ПСАД 2020. ВМК. Примеры исследовательских задач

Данные доступны по ссылкам:
https://yadi.sk/d/3-VrMcUn3GB4MN
https://yadi.sk/d/Y0F3614a3ExSYe

1. Афанасьев Кирилл

Качество воды в Миннесоте. Для 895 источников воды в Миннесоте известны водоносный горизонт, водоём, уровень и химические свойства воды (pH, щёлочность, содержание алюминия, мышьяка, хлора и свинца).

Задача. Сравнить свойства воды из разных водоёмов.

Данные. MnGroundwater.csv

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

Задача. Построить модель вероятности ядовитости гриба, оценить вклад факторов.

Данные. mushroom.csv

2. Голдобина Ольга

General Social Survey. General Social Survey - ежегодный социологический опрос нескольких тысяч граждан США. На сайте **https://gssdataexplorer.norc.org/** доступны все данные с 1972 по 2014 год (GSS.xls).

Задача. Для опрошенных **2014** года исследовать связь суммарного количества правильных ответов на 11 вопросов на знание базовых научных фактов с демографическими признаками (пол, возраст, раса, семейное положение, количество детей, образование, сексуальная ориентация, занятость, доход).

Данные. GSS.xls

Сейсмическая опасность. 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.

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

Данные. seismic.xlsx

3. Демин Георгий

Годовой заработок. Опрос US Bureau of Labor Statistics 2002 года содержит данные о годовом заработке 55729 участников; известны также их пол (1 = male, 2 = female), возраст, уровень образования (1 = no high school, 2 = some high school, 3 = high school diploma, 4 = some college, 5 = bachelor's degree, 6 = postgraduate degree) и тип работы (5 = private sector, 6 = government, 7 = self-employed).

Задача. Оценить влияние образования, пола и типа работы на годовой заработок.

Данные. workers.xls

Global Longitudinal Study of Osteoporosis in Women. 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.

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

Данные. GL0W500.txt

4. Еремеев Максим

Засеивание облаков и уровень осадков. Исследовалось воздействие засеивания облаков на обилие дождей. Измерения проводились в течение 108 периодов на пяти участках земли в Тасмании - участки обозначены в файле как западный, восточный, южный, северный и северо-восточный. В выборке содержатся данные об уровне осадков (в миллиметрах) на каждом из пяти участков, о времени года, к которому относится период, и о том, проводилось ли засеивание, проверить, как засеивание облаков повлияло на уровень осадков отдельно по каждому из пяти экспериментальных участков.

Задача. Проверить, одинаково ли проявляется эффект засеивания на каждом из них, или, возможно, он как-то зависит от исходного уровня осадков на участке?

Данные. cloudseeding.txt

Автомобильные кражи. 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.

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

Данные. crimes.xlsx

5. Коптелов Дмитрий

Допустимость наказаний. Известно мнение двенадцати родителей о допустимости наказания их детей по результатам оценки в психогенном эксперименте; допустимость выражается в баллах. Чем ниже балл, тем менее допустимым участник исследования считает наказание. Имеются результаты о наказании самим родителем, бабушкой и учителем ребёнка.

Задача. Как зависит оценка допустимости наказания от наказывающего?

Данные. punishment.txt

Успеваемость учеников. 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.

Задача. 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

6. Кормаков Георгий

Продажи платьев. 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.

Задача. Исследовать, как каждый из признаков по отдельности влияет на уровень продаж.

Данные. aliexpress _ dress _ data.csv

Оценка веса. Для 247 мужчин и 260 женщин измерены две группы антропометрических показателей - легко измеримые характеристики скелета и обхваты, всего 21 признак. Указаны возраст, пол, вес и рост.

Задача. 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

7. Королев Николай

Пожертвования на благотворительность. 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.

Задача. Какие признаки отличают людей, совершающих пожертвования?

Данные. charity.xlsx

Платеж по кредиту. Для 30000 клиентов тайваньского банка известны сумма кредита, демографические показатели и история платежей по кредитам за последние пять месяцев (факт

просрочки, сумма необходимой выплаты, сумма платежа).

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

Данные. default.xls

8. Крыжановская Светлана

Дома престарелых Нью-Мексико. Для 52 лицензированных домов престарелых Нью-Мексико известны: число коек, суммарное годовое число дней в стационаре и койко- дней (в сотнях), суммарные годовые расходы на уход за пациентами, зарплату медсестёр и инфраструктуру (в сотнях долларов).

Задача. Есть ли различия между сельскими и городскими домами престарелых? По каким признакам?

Данные. nursing_homes.txt

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

Задача. Построить диагностическую модель хронической болезни почек, оценить вклад факторов.

Данные. chronic_kidney_disease. xlsx

9. Лебедь Федор

Maryland's Pick-3 Lottery. Даны результаты розыгрыша лотереи Maryland's Pick-3 Lottery за 218 подряд идущих дней. Результатом является трёхзначное число.

Задача. Можно ли считать розыгрыш случайным?

Данные. lottery.txt

Открытие депозита. 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.

Задача. Какие признаки определяют готовность клиента открыть депозит по результатам обзвона?

Данные. deposit.xlsx

10. Лукьянов Павел

Задержка авиарейсов. Для 4029 рейсов, вылетавших из Нью-Йоркского аэропорта Ла-Гуардия, известны название авиакомпании-перевозчика, аэропорт назначения, диапазон планируемого времени вылета, день недели, месяц, продолжительность полёта и время задержки вылета.

Задача. Есть ли закономерности в задержках вылетов?

Данные. FlightDelays.csv

Мечехвосты. Изучалось влияние внешних характеристик самок морских ракообразных мечехвостов 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.

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

Данные. horseshoe_crab.txt

11. Находнов Максим

Интеллект и размер головного мозга. 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.

Задача. Исследовать взаимосвязи между коэффициентами интеллекта и биологическими характеристиками испытуемых (пол, рост, вес, объём мозга).

Данные. brain.xlsx

Оценка массовой доли жира. 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.)

Задача. Построить функцию, оценивающую массовую долю жира по легко измеряемым

антропометрическим признакам.

Данные. fat.xls

12. Петренко Дарья

Обучение родителей воспитанию детей. 975 родителей участвовало в программе обучения воспитанию. Было проведено три опроса, в ходе которых родители отвечали на вопрос: «За последние несколько недель обращались ли дети к вам с проблемой или вопросом, который их беспокоил?» Первый опрос был проведён до начала обучения, второй - сразу после, и третий - по прошествии 6-8 недель после окончания обучения. Известен также уровень образования родителя.

Задача. Стали ли родители больше общаться с детьми в результате обучения? Проанализировать с учётом уровня образования родителей.

Данные. education.txt

Вспышки на солнце. 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.

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

Данные. solar flares.xls

13. Попов Дмитрий

Продолжительность жизни больных онкологическими заболеваниями. Выборка состоит из 64 пациентов, у которых был диагностирован неизлечимый рак какого- либо органа. Всем им в качестве поддерживающей терапии был назначен к приёму витамин С (считалось, что он может способствовать выздоровлению раковых больных). Приведены данные об остаточной продолжительности жизни пациентов в днях.

Задача. Исследовать связь между остаточной продолжительностью жизни и типом рака.

Данные. cancer.txt

Количество комментариев. 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.

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

Данные. blog_feedback.xlsx

14. Саенко Иван

Линька крабов. У 472 самок крабов metacarcinus magister измерена ширина панциря до и после линьки. Измерения были получены двумя способами: 1) 12000 крабов измеряли, помечали сигнальными маячками и выпускали обратно в естественную среду перед периодом линьки, затем часть крабов за вознаграждение возвращалась в лабораторию рыбаками, выловившими их с помощью стандартных ловушек; 2) крабов, выловленных на суше во время спаривания непосредственно перед линькой, приносили в лабораторию, измеряли, затем, через несколько дней после линьки, измеряли снова. Для второй категории известен год вылова.

Задача. Исследовать различия между изменениями размеров панциря особей, линька которых проходила в лабораторных условиях и в естественных. Для последних оценить влияние года вылова.

Данные. crabs.csv

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

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

Данные. retinopathy.xlsx

15. Ташевцев Артем

Оптимальные условия размножения штаммов золотистого стафилококка. 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.

Задача. Оценить зависимость итогового числа КОЕ каждого штаммов стафилококка от внешних условий; одинакова ли эта зависимость?

Данные. Staphylococcus aureus.txt.

Преступления в США. 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.

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

Данные. crimes.xlsx

16. Устюжанин Александр

Белки в коре мозга мышей. В 1080 образцах коры мозга мышей измерен уровень экспрессии 77 белков. Часть образцов взята от трисомных мышей (лабораторная модель синдрома Дауна), часть — от здоровых; в эксперименте перед получением образцов некоторые мыши получали стимул к обучению, а некоторые — нет; наконец, части мышей вводился Мемантин, а части — физраствор. Цель эксперимента — проверить, восстанавливает ли Мемантин способность к обучению у трисомных мышей.

Задача. Отличается ли экспрессия белков у здоровых и трисомных мышей в каких- нибудь из экспериментальных подгрупп?

Данные. memantine.xls.

Рейтинг товаров и продажи. 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.

Задача. Оценить влияние рейтинга товаров на продажи с учётом остальных факторов.

Данные. aliexpress_dress_data.csv

17. Холодов Александр

Заживление ран. На 26 пациентах было испытано экспериментальное лекарство, способствующее заживлению ран; для сравнения ещё к 26 пациентам применялась стандартная терапия. Измерялась площадь раны до начала терапии, после курса лечения и на заключительном визите через длительное время после завершения лечения. Кроме того, приведена субъективная оценка изменения состояния раны пациентом и врачом.

Задача. Отличается ли эффективность экспериментального лекарства от эффективности стандартного?

Данные. wounds.csv

Вес новорожденных. 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.

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

Данные. birthweight.csv

18. Чернышёв Александр

Массовая доля жира в организме. Массовая доля жира, важная характеристика здоровья, рассчитывается через плотность тела, измеряемую при помощи взвешивания в воде. Для 252 мужчин проведены такие расчёты.

Задача. Имеются также данные антропометрии (возраст, рост, вес, обхват грудной клетки и т.д.) как связаны простые антропометрические показатели (возраст, рост, вес, ИМТ) с обхватами?

Данные. fat.xls

Вакцинация 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.

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

Данные. gardasil.xls