ГЛАВА 2

Решение задачи

2.1 Подготовка данных

Загружаем во фрейм данных исходную таблицу и просматриваем её. (Рис. 2.1)

	number	days	gender	age	afftype	melanch	inpatient	edu	marriage	work	madrs1	madrs2
0	condition_1	11	2	35-39	2.0	2.0	2.0	6-10	1.0	2.0	19.0	19.0
1	condition_2	18	2	40-44	1.0	2.0	2.0	6-10	2.0	2.0	24.0	11.0
2	condition_3	13	1	45-49	2.0	2.0	2.0	6-10	2.0	2.0	24.0	25.0
3	condition_4	13	2	25-29	2.0	2.0	2.0	11-15	1.0	1.0	20.0	16.0
4	condition_5	13	2	50-54	2.0	2.0	2.0	11-15	2.0	2.0	26.0	26.0
5	condition_6	7	1	35-39	2.0	2.0	2.0	6-10	1.0	2.0	18.0	15.0
6	condition_7	11	1	20-24	1.0	NaN	2.0	11-15	2.0	1.0	24.0	25.0
7	condition_8	5	2	25-29	2.0	NaN	2.0	11-15	1.0	2.0	20.0	16.0

Рисунок 2.1 – Исходная таблица

Сразу привлекает внимание разнородность измерений: некоторые представлены в виде строк, другие типом int, а третьи типом float. Однако, для начала проверим нет ли пропущенных значений.

Отсутствует несколько в первой половине таблицы: где находятся пациенты с депрессией, а также большинство параметров во второй половине: вероятно, у здоровых людей эти показатели не измерялись.

Разделим фрейм на два - по наличию поставленного диагноза и сразу добавим для первого фрейма столбец с разностью значений оценок MADRS до и после периода измерения активности.

В первую очередь будем рассматривать выборку пациентов с наличием поставленного диагноза. Еще раз проверим, есть ли пропуски в данных: пропущены три значения в столбце melanch. В данном столбце у подавляющего большинства пациентов одно и то же значение, поэтому велика вероятность, что и пропущенные будут такими же, заменяем NaN на средние значения по

столбцу и меняем тип на category, так как вариативность допустимых значений строго ограничена двухэлементным множеством {1,2}.

Возраст, пол и образование по аналогичной причине также преобразуем в category. Значения в столбцах afftype, inpatient, marriage и work принадлежат к типу float, поэтому преобразуем их сначала в тип int и только потом в category.

В столбце с информацией об образовании отсутствуют данные об одном пациенте, исходя из прочих, невозможно сделать выводы о годе, когда он получил образование: он мог обучаться во время сбора данных, мог забыть указать информацию, либо не получать образование вовсе. Не будем учитывать данное значение при рассмотрении взаимосвязей, заменим его текстом 'miss'.

Просматриваем очищенные данные и убеждаемся, что теперь все в порядке. Следующим шагом разделяем числовые и категориальные признаки: в числовые выделяем оценки MADRS, а также их разность и количество дней, которые человек носил на руке актиграф, в категориальные - все остальные.

2.2 Числовые характеристики

Просмотрим гистограмму изменения оценок MADRS.(Рис. 2.2)

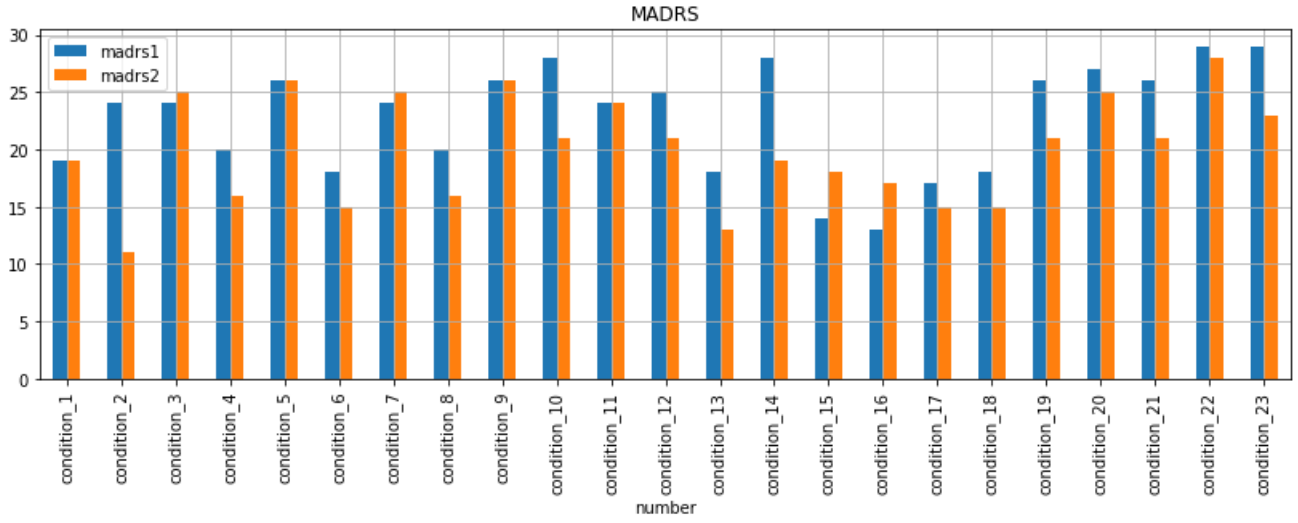

Рисунок 2.2 – Гистограмма изменения оценок MADRS

Можно заметить, что у подавляющего большинства пациентов данная оценка стала ниже, из чего делаем вывод о том, что в процессе исследования пациенты проходили терапию у своего психиатра, либо принимали назначенные препараты.

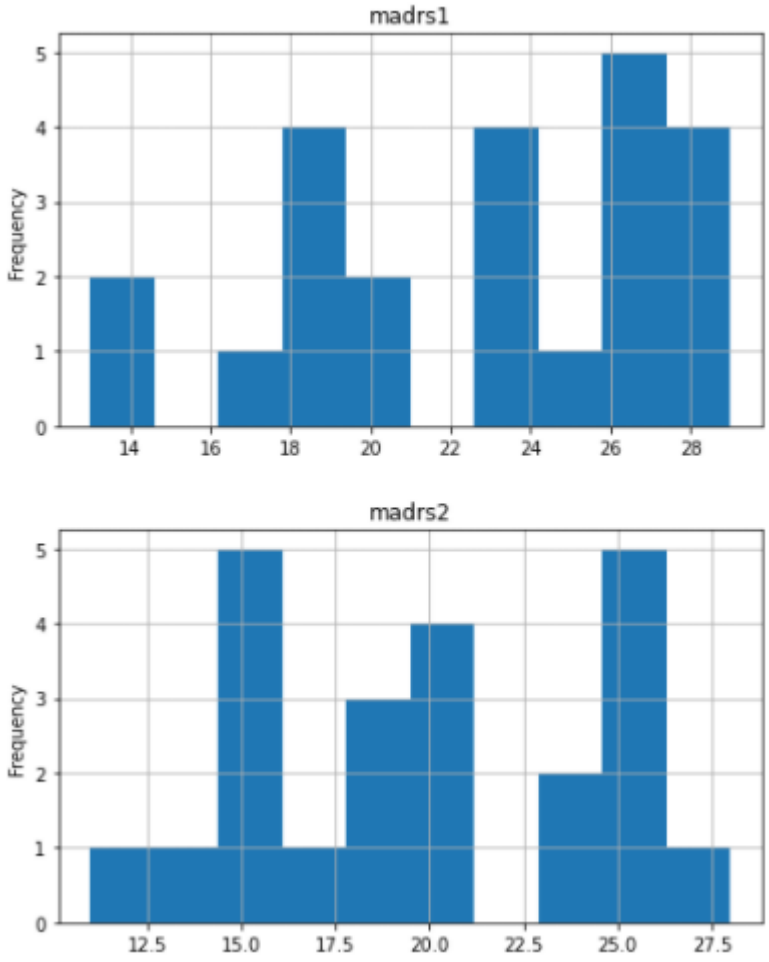

Рисунок 2.3 – Распределение оценок MADRS

Среди наблюдаемых нет людей с оценкой, соответствующей здоровому состоянию психики, а также ни один из них не выздоровел к моменту окончания периода измерения активности.(Рис. 2.3)

2.3 Категориальные характеристики

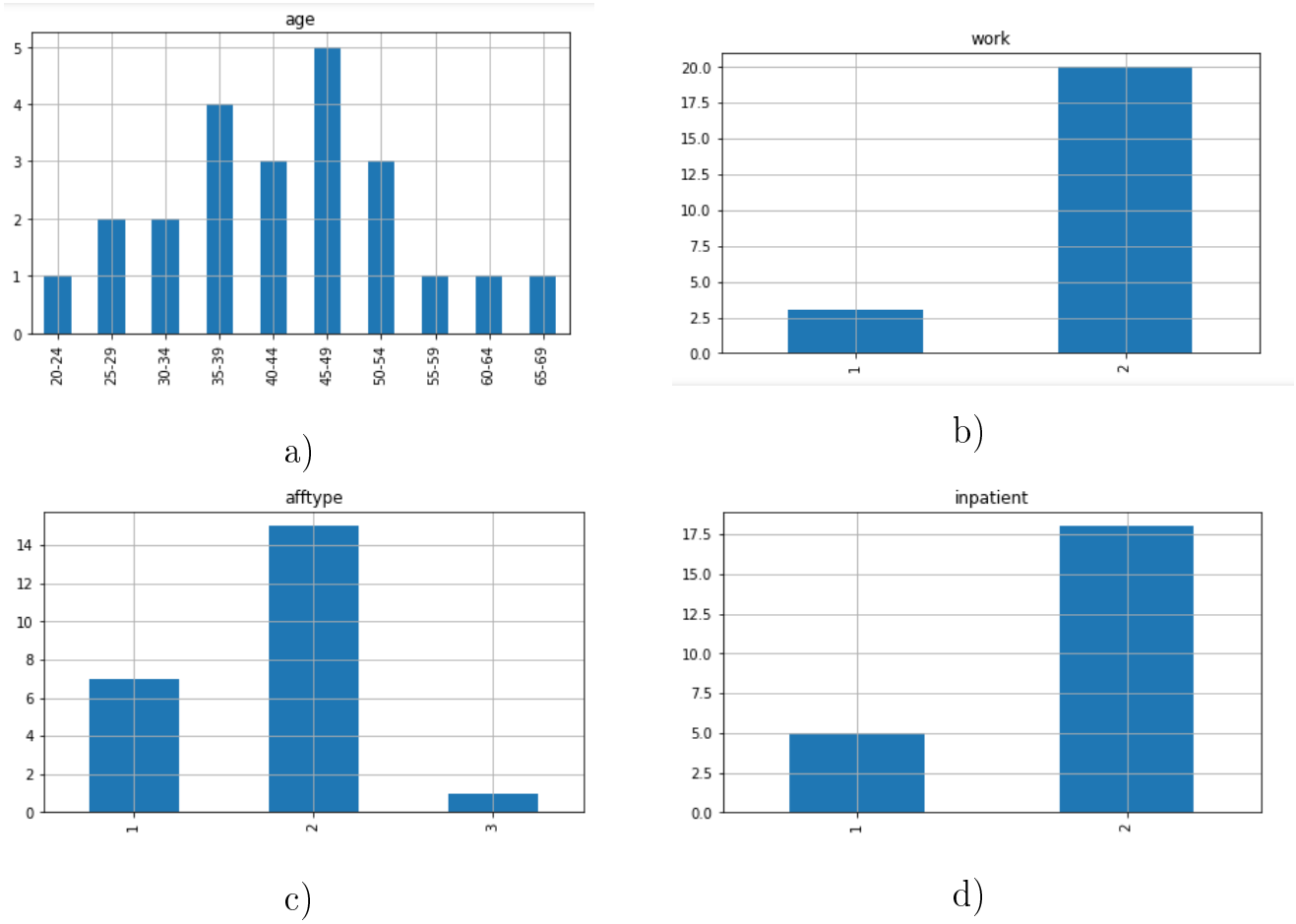

Рисунок 2.4 – Диаграммы распределения категориальных признаков

Судя по диаграммам распределения категориальных признаков (Рис 2.4), в исследовании в основном участвовали безработные люди в возрасте от 25 до 54 лет с униполярной депрессией. Всего 5 человек находились в стационаре, остальные - амбулаторные пациенты.

Рассмотрим влияние признаков на оценки MADRS. У людей в возрасте 30-34 года наблюдаются более высокие оценки депрессивного состояния. (Рис 2.5) На оценку сильно влияет семейное положение наблюдаемого (Рис 2.6) и наличие у него работы. (Рис 2.7) Безработные и одинокие показывают наиболее высокие оценки.

Делаем вполне обоснованный вывод, что у всех, стационарных пациентов наличествует умеренный депрессивный эпизод, характеризуемый высокими показателями MADRS. (Рис 2.8)

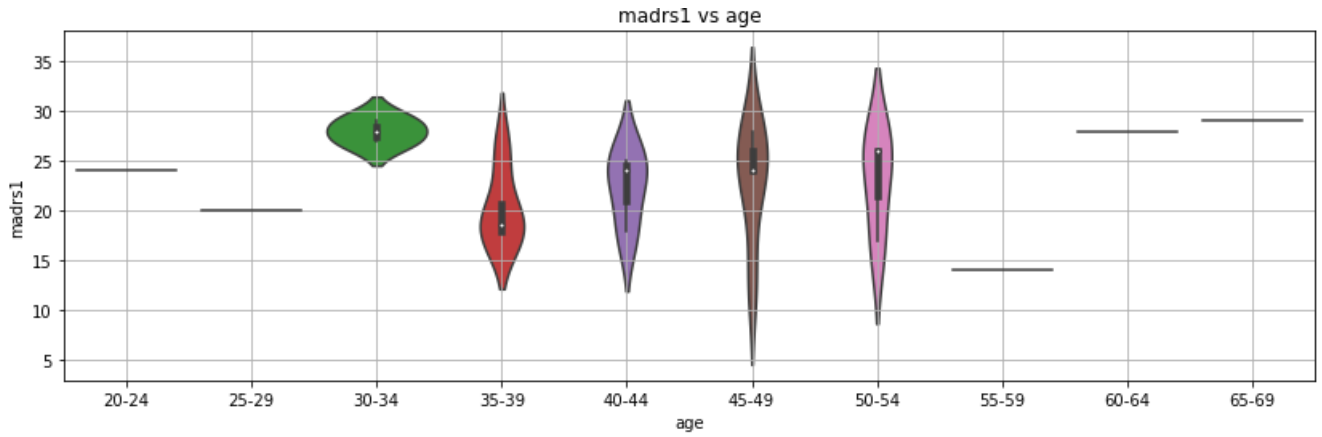

Рисунок 2.5 – Влияние возраста на оценку MADRS

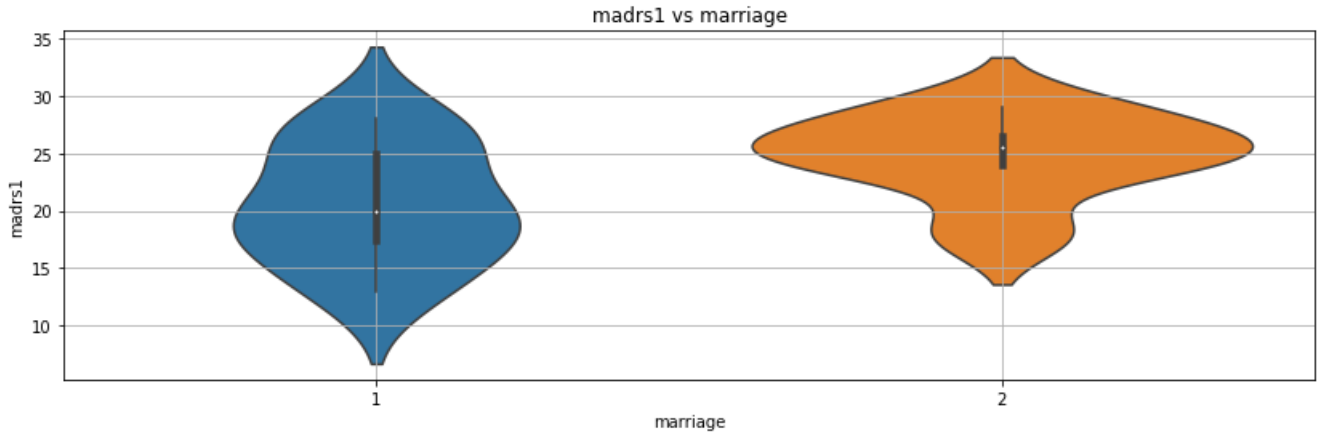

Рисунок 2.6 – Влияние семейного положения на оценку MADRS

В течение наблюдения снижение оценки произошло в большей степени также у стационарных пациентов, что возвращает нас к идее того, что они проходили терапию.

Изучим таблицу со здоровыми людьми. Здесь представленно гораздо меньше параметров. Можно сделать выводы лишь о половой принадлежности и возрасте: в исследовании в большинстве своем принимали участие женщины в возрасте от 20 до 54 лет.

2.4 Сравнение категорий исследуемых

На примере одного из многочисленных файлов с данными об активности пациентов, замечаем, что первый и последний дни неполные, а также нали-

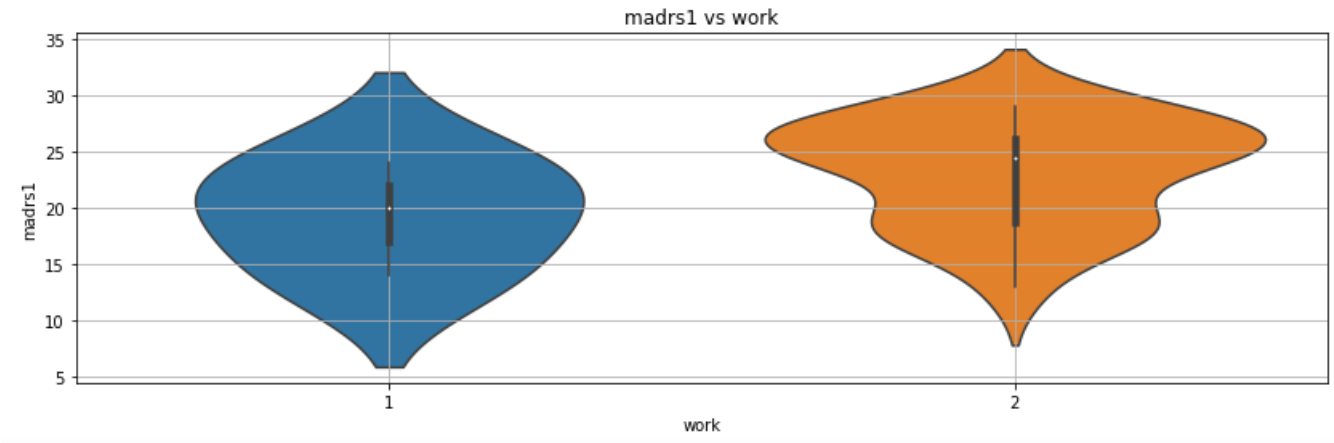

Рисунок 2.7 – Влияние наличия работы на оценку MADRS

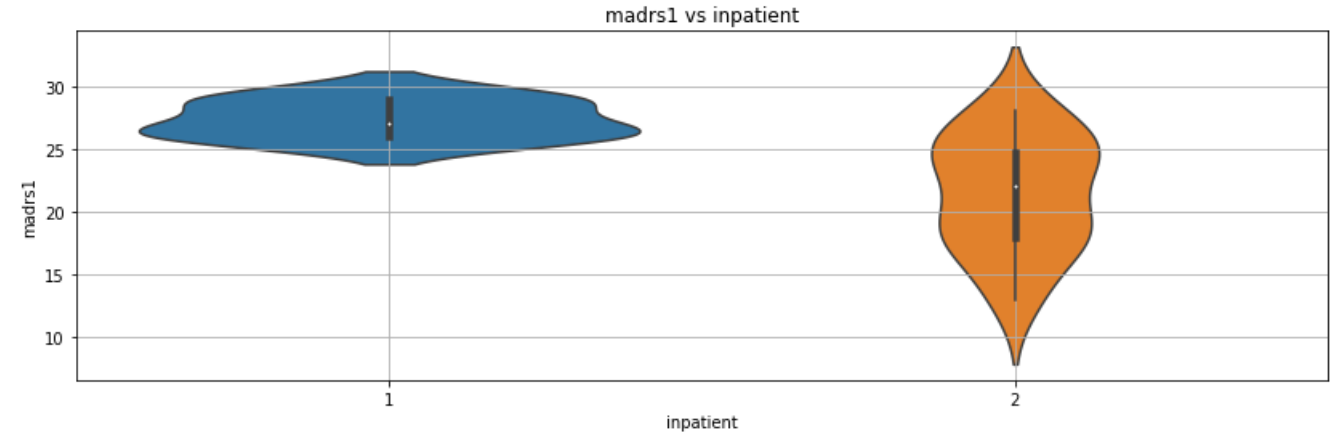

Рисунок 2.8 – Влияние типа лечения на оценку MADRS

чие дней, когда активность была чрезвычайно мала, либо не фиксировалась вовсе. Вероятно в эти дни прибор был отключен, либо не работал должным образом. Полученная информация позволит в дальнейшем при рассмотрении остальных файлов установить пороговое значение среднедневной активности и исключить дни с крайне низкой активностью.

Неполные дни - первый и последний - также исключим из исследования, в эти дни датчик устанавливали и снимали, поэтому измерение активности в первый день начиналось слишком поздно, а в последний - прерывалось задолго до окончания суток.

При помощи цикла по файлам папки извлекаем необходимую информацию: среднее значение среднедневной активности, стандартное отклонение среднедневной активности, среднее значение 99-го процентиля и стандартное отклонение 99-го процентиля дневной активности. Результаты сохраняем во фрейме данных и повторяем процесс с файлами из папки с данными об ак-

тивности контрольной группы здоровых.

Сравним полученные показатели.

	Mean_MeanAct	Mean_Q99Act	Std_MeanAct	Std_Q99Act	CV_MeanAct	CV_Q99Act
count	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000
mean	263.666975	1696.860976	81.291596	410.943519	0.310272	0.230448
std	71.689552	405.953265	31.768637	309.473258	0.084625	0.118904
min	139.782917	915.619333	43.967642	134.236127	0.144795	0.083808
25%	201.082937	1401.476008	54.296764	213.154471	0.250760	0.140532
50%	262.758110	1729.681104	70.213031	307.159779	0.318632	0.197742
75%	315.547474	1867.617679	119.680854	497.436726	0.354754	0.284825
max	407.458697	2635.942143	137.730227	1520.288118	0.537699	0.594511

Рисунок 2.9 – Показатели здоровых людей

	Mean_MeanAct	Mean_Q99Act	Std_MeanAct	Std_Q99Act	CV_MeanAct	CV_Q99Act
count	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000
mean	177.359263	1293.282746	60.221801	318.413688	0.327726	0.248803
std	75.725958	335.760931	39.333135	168.601320	0.135601	0.122393
min	65.370585	665.491429	9.301415	96.547265	0.130202	0.106936
25%	120.196612	1030.523544	28.137077	190.752177	0.229600	0.158577
50%	172.622371	1300.538571	60.377039	287.349708	0.286429	0.199776
75%	235.399106	1574.730899	73.919049	389.201044	0.432363	0.349973
max	296.403373	1804.810000	166.280241	745.719621	0.581268	0.572878

Рисунок 2.10 – Показатели людей с депрессией

Контрольная группа (Рис 2.9) показывает более высокие показатели активности, что является подтверждением того факта, что люди, находящиеся в депрессии(Рис 2.10) испытывают вялость и сонливость в течение всего дня. Объединим статистики в общий фрейм данных, сравним показатели на графиках и заметим, что графическое представление полностью соответствует выводам, сделанным на основе изучения таблиц.

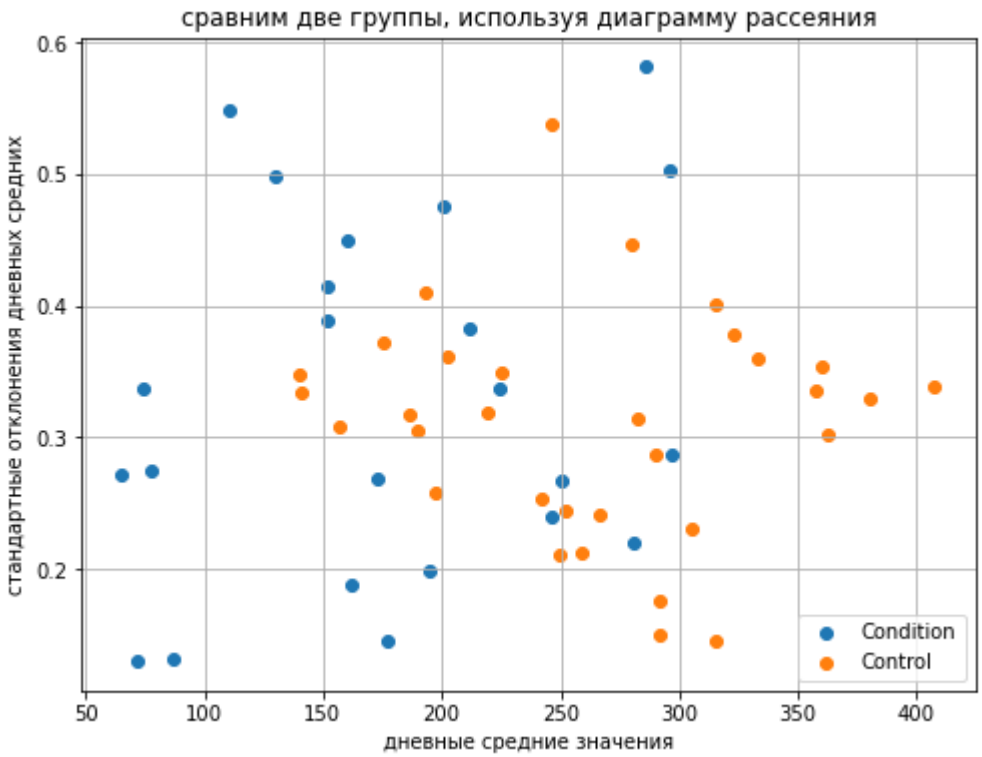

Рисунок 2.11 – Диаграмма рассеяния сравнения двух категорий

На диаграмме рассеяния(Рис 2.11) видим людей из контрольной группы с достаточно низкими показателями, а также пациентов с депрессией, выдающих активность на уровне здорового человека. Можно сделать вывод о том, что первые вероятно скоро станут посещать врача гораздо чаще, чем планировали, в то время, когда прогноз вторых более утешителен - они находятся в паре шагов от выздоровления.

Также на диаграмме рассеяния(Рис 2.12) проверим, влияет ли на показатели активности тип депрессии.

Нет - активность ни в коей мере не зависит от типа поставленного диагноза.

2.5 Поиск зависимостей между показателями активности и наличием депрессии

При помощи цикла извлекаем необходимые данные из файлов с данными об активности пациентов с депрессией, кроме рассмотренных ранее добавляем максимальную дневную активность и среднее значение из максимальных дневных. Формируем фрейм данных, в котором вычисляем среднее значение

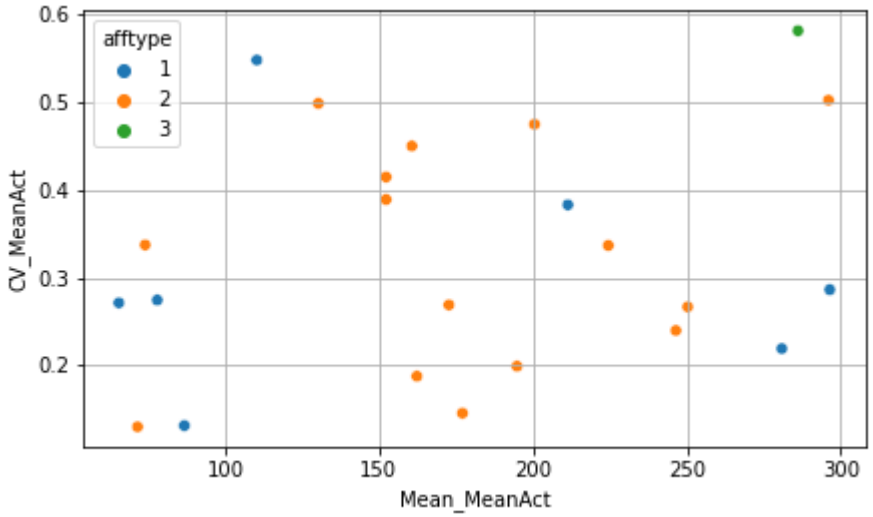

Рисунок 2.12 – Тип депрессии и активность

оценки MADRS.

Аналогичным образом формируем фрейм из данных об активности контрольной группы. Средней оценкой MADRS для них берем среднее из диапазона оценок, характеризующих отсутствие депрессии:$[0; 10]$.

Объединяем фреймы, а средние значения оценок выделяем отдельным массивом и удаляем из объединенного фрейма лишние столбцы: целевой и столбец с наименованиями.

Разделим тестовую и обучающую выборки с долей 0.6. Набор данных достаточно мал, поэтому есть смысл применить линейную регрессию, однако, некоторые значения могут коррелировать друг с другом, что при линейной регрессии приводит к низкой точности результата, поэтому здесь находит свое применение регрессия типа LASSO.

Обучаем модель и проверяем коэффициенты.

$[-0.10155681 - 0.007147650.06633724 - 0.01049892 - 0.001931650.01365076]$

Наибольшее влияние на итоговую оценку оказывает стандартное отклонение среднедневной активности, а также среднее значение максимальной дневной активности.

Средняя квадратическая ошибка(инв) $MAE = -77.02457150674981$

Средняя абсолютная ошибка(инв) $MSE = -6.7155152419334705$

Коэффициент детерминации $R^2 = 0.14133074625763486$

2.6 Классификация

Для классификации используем полиномиальную логистическую регрессию. Выбор метода обоснован малым набором данных, к тому же путем LASSO-регрессии, мы выяснили, что у данных присутствует линейная зависимость.

Разделим тестовую и обучающую выборки с долей 0.57 и обучим классификатор, проверяя его точность на тестовых данных путем рассмотрения матрицы ошибок. $\begin{pmatrix} 12 & 0 & 1 \\ 0 & 7 & 2 \\ 0 & 2 & 0 \end{pmatrix}$

А также при помощи коэффициента точности обучения, который оказался равен 0.7916666666666666

На графиках рассеяния смотрим различия между истинными классами (Рис 2.13) и классификацией предсказанных.(Рис 2.14)

Они в достаточной степени схожи, из чего можно сделать вывод о возможности использования данной модели классификации на предсказанных значениях.

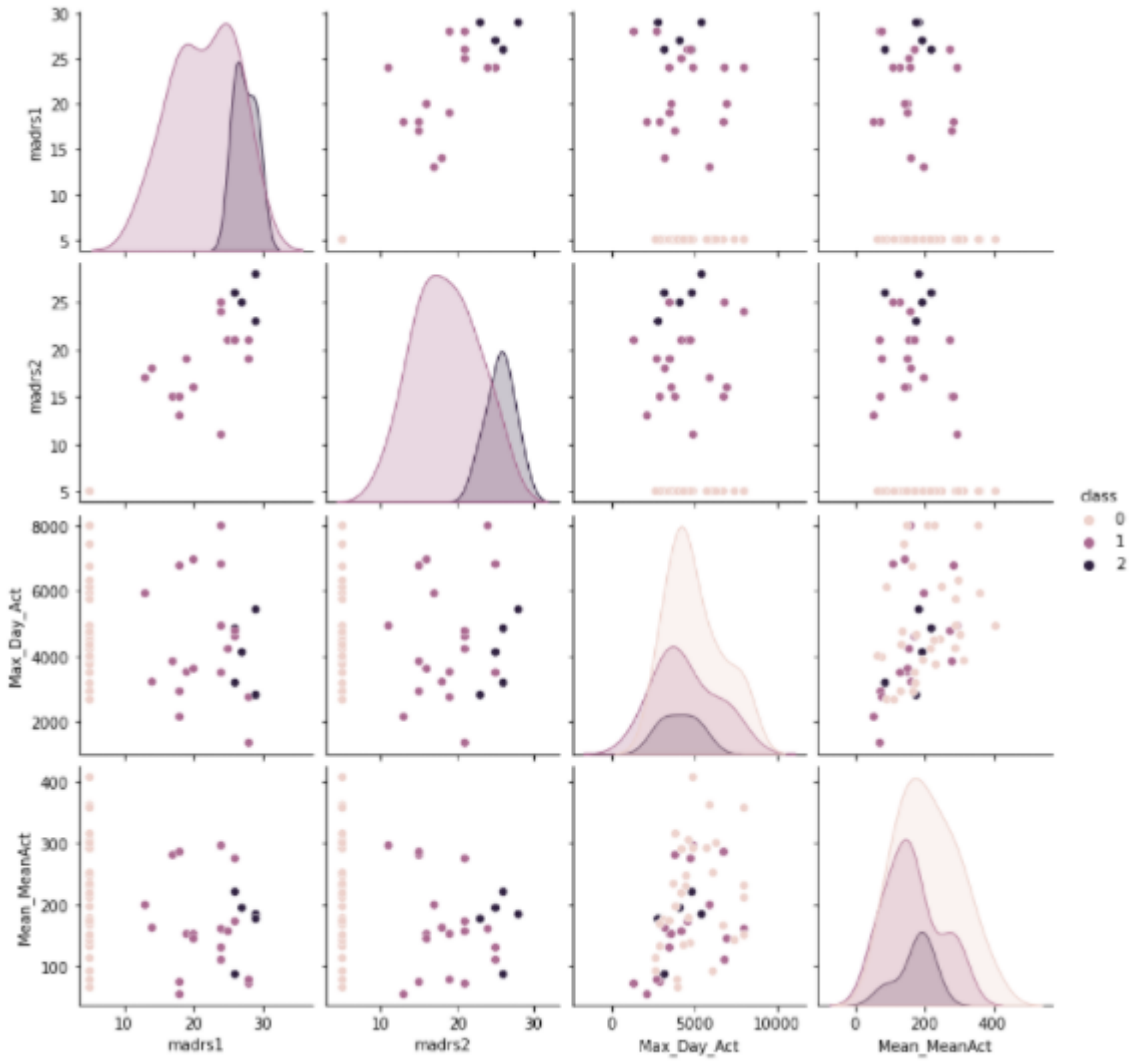

Рисунок 2.13 – Истинные классы

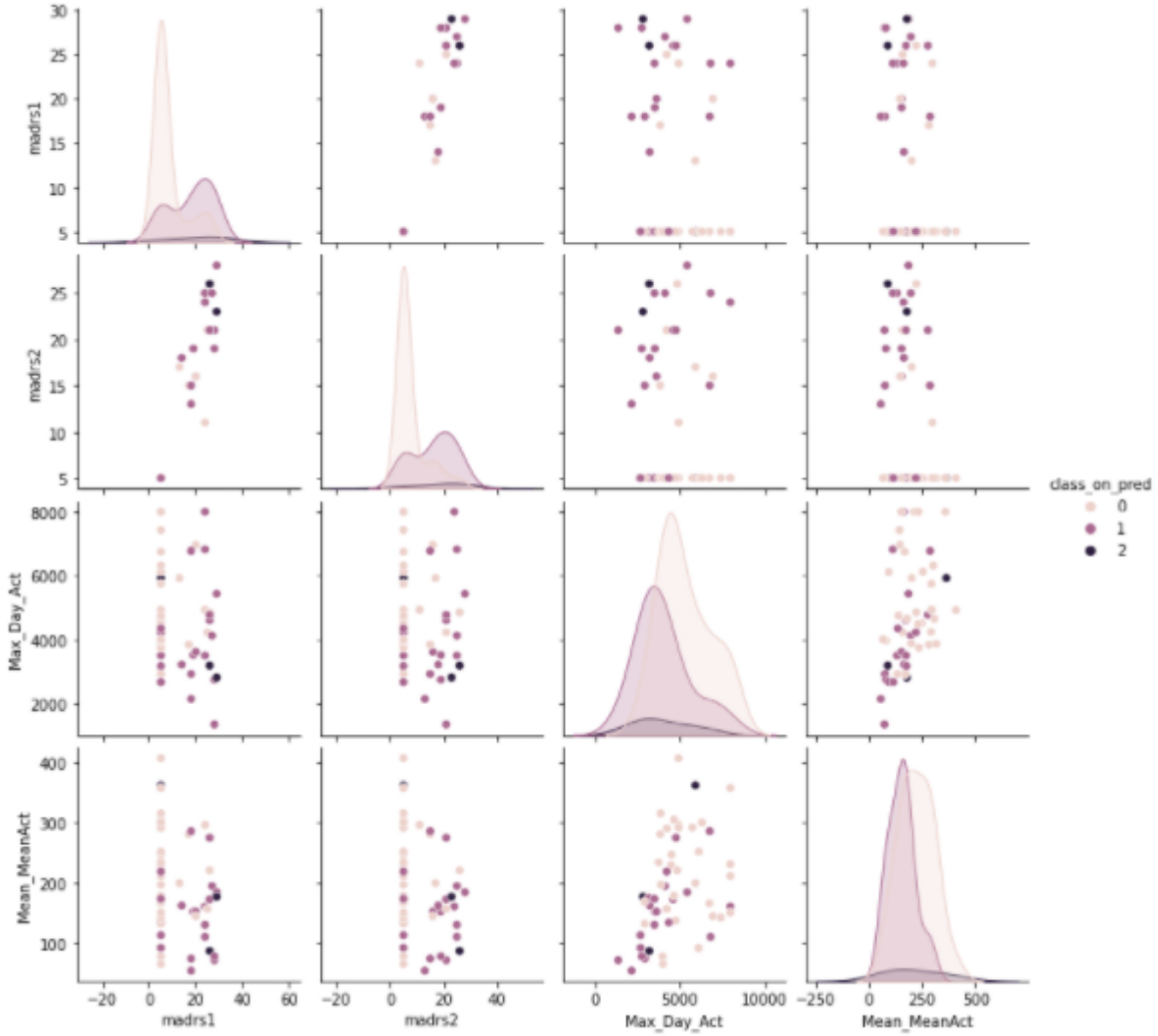

Рисунок 2.14 – Классификация предсказанных результатов