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»

Отчёт по лабораторной работе №7 по дисциплине "Методы машинного обучения"

Тема Анализ социологического исследования

Студент Варламова Е. А.

Группа ИУ7-23М

Оценка (баллы) _____

Преподаватели Солодовников Владимир Игоревич

Москва — 2024 г.

СОДЕРЖАНИЕ

1 | Теоретическая часть

Исследование и прогнозирование общей удовлетворенности жизнью людей является важной задачей, поскольку уровень удовлетворенности жизнью напрямую влияет на качество жизни и благополучие общества в целом. Понимание факторов, влияющих на удовлетворенность жизнью, позволяет разрабатывать эффективные стратегии и политики для улучшения качества жизни населения. Кроме того, прогнозирование уровня удовлетворенности жизнью помогает предсказывать возможные изменения и риски, связанные с социально-экономическими и политическими процессами. Таким образом, изучение общей удовлетворенности жизнью является важным инструментом для создания устойчивого и процветающего общества.

Целью данной лабораторной работы является предсказание интегральной оценки счастья с помощью значений признаков-причин, влияюших на состосние человека, которые [состояния] в свою очерель влияют на интегральную оценку счастья. Для этого необходимо решить следующие задачи:

— описать набор данных и предметную область;

— привести постановку задачи;

— проанализировать существующие модели регрессии и классификации для решения задачи;

— разработать метод предсказания оценки счастья;

— оценить точность, полноту, F-меру полученного классификатора; построить матрицы ошибок.

1.1 Описание набора данных и предметной области

В Файле данные содержатся результат 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

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. Также в данных присутствует интегральная характеристика удовлетворенности человека жизнью – для ее описания используется шкала Кантрила (рисунок 1.1).

Рис. 1.1: Шкала Кантрила

Шкала Кантрила (The Cantril Scale) – простая визуальная шкала, которая позволяет оценить общую удовлетворенность жизнью.

— Prospering – Процветающий/благоденствующий
— Thriving – Преуспевающий
— Blooming – Расцветающий
— Doing well – Дела идут хорошо
— Just Ok – Просто нормально
— Coping – Справляющийся

— Struggling – Столкнувшийся с трудностями/борющийся

— Suffering – Страдающий

— Depressed – Депрессивный

— Hopeless – Безнадежный

1.2 Постановка задачи

— Определить какие из признаков состояния наиболее сильно связаны с интегральной оценкой счастья (благополучия) респондента.

— Определить, как влияют признаки причины на наиболее важные признаки состояния.

— Пользуясь найденными закономерностями спрогнозировать попадание респондентов, у которых интегральная характеристика отмечена как «Неизвестно», в укрупненные группы шкалы Кантрила.

1.3 Анализ существующих моделей регрессии и классификации

1.3.1 Линейная модель

Линейная модель может быть описана следующей формулой:

$$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \tag{1.1}$$

где $\theta_0, \theta_1, \ldots, \theta_n$ – параметры модели, которые необходимо настроить по обучающим данным.

1.3.2 Случайный лес

Случайный лес (Random Forest) – это ансамблевый метод машинного обучения, основанный на построении множества деревьев решений в процессе обучения. Каждое дерево строится независимо и случайным образом, а итоговое предсказание получается путем усреднения предсказаний всех деревьев.

Модель случайного леса может быть описана следующим образом:

1. Для построения случайного леса необходимо определить количество деревьев T и размер подвыборки признаков m.

2. Для каждого дерева $t = 1, 2, ..., T$ строится дерево решений на основе случайной подвыборки данных размера m. При построении каждого узла дерева выбирается случайное подмножество признаков размером m, и разбиение узла происходит наилучшим образом по одному из этих признаков.

3. После построения всех деревьев случайного леса, для предсказания нового объекта x происходит усреднение предсказаний всех деревьев:

$$y'(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{1.2}$$

где $f_t(x)$ - предсказание отдельного дерева.

Таким образом, случайный лес комбинирует предсказания множества деревьев, что позволяет улучшить качество предсказаний и снизить переобучение.

1.3.3 Метод k ближайших соседей

сновная идея метода заключается в том, что предсказание для нового объекта делается на основе значений целевой переменной ближайших к нему соседей.

Модель k-NN может быть описана следующим образом:

— Основным параметром модели k-NN является число соседей k, которые будут использоваться для предсказания.

— Предсказание для нового объекта x_new вычисляется путем усреднения (или взвешенного усреднения) значений целевой переменной ближайших к нему k соседей.

— Предсказание для нового объекта x_new вычисляется путем усреднения (или взвешенного усреднения) значений целевой переменной ближайших к нему k соседей.

$$y'\_new = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{1.3}$$

где:

- y'_new - предсказанное значение целевой переменной для нового объекта;
- y_i - значение целевой переменной i-го ближайшего соседа объекта x_new,
- k - количество соседей, используемых для предсказания.

— В случае взвешенного усреднения можно использовать веса, зависящие от расстояния между новым объектом и его соседями.

$$y'\_new = \sum_{i=1}^{k} w_i \frac{y_i}{\sum_{i=1}^{k} w_i} \tag{1.4}$$

где:

— w_i - вес, присвоенный i-му соседу на основе расстояния до нового объекта.

Таким образом, модель k-NN для задачи классификации (регрессии) предсказывает класс (значение целевой переменной) для нового объекта на основе классов (значений целевой переменной) ближайших к нему соседей, используя усреднение или взвешенное усреднение.

2 | Практическая часть

2.1 Выбор средств разработки

В качестве языка программирования был использован язык Python, поскольку этот язык кроссплатформенный и для него разработано огромное количество библиотек и модулей, решающих разнообразные задачи.

В частности, имеются библиотеки, включающие в себя модели регрессии и классификации в библиотеке [1].

2.2 Исследование ПО

2.3 Определение состояний, влияющих на счастье

Необходимо определить, какие из признаков состояния наиболее сильно связаны с интегральной оценкой счастья (благополучия) респондента.

Для этого на обучающих данных (данных, в которых оценка счастья известна) были построены модели классификации (линейная, k-ближайших соседей и случайный лес). Оценена их точность:

Таблица 2.1: Точность моделей классификации

Алгоритм классификации	Целевая переменная	Точность (accuracy)	Точность (precision)	Полнота	f-мера
линейная	Ощущаемое счастье	0.876	0.877	0.876	0.877
k-соседей	Ощущаемое счастье	0.650	0.663	0.650	0.650
случайный лес	Ощущаемое счастье	0.962	0.962	0.962	0.962

Видно, что лучшая модель по точности: случайный лес. Для неё были определены состояния, наиболее связанные с интегральной оценкой счастья.

Выбранные состояния: 'Оценка социальной поддержки', 'Оценка риска безработицы', 'Индекс кредитного оптимизма', 'Индекс семьи', 'Чувство технологического прогресса'.

Визуализация представлена на рисунке 2.1.

Доля выбранных состояний: 0.5.

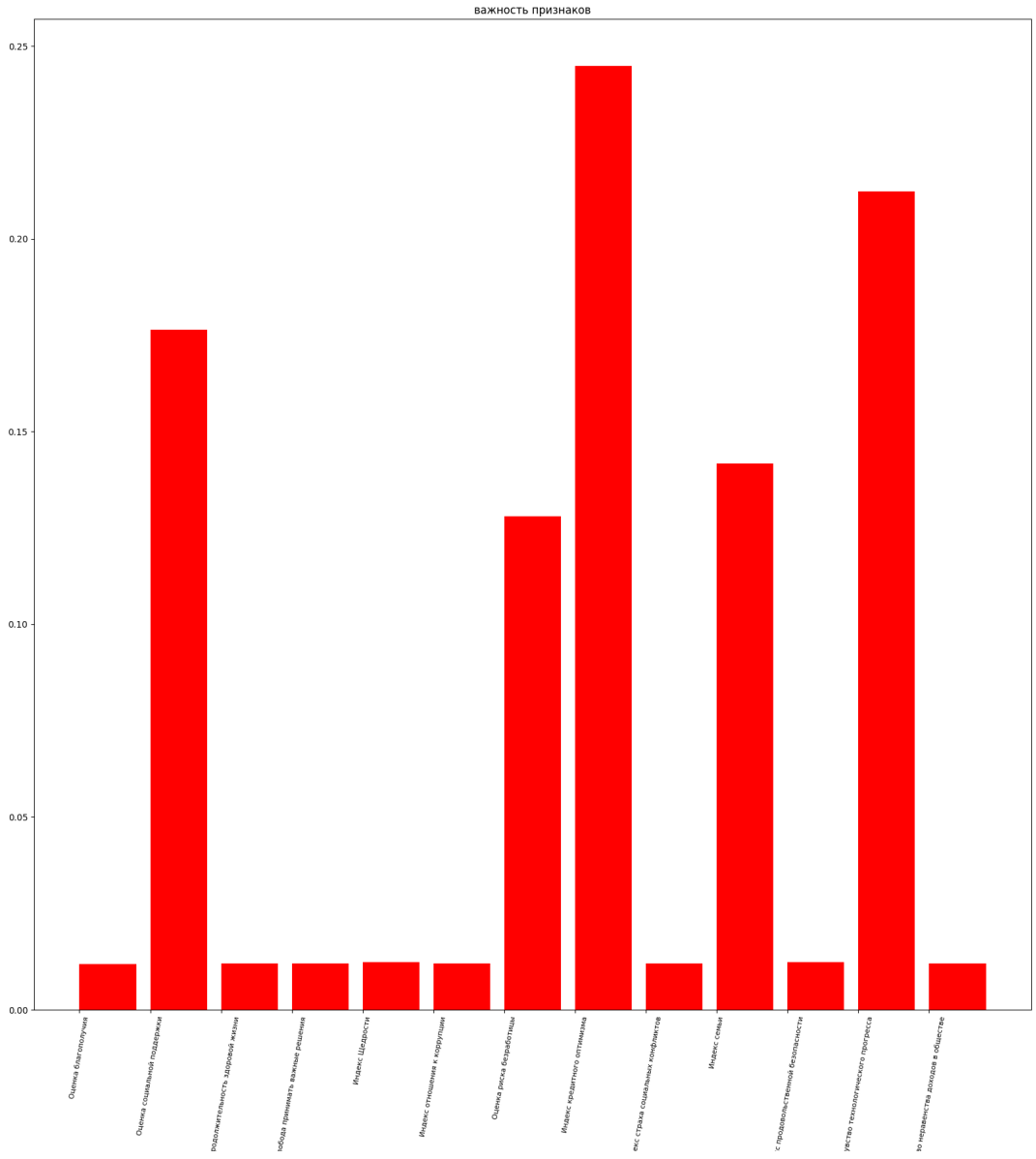

Рис. 2.1: Ощущаемое счастье, случайный лес

Ошибка выбранной модели при наличии (в признаках модели) только выбранных состояний:

Таблица 2.2: Точность классификации

Алгоритм классификации	Целевая переменная	Точность (accuracy)	Точность (precision)	Полнота	f-мера
случайный лес	Ощущаемое счастье	0.967	0.967	0.967	0.967

Матрицы ошибок моделей приведены ниже:

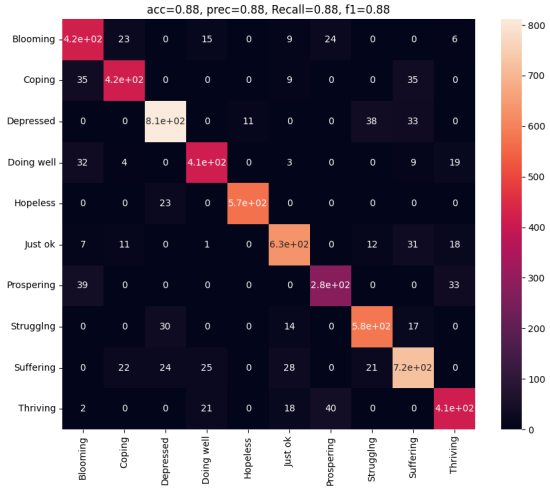

Рис. 2.2: Результат линейной модели классификации

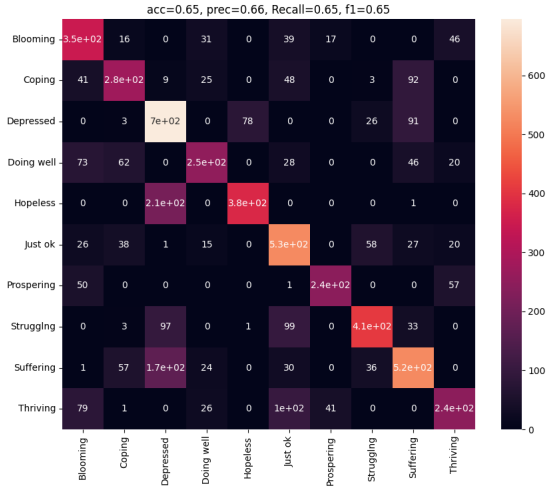

Рис. 2.3: Результат k-соседей

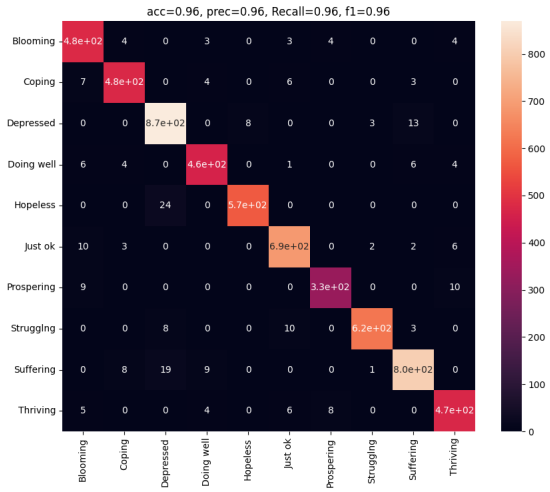

Рис. 2.4: Результат случайного леса по всем признакам состояний

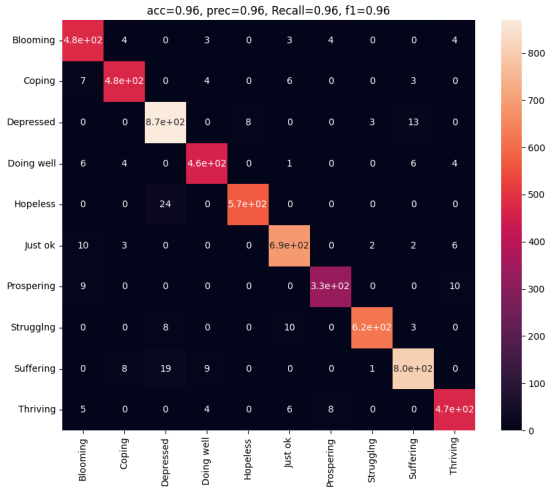

Рис. 2.5: Результат случайного леса по выбранным признакам состояний

2.4 Определение признаков причин, важных для выбранных состояний

Было выбрано 5 состояний: 'Оценка социальной поддержки', 'Оценка риска безработицы', 'Индекс кредитного оптимизма', 'Индекс семьи', 'Чувство технологического прогресса'.

Для определения по признакам причинам значения признаков состояний были использованы модели регрессии (линейная, k-ближайших соседей и случайный лес). Для каждого состояния:

1. были оценены ошибки регрессии для предсказания обрабатываемого при-

знака состояния по всем признакам причинам, по ошибкам выбрана наилучшая модель;

2. по выбранной модели были вычислены наиболее важные признаки причины для обрабатываемого признака состояния;

3. определена доля выбранных признаков причин относительно всех признаков причин;

4. посчитана ошибка регрессии для предсказания состояния только по выбранным признакам-причинам.

2.4.1 Оценка социальной поддержки

ошибки моделей:

1. ошибки моделей: лучшая модель по ошибкам: линейная

Таблица 2.3: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Оценка социальной поддержки	0.549
k-соседей	Оценка социальной поддержки	5.835
случайный лес	Оценка социальной поддержки	1.092

2. Выбранные причины: 'V алкоголя в год', 'Количество лет образования', 'Доля дохода семьи на продовольствие', 'Коэффициент Джини сообщества', 'Охват беспроводной связи в сообществе', 'Количество смертей от заболеваний в сообществе', 'Волатильность цен в сообществе'

Визуализация:

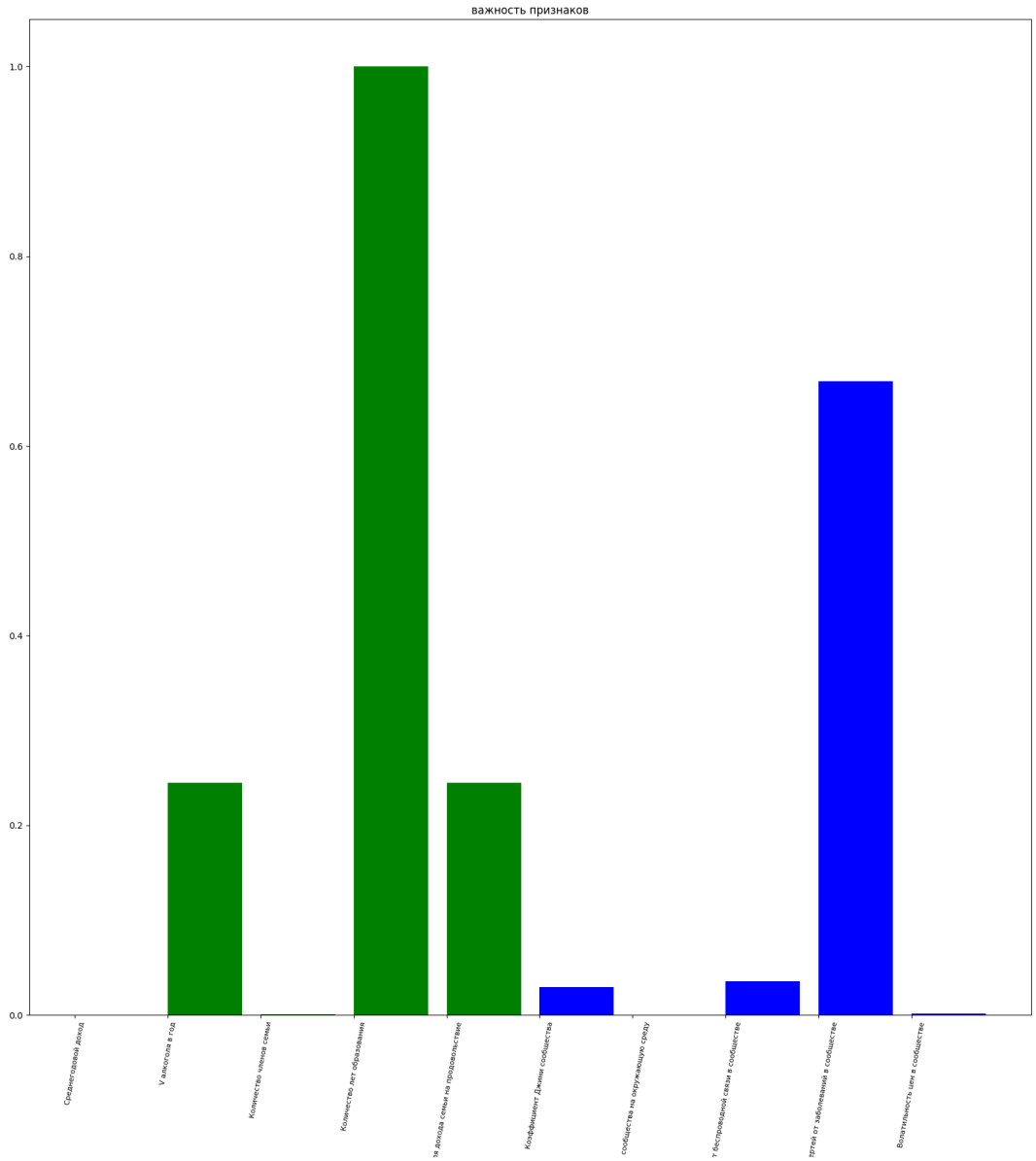

Рис. 2.6: Важность причин для состояния

3. Доля выбранных причин: 0.7;

4. Ошибка на выбранных причинах:

Таблица 2.4: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Оценка социальной поддержки	0.549

2.4.2 Оценка риска безработицы

ошибки моделей:

1. ошибки моделей: лучшая модель по ошибкам: линейная

Таблица 2.5: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Оценка риска безработицы	0.602
k-соседей	Оценка риска безработицы	15.861
случайный лес	Оценка риска безработицы	1.595

2. Выбранные причины: 'Количество членов семьи', 'Количество лет образования', 'Доля дохода семьи на продовольствие', 'Коэффициент Джини сообщества', 'Охват беспроводной связи в сообществе'.

Визуализация:

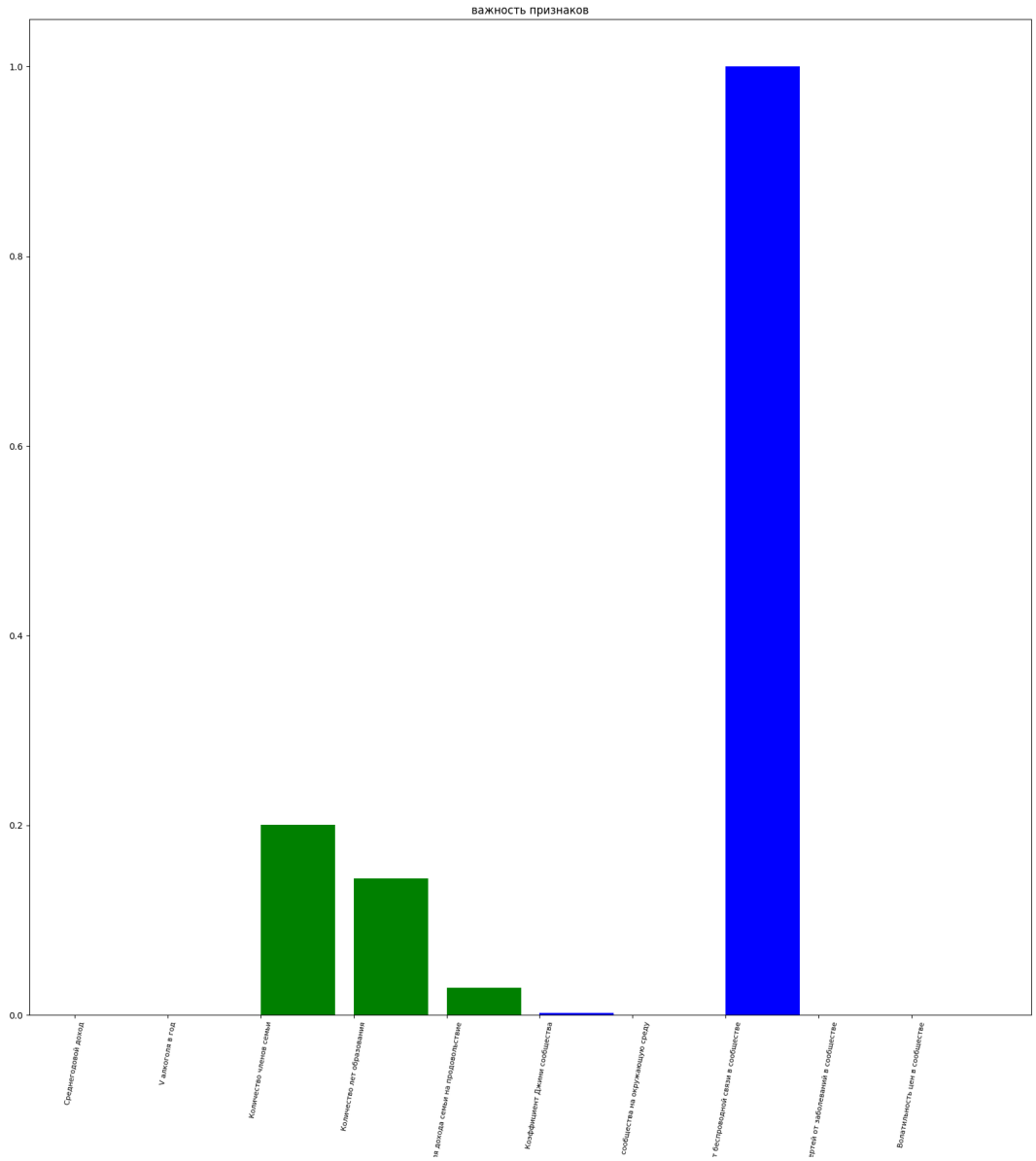

Рис. 2.7: Важность причин для состояния

3. Доля выбранных причин: 0.5;

4. Ошибка на выбранных причинах:

Таблица 2.6: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Оценка риска безработицы	0.602

2.4.3 Индекс кредитного оптимизма

ошибки моделей:

1. ошибки моделей: лучшая модель по ошибкам: линейная

Таблица 2.7: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Индекс кредитного оптимизма	0.638
k-соседей	Индекс кредитного оптимизма	10.850
случайный лес	Индекс кредитного оптимизма	1.720

2. Выбранные причины: 'Количество членов семьи', 'Количество лет образования', 'Коэффициент Джини сообщества', 'Издержки сообщества на окружающую среду', 'Количество смертей от заболеваний в сообществе'.

Визуализация:

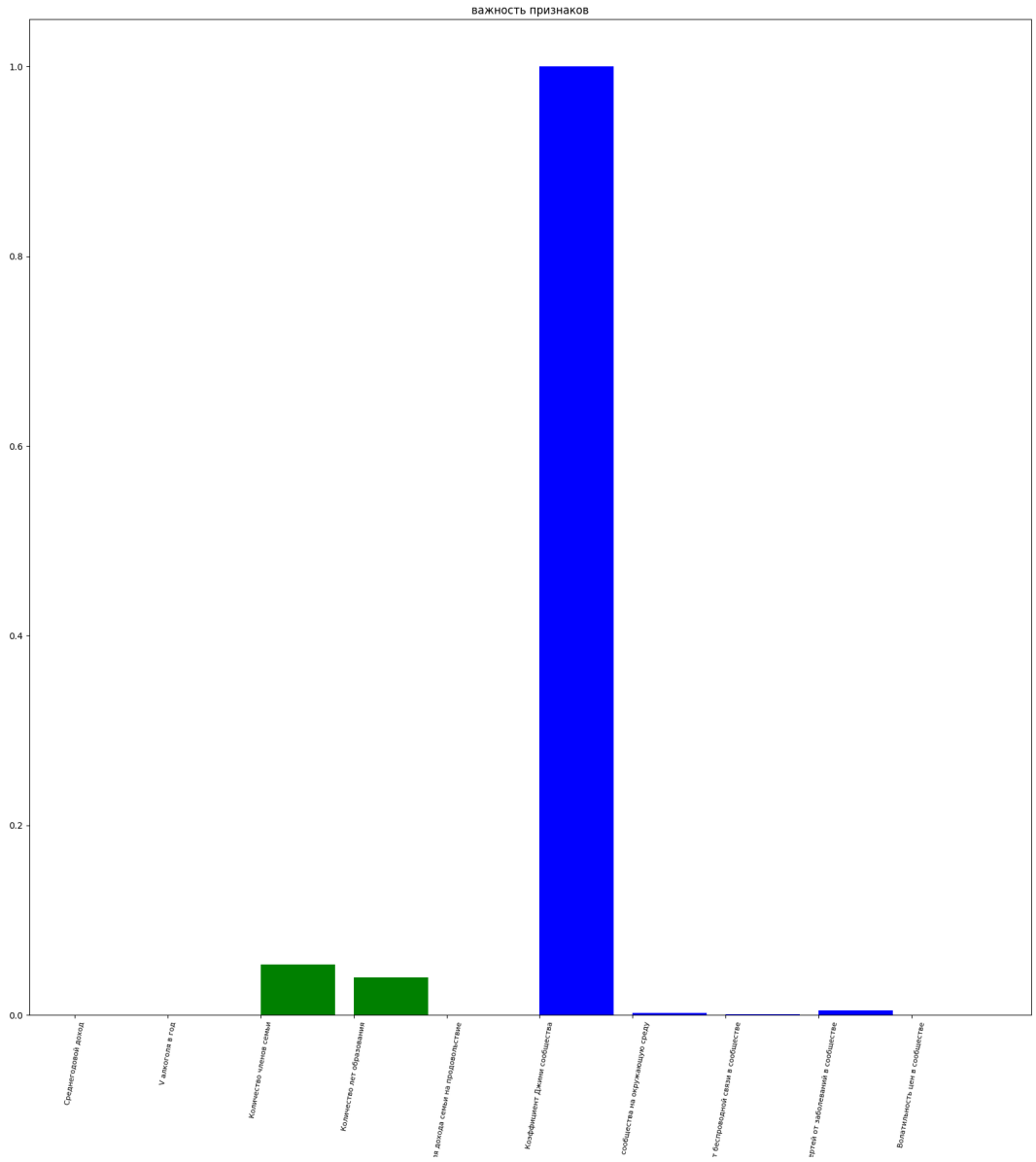

Рис. 2.8: Важность причин для состояния

3. Доля выбранных причин: 0.5;

4. Ошибка на выбранных причинах:

Таблица 2.8: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Индекс кредитного оптимизма	0.638

2.4.4 Индекс семьи

ошибки моделей:

1. ошибки моделей: лучшая модель по ошибкам: линейная

Таблица 2.9: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Индекс семьи	1.198
k-соседей	Индекс семьи	5.408
случайный лес	Индекс семьи	1.853

2. Выбранные причины: 'Среднегодовой доход', 'V алкоголя в год', 'Коэффициент Джини сообщества', 'Охват беспроводной связи в сообществе', 'Волатильность цен в сообществе'.

Визуализация:

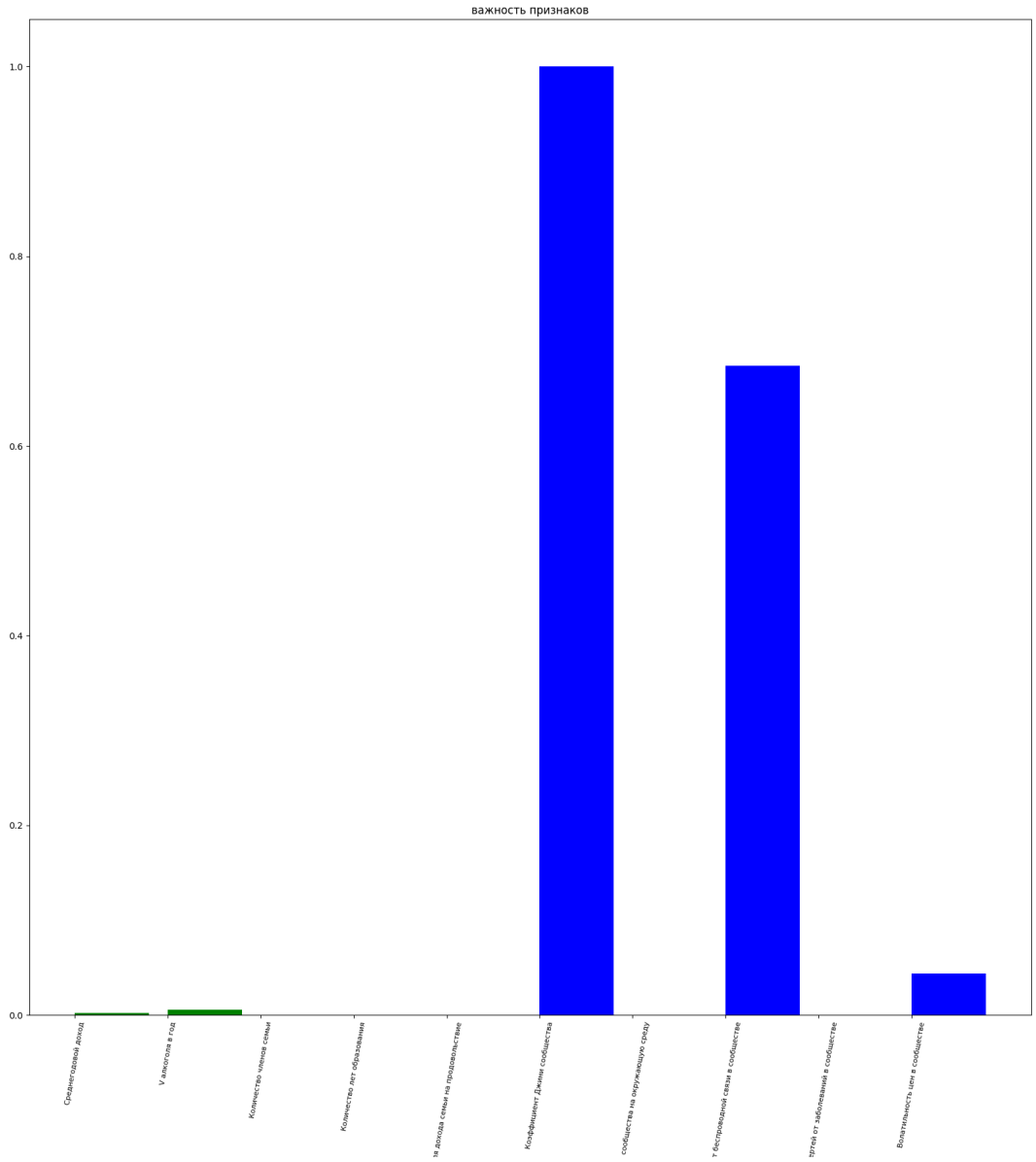

Рис. 2.9: Важность причин для состояния

3. Доля выбранных причин: 0.5;

4. Ошибка на выбранных причинах:

Таблица 2.10: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Индекс семьи	1.198

2.4.5 Чувство технологического прогресса

ошибки моделей:

1. ошибки моделей: лучшая модель по ошибкам: линейная

Таблица 2.11: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Чувство технологического прогресса	0.714
k-соседей	Чувство технологического прогресса	10.069
случайный лес	Чувство технологического прогресса	1.190

2. Выбранные причины: 'Среднегодовой доход', 'Количество лет образования', 'Коэффициент Джини сообщества', 'Издержки сообщества на окружающую среду', 'Охват беспроводной связи в сообществе', 'Волатильность цен в сообществе'.

Визуализация:

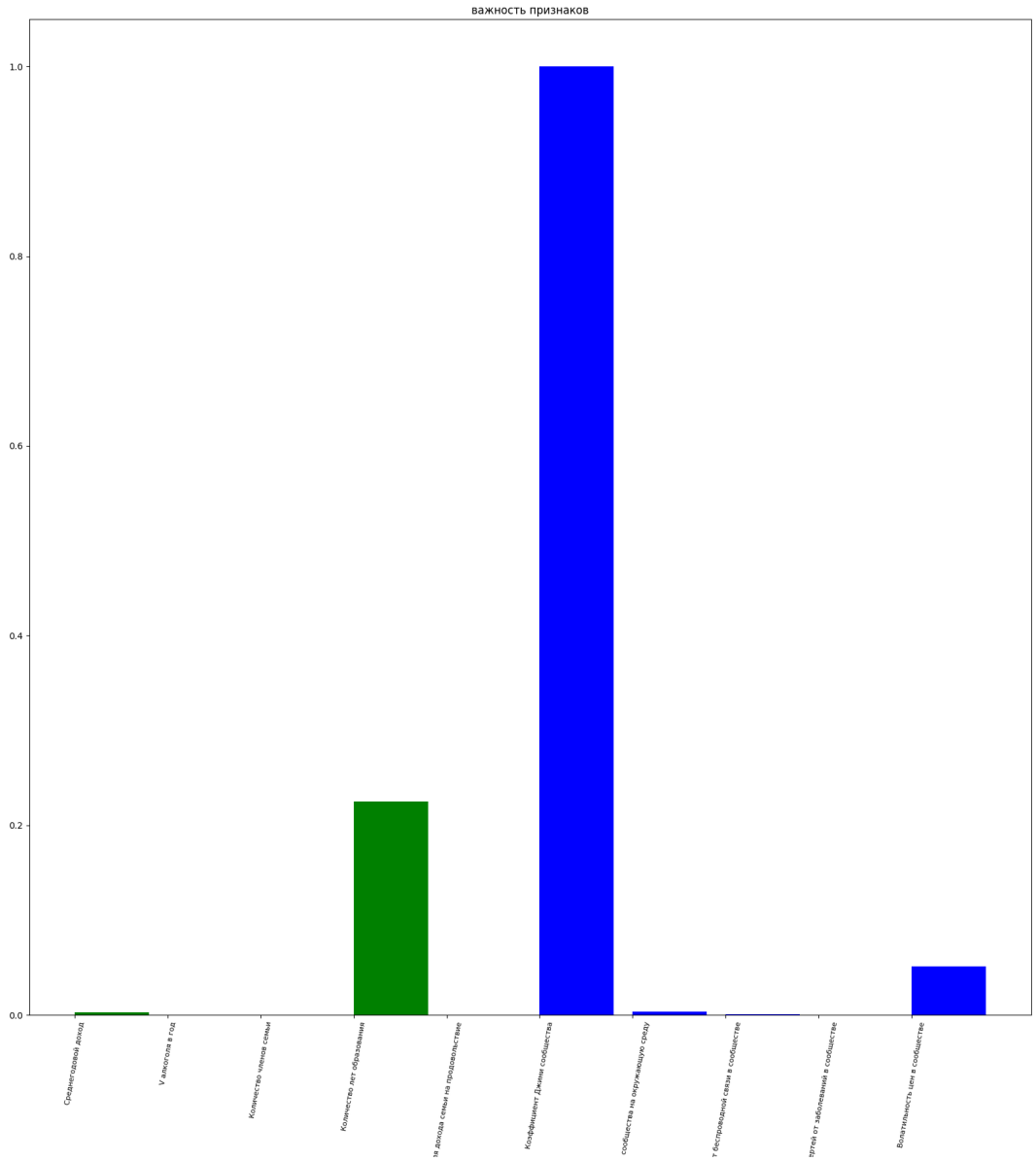

Рис. 2.10: Важность причин для состояния

3. Доля выбранных причин: 0.6;

4. Ошибка на выбранных причинах:

Таблица 2.12: Ошибки моделей

Алгоритм регрессии	Целевая переменная	Ошибка (MAPE))
линейная	Чувство технологического прогресса	0.714

2.5 Прогнозирование интегральной характеристики

Пользуясь найденными закономерностями спрогнозировать попадание респондентов, у которых интегральная характеристика отмечена как «Неизвестно», в укрупненные группы шкалы Кантрила.

Так, была сконструирована двухуровневая модель классификации целевого параметра (ощущаемого счастья):

1. По важным признакам причинам были предсказаны соотвтствующие признаки состояний с помощью сконструированнхы моделей регрессии;

2. По предсказанным состояниям был предсказан целевой параметр с помощью сконструированной модели классификации.

Описанные действия были проделаны на обучающей выборке (где интегральная характеристика известна). Были получены следующие результаты.

Видно, что точность предсказания близка к 97%.

Аналогичные действия были проделаны с данными, в которых интегральная характеристика неизвестна, а затем полученные значения укрупнены.

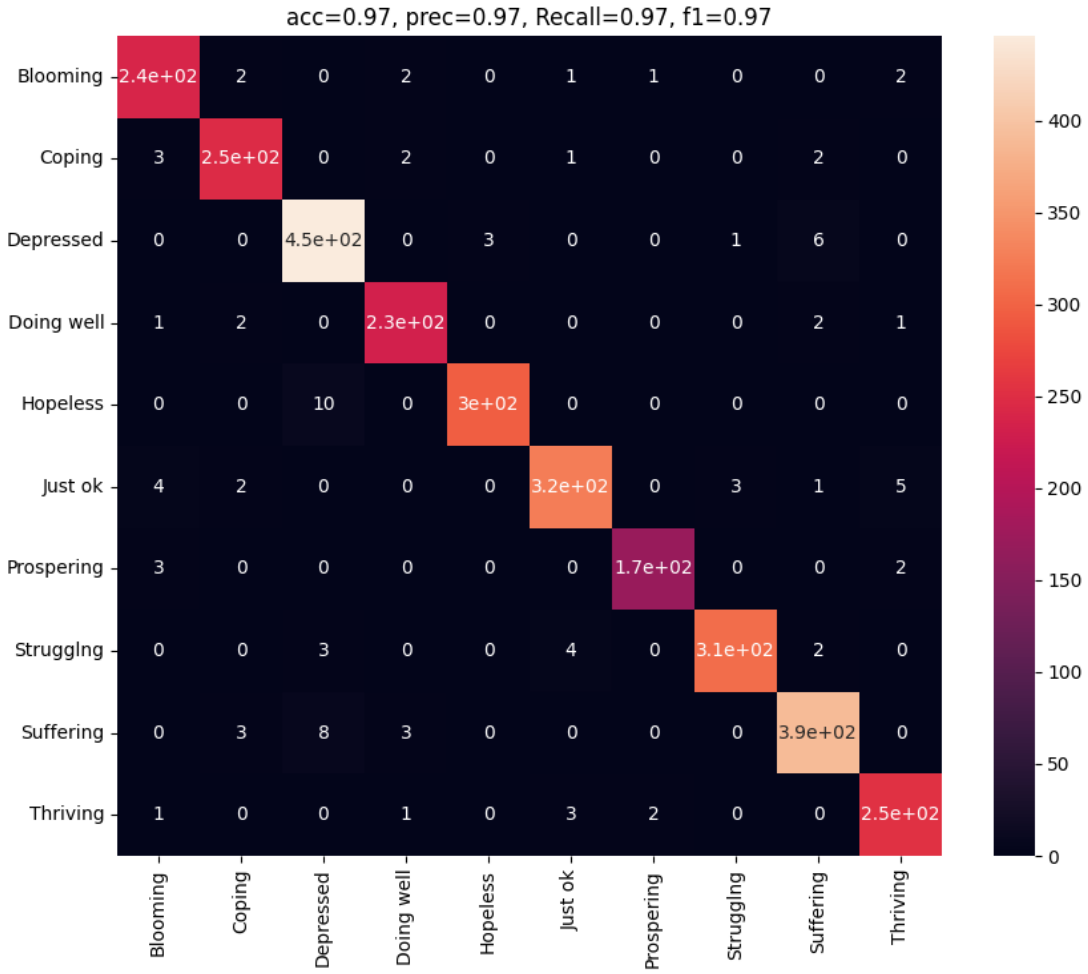

Рис. 2.11: Результат

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Virtanen P.*, *Gommers R.*, *Oliphant T. E.* SciPy: Fundamental Algorithms for Scientific Computing in Python. — 2020. — DOI: 10.1038/s41592-019-0686-2.