МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА
Институт информационных технологий
Кафедра прикладной математики

КУРСОВАЯ РАБОТА

по дисциплине Языки программирования для статистической обработки данных
(наименование дисциплины)

Тема курсовой работы: Факторный анализ заболевания раком лёгких

Студент группы: ИМБО-01-21 Малкина Влада Валерьевна _____
(учебная группа, фамилия, имя, отчество) (подпись студента)

Руководитель курсовой работы: Доцент, к. п. н. Митина Ольга Алексеевна _____
(должность, звание, ученая степень, фамилия, имя, отчество) (подпись руководителя)

Рецензент (при наличии)**:** _____ _____
(должность, звание, ученая степень, фамилия, имя, отчество) (подпись рецензента)

Работа предоставлена к защите до « 19 » июня 2023 г.

Допущен к защите до « 19 » июня 2023 г.

Москва 2023 г.

МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий
Кафедра прикладной математики

Утверждаю

Заведующий кафедрой: _____
(Подпись)
Дзержинский Роман Игоревич
(фамилия, имя, отчество)

ЗАДАНИЕ
на выполнение курсовой работы по дисциплине
«Языки программирования для статистической обработки данных»

Студент: Малкина Влада Валерьевна **Группа:** ИМБО-01-21
Тема: Факторный анализ заболевания раком лёгких

Исходные данные: данные о пациентах, заболевших или имевших подозрение на заболевание раком лёгких

Перечень вопросов, подлежащих разработке, и обязательного графического материала:

1. Выявление зависимостей, построение графиков для данных о заболевании раком лёгких
2. Анализ графического материала
3. Построение модели-классификатора для прогноза заболевания раком лёгких

Срок представления к защите курсовой работы: до « 19 » июня 2023 г.

Задание на курсовую работу выдал _____ (Митина Ольга Алексеевна)
(подпись руководителя) (фамилия, имя, отчество)

Задание на курсовую работу получил: до « 14 » 02 2023 г.

_____ (Малкина Влада Валерьевна)
(подпись студента) (фамилия, имя, отчество)

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Рак — вторая по распространенности причина смерти в России. В настоящее время с появлением новых взглядов на лечение, проблемы диагностики онкологических заболеваний приобретают все большее значение. Одновременно успешное развитие новых методов искусственного интеллекта, а точнее, одной из его составляющих, машинного обучения, в сочетании с повышением производительности средств вычислительной техники, привели к стремительному росту интереса к этой области со стороны ученых, инженеров и исследователей. Результатом такого интереса явилось большое количество новых разработок, связанных с созданием интеллектуальных систем диагностики (ИСД) онкологических заболеваний, ориентированных, прежде всего, на их раннее выявление.

Актуальность темы исследования обусловлена тем, что искусственный интеллект (ИИ) расширяет клинические возможности в диагностике и терапии злокачественных новообразований.

Цель данной курсовой работы — провести факторный анализ и построить модель, исследующую зависимость заболевания раком от различных факторов.

Задачи, решаемые в данной курсовой работе:

- проведение анализа данных обследования рака легких;
- проведение факторного анализа, выявление главных факторов;
- построение модели-классификатора для выявления рака у пациента;
- исследование данных с использованием языка для статистической обработки данных — R.

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Факторный анализ

Факторный анализ — методика комплексного и системного изучения и измерения воздействия факторов на величину результативного показателя.

Цели факторного анализа. К примеру, анализируя оценки, полученные по нескольким шкалам, исследователь отмечает, что они сходны между собой и имеют высокий коэффициент корреляции, в этом случае он может предположить, что существует некоторая латентная переменная, с помощью которой можно объяснить наблюдаемое сходство полученных оценок. Такую латентную переменную называют фактором, который влияет на многочисленные показатели других переменных, что приводит к возможности и необходимости отметить его как наиболее общий, более высокого порядка.

Так можно выделить две цели факторного анализа:

1. Определение взаимосвязей между переменными, их классификация, т. е. «объективная R-классификация».
2. Сокращение числа переменных.

Для выявления наиболее значимых факторов и, как следствие, факторной структуры, наиболее оправданно применять метод главных компонентов. Суть данного метода состоит в замене коррелированных компонентов некоррелированными факторами. Другой важной характеристикой метода является возможность ограничиться наиболее информативными главными компонентами и исключить остальные из анализа, что упрощает интерпретацию результатов. Достоинство данного метода также в том, что он — единственный математически обоснованный метод факторного анализа.

Существуют следующие типы факторного анализа:

- детерминированный (функциональный) — результативный показатель представлен в виде произведения, частного или алгебраической суммы факторов;
- стохастический (корреляционный) — связь между результативным и факторными показателями является неполной или вероятностной;
- прямой (дедуктивный) — от общего к частному;
- обратный (индуктивный) — от частного к общему;
- одноступенчатый и многоступенчатый;
- статический и динамический;
- ретроспективный и перспективный.

Также факторный анализ может быть разведочным — он осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках и конфирматорным, предназначенным для проверки гипотез о числе факторов и их нагрузках. Практическое выполнение факторного анализа начинается с проверки его условий.

Обязательные условия факторного анализа:

- все признаки должны быть количественными;
- число признаков должно быть в два раза больше числа переменных;
- выборка должна быть однородна;
- исходные переменные должны быть распределены симметрично;
- факторный анализ осуществляется по коррелирующим переменным. [1.1]

При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. После объединения коррелированность компонент внутри каждого фактора между собой будет выше, чем их

коррелированность с компонентами из других факторов. Эта процедура также позволяет выделить латентные переменные, что бывает особенно важно при анализе социальных представлений и ценностей.

Как правило, факторный анализ проводится в несколько этапов.

Этапы факторного анализа:

1. Отбор факторов.
2. Классификация и систематизация факторов.
3. Моделирование взаимосвязей между результативным и факторными показателями.
4. Расчет влияния факторов и оценка роли каждого из них в изменении величины результативного показателя.
5. Практическое использование факторной модели (подсчет резервов прироста результативного показателя).

В основе факторного анализа лежит гипотеза о том, что непосредственно наблюдаемые переменные, например, признаки поведения, являются лишь признаками некоторого ненаблюдаемого, латентного явления.

В процедуре факторного анализа отдельные эмпирические данные группируются таким образом, чтобы коррелировать с неким гипотетическим фактором, при этом они сами должны коррелировать друг с другом. На основе матрицы корреляций всех исходных переменных друг с другом происходит поиск такого фактора, при исключении влияния которого взаимокорреляции между исходными переменными, обусловленные только этой переменной, станут равными нулю.

Если это происходит, то полученный фактор считается способным полностью заменить все исходные переменные. Если этого не происходит, т. е. в матрице остаточных корреляции есть еще значения, не равные нулю, тогда выдвигается предположение, что существует еще один фактор, обусловливающий эти остаточные корреляции.

Таким образом, из матрицы корреляций исходных переменных всех исходных переменных последовательно вычитаются значения тех

корреляций, которые приписываются влиянию вновь выделенного фактора, и работа по выделению факторов продолжается уже на основе матрицы остаточных значений корреляций.

На каком этапе стоит остановить этот процесс, т. е. каким числом факторов ограничиться, определяется прежде всего возможностями хорошей содержательной интерпретации факторов. Получаемые коэффициенты корреляции переменной с факторами выступают факторными нагрузками, которые определяют сам фактор. Квадрат каждой факторной нагрузки — это та часть дисперсии, которая объясняется данным фактором.

Так, если переменная имеет нагрузку на фактор 0,83, то это означает, что приблизительно 68 % (0,832) его дисперсии отражается этим фактором. При конструировании личностных опросников факторный анализ используют прежде всего для определения его внутренней согласованности — в оптимальном варианте пункты, образующие одну шкалу, должны образовывать один фактор. Но факторный анализ часто используется как дополнение или даже замена для формулирования личностного конструкта самого опросника. Объединение различных, например поведенческих, признаков в единый фактор позволяет рассматривать предположение о том, что их объединяет какая-то внутренняя латентная черта.

Существует два основных метода факторного анализа: в одном из них выделяются коррелированные факторы, в другом — некоррелированные (ортогональное решение). По мнению Дж. П. Гилфорда, желательным является выделение некоррелированных факторов, когда появляются составные, но не связанные друг с другом факторы, в отличие от того, когда появляются простые, но связанные. Для получения ортогонального решения используется метод главных компонент. [1.2]

По характеру взаимосвязи между показателями различают методы детерминированного стохастического факторного анализа.

Детерминированный факторный анализ представляет собой методику исследования влияния факторов, связь которых с результативным

показателем носит функциональный характер, т. е. когда результативный показатель факторной модели представлен в виде произведения, частного или алгебраической суммы факторов.

Методы детерминированного факторного анализа:

- метод цепных подстановок;
- метод абсолютных разниц;
- метод относительных разниц;
- интегральный метод;
- метод логарифмирования.

Данный вид факторного анализа наиболее распространен, поскольку, будучи достаточно простым в применении (по сравнению со стохастическим анализом), позволяет осознать логику действия основных факторов развития предприятия, количественно оценить их влияние, понять, какие факторы, и в какой пропорции возможно и целесообразно изменить для повышения эффективности производства.

Стохастический анализ представляет собой методику исследования факторов, связь которых с результативным показателем в отличие от функциональной является неполной, вероятностной (корреляционной). Если при функциональной (полной) зависимости с изменением аргумента всегда происходит соответствующее изменение функции, то при корреляционной связи изменение аргумента может дать несколько значений прироста функции в зависимости от сочетания других факторов, определяющих данный показатель.

Методы стохастического факторного анализа:

- способ парной корреляции;
- множественный корреляционный анализ;
- матричные модели;
- математическое программирование;
- метод исследования операций;
- теория игр.

Необходимо также различать статический и динамический факторный анализ. Первый вид применяется при изучении влияния факторов на результативные показатели на соответствующую дату. Другой вид представляет собой методику исследования причинно-следственных связей в динамике.

И, наконец, факторный анализ может быть ретроспективным, который изучает причины прироста результативных показателей за прошлые периоды, и перспективным, который исследует поведение факторов и результативных показателей в перспективе. [1.3]

Основное требование к исходным данным для факторного анализа — это то, что они должны подчиняться допущению о многомерном нормальном распределении в совокупности. Для проверки этой гипотезы используют тест «сферичности» распределения данных Бартлетта, где оценивается предположение о диагональности матрицы корреляций.

Данный вид анализа позволяет исследователю решить две основные задачи: описать предмет измерения компактно и в то же время всесторонне. С помощью факторного анализа возможно выявление факторов, отвечающих за наличие линейных статистических связей корреляций между наблюдаемыми переменными. [1.4]

Факторный анализ часто используется для снижения размерности данных, чтобы найти небольшое число факторов, которые объясняют большую часть дисперсии, наблюденной для значительно большего числа явных переменных. Факторный анализ может также использоваться для формирования гипотез относительно механизмов причинных связей или с целью проверки переменных перед дальнейшим анализом (например, чтобы выявить коллинеарность перед проведением линейного регрессионного анализа). [1.5]

Таким образом, целью факторного анализа является выявление скрытых переменных или факторов, объясняющих структуру корреляций внутри набора наблюденных переменных, сокращение числа переменных на

основе их классификации и определения структуры взаимосвязей между ними. Благодаря сокращению числа переменных вместо исходного набора переменных появляется возможность анализировать данные по выделенным факторам, число которых значительно меньше исходного числа взаимосвязанных переменных.

1.2 Детерминированный анализ. Метод главных компонент

Детерминированный факторный анализ — это процесс выявления причинно-следственных связей между экономическими явлениями. Детерминированный факторный анализ носит функциональный характер. Данный метод анализа исследует только один показатель, при этом изучает влияние на него ряда факторов.

Детерминированный факторный анализ предполагает применение различных методов манипулирования действующими факторами. Как правило, он позволяет методом исключения оставлять один фактор и исследовать его влияние на функцию. Для этого специалисты могут использовать методы цепной подстановки, абсолютные и относительные разницы, индексный метод, метод долевого участия и другое.

Факторный анализ методом цепных подстановок представляет собой процесс анализа, в ходе которого определяется ряд промежуточных значении обобщающего показателя в ходе последовательной замены базовых значении (плановых) факторов на отчетные (фактические). [1.6]

Индексный метод применяется для того, чтобы определить, насколько фактический показатель соответствует запланированным значениям, и отслеживает показатель в динамике. Этот прием определяет фактическое значение того или иного индикатора в одном периоде по сравнению с его же уровнем в базисном периоде.

Факторный анализ способом абсолютных разниц схож с методом цепных подстановок и так же сводиться к исключению факторов. Отличием данного способа является, то, что в начале вычисляется динамика каждого фактора, а затем производится замена данных на эту динамику.

Метод относительных разниц применяется при детерминированном факторном анализе, чтобы оценить влияние конкретного фактора на прирост результативных показателей. Самым главным достоинством рассматриваемого метода является его простота.

В основе этого пропорционального деления факторного анализа лежит следующее: необходимо определить долю воздействия каждого параметра в общей структуре изменения результативного показателя, а затем умножить полученный результат на общий прирост конечного показателя.

Интегральный метод применяется для определения влияния факторов в мультипликативных, кратных и смешанных моделях кратно-аддитивного вида. [1.7]

Метод главных компонент — один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Изобретен К. Пирсоном в 1901 году. Применяется во многих областях, таких как распознавание образов, компьютерное зрение, сжатие данных и т.п. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных или к сингулярному разложению матрицы данных. Иногда метод главных компонент называют преобразованием Карунена-Лоэва или преобразованием Хотеллинга.

Цель метода главных компонент — извлечение из этих данных нужной информации. Что является информацией, зависит от сути решаемой задачи. Данные могут содержать нужную нам информацию, они даже могут быть избыточными. Однако, в некоторых случаях, информации в данных может не быть совсем.

Размерность данных — число образцов и переменных — имеет большое значение для успешной добычи информации. Лишних данных не бывает — лучше, когда их много, чем мало. На практике это означает, что если получен спектр какого–то образца, то не нужно выбрасывать все точки, кроме нескольких характерных длин волн, а использовать их все, или, по крайней мере, значительный кусок.

Данные всегда (или почти всегда) содержат в себе нежелательную составляющую, называемую шумом. Природа этого шума может быть различной, но, во многих случаях, шум — это та часть данных, которая не содержит искомой информации. Что считать шумом, а что — информацией, всегда решается с учетом поставленных целей и методов, используемых для ее достижения.

Шум и избыточность в данных обязательно проявляют себя через корреляционные связи между переменными. Погрешности в данных могут привести к появлению не систематических, а случайных связей между переменными. Понятие эффективного ранга и скрытых, латентных переменных, число которых равно этому рангу, является важнейшим понятием в анализе.[1.8]

Цель метода главных компонент — уменьшение количества измерений набора данных.

Современные наборы данных часто содержат очень большое количество переменных. Это затрудняет проверку каждой из переменных по отдельности из-за практического факта, что человеческий разум не может проанализировать данные в таком большом масштабе.

Когда набор данных содержит большое количество переменных, между этими переменными часто существует серьезное перекрытие.

Компоненты, найденные с помощью метода главных компонент, упорядочены от самого высокого содержания информации до самого низкого содержания информации.

Метод главных компонент — это статистический метод, который позволяет перегруппировать переменные в меньшее количество переменных, называемых компонентами. Эта перегруппировка выполняется на основе вариации, которая является общей для нескольких переменных.

Перегруппировка переменных происходит таким образом, чтобы первый (вновь созданный) компонент содержал максимум вариаций. Второй компонент содержит второе по величине количество вариаций и т. д. Последний компонент логически содержит наименьшее количество вариаций.

Благодаря такому порядку компонентов становится возможным сохранить только несколько вновь созданных компонентов, сохраняя при этом максимальное количество вариаций. Затем мы можем использовать компоненты, а не исходные переменные для исследования данных.

Первый этап этого метода — математическая модель — максимизация дисперсии новых компонентов.

Математическое определение проблемы этого метода состоит в том, чтобы найти линейную комбинацию исходных переменных с максимальной дисперсией. Это также можно описать как применение матричной декомпозиции к корреляционной матрице исходных переменных. Метод эффективен в поиске компонентов, которые максимизируют дисперсию. [1.9]

Анализ основных компонентов — это метод, который устраняет зависимость или избыточность в данных, удаляя те функции, которые содержат ту же информацию, что и другие атрибуты и производные компоненты независимы друг от друга.

Подход данного метода к сокращению ненужных функций, присутствующих в данных, заключается в создании или получении новых измерений (или также называемых компонентами). Эти компоненты представляют собой линейную комбинацию исходных переменных. Таким образом, метод преобразует большее количество коррелированных

переменных (т. е. разбивает данные) на меньший набор некоррелированных переменных.

Основным компонентом набора данных является направление с наибольшей дисперсией. Технически метод главных компонент делает это путем вращения осей каждой из переменных. Оси поворачиваются так, чтобы поглощать всю информацию или разброс, доступный в переменной.

Итак, теперь каждая из осей представляет собой новое измерение или главный компонент. Компонент определяется как направление набора данных, объясняющее наибольшую дисперсию, которая подразумевается собственным значением этого компонента. Вращение оси изображается графически.

Следующий этап метода главных компонент — вычисление ковариационной матрицы.

Цель этого шага — понять, как переменные набора входных данных отличаются от среднего по отношению друг к другу, или, другими словами, увидеть, есть ли какая-либо связь между ними. Потому что иногда переменные сильно коррелированы таким образом, что содержат избыточную информацию. Итак, чтобы идентифицировать эти корреляции, мы вычисляем ковариационную матрицу.

Метод главных компонент — это технология многомерного статистического анализа, используемая для сокращения размерности пространства признаков с минимальной потерей полезной информации.

С математической точки зрения метод главных компонент представляет собой ортогональное линейное преобразование, которое отображает данные из исходного пространство признаков в новое пространство меньшей размерности.

Смысл метода заключается в том, что с каждой главной компонентой связана определённая доля общей дисперсии исходного набора данных (её называют нагрузкой). В свою очередь, дисперсия, являющаяся мерой изменчивости данных, может отражать уровень их информативности.

Задача метода главных компонент заключается в том, чтобы построить новое пространство признаков меньшей размерности, дисперсия между осями которой будет перераспределена так, чтобы максимизировать дисперсию по каждой из них. Для этого выполняется последовательность следующих действий:

1. Вычисляется общая дисперсия исходного пространства признаков; это нельзя сделать простым суммированием дисперсий по каждой переменной, поскольку они, в большинстве случаев, не являются независимыми, поэтому суммировать нужно взаимные дисперсии переменных, которые определяются из ковариационной матрицы.
2. Вычисляются собственные векторы и собственные значения ковариационной матрицы, определяющие направления главных компонент и величину связанной с ними дисперсии.
3. Производится снижение размерности: диагональные элементы ковариационной матрицы показывают дисперсию по исходной системе координат, а её собственные значения — по новой, тогда разделив дисперсию, связанную с каждой главной компонентой на сумму дисперсий по всем компонентам, получаем долю дисперсии, связанную с каждой компонентой; после этого отбрасывается столько главных компонент, чтобы доля оставшихся составляла 80 %. [1.10]

Таким образом, детерминированный анализ представляет собой методику исследования влияния факторов, связь которых с результативным показателем носит функциональный характер, т.е. когда результативный показатель представлен в виде произведения, частного или алгебраической суммы факторов. Детерминированный факторный анализ и его методы позволяют оценивать влияние факторов на конечный результат.

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Факторный анализ обследования рака легких

Инструментом решения данной задачи является RStudio — свободная среда разработки программного обеспечения.

В курсовой работе будут рассмотрены данные обследования рака легких (Рисунок 2.1). Для выявления заболевания раком легких и прогнозирования вероятности его появления проведем факторный анализ данных пациентов и выявим факторы, оказывающие наибольшее влияние на заболевание раком легких.

Данные пациентов содержат следующие признаки (https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer):

1. Gender — пол пациента.
2. Age — возраст пациента.
3. Smoking — курение.
4. Yellow fingers — у пациента жёлтые пальцы.
5. Anxiety — наличие тревожности.
6. Peer_pressure — давление окружения.
7. Chronic Disease — наличие хронических заболеваний.
8. Fatigue — присутствие усталости.
9. Allergy — наличие аллергии.
10. Wheezing — свистящее дыхание.
11. Alcohol — употребляет алкоголь.
12. Coughing — присутствует кашель.
13. Shortness of Breath — одышка.
14. Swallowing Difficulty — затрудненное глотание.
15. Chest pain — боль в груди.
16. Lung Cancer — рак лёгких.

survey lung cancer

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO

Рисунок 2.1 — Данные исследования рака легких

В процессе проведения факторного анализа рассчитываются и анализируются следующие показатели:

- корреляционная матрица — матрица, включающая в себя все возможные коэффициенты корреляций r между анализируемыми переменными (Рисунок 2.2);

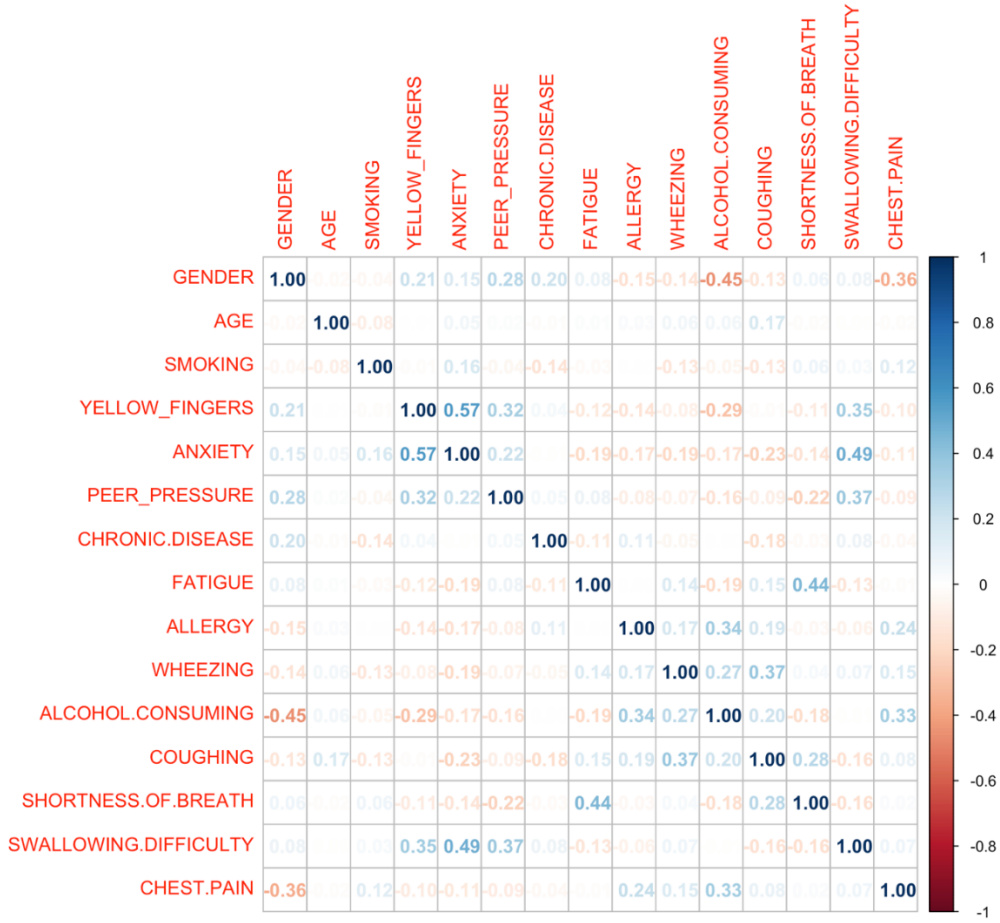

Рисунок 2.2 — Корреляционная матрица

Сильная положительная корреляция: когда значение одной переменной увеличивается, значение другой переменной увеличивается аналогичным образом. Наблюдается сильная положительная корреляция между желтыми пальцами и тревожностью, тревожностью и затрудненным глотанием, усталости и одышкой.

Сильная отрицательная корреляция: когда значение одной переменной увеличивается, значение другой переменной имеет тенденцию к уменьшению. Наблюдается сильная отрицательная корреляция между полом и употреблением алкоголя, полом и болью в груди

Слабая корреляция: между переменными очень слабая связь. Рассматривая остальные комбинации параметров, наблюдается слабая корреляция. [2.1]

- кмо — мера адекватности выборки Кайзера-Мейера-Олкина — величина, используемая для оценки применимости факторного анализа, значения от 0,5 до 1 говорят об адекватности факторного анализа, значения до 0,5 указывают на то, что факторный анализ неприменим к выборке (Рисунок 2.3);

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(X))
Overall MSA =  0.61
MSA for each item =
               GENDER                   AGE
                 0.71                  0.44
              SMOKING        YELLOW_FINGERS
                 0.48                  0.64
              ANXIETY         PEER_PRESSURE
                 0.63                  0.58
     CHRONIC.DISEASE               FATIGUE
                 0.42                  0.52
              ALLERGY              WHEEZING
                 0.74                  0.61
   ALCOHOL.CONSUMING              COUGHING
                 0.68                  0.55
  SHORTNESS.OF.BREATH SWALLOWING.DIFFICULTY
                 0.49                  0.63
          CHEST.PAIN
                 0.70
```

Рисунок 2.3 — Оценка применимости факторного анализа

Итоговая КМО равна 0.61, из чего делаем вывод об адекватности факторного анализа.

- критерий сферичности Бартлетта — показатель, с помощью которого проверяют, отличаются ли корреляции от 0; если г

близко к нулю, то выбранная переменная не взаимосвязана с другими, значимость меньше 0,05 указывает на то, что проведение факторного анализа приемлемо (Рисунок 2.4);

```
$p.value
[1] 2.330287e-125
```

Рисунок 2.4 — Критерий сферичности Бартлетта

Получили небольшое значение (меньше 0,05) уровня значимости, которое указывает на то, что факторный анализ может быть полезен для наших данных.

Вычислим детерминант (Рисунок 2.5).

```
> det(cor(X))
[1] 0.05189055
```

Рисунок 2.5 — Детерминант

Получили положительный определитель, следовательно, факторный анализ может быть проведен.

- графическое изображение критерия «каменистой осыпи» — график собственных значений факторов, расположенных в порядке убывания, используется для определения достаточного числа факторов (Рисунок 2.6). На рисунке по оси x отмечено количество факторов, по оси y указано начальное собственное значение [2.2]

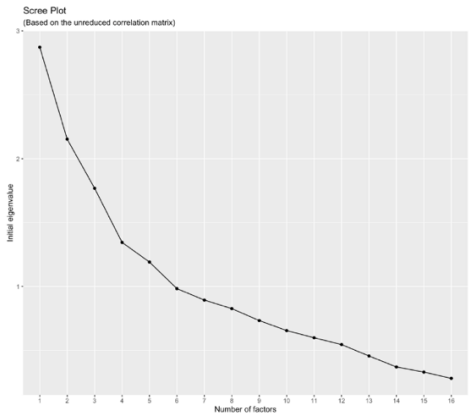

Рисунок 2.6 — Графическое изображение критерия «каменистой осыпи»

В результате получаем 6 главных компонент, у которых собственные значения больше 1.

Применим функцию parallel() для выполнения параллельного анализа (Рисунок 2.7). По оси X указано количество факторов, по оси Y указаны собственные значения компонент.

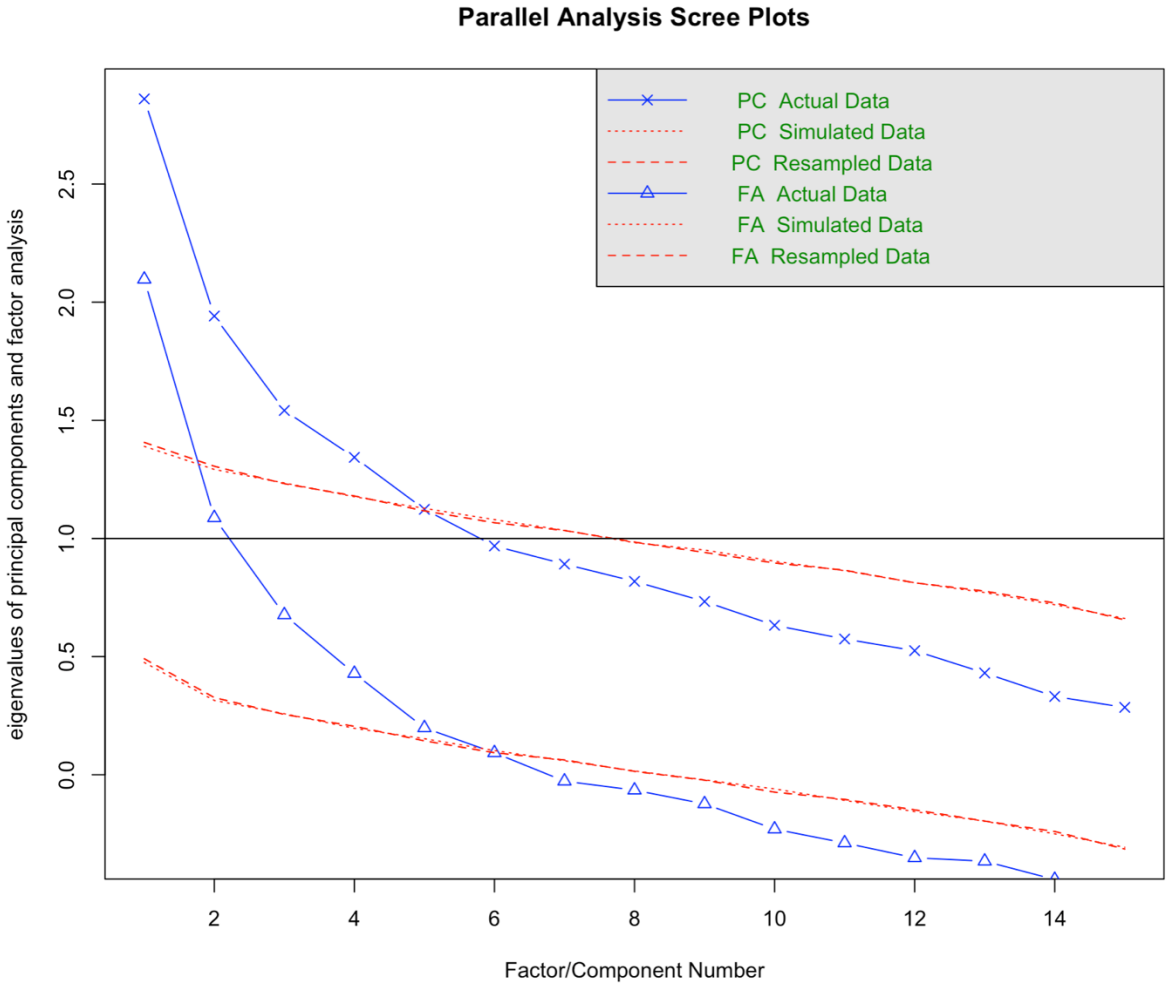

Рисунок 2.7 — Параллельный анализ

Параллельный анализ предполагает, что количество факторов равно 6.

Проведем факторный анализ, вызвав функцию factanal() (Рисунок 2.8).

Первый фрагмент представляет «Уникальность», которая варьируется от 0 до 1. Обратим внимание на высокие числа. Высокая уникальность переменной означает, что она не войдет в наши факторы. Если мы вычтем высокую уникальность из 1, мы получим количество, называемое общностью. Общность — это доля дисперсии i-й переменной, обусловленная

m общими факторами. Нам нужны факторы с низкой уникальностью или высокой общностью.

```
Call:
factanal(x = X, factors = 8)

Uniquenesses:
               GENDER                   AGE              SMOKING        YELLOW_FINGERS              ANXIETY         PEER_PRESSURE       CHRONIC.DISEASE
                0.659                 0.958                0.894                0.532                0.103                0.005                0.071
              FATIGUE              ALLERGY             WHEEZING     ALCOHOL.CONSUMING             COUGHING   SHORTNESS.OF.BREATH SWALLOWING.DIFFICULTY
                0.637                0.822                0.005                0.005                0.005                0.219                0.601
           CHEST.PAIN
                0.005

Loadings:
                    Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8
GENDER               0.122  -0.406           0.189  -0.261                    0.205
AGE                                                          0.193
SMOKING              0.151                                  0.118  -0.135  -0.120  -0.163
YELLOW_FINGERS       0.578  -0.279           0.143           0.157
ANXIETY              0.918          -0.112                           -0.140
PEER_PRESSURE        0.250  -0.121           0.956
CHRONIC.DISEASE                                                              0.958
FATIGUE             -0.143  -0.125   0.538   0.129                    0.119
ALLERGY             -0.117   0.298                   0.176   0.140           0.128
WHEEZING                     0.167                           0.192   0.957
ALCOHOL.CONSUMING            0.964  -0.163                   0.127
COUGHING            -0.170           0.226                   0.932   0.155
SHORTNESS.OF.BREATH                  0.852  -0.191
SWALLOWING.DIFFICULTY  0.559                 0.241                    0.106
CHEST.PAIN                   0.262                   0.959

                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8
SS loadings       1.668   1.409   1.150   1.090   1.069   1.043   1.026   1.025
Proportion Var    0.111   0.094   0.077   0.073   0.071   0.070   0.068   0.068
Cumulative Var    0.111   0.205   0.282   0.354   0.426   0.495   0.564   0.632

Test of the hypothesis that 8 factors are sufficient.
The chi square statistic is 14.04 on 13 degrees of freedom.
The p-value is 0.371
```

Рисунок 2.8 — Результаты применения функции factanal()

Следующий раздел — «Нагрузки», которые варьируются от -1 до 1. Нам нужны группы больших чисел. Для некоторых событий нет записей. Это потому, что R не печатает загрузки меньше 0,1.

Следующий раздел технически является частью загрузок. Цифры здесь суммируют факторы. Строка «Cumulative Var» говорит о кумулятивной пропорции объясненной дисперсии, поэтому эти числа варьируются от 0 до 1. [2.3]

В последнем разделе представлены результаты проверки гипотезы. Ноль этого теста в том, что для нашей модели достаточно 5 факторов. Низкое значение p-value заставляет нас отвергнуть гипотезу.

Проверим следующие гипотезы:

- для нашей модели достаточно 6 факторов (Рисунок 2.9);

```
The p-value is 0.000515
```

Рисунок 2.9 — Значение p-value для гипотезы с 6 факторами

- для нашей модели достаточно 7 факторов (Рисунок 2.10);

```
The p-value is 0.00566
```

Рисунок 2.10 — Значение p-value для гипотезы с 7 факторами

- для нашей модели достаточно 8 факторов (Рисунок 2.11).

```
The p-value is 0.371
```

Рисунок 2.11 — Значение p-value для гипотезы с 8 факторами

Высокое значение p-value заставляет нас принять последнюю гипотезу. [2.5]

Для визуализации результатов и отбора факторов построим тепловую карту (Рисунок 2.12).

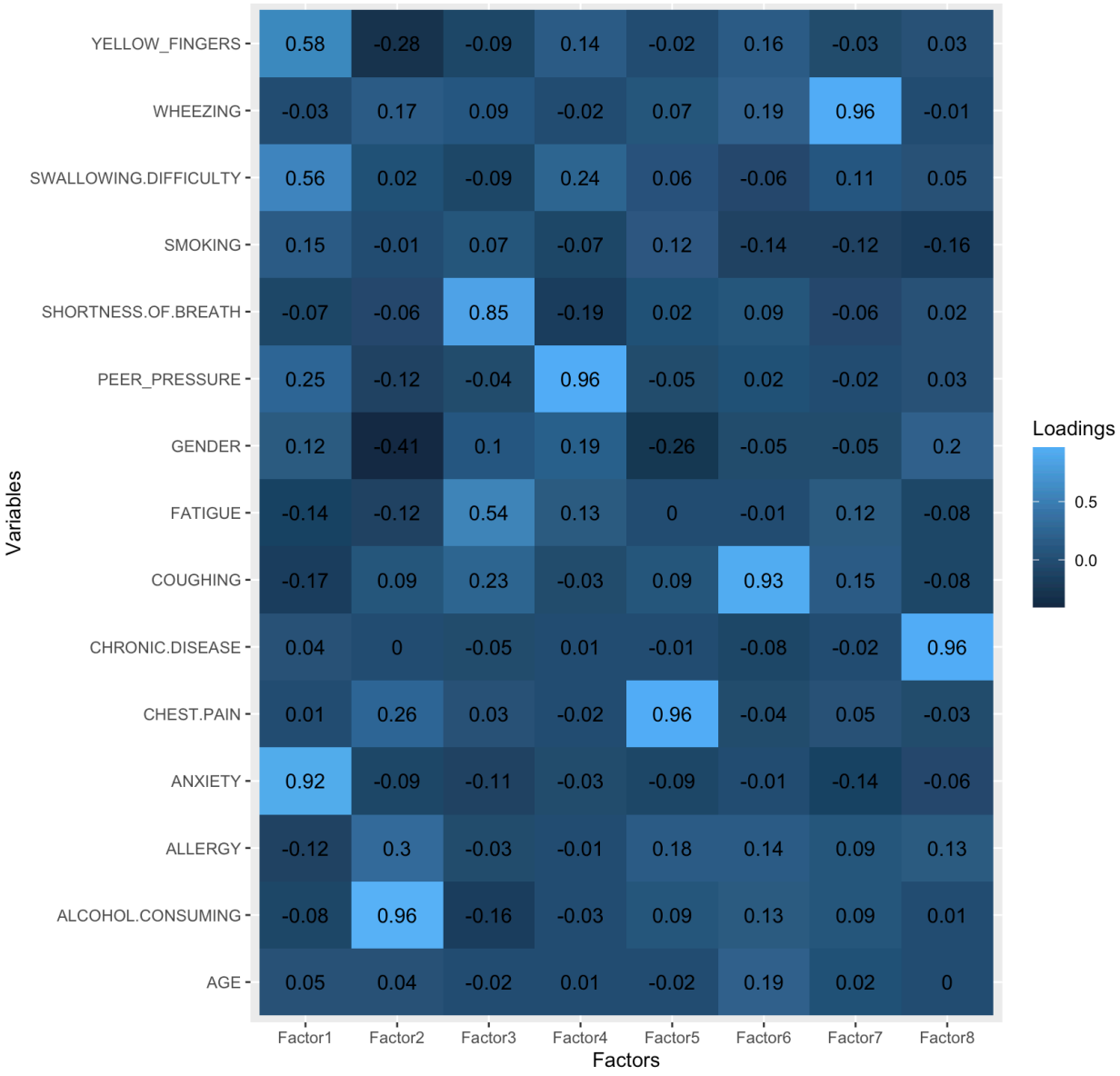

Рисунок 2.12 — Тепловая карта

Отберем факторы для обучения модели: хронические заболевания, тревожность, употребление алкоголя, одышка, давление окружения, боль в груди, кашель, свистящее дыхание.

В результате данные проверены на допустимость применения факторного анлиза. Построена матрица корреляции, проведены тесты Бартлетта и Кайзера-Мейера-Олкина. С помошью факторного анализа обнаружены факторы, оказывающие наибольшее влияние на прогнозируемый признак.

2.2 Построение байесовского классификатора

Классификация — это крупнейшая задача машинного обучения, которая ставит своей целью назначить метку класса наблюдениям из предметной области. [2.4]

Выделяют 4 основных типа классификации:

1. Мультиклассовая классификация.
2. Классификация по нескольким меткам.
3. Несбалансированная классификация.
4. Бинарная классификация.

Для наших данных используем бинарную классификацию. Задачи такого типа включают один класс, который является нормальным состоянием, и другой, который является ненормальным. Модель в этом случае предсказывает распределение вероятностей Бернулли для каждого примера. Популярные алгоритмы, которые можно использовать для двоичной классификации:

- логистическая регрессия;
- метод К-ближайших соседей;
- дерево решений;
- метод опорных векторов;

- наивный байесовский классификатор.

Байесовский классификатор — широкий класс алгоритмов классификации, основанный на принципе максимума апостериорной вероятности. Для классифицируемого объекта вычисляются функции правдоподобия каждого из классов, по ним вычисляются апостериорные вероятности классов.

Байесовский классификатор в машинном обучении — семейство простых вероятностных классификаторов, основанных на использовании теоремы Байеса и «наивном» предположении о независимости признаков классифицируемых объектов. [2.5]

Построим байесовский классификатор для исследования рака легких, предварительно разбив данные на обучающую и тестовую выборки.

Для оценки качества модели, вызовем функцию confusionMatrix (Рисунок 2.13).

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1  0  0
         2  8 23

               Accuracy : 0.7419
                 95% CI : (0.5539, 0.8814)
    No Information Rate : 0.7419
    P-Value [Acc > NIR] : 0.59359

                  Kappa : 0

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.7419
             Prevalence : 0.2581
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

Рисунок 2.13 — Результаты вызова функции confusionMatrix()

Проведя факторный анализ, мы выяснили какие факторы оказывают наибольшее влияние на развитие рака легких, а именно: хронические заболевания, тревожность, употребление алкоголя, одышка, давление окружения, боль в груди, кашель, свистящее дыхание. Построили модель-классификатор для прогнозирования рака легких у пациентов. Получили высокое значение accuracy (точность модели).

ЗАКЛЮЧЕНИЕ

Высокий уровень рецидивов и смертности от рака сочетается с длительным и дорогостоящим лечением. Снижение эффективности медицинской помощи на поздних стадиях заставляет исследователей и врачей искать новые способы раннего обнаружения опухолей.

Факторный анализ данных и машинное обучение повышают точность обнаружения и прогнозирования предрасположенности к онкологическим заболеваниям. Они необходимы для интерпретации медицинских исследований и являются важнейшим этапом изучения клинических, диагностических, лечебных и профилактических мероприятий.

Целью факторного анализа данных обследования рака легких является получение ключевых (наиболее информативных) факторов, дающих объективное и наиболее точное представление о влиянии факторов на заболевание раком. Более того, на основании результатов факторного анализа была построена модель прогнозирования, оценивающая вероятность заболевания раком легких.

Цель — провести факторный анализ и построить модель, исследующую зависимость заболевания раком от различных факторов — достигнута.

В ходе данной курсовой работы выполнены следующие задачи:

- проведен анализ данных обследования рака легких;
- проведен факторный анализ, выявлены главные факторы;
- построена модель-классификатор для выявления рака у пациента;
- исследованы данные с использованием языка для статистической обработки данных — R.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1. Макшанов А.В. Технологии интеллектуального анализа данных: учебное пособие / Макшанов А. В., Журавлев А. Е. — Москва: Издательство «Лань», 2019 — 212 с.

1.2. Овсянников Г. Н. Факторный анализ в доступном изложении: Изучение многопараметрических систем и процессов/ Овсянников Г.Н. — Москва: Издательство "Стереотип", 2022 — 176 с.

1.3. Гусева Е.Н. Теория вероятностей и математическая статистика: учебное пособие/ Гусева Е. Н. — Москва: Издательство «ФЛИНТА», 2021 — 220 с.

1.4. Гитис Л.Х. Факторный анализ/ Гитис Л.Х. — Москва: Издательство "Горная книга", 2017 — 22 с.

1.5. Гайдуков А.А. Методика оценки влияния структурного фактора в детерминированном факторном анализе с использованием интегрального метода/ Гайдуков А. А. — Минск: Вестник Белорусской государственной сельскохозяйственной академии, 2022 — 205 с.

1.6. Малахова О.А. Методы анализа: учебное пособие/ Малахова О.А., Зайцев В.В. — Самара: Самарский государственный аграрный университет, 2022 — 126 с.

1.7. Фарус О.А. Инструментальные методы анализа: учебно-методическое пособие/ Фарус О.А, Якушева Г.И. — Оренбург: Оренбургский государственный педагогический университет, 2021 — 114 с.

1.8. Иванов Б. Н. Теория вероятностей и математическая статистика: учебное пособие/ Иванов Б. Н. — Москва: Издательство «Лань», 2022 — 224 с.

1.9. Салкинд Н.Дж. Статистика для тех, кто (думает, что) ненавидит статистику/ Салкинд Н.Дж. — Москва: Издательство «ДМК Пресс», 2020 — 502 с.
1.10. Горелов Г. Н. Высшая математика. Правктикум для студентов технических и экономических специальностей: Учебное пособие для вузов/ Горелов Г. Н.. Горлач Б.А., Додонова Н.Л., Ефимов Е.А., Подклетнова С.В., Ростова Е.П. — Москва: Издательство «Лань», 2023 — 676 с.

ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1. РБК. Тренды [Электронный ресурс] / Как машинное обучение помогает в борьбе с онкологией. — Режим доступа: https://trends.rbc.ru/trends/industry/610032979a7947c814cd616c
2.2. ISA [Электронный ресурс] / Обзор методов машинного обучения в диагностике рака легкого. — Режим доступа: http://www.isa.ru/aidt/images/documents/2018-03/28_38.pdf
2.3. SBER MED AI [Электронный ресурс] / Как нейросети помогают обнаружить и лечить рак. — Режим доступа: https://sbermed.ai/ii-v-lechenii-raka/#yak1
2.4. НАФИ AI [Электронный ресурс] / Факторный анализ. — Режим доступа: https://nafi.ru/upload/spss/Lection_8.pdf
2.5. TWD [Электронный ресурс] / Exploratory Factor Analysis in R. — Режим доступа: https://towardsdatascience.com/exploratory-factor-analysis-in-r-e31b0015f224
2.6. Документация Loginom AI [Электронный ресурс] / Байесовский классификатор. — Режим доступа: https://wiki.loginom.ru/articles/bayesian_classifier.html
2.7. ИТМО [Электронный ресурс] / Метод главных компонент(PCA). — Режим доступа:

https://neerc.ifmo.ru/wiki/index.php?title=Метод_главных_компонент_(PCA)

2.8. Kaggle [Электронный ресурс] / Lung Cancer. — Режим доступа: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

2.9. Towards Data Science [Электронный ресурс] / Exploratory Factor Analysis in R. — Режим доступа: https://towardsdatascience.com/exploratory-factor-analysis-in-r-e31b0015f224

2.10. Data Flair [Электронный ресурс] / Principal Components and Factor Analysis in R – Functions & Methods. — Режим доступа: https://data-flair.training/blogs/principal-components-and-factor-analysis-in-r/

ПРИЛОЖЕНИЯ

Приложение А — Программный код

Приложение А

Код курсовой работы:

```
library(gmodels)
library(dplyr)
library(e1071)
library(psych)
library(corrplot)
library("psych")
library(ggplot2)
library(jsonlite)
library(data.table)
# Получение данных: считывание из файла
df <- read.csv("survey lung cancer.csv")
str(df)
# Замена значений в колонке GENDER
df$GENDER[df$GENDER == 'M'] <- '1'
df$GENDER[df$GENDER == 'F'] <- '2'
# Замена значений в колонке LUNG_CANCER
df$LUNG_CANCER[df$LUNG_CANCER == 'YES'] <- '2'
df$LUNG_CANCER[df$LUNG_CANCER == 'NO'] <- '1'
# Преобразование типов колонок
df$GENDER <- as.numeric(as.character(df$GENDER))
# Преобразование типов колонок
df$LUNG_CANCER <- as.numeric(as.character(df$LUNG_CANCER))
str(df)
# Проведем факторный анализ
# Описание данных
describe(df)
# Размерность датафрейма
dim(df)
# Построим корреляционную матрицу
datamatrix <- cor(df[,c(-16)])
dev.off()
corrplot(datamatrix, method="number", addCoef.col = 30)
```

```
# Факторизуемость данных
X <- df[,-c(16)]
Y <- df[,16]
# Кайзер-Мейер-Олкин (KMO)
KMO(r=cor(X))
# Тест Бартлетта на сферичность
cortest.bartlett(X)
det(cor(X))
# Количество факторов для извлечения
# Scree Pilot
fafitfree <- fa(df,nfactors = ncol(X), rotate = "none")
n_factors <- length(fafitfree$e.values)
scree     <- data.frame(
  Factor_n =  as.factor(1:n_factors),
  Eigenvalue = fafitfree$e.values)
ggplot(scree, aes(x = Factor_n, y = Eigenvalue, group = 1))
+
  geom_point() + geom_line() +
  xlab("Number of factors") +
  ylab("Initial eigenvalue") +
  labs( title = "Scree Plot",
        subtitle = "(Based on the unreduced correlation
matrix)")
# Проведем параллельный анализ
parallel <- fa.parallel(X)
# Факторный анализ с использованием метода factanal
factanal.none   <-   factanal(X,   factors=5,   scores   =
c("regression"), rotation = "varimax")
print(factanal.none)
factanal.none   <-   factanal(X,   factors=6,   scores   =
c("regression"), rotation = "varimax")
print(factanal.none)
factanal.none   <-   factanal(X,   factors=7,   scores   =
c("regression"), rotation = "varimax")
print(factanal.none)
```

```
    factanal.none <- factanal(X, factors=8, scores =
c("regression"), rotation = "varimax")
    print(factanal.none)
    # расчет главных компонент
    fit <- princomp(df)
    # Выводим результаты анализа
    summary(fit)
    # расчет главных компонент с собственными значениями более 1
    NumOfFactors <- sum(fit$sdev > 1)
    # вывод результата
    cat("Количество факторов, определенное методом Кайзера:",
NumOfFactors)
    # Применяем факторный анализ  с ограничением до 8 факторов
    fa_result <- factanal(df, factors = 8)
    # Выводим результаты анализа
    print(fa_result)
    #Получим вектор относительных весов
    loadings<-fa_result$loadings
    # Определяем данные
    dframe <- data.frame(var = rownames(loadings), factor =
rep(colnames(loadings), each = nrow(loadings)), value =
as.vector(loadings))
    # Строим тепловую карту
    ggplot(dframe, aes(x = factor, y = var, fill = value)) +
      geom_tile() +
      geom_text(aes(label=round(value, 2))) +
      labs(x = "Factors", y = "Variables", fill = "Loadings")
    fa_result$loadings
    factor1 = df$ALCOHOL.CONSUMING
    factor2 = df$SWALLOWING.DIFFICULTY
    factor3 = df$YELLOW_FINGERS
    factor4 = df$SHORTNESS.OF.BREATH
    factor5 = df$COUGHING
    factor6 = df$FATIGUE
    factor7 = df$CHRONIC.DISEASE
    factor8 = df$ANXIETY
```

```
new_df = data.frame(factor1, factor2, factor3, factor4,
factor5, factor6, factor7, factor8, df$LUNG_CANCER)
# Построение классификатора
# Вывод первых 6 строк датасета
head(new_df)
# Суммируем значения в датасете:
summary(new_df)
#Разделим данные на обучающую и тестовую выборки
index = sample(2,nrow(new_df),prob =
c(0.9,0.1),replace=TRUE)
set.seed(1234)
train = new_df[index==1,]
test = new_df[index==2,]
test_data = test[,1:5]
test_label = test[,6]
# Обучим модель:
train
model=naiveBayes(train$df.LUNG_CANCER~train$factor1+train$fa
ctor2+train$factor3+train$factor4+train$factor5+train$factor6+tr
ain$factor7+train$factor8,data = train, family = 'binomial')
model
# Получим предсказания модели:
test_result=predict(model,test_data)
test_result
# Оценка модели:
caret::confusionMatrix(factor(test_result),factor(test_label
))
```