Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

Лабораторная работа № 4 по дисциплине «Методы машинного обучения»

Тема Анализ социологического исследования

Студент Сапожков А.М,

Группа ИУ7-23М

Преподаватель Солодовников В.И.

Москва, 2025

Содержание

ВВЕДЕНИЕ

В современных социологических исследованиях часто возникает проблема неполноты данных, связанная как с отказом респондентов отвечать на отдельные вопросы, так и с техническими сложностями при сборе информации. Одновременно с этим, понимание взаимосвязей между различными социальными показателями представляет существенный интерес для исследователей. Методы машинного обучения предоставляют эффективный инструментарий для решения обеих задач: выявления корреляционных зависимостей между переменными и восстановления пропущенных значений на основе имеющихся данных.

Даны результаты опроса населения о его условиях существования. Переменные разбиты на 2 класса — «Признаки состояния» — это субъективная оценка населения своего бытия и «Признаки причины» — объектные количественные признаки оценивающие жизнедеятельность индивида и социума, в котором он проживает.

К признакам состояния относятся:

1. Оценка благополучия.
2. Оценка социальной поддержки.
3. Ожидаемая продолжительность здоровой жизни.
4. Свобода граждан самостоятельно принимать жизненно важные решения.
5. Индекс Щедрости.
6. Индекс отношения к коррупции.
7. Оценка риска безработицы.
8. Индекс кредитного оптимизма.
9. Индекс страха социальных конфликтов.
10. Индекс семьи.
11. Индекс продовольственной безопасности.
12. Чувство технологического прогресса.
13. Чувство неравенства доходов в обществе.

К индивидуальным признакам причины относятся:

1. Среднегодовой доход, тыс. $.
2. Объем потребленного алкоголя в год, л.
3. Количество членов семьи.
4. Количество лет образования.
5. Доля от дохода семьи, которая тратится на продовольствие, %.

К общественным признакам причины относятся:

1. Коэффициент Джини сообщества — показатель степени расслоения общества по какому-либо социальному признаку. Одними из ключевых признаков, по которым рассчитывается коэффициент Джини, является уровень доходов и активов домохозяйств. Показатель может варьироваться в диапазоне от 0 до 1, и чем больше его значение, тем большее расслоение общества он отражает.

2. Издержки сообщества на окружающую среду, млн. $.

3. Охват беспроводной связи в сообществе, %.

4. Количество смертей от вирусных и респираторных заболеваний в сообществе, тыс. человек.

5. Волатильность потребительских цен в сообществе.

6. Индивидуальные показатели характеризуют непосредственно индивида, общественные - сообщество в котором он проживает. В выборке могут присутствовать по несколько человек из одного сообщества. Все их общественные характеристики таким образом будут совпадать. В данных, относящихся к признакам состояния, присутствуют пропуски.

Целью данной лабораторной работы является освоение практических навыков корреляционного анализа.

Задачи данной лабораторной работы:

1) определить влияние признаков причины на признаки состояния;
2) выявить корреляционные зависимости;
3) заполнить пропуски в данных.

1 Аналитическая часть

1.1 Корреляция

Корреляция (от лат. correlatio «соотношение, взаимосвязь»), или корреляционная зависимость — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Значительная корреляция между двумя случайными величинами всегда является свидетельством существования некоторой статистической связи в данной выборке, но эта связь не обязательно должна наблюдаться для другой выборки и иметь причинно- следственный характер.

— Рассматривая пожары в конкретном городе, можно выявить весьма высокую корреляцию между ущербом, который нанёс пожар, и количеством пожарных, участвовавших в ликвидации пожара, причём эта корреляция будет положительной.

— Из этого не следует вывод «увеличение количества пожарных приводит к увеличению причинённого ущерба», и тем более не будет успешной попытка минимизировать ущерб от пожаров путём ликвидации пожарных бригад. Отсутствие корреляции между двумя величинами ещё не значит, что между ними нет никакой связи. Например, зависимость может иметь сложный нелинейный характер, который корреляция не выявляет.

Некоторые виды коэффициентов корреляции могут быть положительными или отрицательными. В первом случае предполагается, что мы можем определить только наличие или отсутствие связи, а во втором — также и её направление.

Если предполагается, что на значениях переменных задано отношение строгого порядка, то в этом случае:

— Отрицательная корреляция — корреляция, при которой увеличение одной переменной связано с уменьшением другой.

— Положительная корреляция в таких условиях — это такая связь, при которой увеличение одной переменной связано с увеличением другой переменной.

Возможна также ситуация отсутствия статистической взаимосвязи — например, для независимых случайных величин.

Ограничения корреляционного анализа:

— Применение возможно при наличии достаточного количества наблюдений для изучения.

— Коэффициент корреляции отражает «зашумлённость» линейной зависимости (верхняя строка)

— Коэффициент корреляции не описывает наклон линейной зависимости (средняя строка)

— Коэффициент корреляции совсем не подходит для описания сложных, нелинейных зависимостей (нижняя строка).

— Для распределения, показанного в центре рисунка, коэффициент корреляции не определен, так как дисперсия y равна нулю.

1.2 Регрессионный анализ

Набор методов моделирования измеряемых данных и исследования их свойств, относится к разделам математической статистики и машинного обучения. Осуществляет исследование влияния одной или нескольких независимых переменных $X_1, X_2, \ldots, X_p$ на зависимую переменную Y.

Независимые переменные называют регрессорами, предикторами или объясняющими переменными, а зависимая переменная является результирующей, критериальной или регрессантом. Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно- следственные отношения.

Наиболее распространенный вид регрессионного анализа — линейная регрессия, когда находят линейную функцию, которая, согласно определённым математическим критериям, наилучшим образом соответствует данным.

Регрессионный анализ очень тесно связан с корреляционным анализом. В корреляционном анализе исследуется направление и теснота связи между количественными переменными. В регрессионном анализе исследуется форма зависимости между количественными переменными.

Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

2 Технологическая часть

2.1 Средства реализации

В качестве языка программирования для реализации алгоритмов был выбран язык программирования Python ввиду наличия библиотек для обучения регрессионных моделей, таких как sklearn и numpy.

2.2 Реализация алгоритмов

На листинге 2.1 представлена реализация алгоритма анализа результатов социологического исследования.

Листинг 2.1 — Анализ результатов социологического исследования

```
mport numpy as np
import pandas as pd
import seaborn as sea
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, r2_score
import re
import math

pd.options.mode.copy_on_write = True

from google.colab import drive
drive.mount('/content/drive')

dataset = pd.read_excel('/content/drive/MyDrive/Colab Notebooks/
   ml_lab_04/ММО_ЛР4_Исходные
   данные.xlsx')
dataset

test_dataset = pd.read_excel('/content/drive/MyDrive/Colab Notebooks/
   ml_lab_04/ММО_ЛР4_Данные для
   проверки.xlsx')
test_dataset

test_dataset['Сообщество'] = test_dataset['Сообщество'].apply(lambda it:
    re.findall(r'\b\d+\b', it)[0])
test_dataset['Респондент'] = test_dataset['Респондент'].apply(lambda it:
    re.findall(r'\b\d+\b', it)[0])
```

```
dataset['Сообщество'] = dataset['Сообщество'].apply(lambda it: re.
    findall(r'\b\d+\b', it)[0])
dataset['Респондент'] = dataset['Респондент'].apply(lambda it: re.
    findall(r'\b\d+\b', it)[0])
dataset

unknowns_dataset = dataset[pd.isnull(dataset['Оценка.благополучия'])]
unknowns_dataset

knowns_dataset = dataset[pd.notnull(dataset['Оценка.благополучия'])]
knowns_dataset

unknowns_data = test_dataset[test_dataset.index.isin(unknowns_dataset
    .index)]
unknowns_data

plt.figure(figsize=(15,5))
src = [
  'Среднегодовой.доход,.тыс..$',
  'Объем.потребленного.алкоголя.в.год,.л.',
  'Количество.членов.семьи',
  'Количество.лет.образования',
  'Доля.от.дохода.семьи.которая.тратится.на.продовольствие,.%',
]
dst = [
  'Оценка.благополучия',
  'Оценка.социальной.поддержки',
  'Ожидаемая.продолжительность.здоровой.жизни',
  'Свобода.граждан.самостоятельно.принимать.жизненно.важные.решения',
  'Индекс.Щедрости',
  'Индекс.отношения.к.коррупции',
  'Оценка.риска.безработицы',
  'Индекс.кредитного.оптимизма',
  'Индекс.страха.социальных.конфликтов',
  'Индекс.семьи',
  'Индекс.продовольственной.безопасности',
  'Чувство.технологического.прогресса',
  'Чувство.неравенства.доходов.в.обществе',
]
correlation = knowns_dataset.corr().round(decimals=3).loc[src,dst].
    head(5)
```

```
#correlation
sea.heatmap(correlation, annot=True, linewidths=1)

plt.figure(figsize=(30,25))
correlation = knowns_dataset.corr().round(decimals=3)
#correlation
sea.heatmap(correlation, annot=True, linewidths=1)

X_train = knowns_dataset[src]
y_train = knowns_dataset[dst]
X_test = unknowns_data[src]
y_test = unknowns_data[[elem.replace('.', ' ') for elem in dst]]
model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
print(f'Средняя абсолютная ошибка: {mae:.2f}')

dataset[dst] = dataset[dst].fillna(pd.DataFrame(y_pred, columns=dst))
dataset.head(15)

errors, r2s, correlations = [], [], []

print('Средняя абсолютная ошибка определения признаков состояния / R2 / суммарная
   корреляция с признаками причины\n')
for state_feature in dst:
  X_train = knowns_dataset[src]
  y_train = knowns_dataset[state_feature]
  X_test = unknowns_data[src]
  y_test = unknowns_data[state_feature.replace('.', ' ')]
  model = LinearRegression()
  model.fit(X_train, y_train)
  y_pred = model.predict(X_test)
  mae = mean_absolute_error(y_test, y_pred)
  r2 = r2_score(y_test, y_pred) * 100
  corr = correlation[state_feature].abs().sum()
  print(f'{state_feature}: {mae:.2f} / {r2:.2f} / {corr:.2f}')
  errors.append(mae)
  r2s.append(r2)
  correlations.append(corr)
  dataset[state_feature] = dataset[state_feature].fillna(pd.Series(
```

```
      y_pred))

dataset.head(15)

plt.figure(figsize=(10,7))
plt.plot(errors, label='MAE')
plt.plot(r2s, label='R2 (x100)')
plt.plot(correlations, label='Abs. correlation sum')
plt.legend(loc='upper right')
plt.show()

src = [
  'Среднегодовой.доход,.тыс..$',
  'Количество.членов.семьи',
  'Доля.от.дохода.семьи.которая.тратится.на.продовольствие,.%',
]
dst = 'Индекс.отношения.к.коррупции'
X_train = knowns_dataset[src]
y_train = knowns_dataset[dst]
X_test = unknowns_data[src]
y_test = unknowns_data[dst.replace('.', ' ')]
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
print(f'Средняя абсолютная ошибка: {mae:.2f}')
dataset[dst].head(15)

src = [
  'Среднегодовой.доход,.тыс..$',
  'Объем.потребленного.алкоголя.в.год,.л.',
  'Количество.лет.образования',
  'Доля.от.дохода.семьи.которая.тратится.на.продовольствие,.%',
]
dst = 'Индекс.семьи'

X_train = knowns_dataset[src]
y_train = knowns_dataset[dst]
X_test = unknowns_data[src]
y_test = unknowns_data[dst.replace('.', ' ')]
model = LinearRegression()
```

```
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
print(f'Средняя абсолютная ошибка: {mae:.2f}')

dataset[dst].head(15)
```

3 Исследовательская часть

3.1 Среда для тестирования

Для тестирования разработанного алгоритма применялась облачная платформа Google Colab, не требующая установки ПО на локальный компьютер.

3.2 Анализ результатов социологического исследования

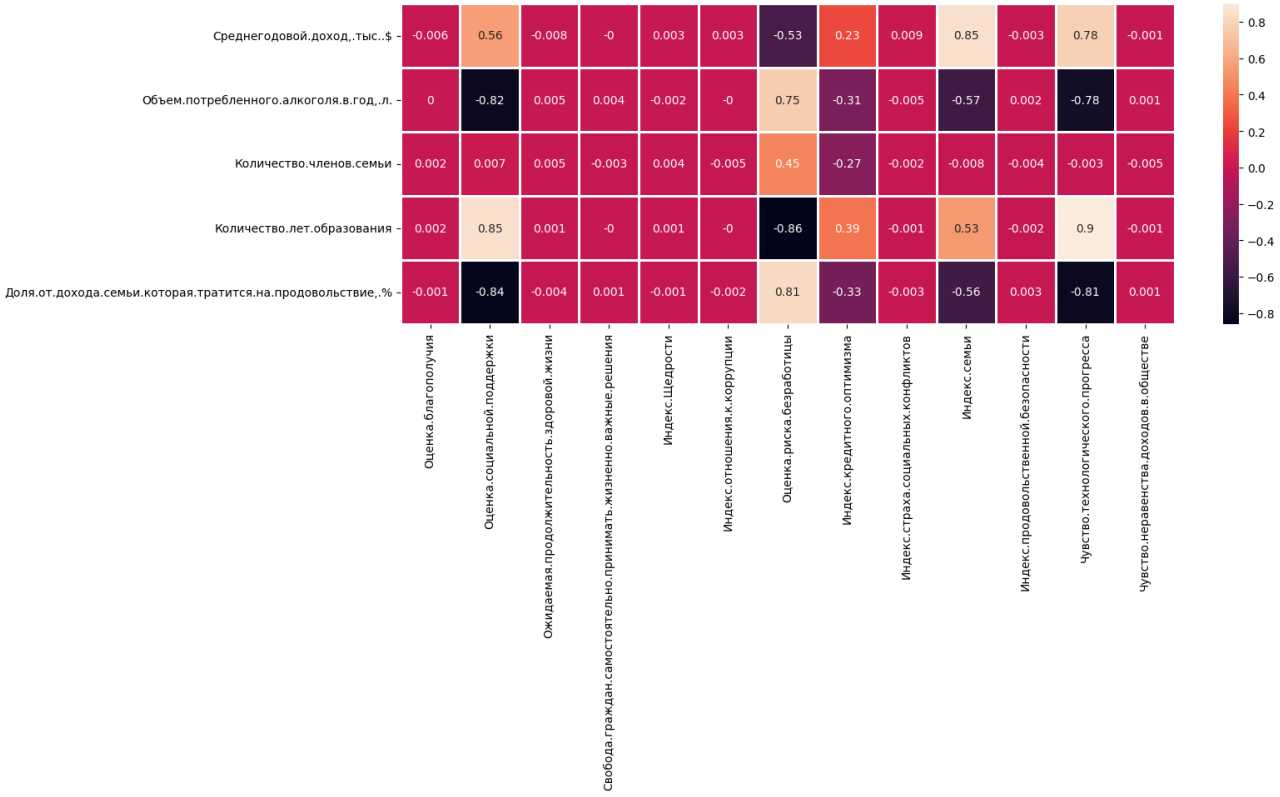

Рисунок 3.1 — Корреляция признаков причины и признаков состояния

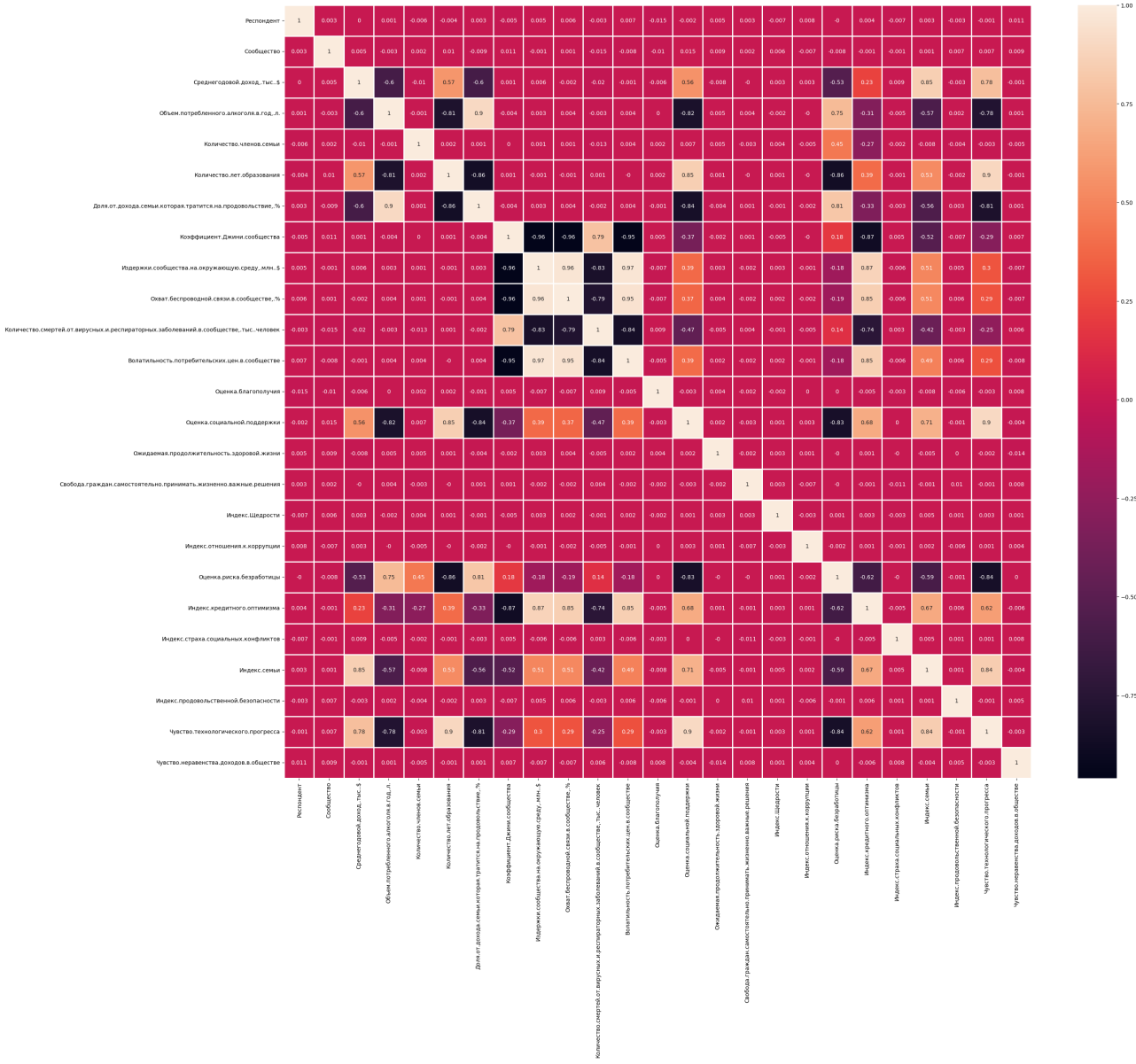

Рисунок 3.2 — Корреляционные зависимости между всеми признаками

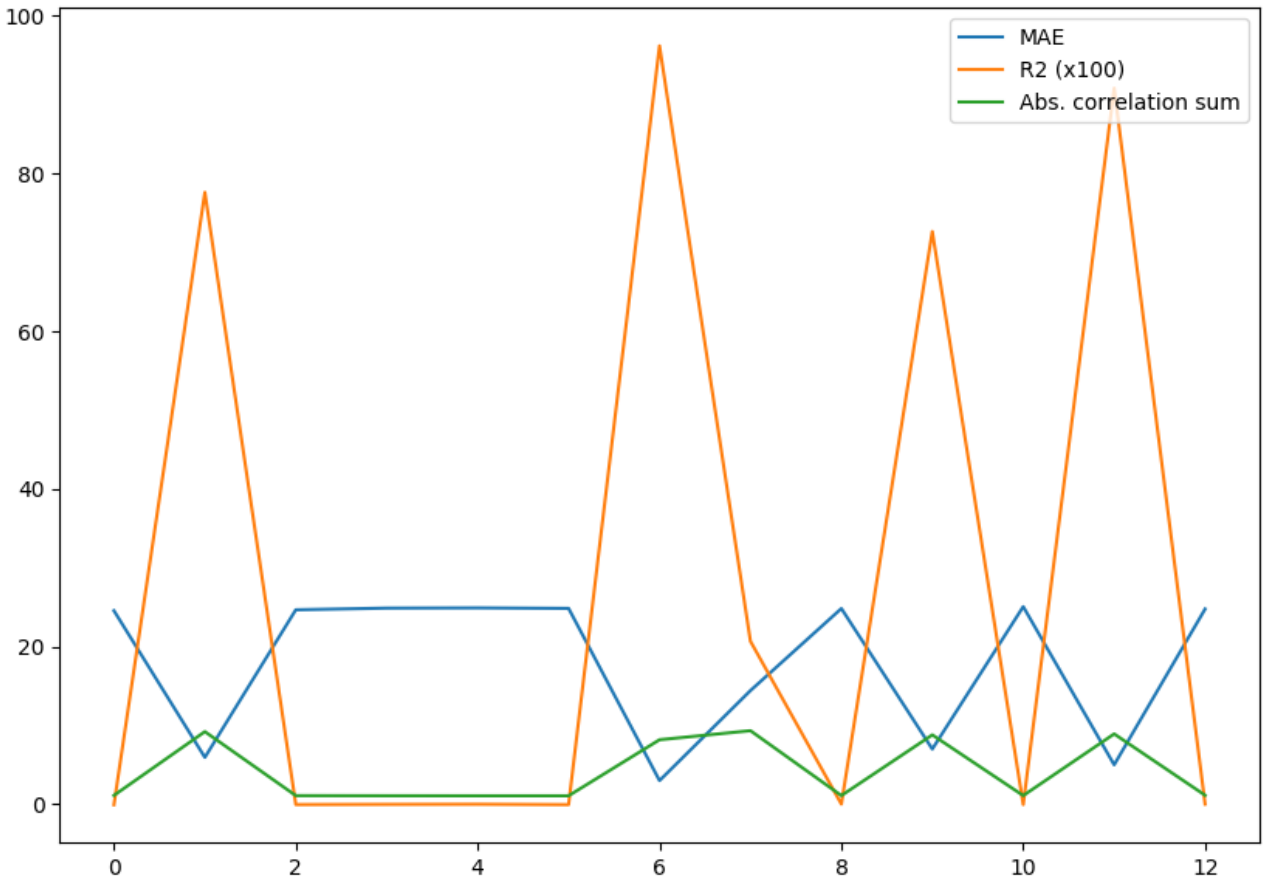

Рисунок 3.3 — Зависимость точности заполнения пропущенных значений признаков состояния от их корреляции с признаками причины

ЗАКЛЮЧЕНИЕ

В рамках лабораторной работы были освоены практические навыки корреляционного анализа на примере анализа результатов социологического исследования.

1. Определено влияние признаков причины на признаки состояния.
2. Выявлены корреляционные зависимости.
3. Заполнены пропуски в данных.

Результаты обучения регрессионных моделей для заполнения пропусков в данных показали, что зависимость точности обучения от степени корреляции целевой переменной с входными параметрами является обратной.