температура вспышки, С\_2', 'Поверхностная плотность, г/м2'
2. 'Прочность при растяжении, МПа', матрица корреляции демонстрирует связь со следующими показателями: 'Соотношение матрица-наполнитель', 'модуль упругости, ГПа' Иные показатели демонстрируют низкий уровень корреляции

R2 принимает отрицательные значения, что говорит об отсутствии как таковой линейной взаимосвязи модели. Необходимо опробовать иные подходы для прогнозирования

Ни одна из рассмотренных моделей не позволяет прогнозировать данные по прочности со сколько-нибудь достаточной достоверностью. В соответствии с данным Wikipedia (<https://en.wikipedia.org/wiki/Ultimate_tensile_strength>). Предел прочности при растяжении (UTS), часто сокращаемый до предела прочности при растяжении (TS), предел прочности или F(tu) в пределах уравнений, это максимальное напряжение, которое материал может выдержать при растяжении или растяжении перед разрывом. В хрупких материалах предел прочности при растяжении близок к пределу текучести, тогда как в пластичных материалах предел прочности при растяжении может быть выше. Предел прочности при растяжении обычно определяется путем проведения испытания на растяжение и регистрации зависимости инженерного напряжения от деформации. Самая высокая точка кривой напряжение–деформация является пределом прочности при растяжении и имеет единицы напряжения. Эквивалентная точка для случая сжатия, а не растяжения, называется прочностью на сжатие. В общем виде все материалы проходят через следующие стадии (материал получен с ресурса: <https://101benzopila.ru/predel-procnosti-stali-pri-szatii-i-rastazenii/>):



«Техническая» (красный) и «истинная» (синяя) кривая зависимости напряжения от деформации, типичная для конструкционной стали.

1: Абсолютная сила

2: Предел текучести (предел текучести)

3: Разрыв

4: Область деформационного упрочнения

5: область шеи

A: Видимое напряжение ( F / A )

B: Фактическое напряжение ( F / A )

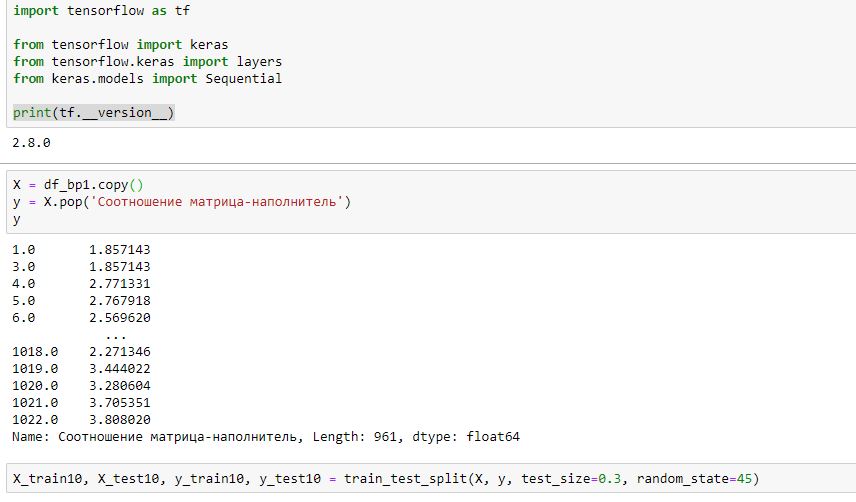
Многие материалы могут демонстрировать линейное упругое поведение , определяемое линейной зависимостью напряжения от деформации, как показано на рисунке 1 до точки 3. Упругое поведение материалов часто распространяется в нелинейную область, представленную на рисунке 1 точкой 2 ( «предел текучести»), до которого деформации полностью восстанавливаются при снятии нагрузки; то есть образец, нагруженный упруго при растяжении, удлиняется, но при разгрузке возвращается к своей первоначальной форме и размеру.

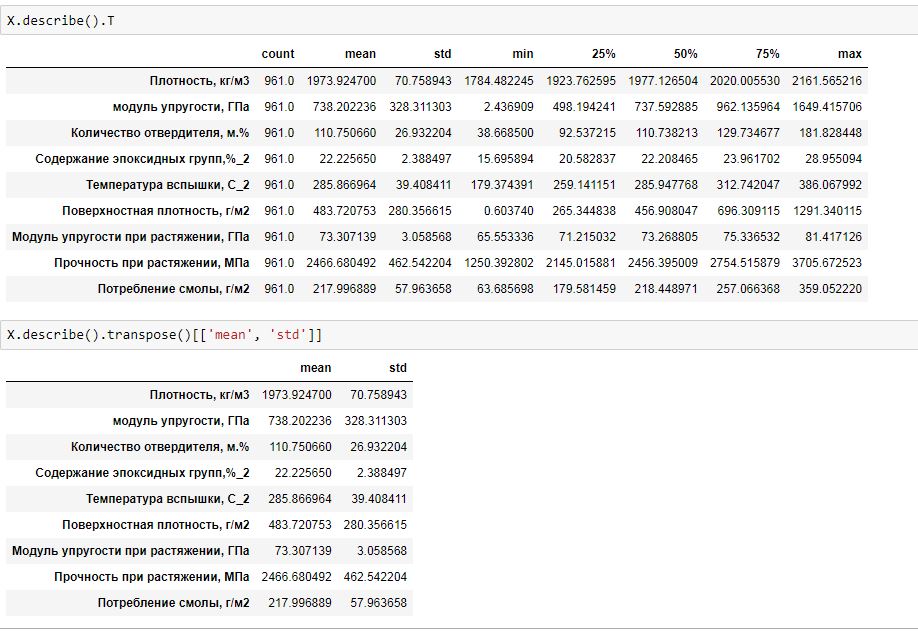
За пределами этой упругой области для пластичных материалов, таких как сталь, деформации пластичны. Пластически деформированный образец не возвращается полностью к своим первоначальным размерам и форме при разгрузке. Для многих приложений пластическая деформация недопустима и используется в качестве конструктивного ограничения.

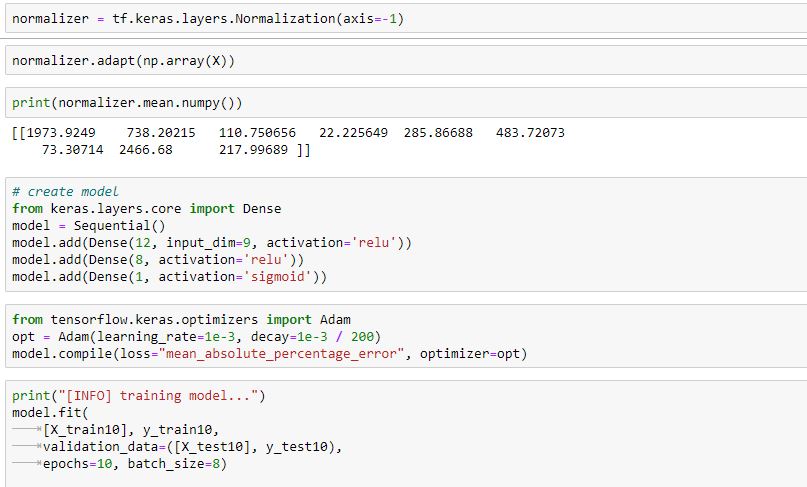
Предел прочности на растяжение не используется при проектировании пластичных статических элементов, поскольку методы проектирования диктуют использование предела текучести . Однако он используется для контроля качества из-за простоты тестирования. Он также используется для приблизительного определения типов материалов для неизвестных образцов.

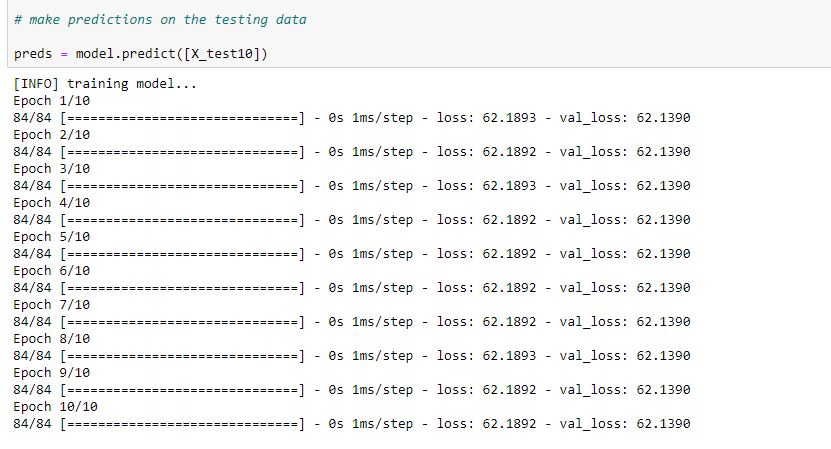
В общем виде прочность волокнистых композитных материлов определяется свойствами волокон; матрица, в основном, должна перераспределять напряжения между армирующими элементами. Поэтому прочность и модуль упругости волокон должны быть значительно больше, чем прочность и модуль упругости матрицы. Жесткие армирующие волокна воспринимают напряжения, возникающие в композиции при нагружении, придают ей прочность и жесткость в направлении ориентации волокон. Рассматриваемая нами модель, прогнозирующая качество прочности материла при растяжении на основании соотношения матрица-наполнитель и модуля упругости в общем виде должна иметь некоторую зависимость, но проведенный анализ данных ее не показывает, что может говорить в том числе о излишней зашумленности данных. Попробуем исключить второй файл, связанный с нашивкой из прогноза.

* 1. Написать нейронную сеть, которая будет рекомендовать соотношение матрица









**3. Вывод**

Как и при прогнозировании прочности, прогноз упругости демонстрирует невозможность достоверного прогнозирования от показателей, используемых в модели.

Модуль упругости в процедурах испытания новых материалов рассчитывается как:

Ep=(F2-F1)*L0/S0*(Dl2-Dl1)

где F1, F2 – значения нагрузок, соответствующих относительному удлинению 0,1% и 0,3%, Н;

Dl1, Dl2 – удлинение при нагрузках F1, F2 соответственно, мм.

За результат измерения прочности, относительного удлинения и модуля упругости принимают среднее арифметическое значение для всех образцов.

Базой для оценки упругости является модуль Юнга определяющийся как численно выраженная жесткость тела помноженная на первоначальную длину стержня, деленые на площадь поперечного сечения стержня. В общем виде график, представленный для расчета прочности имеет отношение и к упругости материала. Упругость как характеристика востановимость материал, как часть прочностной характеристикпи должна иметь вляние на характристику прочность, но представленные материалы видимой зависимости не демонстрируют.

В целях проверки гипотезы о возможности прогнозирования поробуем использовать нейронные сети.

Анализ рещультатов работы нейронной модели показывает неудовлетворительные результаты. Ошибка в 62% является значительной при прогнозировании качества композитного материла. Все предложенные характеристики так или иначе имеют возможность влиять на потенциальную прочность, упругость материала. Наполнение матрица/наполнитель является базовым, но не единственным способом, позволяющим прогнозировать конечные свойства материла. К сожалению, помимо них значительное влияние имеют температра, время обработки, давление, иные характеристики как матрицы, так и наполнителя, что в значительной мере осложняет прогнозирование. Также сложности к построению прогноза вносит отсутствие корреляции меж данными, предоставленными к анализу, что в значительной степени осложняет обучение модели.