ПСАД. ВМК. Практическое задание №3. Линейная и обобщенная линейная регрессия

1 Формулировки задания

Необходимо сдать: Rmd и сгенерированный по нему html/pdf-файл с подробным отчётом по проведённому исследованию, содержащий визуализацию исходных данных, описания и выводы каждого этапа анализа — используемые методы, обоснование их применимости, графики.

Дедлайн: 26 апреля 23:59

Все данные доступны по ссылке: https://yadi.sk/d/jxWFWgyv3Gu6BS

1.1 Ковалёва Валерия

Для 8416 грибов задано признаковое описание согласно справочнику The Audubon Society Field Guide to North American Mushrooms.

Построить модель вероятности ядовитости гриба, оценить вклад факторов.

Данные: `mushroom.csv`

1.2 Цветкова Ольга

1055 химических молекул описаны с помощью 41 признака (число атомов кислорода, нитратных групп, донорных связей с водородом, потенциал ионизации и т.д.); 355 из них биоразложимы.

Какие свойства молекул влияют на их биоразлагаемость?

Данные: `biodeg.xlsx`

1.3 Шлёнский Владислав

Собраны данные мониторинга сейсмической активности в польских угольных шахтах столбовой системы разработки. При сейсмической опасности существует серьёзный риск обрушения; в этом случае необходимо отозвать рабочих или использовать направленные взрывы для нейтрализации напряжения породы. Для каждого измерения известен бинарный индикатор сейсмической опасности — наличия в следующую восьмичасовую смену сейсмических толчков с энергией выше 10^4 Джоулей.

Построить модель сейсмической опасности, дать интерпретацию вклада показателей сейсмической активности.

Данные: `seismic.xlsx`

1.4 Стельмах Иван

Для 500 участниц исследования Global Longitudinal Study of Osteoporosis in Women (Center for Outcomes Research, the University of Massachusetts/Worcester) измерены возраст, вес, рост, ИМТ, бинарные признаки: курение, индикатор наступления менопаузы до 45 лет, индикатор необходимости помощи при подъёме из сидячего положения, перелом шейки бедра в прошлом (был/не было), перелом шейки бедра у матери (был/не было), а также самостоятельная субъективная оценка вероятности перелома (меньше/такая же/больше, чем у сверстниц). Известно, у кого из участниц в первый год исследования произошёл перелом шейки бедра.

Построить модель вероятности перелома с учётом имеющихся признаков, дать интерпретацию.

Данные: `GLOW500.txt`

1.5 Горишний Юрий

Данные собраны из переписи населения США 1990 года, отчёта ФБР о преступности за 1995 год и опроса сотрудников полиции LEMAS за 1990 год. По 2215 округам собрана статистика преступлений и 125 демографических показателей.

Построить функцию, оценивающую абсолютное число автомобильных краж по демографическим показателям, дать интерпретацию коэффициентов модели.

Данные: `crimes.xlsx`

1.6 Сафин Камиль

Полихлорированные дифенилы — органические соединения, активно использовавшиеся в промышленности до 1970 годов, когда была показана их токсичность. Накопление ПХБ в организме приводит к подавлению иммунитета, провоцирует развитие рака, поражений печени, почек, нервной системы, кожи, способствуют развитию детской патологии. Из-за накопления ПХБ в озёрах США некоторые виды рыб в некоторых областях запрещены к употреблению в пищу. Для своевременного обновления таких запретов необходимо периодически проводить мониторинг ПХБ. К сожалению, существует 209 различных разновидностей ПХБ, концентрация каждой из которых измеряется отдельным тестом. Для 69 видов рыбы известны концентрации семи соединений ПХБ (в миллионных долях), а также суммарная концентрация всех разновидностей ПХБ, их токсическая эквивалентность (TEQ) и суммарная токсическая эквивалентность образца, определяемая также вкладом диоксинов и фуранов.

Насколько точно токсичность рыбы можно предсказывать по концентрации только нескольких ПХБ? Концентрации какого минимального количества соединений ПХБ нужно измерить, чтобы достаточно точно предсказать суммарную токсичность, или хотя бы токсичность только совокупности ПХБ?

Данные: `pcb.txt`

1.7 Павлов Сергей

Для 649 учеников старших классов двух португальских школ известны ряд демографических показателей и показателей успеваемости; для каждого студента известны также уровень потребления алкоголя по выходным и будним дням в пятибалльной шкале от очень низкого до очень высокого и финальная оценка по португальскому языку.

Смоделировать финальную оценку как функцию от всех показателей, кроме итоговых оценок по промежуточным семестрам; оценить влияние уровня потребления алкоголя на неё.

Данные: `student-por.xlsx`

1.8 Самсонов Никита

Для 30000 клиентов тайваньского банка известны сумма кредита, демографические показатели и история платежей по кредитам за последние пять месяцев (факт просрочки, сумма неоходимой выплаты, сумма платежа).

Построить модель, предсказывающую вероятность просрочки следующего платежа, оценить вклад факторов.

Данные: `default.xls`

1.9 Сноровихина Виктория

Госпиталь города Карайкуди, Тамилнад, Индия, собрал данные анализов 250 пациентов с хронической болезнью почек и 150 пациентов без неё.

Построить диагностическую модель хронической болезни почек, оценить вклад факторов.

Данные: `chronic_kidney_disease.xlsx`

1.10 Володин Сергей

Мерой надёжности шарикоподшипников служит величина L_{10} — максимальное число оборотов, которое выдерживает 90% одинаковых подшипников. Имеются данные измерений надёжности по шарикоподшипникам

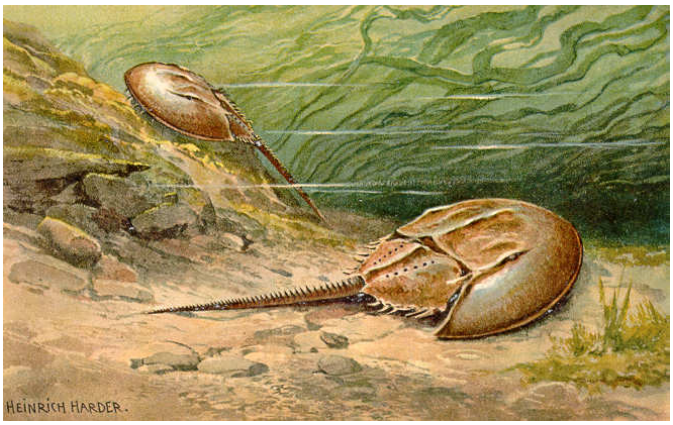

Рис. 1: Мечехвост

трёх производителей (для одного из производителей исследовано три вида подшипников), для каждого испытания указаны диаметр и число шаров в подшипнике, нагрузка и величина L_{10}.

Построить функцию, оценивающую L_{10} по имеющимся признакам, оценить вклад признаков.

Данные: `bearing.xlsx`

1.11 Катугина Татьяна

Имеются результаты обзвона 4119 клиентов португальского банка, которым предлагалось завести депозит. Известны социально-демографические характеристики клиентов, история предыдущих коммуникаций, социально-экономические показатели на момент совершения звонка.

Какие признаки определяют готовность клиента открыть депозит по результатам обзвона?

Данные: `deposit.xlsx`

1.12 Гончаренко Владислав

Изучалось влияние внешних характеристик самок морских ракообразных мечехвостов (Рис. 1) на их привлекательность для самцов. Выборка состоит из данных о наблюдениях над 173 особями и содержит закодированные данные о размере самок, их весе, цвете, состоянии панциря, а также о количестве спутников.

Построить функцию, по внешним параметрам самки предсказывающую количество спутников у самки, оценить значимость каждого фактора.

Данные: `horseshoe crab.txt`

1.13 Федоряка Дмитрий

Массовая доля жира, важная характеристика здоровья, рассчитывается через плотность тела, измеряемую при помощи взвешивания в воде. Для 252 мужчин проведены такие расчёты. Имеются также данные антропометрии (возраст, рост, вес, обхват грудной клетки и т.д.)

Построить функцию, оценивающую массовую долю жира по легко измеряемым антропометрическим признакам.

Данные: `fat.xls`

1.14 Фокина Дарья

Имеется 1066 наблюдений над различными участками поверхности Солнца. Известны: класс участка, размер максимального пятна на участке, распределение пятен, относительная активность, тип эволюции участка, код активности в предыдущие 24 часа, площадь участка. Известны также сложность участка в наблюдавшемся прошлом и при последнем повороте вокруг Солнца. Известно также число вспышек на каждом участке в течение 24 часов после начала наблюдения, причём вспышки разделены на три категории по мощности.

Построить модель, по свойствам участка предсказывающую суммарное число вспышек в последующие 24 часа, дать интерпретацию коэффициентов.

Данные: `solar flares.xls`

1.15 Водопьян Даниил

Для 60021 постов в блогах, опубликованных не более, чем за 72 часа до базового времени, собрана информация о количестве комментариев, времени публикации, длине и количестве каждого из 200 часто встречающихся слов.

Построить модель, предсказывающую количество новых комментариев за следующие 24 часа.

Данные: `blog_feedback.xlsx`

1.16 Колоскова Анастасия

Имеются результаты обработки 1147 изображений сетчаток. По изображениям рассчитаны значения 17 признаков; записаны также результаты предварительного скрининга на наличие диабетической ретинопатии и окончательный диагноз.

Построить модель, оценивающую вероятность наличия диабетической ретинопатии, дать интерпретацию коэффициентов.

Данные: `retinopathy.xlsx`

1.17 Парубченко Александр

Данные собраны из переписи населения США 1990 года, отчёта ФБР о преступности за 1995 год и опроса сотрудников полиции LEMAS за 1990 год. По 2215 округам собрана статистика преступлений и 125 демографических показателей.

Построить функцию, оценивающую число ненасильственных преступлений на сто тысяч населения по демографическим показателям, дать интерпретацию коэффициентов модели.

Данные: crimes.xlsx

1.18 Белозёрова Анастасия

Имеются данные по продажам 479 платьев на сайте aliexpress.com за полтора месяца осени 2013 года. Для каждого из платьев известны также стиль, ценовая категория, рейтинг, размер, сезон, ряд характеристик внешнего вида и индикатор участия в программе рекомендаций.

Оценить влияние рейтинга товаров на продажи с учётом остальных факторов.

Данные: `aliexpress_dress_data.csv`

1.19 Стрельцов Федор

Имеется выборка из 1009 детей, родившихся в Северной Каролине в 2004 году; известны пол ребёнка, вес при рождении, период вынашивания, возрастная группа матери, а также курила ли мать во время беременности и употребляла ли алкоголь

Как вес ребёнка зависит от курения и употребления алкоголя (после поправки на остальные признаки)?

Данные: `birthweight.csv`

1.20 Багаев Рамазан

Собраны данные по 1413 пациенткам клиник при университете Джона Хопкинса, проходившим с 2006 по 2008 вакцинацию против папилломавируса человека препаратом Гардасил. Рекомендуемый курс — три укола в течение года — был пройдён только 469 пациентками. Производитель препарата исследует, в каких демографических группах и каком способе получения вакцины проведение полного курса наиболее вероятно.

Построить модель вероятности прохождения полного курса вакцинации в течение года, оценить вклад факторов.

Данные: `gardasil.xls`

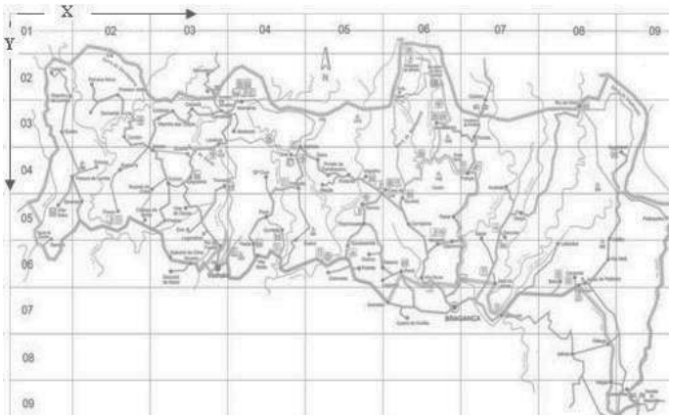

Рис. 2: Парк Монтезинью, разбиение на зоны.

1.21 Малюков Евгений

Имеются данные использования городского велопроката Вашингтона за каждый день 2011-2012 годов; известны также данные о погоде и ряд календарных признаков

Построить модель использования велопроката в зависимости от имеющихся признаков. Достаточно ли использовать дату с точностью до сезона, или месяц позволяет предсказывать значение признака значимо лучше? Есть ли смысл в использовании полной информации о днях недели, или достаточно разделять выходные и рабочие дни?

Данные: `bikeshares.xls`

1.22 Баяндина Анастасия

Данные собраны в 2001-2003 годах в португальском природном парке Монтезинью. Известны: месяц и день недели, температура воздуха, относительная влажность, скорость ветра, число выпавших осадков, значения четырёх метеорологических индексов, координаты зоны, в которой были произведены эти измерения (Рис. 2), а также площадь леса, уничтоженного произошедшим в этот день пожаром (если он был)."

Построить модель, позволяющую оценить по рассматриваемым признакам вероятность пожара.

Данные: `forest_fires.csv`

1.23 Чуйкова Екатерина

Имеется 1066 наблюдений над различными участками поверхности Солнца. Известны: класс участка, размер максимального пятна на участке, распределение пятен, относительная активность, тип эволюции участка, код активности в предыдущие 24 часа, площадь участка. Известны также сложность участка в наблюдавшемся прошлом и при последнем повороте вокруг Солнца. Известно также число вспышек на каждом участке в течение 24 часов после начала наблюдения, причём вспышки разделены на три категории по мощности.

Построить модель, по свойствам участка предсказывающую суммарную вероятность возникновения вспышек любого типа и доверительный интервал для неё.

Данные: `solar flares.xls`

1.24 Погодин Роман

Имеются данные измерений двухсот швейцарских тысячефранковых банкнот, бывших в обращении в первой половине XX века. Сто из банкнот были настоящими, и сто — поддельными. Измерены следующие величины: X_1 — длина банкноты, X_2 — ширина банкноты с левой стороны, X_3 — ширина банкноты с правой стороны, X_4 — расстояние от нижнего края до рамки рисунка, X_5 — расстояние от нижнего края до рамки рисунка, X_6 — длина диагонали рисунка. (Рис. 3)

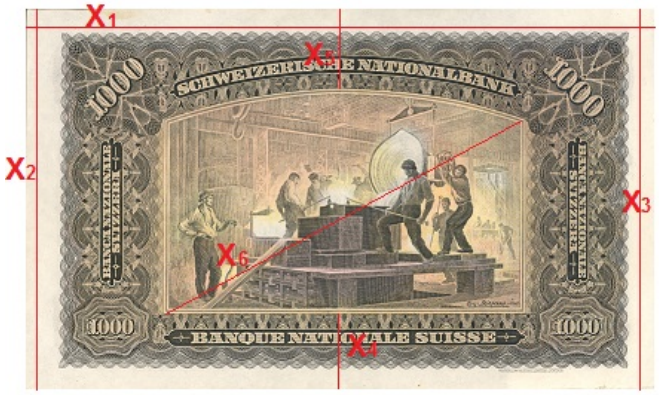

Рис. 3: Банкнота в 1000 швейцарских франков серии, действовавшей в период с 1911 по 1958. Красным обозначены измеренные величины.

Построить функцию, оценивающую по данным измерений вероятность того, что имеющаяся банкнота фальшивая; сравнить эффективность выявления фальшивых купюр по признакам $X_1 - X_3$ и по признакам $X_4 - X_6$, сделать выводы.

Данные: `banknotes.txt`

1.25 Писов Максим

Изучалось влияние внешних характеристик самок морских ракообразных мечехвостов (Рис. 1) на их привлекательность для самцов. Выборка состоит из данных о наблюдениях над 173 особями и содержит закодированные данные о размере самок, их весе, цвете, состоянии панциря, а также о количестве спутников.

Построить функцию, по внешним параметрам самки предсказывающую, будет ли у неё хотя бы один спутник. Оценить значимость каждого фактора.

Данные: `horseshoe crab.txt`

1.26 Чигринский Виктор

Собраны данные по 206 пациентам второго кардиологического отделения московской городской клинической больницы №25. Имеются результаты 14 анализов, а также 8 дополнительных признаков, описывающих пациента (пол, возраст, курение, наличие диабета и т.д.)

Построить функцию, оценивающую вероятность возникновения осложнений у пациента в результате тромболитической терапии по приведённым 22 признакам.

Данные: `cardio.xls`

1.27 Трахов Роман

Благотворительная организация разослала 4268 писем с предложением сделать пожертвование и получила отклик с пожертвованиями от 1707 адресатов. Для каждого адресата известны: индикатор ответа на предыдущее письмо, число недель, прошедших с момента предыдущего пожертвования, размеры текущего, предыдущего и среднего по всем предыдущим пожертвованиям в голландских гульденах, число писем, отправляемых адресату в год, доля писем, в ответ на которые приходят пожертвования.

Построить функцию, оценивающую вероятность получения пожертвования от адресата по историческим данным.

Данные: `charity.xlsx`

1.28 Владимирова Мария

Для 310 испытуемых измерены: наклон и смещение таза, угол изгиба поясницы, наклон плоскости тазовой поверхности крестца, радиус таза, степень смещения позвонков. Каждый из испытуемых либо здоров, либо

Рис. 5: Некоторые из измеренных характеристик скелета.

болен спондилолистезом или межпозвонковой грыжей.

Построить и интерпретировать модель, предсказывающую вероятность наличия заболевания позвоночника.

Данные: `spine.csv`

1.29 Иванов Алексей

Для 247 мужчин и 260 женщин измерены две группы антропометрических показателей — легко измеримые характеристики скелета и обхваты, всего 21 признак (Рис. 4). Указаны возраст, пол, вес и рост.

Построить функцию, оценивающую по наименьшему набору признаков вероятность того, что испытуемый — женщина, и доверительный интервал для этой вероятности.

Данные: `body.xlsx`

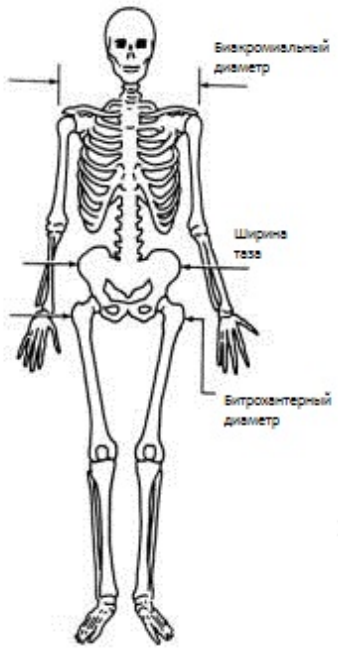

Рис. 4: Некоторые из измеренных характеристик скелета.

1.30 Шульгина Евгения

Для 247 мужчин и 260 женщин измерены две группы антропометрических показателей – легко измеримые характеристики скелета и обхваты, всего 21 признак (Рис. 4). Указаны возраст, пол, вес и рост.

Построить функцию, эффективно оценивающую вес по наименьшему набору признаков; сравнить точность оценки веса при отсутствии информации по обхватам и отсутствии информации по характеристикам скелета.

Данные: `body.xlsx`

1.31 Борзов Артём

357 испытуемым с доброкачественными и 212 со злокачественными опухолями груди была сделана тонкоигольная аспирационная пункция с гистологическим исследованием пунктата. По полученным изображениям (Рис. 5) определялись следующие признаки опухолевых клеток: радиус, однородность текстуры, периметр, площадь, гладкость, компактность, степень вогнутости, доля вогнутых участков контура, симметричность, фрактальная размерность. Для каждого изображения были рассчитаны среднее значение каждого из этих признаков, стандартное отклонение и среднее по трём клеткам с максимальным значением признака.

Оценить вероятность того, что опухоль злокачественная, по набору рассчитанных по изображению признаков. Построить функции, дающие точечную оценку и границы 95% доверительного интервала.

Данные: `breast cancer.xls`

1.32 Рязанов Андрей

Собраны данные по 206 пациентам второго кардиологического отделения московской городской клинической больницы №25. Имеются результаты 14 анализов, а также 8 дополнительных признаков, описывающих пациента (пол, возраст, курение, наличие диабета и т.д.)

Построить функцию, оценивающую вероятность выздоровления пациента в результате тромболитической терапии по приведённым признакам.

Данные: `cardio.xls`

1.33 Змеев Максим

Данные собраны из переписи населения США 1990 года, отчёта ФБР о преступности за 1995 год и опроса сотрудников полиции LEMAS за 1990 год. По 2215 округам собрана статистика преступлений и 125 демографических показателей.

Построить функцию, оценивающую число поджогов на сто тысяч населения по демографическим показателям, дать интерпретацию модели.

Данные: `crimes.xlsx`

1.34 Малькова Александра

Известна продажная стоимость 1198 домов, проданных в 2006-2010, и значения 71 признака, описывающего эти дома.

Построить модель оценки продажной стоимости дома по его признаковому описанию; дать интерпретацию воздействия наиболее сильных факторов

Данные: `houses.xls`

1.35 Рябых Алексей

Имеются помесячные данные о тратах на электроэнергию одного фиксированного домохозяйства на среднем западе США. За каждый месяц 1991-2000 годов приведены затраты на электроэнергию в долларах. Для объяснения колебаний размера счёта приведены следующие переменные: среднемесячная температура по данным последних тридцати лет, погодные индексы CDD и HDD (CDD - Cooling Degree Day - количество градусов, на которые средняя дневная температура больше 65F, взятое суммой за все дни месяца; HDD - Heating Degree Day - аналогично, суммарное количество градусов, на которое средняя дневная температура меньше 65F), число проживающих в доме членов семьи, индикатор установки нового счётчика, индикаторы установки двух новых тепловых насосов, объём потребления электроэнергии в киловатт-часах.

Оценить влияние установки нового оборудования на объём потребления и затраты на электроэнергию.

Данные: `electricity.xls`

1.36 Луканин Артём

Имеются данные измерений состояния атмосферы, произведённых в Нью-Йорке в течение 111 подряд идущих дней. Измерены температура воздуха, скорость ветра, уровень солнечной радиации и концентрация озона.

Построить функцию, по имеющимся признакам оценивающую наиболее вероятное значение концентрации озона и доверительный интервал для него.

Данные: `ozone.txt`

1.37 Копырин Денис

Для 103 образцов раствора бетона известно содержание в кубическом метре семи основных компонент, для каждого образца измерены также осадка, растекание и прочность на сжатие.

Построить функцию, оценивающую прочность бетона на сжатие по всем имеющимся характеристикам, оценить вклад растекания и осадки.

Данные: `concrete.xlsx`

1.38 Рындин Максим

Для 103 образцов раствора бетона известно содержание в кубическом метре семи основных компонент, для каждого образца измерены также осадка, растекание и прочность на сжатие.

Построить функцию, оценивающую растекание бетона по его составу.

Данные: `concrete.xlsx`

1.39 Малыгин Виталий

Исследование проводилось среди студентов психологического факультета крупного университета. Все испытуемые должны были быть правшами, а также не иметь повреждений мозга, эпилепсии, алкоголизма и сердечных заболеваний. Участники предварительного этапа эксперимента прошли несколько IQ-тестов, после чего для дальнейшего участия было отобрано 20 мужчин и 20 женщин, имевших коэффициент интеллекта либо ниже 103, либо выше 130 баллов. Для каждого из отобранных при помощи магнитно-резонансной томографии были получены 18 снимков срезов головного мозга, и общее количество пикселей на всех 18 снимках было принято в качестве меры объёма мозга. Помимо этого, были собраны данные о росте и массе тела испытуемых.

Проанализировать, какие из факторов значимо влияют на объём головного мозга; проверить, по какой из двух групп факторов можно предсказывать объём головного мозга с большей уверенностью — по результатам тестов интеллекта, или по полу, росту и весу.

Данные: `brain.xlsx`

1.40 Комаров Никита

Для 1599 образцов красного и 4898 белого португальского вина Vinho Verde известны оценки (от 0 до 10), выставленные дегустаторами при слепом тестировании, а также значения одиннадцати биохимических показателей, полученных при лабораторном анализе.

Построить модель, оценивающую содержания алкоголя по остальным характеристикам вина.

Данные: `wine.xlsx`

1.41 Григорук Василий

При подозрении на инфекционное заболевание для правильной постановки диагноза часто бывает важно из взятых у пациентов образцов вырастить как можно более многочисленную колонию бактерий, чтобы её было удобнее исследовать. Считается, что оптимальные параметры для размножения штаммов стафилококка в лабораторных условиях следующие: температура 35 градусов, концентрация триптона в питательном растворе 1.0%, время выдержки 24 часа. Для проверки оптимальности этих условий было проведено 30 экспериментов над пятью различными штаммами стафилококка. Для каждого из экспериментов известны время выдержки, температура, концентрация триптона, а также измеренное по окончании выдержки число колониеобразующих единиц (КОЕ) бактерий каждого штамма.

Построить функцию, предсказывающую итоговое суммарное число КОЕ бактерий всех пяти штаммов по времени выдержки, температуре и концентрации триптона в растворе, и определить по ней оптимальные условия размножения стафилококка.

Данные: `Staphylococcus aureus.txt`

1.42 Молибог Игорь

Для 1599 образцов красного и 4898 белого португальского вина Vinho Verde известны оценки (от 0 до 10), выставленные дегустаторами при слепом тестировании, а также значения одиннадцати биохимических показателей, полученных при лабораторном анализе.

Построить модель экспертной оценки по характеристикам вина, оценить влияние содержания алкоголя на экспертную оценку.

Данные: `wine.xlsx`

1.43 Зильберман Лев

Имеются данные о цене и свойствах 53940 бриллиантов. Известны: линейные размеры и признаки, построенные на их комбинациях, вес в каратах, цвет (закодирован буквами латинского алфавита: наиболее чистый цвет — буквой D, менее чистые — буквами E, F, G и т.д., чем ближе к концу алфавита, тем "грязнее"), группа чистоты (отсутствие дефектов, профессиональная оценка, выдаваемая специалистами при исследовании бриллианта в лупу десятикратного увеличения; бриллианты без трещин и включений получают оценку IF ("internally flawless"), далее в порядке убывания чистоты следуют группы VVS1 и VVS2 ("very very slightly imperfect"), VS1 и VS2 ("very slightly imperfect")), стоимость бриллианта в долларах США.

Существует общепринятая система классификации бриллиантов на мелкиenbsp;— до 0.29 карата, среднienbsp;— от 0.30 до 0.99 карата и крупныenbsp;— свыше C карата. Достаточно ли для предсказания цены знать о весе бриллианта только к какому классу он относится, или предсказания с использованием знаний о точном весе значимо лучше?

Данные: `diamonds.xlsx`

1.44 Малышев Иван

Для 398 автомобилей известен расход топлива по городу (в милях на галлон), а также их технические характеристики.

Построить модель расхода бензина в зависимости от характеристик автомобиля.

Данные: `mpg.csv`

1.45 Гарипов Ильяс

Для изучения влияния активности размножения самцов дрозофилы на продолжительность их жизни был организован следующий эксперимент. По 25 самцов в пяти группах содержались в одинаковых условиях, за исключением одного отличия: в первой группе к каждому самцу ежедневно подсаживалась готовая к размножению самка, во второй – восемь готовых к размножению самок, в третьей и четвёртой - соответственно, одна и восемь беременных самок, не готовых к размножению, наконец, к самцам пятой группы не подсаживали никого. Для каждого самца измерена продолжительность жизни, длина грудной клетки и доля времени, проводимого во сне.

Построить функцию, предсказывающую продолжительность жизни самца дрозофилы в зависимости от условий его содержания, дать интерпретацию вклада признаков.

Данные: `fly.txt`

1.46 Леонтьев Семён

Имеются данные о цене и потребительских качествах 308 бриллиантов, продававшихся в Сингапуре в 2000 году. Известны: вес бриллианта в каратах, цвет (закодирован буквами латинского алфавита: наиболее чистый цветnbsp;— буквой D, менее чистыenbsp;— буквами E, F, G и т.д., чем ближе к концу алфавита, тем "грязнее"), группа чистоты (отсутствие дефектов, профессиональная оценка, выдаваемая специалистами при исследовании бриллианта в лупу десятикратного увеличения; бриллианты без трещин и включений получают оценку IF ("internally flawless"), далее в порядке убывания чистоты следуют группы VVS1 и VVS2 ("very very slightly imperfect"), VS1 и VS2 ("very slightly imperfect")), название организации, выдавшей сертификат по группе чистоты (GIAnbsp;— Gemmological Institute of America, IGInbsp;— International Gemmological Institute, HRDnbsp;— Hoge Raad Voor Diamant), стоимость бриллианта в сингапурских долларах.

Построить модель ценообразования бриллиантов, учитывая все особенности имеющихся данных

Данные: `diamonds.xlsx`

1.47 Ратько Мария

Имеются данные о стоимости 804 подержанных автомобилей и их характеристиках: известны пробег, производитель, модель, вид модели, тип кузова, число цилиндров, объём двигателя, число дверей, а также наличие или отсутствие круиз контроля, продвинутой звуковой системы и кожаной обивки сидений.

Построить модель стоимости автомобиля по данному набору признаков.

Данные: `cars.xls`

1.48 Биктайров Роман

Массовая доля жира, важная характеристика здоровья, рассчитывается через плотность тела, измеряемую при помощи взвешивания в воде. Для 252 мужчин проведены такие расчёты. Имеются также данные антропометрии (возраст, рост, вес, обхват грудной клетки и т.д.)

Построить функцию, оценивающую индекс ожирения без использования данных взвешивания.

Данные: `fat.xls`

1.49 Филиппенко Константин

Благотворительная организация разослала 4268 писем с предложением сделать пожертвование и получила отклик с пожертвованиями от 1707 адресатов. Для каждого адресата известны: индикатор ответа на предыдущее письмо, число недель, прошедших с момента предыдущего пожертвования, размеры текущего, предыдущего и среднего по всем предыдущим пожертвованиям в голландских гульденах, число писем, отправляемых адресату в год, доля писем, в ответ на которые приходят пожертвования.

Построить функцию, оценивающую вероятный размер пожертвования от адресата по историческим данным.

Данные: `charity.xlsx`