# ОГЛАВЛЕНИЕ

[**Постановка задачи** 1](#_Toc40462989)

[**Предложенный подход** 2](#_Toc40462990)

[**Отбор информативных признаков с помощью генетического алгоритма (ГА)** 4](#_Toc40462991)

[**Результаты экспериментов** 7](#_Toc40462992)

## **Постановка задачи**

База данных представляет собой две группы онкологических больных пациентов с разным видом терапии наркотическими препаратами для подавления боли. Первая группа включает 45 пациентов подверженных терапии Морфином, вторая группа – 45 пациентов с терапией Фентанилом. Цепочка генотипа представляет собой совокупность из 13 видов генов.

Входными переменными являются следующие 57 признаков (данные пациентов), представленных в таблице 1.

Табл. 1 Описание набора данных

|  |  |
| --- | --- |
| **Количество признаков** | **Название признаков** |
| 6 – Общие: | Пол, Возраст, Стадия, Локализация, Патогенетический вариант боли, Название операции |
| 13 - Цепочка генов: | rs1799971, rs1045642, rs2032582, rs1128503, rs1143627, rs5275, rs1800795, rs2740574, rs7668258, rs12233719, rs7438135, rs35599367, rs776746 |
| 7 | ХБСстарт (интенсивность боли на старте исследования), Суточная доза морфина сульфата на старте включения в исследование, Индекс массы тела 0, Желтуха, Диспепсия, Рак-обусловл. слабость, Асцит, Коморбидность (количество сопутствующей патологии) |
| 2 – Физический статус: | Эког 0, Эког 6 |
| 2 – Психический статус: | Ммса 0, Ммса 6 |
| 2 - Шкала качества жизни: | ESAS 0, ESAS 6 |
| 24 | Гемоглобин 0, Гемоглобин 6, Эритроциты 0, Эритроциты 6, Тромбоциты 0, Тромбоциты 6, Лейкоциты 0, Лейкоциты 6, Лимфоциты 0, Лимфоциты 6, Общий белок 0, Общий белок 6, Общий билирубин 0, Общий билирубин 6, АЛТ 0, АЛТ 6, АСТ 0, АСТ 6, Глюкоза 0, Глюкоза 6, Амилаза 0, Амилаза 6, Скф 0, Скф 6 |

Исходом (выходом) являются:

- *резистентность* (сопротивляемость) пациентов с данным генотипом к терапии, (бинарный показатель, где 0 - эффективное лечение, 1 - резистентное лечение (сопротивление));

- *появление неблагоприятных побочных реакций (НПР)* (бинарный показатель, где 0 – отсутствие реакции, 1 – наличие).

В качестве НПР представлены такие показатели как: *Констипация, Тошнота/Рвота, Седация, Слабость, Сухость во рту, Головокружение, Дезориентация, Кожный зуд, Затрудненное мочеиспускание, Местные реакции.*

По базе данных пациентов, основываясь на результатах примененной для них терапии, необходимо определить существует ли зависимость исхода от комбинации входных учитываемых признаков.

На основе представленной базы данных было решено 7 задач со следующими условиями:

1. Терапия производилась Морфином. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось *бинарное значение резистентности* (сопротивляемость) пациентов с данным генотипом к терапии.
2. Терапия производилась Морфином. В качестве входа использовался весь набор признаков, но уже с исключением показателя *“Суточная доза морфина сульфата на старте включения в исследование”* (56 признаков). В качестве выхода также использовалось *бинарное значение резистентности* (сопротивляемость) пациентов с данным генотипом к терапии.
3. Терапия производилась Морфином. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось *бинарное значение показателя появления НПР в виде констипации.*
4. Терапия производилась Морфином. В качестве входа использовался весь набор из 57 признаков. В качестве выхода *использовалось объединение выходов всех показателей НПР для каждого пациента в бинарный признак*, имеющий значение 0 – в случае отсутствия любой НПР, 1 – присутствует хотя бы одна НПР.
5. Терапия производилась Фентонилом. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось *бинарное значение резистентности* (сопротивляемость) пациентов с данным генотипом к терапии.
6. Терапия производилась Фентонилом. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось *бинарное значение показателя появления НПР в виде седации.*
7. Терапия производилась Фентонилом. В качестве входа использовался весь набор из 57 признаков. В качестве выхода *использовалось объединение выходов всех показателей НПР для каждого пациента в бинарный признак*, имеющий значение 0 – в случае отсутствия любой НПР, 1 – присутствует хотя бы одна НПР.

## **Предложенный подход**

В данной работе предлагается для поиска набора информативных признаков использовать генетический алгоритм многокритериальной оптимизации, в котором два критерия сводятся к одному обобщенному с помощью метода «аддитивной свертки». Сформированная комбинация входных учитываемых признаков, влияющая на исход, используется при построении модели классификатора, и дальнейшей ее оценке.

На рисунке 1 представлена общая схема описанного выше подхода.

Для оценки модели ГА использовался подход, называемый «Стратифицированная k-блочная кросс-валидация». Исходная выборка разбивается на k=5 равных блоков/частей (стратифицированное пропорциональное разбиение). Поочередно каждый блок рассматривается, как тестовая выборка, а остальные k-1 блоков – как обучающая выборка. Модель, используемая при оценке пригодности индивида (набор текущих признаков) в ГА, обучается на k-1 блоках и проверяется на тестовом блоке. Результат модели оценивается с помощью показателя F1- меры. Процесс повторяется k раз, и мы получаем k оценок, для которых рассчитывается среднее значение, являющееся итоговой оценкой модели.

По окончанию работы алгоритма ГА формирует лучший набор признаков, который в дальнейшем используется для обучения модели классификатора и его валидации. Для обучение данной модели производится разбиение исходной выборки на обучающую (Train: 70% от исходного объема выборки) и тестовую (Test: 30% от исходного объема выборки).

Обучение модели производится на всем обучающем множестве *Train*. Следующий шаг состоит в проверке качества модели и ее обобщающей способности на реальных данных, не участвующих в обучении, то есть на тестовом наборе. Соответственно оценка модели производится на тестовой выборке *Test.*

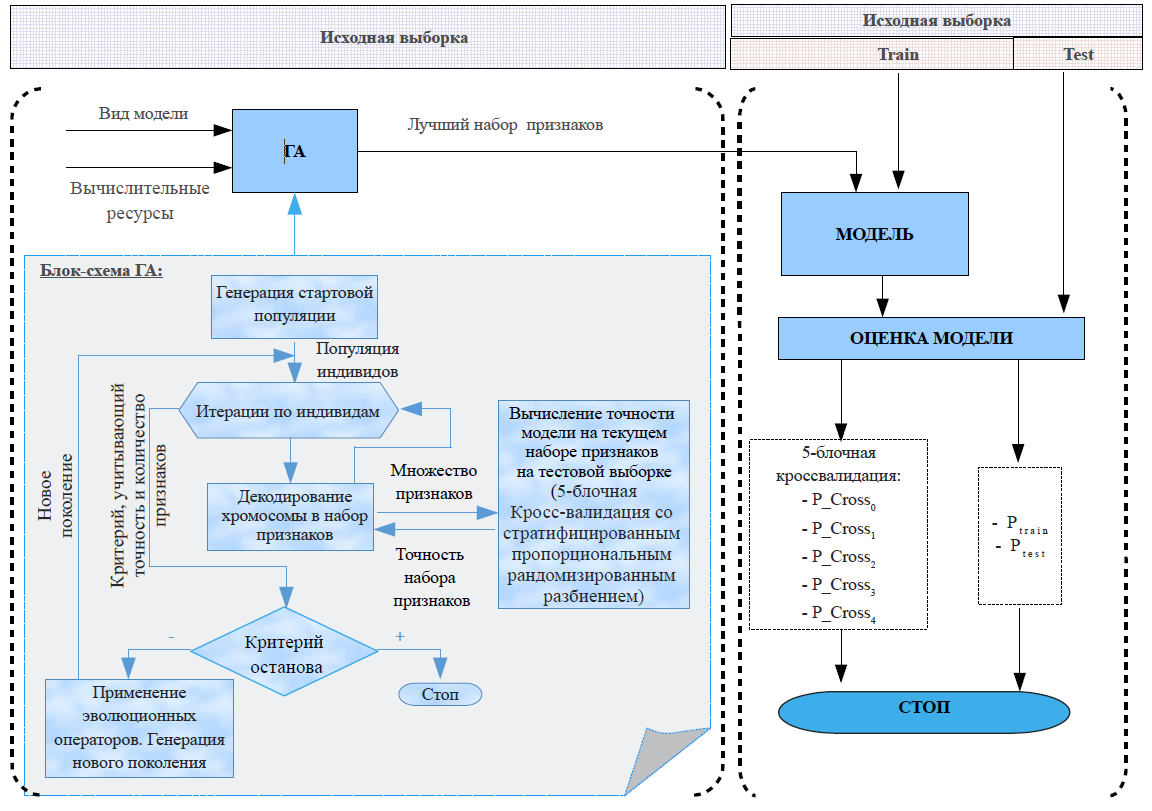


Рис.1 Общая схема подхода для решения задачи исследования влияния на исход с процедурой отбора информативных признаков

Также на рисунке 1 представлена оценка модели с наилучшим набором признаков, где Ptrain - точность модели с использованием лучшего набора признаков, полученных на основе ГА на обучающей выборке, соответственно Ptest – на тестовой.

Далее также использовался подход «Стратифицированная k-блочная кросс-валидация» но только для обучающей выборки. Обучающая выборка была разделена на 5 равных блоков (частей), 4 из которых отводились для обучения модели и 1 блок предназначался для ее валидации. И полученные результаты были оценены и усреднены, чтобы убедиться, что качество нашей модели не зависит от конкретного набора данных. На рисунке 1 *P\_Crossk* - это точность на *k*-м блоке, когда он является тестовым. Затем происходит оценка точности модели на множестве *Test.*

Критерием эффективности модели также является F1-мера.

## **Отбор информативных признаков с помощью генетического алгоритма (ГА)**

Генетические алгоритмы — это адаптивные методы поиска, которые широко используются для решения задач оптимизации. В них используются как аналог механизма генетического наследования, так и аналог естественного отбора. При этом сохраняется биологическая терминология в упрощенном виде и основные понятия линейной алгебры. Поскольку генетические алгоритмы используют биологические аналогии, то и применяющаяся терминология напоминает биологическую. Так, одно пробное решение, записанное в двоичной форме, мы будем называть особью или хромосомой, а набор всех пробных решений – популяцией.

Такие термины как ген, генотип, хромосома в этой статье встречаются в двух смыслах – гены как значение генов определенного пациента, и гены как традиционно сложившаяся терминология в рамках генетических алгоритмов. Генетические алгоритмы относятся к области искусственного интеллекта и к биологии имеют косвенное отношение. Поэтому в дальнейшем эти термины будут рассматриваться в качестве искусственных (символических) аналогов с использованием методов естественной эволюции, таких как наследование, мутации, отбор и скрещивание.

ГА используется для поиска набора информативных признаков. На рисунке 2 представлено кодирование хромосомы ГА (набора признаков), которое производится в 2 этапа.

*1 этап*: *Представление исходной выборки (таблицы данных) в признаковом пространстве.* Первая и вторая строка на рисунке 2 не являются частью хромосомы, а представляют собой маску, позволяющую производить декодирование хромосомы и представления ее в виде готового набора признаков.

Первая строка представляет собой строку в исходной таблице, длина которой равна исходному количеству признаков (57). Для каждого признака указаны возможные значения переменных. Из рисунка 2 видно, что часть переменных являются количественными, а часть – категориальными.

Вторая строка представляет собой строку в измененном признаковом пространстве, в котором количественные переменные таблицы с исходными данными не изменяются, а категориальные переменные, принимающие только два значения, кодируются 0 и 1. В противном случае для кодирования категориальных признаков отводится столько бит, сколько значений имеет данный показатель.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Пол (М / Ж)** | **Возраст (Ai)** | ***Стадия***  ***(*2a / 2b / 3 / 4*)*** | | | | ***ХБСстарт (*Bi)** | ***Пат. вар. боли (*0 / 1)** | ***Назв. операции (*0 / 1)** | ***rs1799971***  ***(*AA / AG / GG)** | | | **…** | ***Коморбидность (*Сi)** |
| **М** | **Ai** | **2b** | | | | **Bi** | **1** | **0** | **AA** | | | **…** | **Ci** |
| **Пол** | **Возраст** | ***Стадия (2a)*** | ***Стадия (2b)*** | ***Стадия (3)*** | ***Стадия (4)*** | ***ХБСстарт*** | ***Пат. вар. боли*** | ***Назв. операции*** | ***rs1799971 (AA)*** | ***rs1799971 (AG)*** | ***rs1799971 (GG)*** | **…** | ***Коморбидность (*Сi)** |
| **М** | **60** | **0** | **1** | **0** | **0** | **9** | **1** | **0** | **1** | **0** | **0** | **…** | **6** |
| ***Перевод фенотипа в генотип*** | | | | | | | | | | | | | |
| **0** | **1** | **1** | **0** | **1** | **0** | **0** | **0** | **1** | **0** | **1** | **1** | **…** | **0** |

Рис.2 Способ кодирования комбинации генов для определения исхода в ГА

Например, для переменной “rs1799971”, принимающей одно из трех значений AA, AG или GG - будет отведено три бита. Соответственно этот признак будет представлен в строке таблицы признакового пространства одной из трех комбинаций: 100 – AA, 010 – AG, 001 – GG.

Рассмотрим переменную “Стадия”, которая принимает одно из четырех значений: 2a, 2b, 3, 4. Тогда для этого признака в признаковом пространстве будет отведено четыре бита и в строке таблицы выражено одной из четырех комбинаций: 1000 - 2a, 0100 - 2b, 0010 - 3, 0001 - 4.

Длина второй строки составляет 82 бита в случае использования всех входных признаков, и 81 – в случае исключения показателя *“Суточная доза морфина сульфата на старте включения в исследование”.*

*2 этап*: *Представление признаков в пространстве поиска ГА.* Так как вторая строка указывает количество признаков, на каждый из которых отводится по одному биту (в новом признаковом пространстве), соответственно в третьей строке это отражается таким образом, что значение 1 означает что признак учитывается при построении модели, а 0 – когда не учитывается. Длина хромосомы составляет 82 (81) бита.

Оценка эффективности ГА проводится на основе следующей функции пригодности:

где - точность модели по критерию F1, а – количество признаков, полученных в хромосоме ГА. Такой критерий обусловлен тем, что необходимо обеспечить точность работы модели, за что отвечает 1 слагаемое, а также минимизировать кол-во признаков, за что отвечает 2 слагаемое. Наибольшее значение критерий достигает при одновременном достижении высокой точности модели при минимальном наборе признаков.

Для определения полноты и точности введем вспомогательные понятия:

1. Ttrue positive (TP): Количество истинно положительных результатов, т.е. таких объектов, которые относятся к положительному классу и модель отнесла их к положительному классу.
2. True negative (TN): Количество истинно отрицательных результатов, т.е. таких объектов, которые относятся к отрицательному классу, и модель их отнесла к отрицательному классу.
3. False positive (FP): Количество ложно положительных результатов, т.е. таких объектов, которые относятся к отрицательному классу, но модель их отнесла к положительному классу.
4. False negative (FN): Количество ложно отрицательных результатов, т.е. таких объектов, которые относятся к положительному классу, но модель их отнесла к отрицательному классу.

*Матрица ошибок (Confusion matrix)* — это способ разбить объекты на четыре категории в зависимости от истинного класса и класса, к которому модель отнесла объект выборки. В таблице 2 приводится пример матрицы ошибок.

Табл. 2 Пример матрицы ошибок

|  |  |  |  |
| --- | --- | --- | --- |
| N=165 | **Predicted: No** | **Predicted: No** |  |
| **Actual: No** | TN = 50 | FP = 10 | 60 |
| **Actual: Yes** | FN = 5 | TP = 100 | 105 |
|  | 55 | 110 |  |

Точность () определяется как отношение числа корректно классифицированных элементов к общему числу элементов:

Полнота () это отношение числа найденных элементов целевого класса, к общему числу элементов целевого класса:

F-мера (F − measure, F1 − score) позволяет объединить точность и полноту в одной усреднённой величине. F-мера представляет собой гармоническое среднее между точностью и полнотой, которая стремится к нулю, если точность или полнота стремится к нулю:

В качестве моделей были использованы следующие алгоритмы классификации, построенные на основе библиотеки «Scikit-learn» (Python):

1. Логистическая регрессия (LR)
2. Алгоритм k – ближайших соседей для задачи классификации (KNC).
3. Коллектив деревьев принятия решений методом «случайного леса» для задачи классификации (RFC).
4. Коллектив деревьев решений методом градиентного бустинга для задачи классификации (GBС).
5. Деревья решений для задачи классификации (DTC). Пример дерева представлен на рисунке 4.
6. Искусственная нейронная сеть (многослойный персептрон) для задачи классификации (MLPC).
7. Линейный метод опорных векторов для задачи классификации (LSVC).
8. Метод опорных векторов для задачи классификации (SVC).

На рисунке 4 представлен пример визуализации дерева решений для задачи исследования влияния на исход с процедурой отбора информативных признаков. Во всех узлах, кроме листьев (узлы без исходящих связей), содержится 4 части:

1. *Условие выбора поддерева.*  Если условие истинно, то переход на левую ветку (True), иначе на правую (False). Это точка разделения узла, в зависимости от ответа определяется, в каком направлении вниз по дереву продвинется образец данных.
2. *Gini*: Gini - метрика качества. Этот показатель должен уменьшаться по мере того, как мы движемся вниз по дереву.
3. *Samples*: количество прошедших через этот узел образцов.
4. *Value*: отношение классов, прошедших через этот узел, выраженное в абсолютных числах. К примеру, верхний узел выделил 16 образцов класса 0 и 13 образцов класса 1.

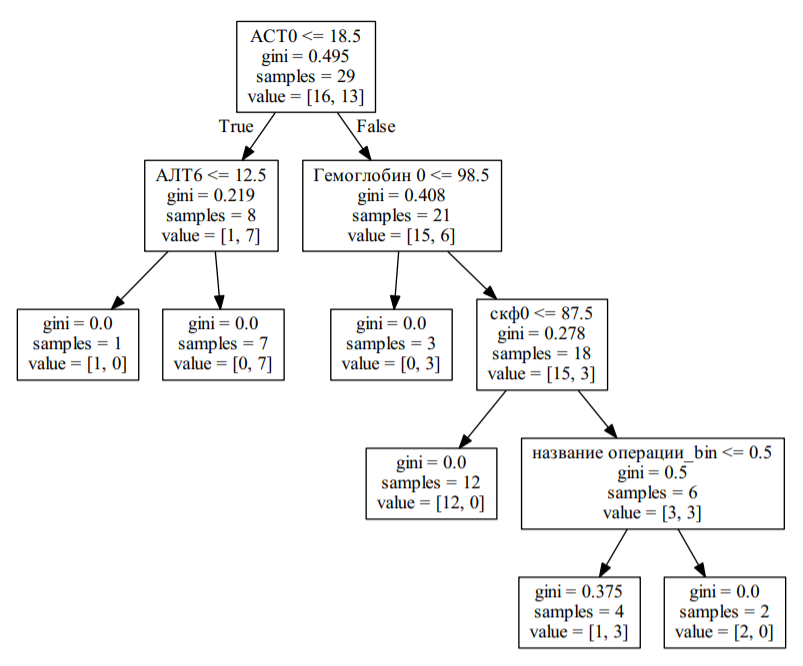


Рис. 4 Пример дерева решений для определения исхода (показатель – констипация) при терапии Морфином

Листья не содержат вопроса, так как являются финальными прогнозируемыми значениями классификации. Чтобы обработать новый элемент набора данных, нужно просто двигаться вниз по дереву, используя параметры элемента для ответов на вопросы.

## **Результаты экспериментов**

ГА для отбора информативных признаков запускался со следующими параметрами: равномерное скрещивание, турнирная селекция (размер турнира - 5), средняя мутация, элитизм.

Этапы ГА моделируют «эволюционный процесс», продолжающийся несколько жизненных циклов (поколений), пока не будет выполнен критерий остановки алгоритма. В данной работе критерием остановки работы ГА являлось окончание числа поколений, отпущенных на эволюцию. В данной работе количество поколений – 50, количество индивидов – 70. Все результаты усреднялись.

Рассмотрим результаты решения задачи №1, в которой терапия производилась Морфином. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось *бинарное значение резистентности* (сопротивляемость) пациентов с данным генотипом к терапии.

В таблице 3 представлены результаты вычисления следующих характеристик: среднее значение по 50 прогонам (*mean*) и среднеквадратическое отклонение по 50 прогонам (*std*), определяемых по выборкам для каждой переменной отдельной моделью:

1. *n\_feature\_mean* – среднее число отобранных значимых признаков в результате работы алгоритма по 50 прогонам;
2. *n\_feature\_std* – среднеквадратическое отклонение числа отобранных значимых признаков в результате работы алгоритма по 50 прогонам;
3. *Ptrain\_mean*– среднее значение точности алгоритма отбора значимых признаков, вычисленная на обучающей выборке по 50 прогонам;
4. *Ptrain\_std*– среднеквадратическое отклонение значения точности алгоритма отбора значимых признаков, вычисленная на обучающей выборке по 50 прогонам;
5. *Ptest\_mean*– среднее значение точности алгоритма отбора значимых признаков, вычисленная на тестовой выборке по 50 прогонам;
6. *Ptest\_std*– среднеквадратическое отклонение значения точности алгоритма отбора значимых признаков, вычисленная на тестовой выборке по 50 прогонам;
7. *P\_Cross\_mean* – среднее значение по 5-блочной кросс-валидации по 50 прогонам;
8. *P\_Cross\_std* – среднеквадратическое отклонение среднего значения по 5-блочной кросс-валидации по 50 прогонам.

Табл. 3 Описательные статистики (задача №1)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | n\_feature\_mean | n\_feature\_std | Ptrain\_mean | Ptrain\_std | Ptest\_mean | Ptest\_std | P\_Cross\_mean | P\_Cross\_std |
| DTC | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| GBC | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| KNC | 1,2 | 1,41 | 0,91 | 0,01 | 0,91 | 0 | 0,92 | 0 |
| LR | 3,78 | 0,93 | 1 | 0 | 0,99 | 0,05 | 1 | 0 |
| LSVC | 3,34 | 0,94 | 1 | 0 | 0,99 | 0,04 | 1 | 0,01 |
| MLPC | 24,74 | 3,86 | 0,83 | 0,25 | 0,83 | 0,23 | 0,82 | 0,1 |
| RFC | 12,2 | 4,65 | 1 | 0,01 | 0,89 | 0,14 | 0,93 | 0,06 |
| SVC | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

По результатам, представленным в таблице 3 можно сделать вывод о том, что модели DTC, GBC, SVC решают поставленную задачу с точностью равной 1 и используют при этом всего один признак. Как показано в таблице 3 между этими моделями нет статистически значимого различия. Однако выделяемый признак – доза морфина. Такой эффект получается не из прямой связи, что резистентность определяется дозой, а наоборот, что врач назначает дозу с учетом признаков резистентности. Поэтому, такая модель не является полезной. Для исключения этого эффекта уберем из выборки признак «Доза морфина» (результаты представлены в задаче №2).

Рассмотрим результаты решения задачи №2, в которой терапия производилась Морфином, представленные в таблицах 4 и 5. В качестве входа использовался весь набор признаков, но уже с исключением показателя “Суточная доза морфина сульфата на старте включения в исследование” (56 признаков). В качестве выхода также использовалось бинарное значение резистентности (сопротивляемость) пациентов с данным генотипом к терапии.

Табл. 4 Описательные статистики (задача №2)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | n\_feature\_mean | n\_feature\_std | Ptrain\_mean | Ptrain\_std | Ptest\_mean | Ptest\_std | P\_Cross\_mean | P\_Cross\_std |
| DTC | 16,58 | 6,37 | 1 | 0 | 0,75 | 0,15 | 0,82 | 0,05 |
| GBC | 9,0 | 2,59 | 1 | 0 | 0,62 | 0,11 | 0,86 | 0,03 |
| KNC | 5,24 | 1,59 | 0,87 | 0,05 | 0,63 | 0,09 | 0,84 | 0,01 |
| LR | 14,12 | 2,75 | 0,99 | 0,02 | 0,66 | 0,16 | 0,91 | 0,03 |
| LSVC | 31,96 | 3,75 | 0,83 | 0,1 | 0,62 | 0,15 | 0,73 | 0,09 |
| MLPC | 32,3 | 4,16 | 0,83 | 0,25 | 0,55 | 0,14 | 0,63 | 0,11 |
| RFC | 29,86 | 5,81 | 0,98 | 0,03 | 0,49 | 0,1 | 0,63 | 0,09 |
| SVC | 5,12 | 3,54 | 0,91 | 0,05 | 0,64 | 0,16 | 0,76 | 0,07 |

Для оценки статистической значимости полученных результатов каждой моделью используется непараметрический статистический тест - U-критерий Манна-Уитни (уровень значимости – 0.05). В таблице 5 представлены результаты сравнительного анализа критерия U-тест Манна-Уитни. Критерий возвращает значение *p*, которое говорит о том, на сколько выборочные распределения переменной *Ptest* отличаются. Чем ближе *p* к единице, тем меньше сила различия. Прочерк в таблице означает, что в полученных выходах модели отсутствует разброс.

Табл. 5 Проверка статистической значимости результатов между моделями (задача №2)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *U-тест Манна-Уитни (р)* | | | | *n\_feature\_mean* | | | | | | | | | | | | | | | |
| **29,86** | | **16,58** | | **5,12** | | **31,96** | | **9,0** | | **14,12** | | **32,3** | | **5,24** | |
| **RFC** | | **DTC** | | **SVC** | | **LSVC** | | **GBC** | | **LR** | | **MLPC** | | **KNC** | |
| *n\_feature\_mean* | **29,86** | **RFC** | 1,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,18 | | 0,0 | |
| **16,58** | **DTC** | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | |
| **5,12** | **SVC** | 0,0 | | 0,0 | | 1,0 | | 0,19 | | 0,05 | | 0,49 | | 0,0 | | 0,09 | |
| **31,96** | **LSVC** | 0,0 | | 0,0 | | 0,19 | | 1,0 | | 0,27 | | 0,25 | | 0,01 | | 0,31 | |
| **9,0** | **GBC** | 0,0 | | 0,0 | | 0,05 | | 0,27 | | 1,0 | | 0,09 | | 0,01 | | 0,17 | |
| **14,12** | **LR** | 0,0 | | 0,0 | | 0,49 | | 0,25 | | 0,09 | | 1,0 | | 0,01 | | 0,14 | |
| **32,3** | **MLPC** | 0,18 | | 0,0 | | 0,0 | | 0,01 | | 0,01 | | 0,01 | | 1,0 | | 0,0 | |
| **5,24** | **KNC** | 0,0 | | 0,0 | | 0,09 | | 0,31 | | 0,17 | | 0,14 | | 0,0 | | 1,0 | |

Результаты таблиц 4 и 5 можно структурировать следующим образом. Наилучшей моделью по точности на тестовой выборке является DTC. При этом результат модели DTC имеет статистически значимое отличие от результатов всех остальных моделей. Далее идет группа моделей GBC, KNC, LR, LSVC, SVС, точности работы которых примерно одинаковые. На основании таблицы 5 можем сделать вывод о том, что различия между этими моделями являются статистически незначимыми и объясняются действиями случайных факторов. Хуже всех справились модели MLP и RFC. Эти модели значимо отличаются от всех остальных и не значимо отличаются между собой. В качестве результирующей модели принимается модель DTC. Кроме высокой точности эта модель обладает еще одним существенным преимуществом – работа данной модели легко интерпретируется на языке рассуждения эксперта (врача) и может быть оценена и скорректирована экспертом вручную. Отдельное внимание заслуживают модели SVC и KNC. Эти модели используют наименьшее количество признаков. При этом разброс по количеству признаков у KNC меньше. При этом KNC тоже поддается интерпретации на естественном языке рассуждения эксперта, т.к. строит свой вывод на основании похожих прецедентов из обучающей выборки. Эта модель также заслуживает внимания и может быть использована для быстрой (предварительной) диагностики пациента.

Для оценки значимости признаков, полученных лучшей моделью DTC используется **метод перестановок**. Метод перестановок представляет собой последовательное взятие каждого признака и перемешивание его значений в рамках столбца. Затем смотрится на сколько это перемешивание повлияло на выход.

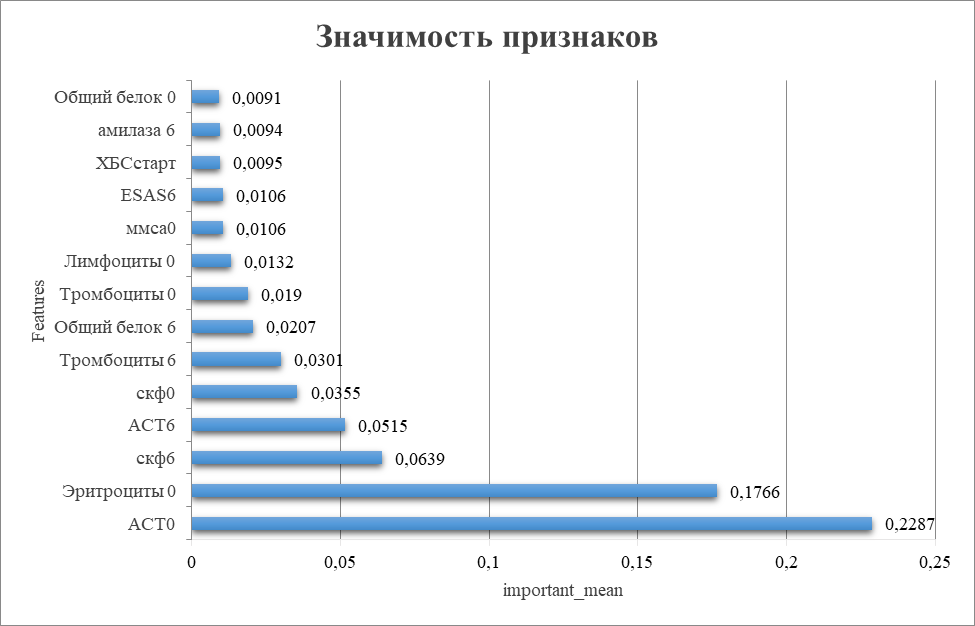


Рис. 5 Значимость признаков, отобранных DTC при решении задачи №2

Табл. 6. Значимость признаков, отобранных DTC при решении задачи №2   
(показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **features** | **important\_mean** |
| DTC | АСТ0 | **0,2287** |
| DTC | Эритроциты 0 | **0,1766** |
| DTC | скф6 | 0,0639 |
| DTC | АСТ6 | 0,0515 |
| DTC | скф0 | 0,0355 |
| DTC | Тромбоциты 6 | 0,0301 |
| DTC | Общий белок 6 | 0,0207 |
| DTC | Тромбоциты 0 | 0,019 |
| DTC | Лимфоциты 0 | 0,0132 |
| DTC | ммса0 | 0,0106 |
| DTC | ESAS6 | 0,0106 |
| DTC | ХБСстарт | 0,0095 |
| DTC | амилаза 6 | 0,0094 |
| DTC | Общий белок 0 | 0,0091 |
| DTC | эког 6\_bin | 0,0081 |
| DTC | амилаза 0 | 0,0061 |
| DTC | rs1800795\_GG | 0,0047 |
| DTC | ESAS0 | 0,0044 |
| DTC | локализация\_1 | 0,0043 |
| DTC | гемоглобин6м | 0,0027 |
| DTC | rs1799971\_AA | 0,0025 |
| DTC | патогенетический вариант боли\_bin | 0,002 |
| DTC | асцит | 0,0019 |
| DTC | коморбидность | 0,0015 |
| DTC | rs1143627\_AG | 0,0012 |
| DTC | rs1799971\_GG | 0,0011 |
| DTC | rs1045642\_AG | 0,001 |
| DTC | желтуха | 0,0009 |

Наибольший вклад в точность модели обеспечивают признаки АСТ0, Эритроциты 0, скф6, АСТ6, Тромбоциты 6, Общий белок 6, Тромбоциты 0, Лимфоциты 0, ммса 0, ESAS6. При многократных прогонах алгоритма, часть признаков: АСТ0, Эритроциты 0 включаются в модель на каждом прогоне, а остальная часть признаков в разных сочетаниях. В среднем модель использует около 6 признаков.

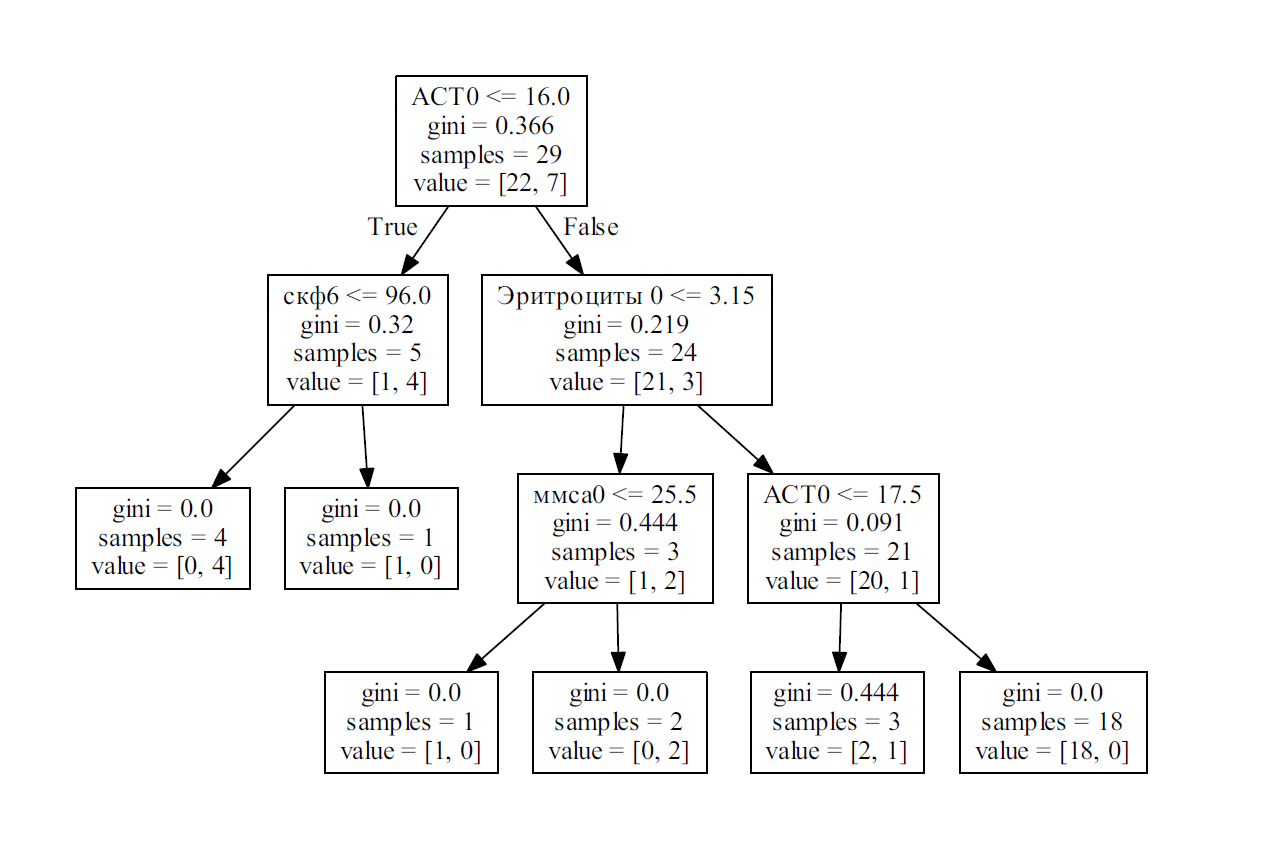


Рис. 6 Пример дерева решений для определения исхода (показатель – резистентность) при терапии Морфином

Рассмотрим результаты решения задачи №3, в которой терапия производилась Морфином, представленные в таблицах 7 и 8. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось бинарное значение показателя появления НПР в виде констипации.

Табл. 7 Описательные статистики (задача №3)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | n\_feature\_mean | n\_feature\_std | Ptrain\_mean | Ptrain\_std | Ptest\_mean | Ptest\_std | P\_Cross\_mean | P\_Cross\_std |
| DTC | 18,62 | 6,68 | 1 | 0 | 0,77 | 0,11 | 0,84 | 0,06 |
| GBC | 21,86 | 4,71 | 1 | 0 | 0,79 | 0,1 | 0,88 | 0,03 |
| KNC | 4,98 | 2,95 | 0,76 | 0,07 | 0,67 | 0,09 | 0,77 | 0,05 |
| LR | 16,04 | 4,49 | 0,98 | 0,03 | 0,67 | 0,1 | 0,91 | 0,02 |
| LSVC | 34,36 | 3,26 | 0,9 | 0,09 | 0,65 | 0,09 | 0,72 | 0,08 |
| MLPC | 29,86 | 4,58 | 0,95 | 0,15 | 0,7 | 0,1 | 0,75 | 0,07 |
| RFC | 28,9 | 4,64 | 0,98 | 0,03 | 0,68 | 0,14 | 0,66 | 0,07 |
| SVC | 12,28 | 4,79 | 0,94 | 0,06 | 0,63 | 0,12 | 0,77 | 0,04 |

В таблице 8 представлены результаты проверки значимости признаков на результирующих моделях.

Табл. 8 Проверка статистической значимости результатов между моделями (задача №3)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *U-тест Манна-Уитни (р)* | | | | *n\_feature\_mean* | | | | | | | | | | | | | | | |
| **28,9** | | **18,62** | | **12,28** | | **34,36** | | **21,86** | | **16,04** | | **29,86** | | **4,98** | |
| **RFC** | | **DTC** | | **SVC** | | **LSVC** | | **GBC** | | **LR** | | **MLPC** | | **KNC** | |
| *n\_feature\_mean* | **28,9** | **RFC** | 1,0 | | 0,0 | | 0,01 | | 0,06 | | 0,0 | | 0,35 | | 0,15 | | 0,21 | |
| **18,62** | **DTC** | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,34 | | 0,0 | | 0,0 | | 0,0 | |
| **12,28** | **SVC** | 0,01 | | 0,0 | | 1,0 | | 0,06 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | |
| **34,36** | **LSVC** | 0,06 | | 0,0 | | 0.06 | | 1,0 | | 0,0 | | 0,16 | | 0,0 | | 0,1 | |
| **21,86** | **GBC** | 0,0 | | 0,34 | | 0,0 | | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | |
| **16,04** | **LR** | 0,35 | | 0,0 | | 0,0 | | 0,16 | | 0,0 | | 1,0 | | 0,01 | | 0,5 | |
| **29,86** | **MLPC** | 0,15 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,01 | | 1,0 | | 0,0 | |
| **4,98** | **KNC** | 0,21 | | 0,0 | | 0,0 | | 0,1 | | 0,0 | | 0,5 | | 0,0 | | 1,0 | |

Результаты моделей, представленные в таблицах 7 и 8 можно разбить на три группы.

1. Наилучшим образом справились модели DTC и GBC. Результаты этих моделей различаются статистически не значимо и обе модели могут быть использованы в качестве результирующих.
2. Вторая группа моделей со средним результатом KNC, LR, LSVC, SVC. Различия между ними статистически не значимы, а от других моделей они отличаются значимо. Как и в предыдущей задачи KNC при этом использует наименьшее количество признаков и может быть использован для быстрой диагностики.
3. Наихудший результат снова дали модели MLPC, RFC.

Проведем анализ значимости признаков отдельно для модели DTC И GBC.

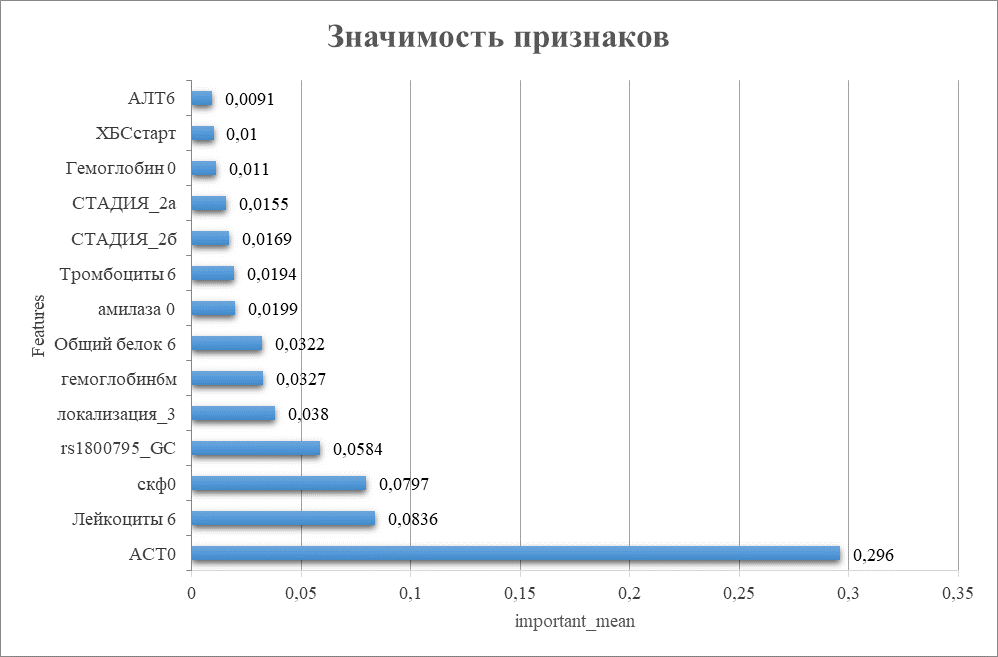


Рис. 7 Значимость признаков, отобранных DTC при решении задачи №3

Табл. 9. Значимость признаков, отобранных DTC при решении задачи №3   
(показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **Features** | **important\_mean** |
| DTC | **АСТ0** | **0,296** |
| DTC | Лейкоциты 6 | 0,0836 |
| DTC | скф0 | 0,0797 |
| DTC | rs1800795\_GC | 0,0584 |
| DTC | локализация\_3 | 0,038 |
| DTC | гемоглобин6м | 0,0327 |
| DTC | Общий белок 6 | 0,0322 |
| DTC | амилаза 0 | 0,0199 |
| DTC | Тромбоциты 6 | 0,0194 |
| DTC | СТАДИЯ\_2б | 0,0169 |
| DTC | СТАДИЯ\_2а | 0,0155 |
| DTC | Гемоглобин 0 | 0,011 |
| DTC | ХБСстарт | 0,01 |
| DTC | АЛТ6 | 0,0091 |
| DTC | Общий билирубин6 | 0,0072 |
| DTC | ммса6 | 0,007 |
| DTC | rs2032582\_GT | 0,0069 |
| DTC | Лимфоциты 0 | 0,0067 |
| DTC | суточная доза морфина 0 | 0,0049 |
| DTC | rs1143627\_AA | 0,0047 |
| DTC | амилаза 6 | 0,0037 |
| DTC | АЛТ0 | 0,0032 |
| DTC | Общий белок 0 | 0,0029 |
| DTC | название операции\_bin | 0,0028 |
| DTC | Пол\_bin | 0,0026 |
| DTC | Лейкоциты 0 | 0,0026 |
| DTC | rs1800795\_CC | 0,0024 |
| DTC | rs1045642\_GG | 0,0023 |
| DTC | локализация\_1 | 0,0022 |
| DTC | рак-обусловл.слабость | 0,0019 |
| DTC | имт0 | 0,0018 |
| DTC | коморбидность | 0,0018 |
| DTC | диспепсия | 0,0016 |
| DTC | rs1799971\_AG | 0,0014 |
| DTC | rs7438135\_AA | 0,0013 |
| DTC | асцит | 0,0013 |
| DTC | rs1045642\_AG | 0,0013 |
| DTC | rs7668258\_CC | 0,0012 |
| DTC | rs7668258\_TT | 0,0012 |
| DTC | Глюкоза 6 | 0,001 |
| DTC | rs1799971\_AA | 0,0009 |
| DTC | rs1800795\_GG | 0,0008 |
| DTC | rs776746\_bin | 0,0008 |
| DTC | желтуха | 0,0008 |
| DTC | АСТ6 | 0,0008 |
| DTC | патогенетический вариант боли\_bin | 0,0006 |

Основной вклад в точности модели обеспечивает признак АСТ0. В среднем модель использует около 7 признаков. Другие признаки встречаются в разных сочетаниях. При этом перечень признаков достаточно широкий. Влияние признаков с уровнем 0.009 и ниже можно считать незначительным (случайным).

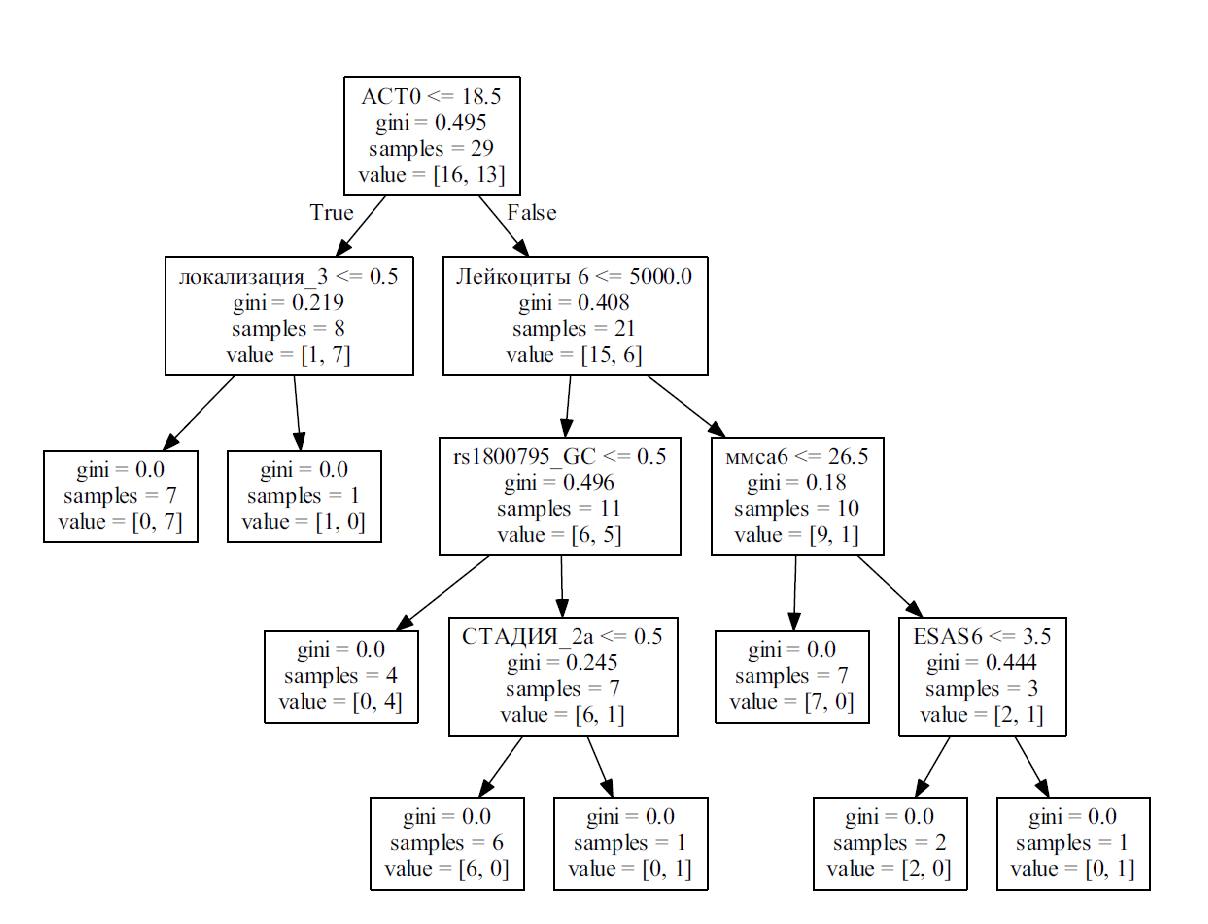


Рис. 8 Пример дерева решений для определения исхода (показатель – констипация) при терапии Морфином

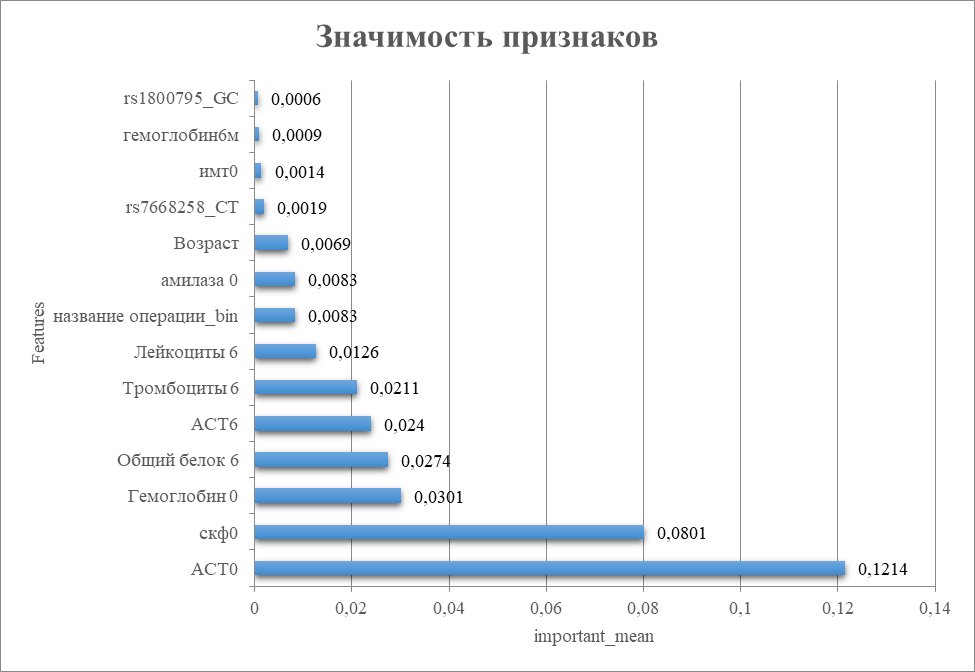


Рис. 9 Значимость признаков, отобранных GBC при решении задачи №3

Табл. 10. Значимость признаков, отобранных GBC при решении задачи №3   
(показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **features** | **important\_mean** |
| GBC | **АСТ0** | **0,1214** |
| GBC | скф0 | 0,0801 |
| GBC | Гемоглобин 0 | 0,0301 |
| GBC | Общий белок 6 | 0,0274 |
| GBC | АСТ6 | 0,024 |
| GBC | Тромбоциты 6 | 0,0211 |
| GBC | Лейкоциты 6 | 0,0126 |
| GBC | название операции\_bin | 0,0083 |
| GBC | амилаза 0 | 0,0083 |
| GBC | Возраст | 0,0069 |
| GBC | rs7668258\_CT | 0,0019 |
| GBC | имт0 | 0,0014 |
| GBC | гемоглобин6м | 0,0009 |
| GBC | rs1800795\_GC | 0,0006 |

Наибольшее влияние на точность модели оказывают признаки АСТ0 и скф0. Кроме этого, значимое влияние оказывают признаки: Гемоглобин 0, Общий белок 6, АСТ6, Тромбоциты 6, Лейкоциты 6, название операции, амилаза 0, Возраст. В среднем модель использует около 4 признаков. Разброс признаков достаточно малый – на 50 прогонах модель использует всего 14 признаков.

Рассмотрим результаты решения задачи №4, в которой терапия производилась Морфином. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось объединение выходов всех показателей НПР для каждого пациента в бинарный признак, имеющий значение 0 – в случае отсутствия любой НПР, 1 – присутствует хотя бы одна НПР.

Табл. 11 Описательные статистики (задача №4)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | n\_feature\_mean | n\_feature\_std | Ptrain\_mean | Ptrain\_std | Ptest\_mean | Ptest\_std | P\_Cross\_mean | P\_Cross\_std |
| DTC | 15 | 4,47 | 1 | 0 | 0,76 | 0,06 | 0,68 | 0,16 |
| GBC | 11,4 | 3,59 | 1 | 0 | 0,78 | 0,07 | 0,77 | 0,11 |
| KNC | 4,9 | 2,68 | 0,6 | 0,1 | 0,71 | 0,11 | 0,31 | 0,18 |
| LR | 16,5 | 4,63 | 0,94 | 0,04 | 0,81 | 0,1 | 0,55 | 0,19 |
| LSVC | 34,88 | 4,5 | 0,89 | 0,09 | 0,66 | 0,07 | 0,7 | 0,17 |
| MLPC | 29,66 | 4,48 | 0,86 | 0,18 | 0,75 | 0,13 | 0,49 | 0,19 |
| RFC | 28,92 | 5,13 | 0,99 | 0,02 | 0,68 | 0,14 | 0,63 | 0,19 |
| SVC | 16,46 | 5,71 | 0,99 | 0,04 | 0,65 | 0,09 | 0,51 | 0,18 |

Табл. 12 Проверка статистической значимости результатов между моделями (задача №4)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *U-тест Манна-Уитни (р)* | | | | *n\_feature\_mean* | | | | | | | | | | | | | | | |
| **28,92** | | **15** | | **16,46** | | **34,88** | | **11,4** | | **16,5** | | **29,66** | | **4,9** | |
| **RFC** | | **DTC** | | **SVC** | | **LSVC** | | **GBC** | | **LR** | | **MLPC** | | **KNC** | |
| *n\_feature\_mean* | **28,92** | **RFC** | 1,0 | | 0,0 | | 0,03 | | 0,1 | | 0,0 | | 0,0 | | 0,0 | | 0,19 | |
| **15** | **DTC** | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,05 | | 0,0 | | 0,26 | | 0,0 | |
| **16,46** | **SVC** | 0,03 | | 0,0 | | 1,0 | | 0,31 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | |
| **34,88** | **LSVC** | 0,1 | | 0,0 | | 0,31 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | | 0,01 | |
| **11,4** | **GBC** | 0,0 | | 0,05 | | 0,0 | | 0,0 | | 1,0 | | 0,01 | | 0,31 | | 0,0 | |
| **16,5** | **LR** | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,01 | | 1,0 | | 0,01 | | 0,0 | |
| **29,66** | **MLPC** | 0,0 | | 0,26 | | 0,0 | | 0,0 | | 0,31 | | 0,01 | | 1,0 | | 0,01 | |
| **4,9** | **KNC** | 0,19 | | 0,0 | | 0,0 | | 0,01 | | 0,0 | | 0,0 | | 0,01 | | 1,0 | |

Результаты решения задачи №4 так же можно группировать. Наилучшей моделью здесь является LR. Она дает достаточно высокую точность и статистически значимо отличается от всех остальных (табл. 12). Хорошее решение так же дают модели DTC, GBC, MLPC. Эти модели значимо отличаются от всех остальных и не значимо отличаются между собой. Следующая группа моделей это RFC, KNC. При этом KNC снова использует меньше всех признаков.

Наихудший результат дают модели LSVC и SVC. В качестве результирующей модели принимается LR. Проведем оценку значимости признаков, полученных лучшей моделью – LR при решении задачи №4.

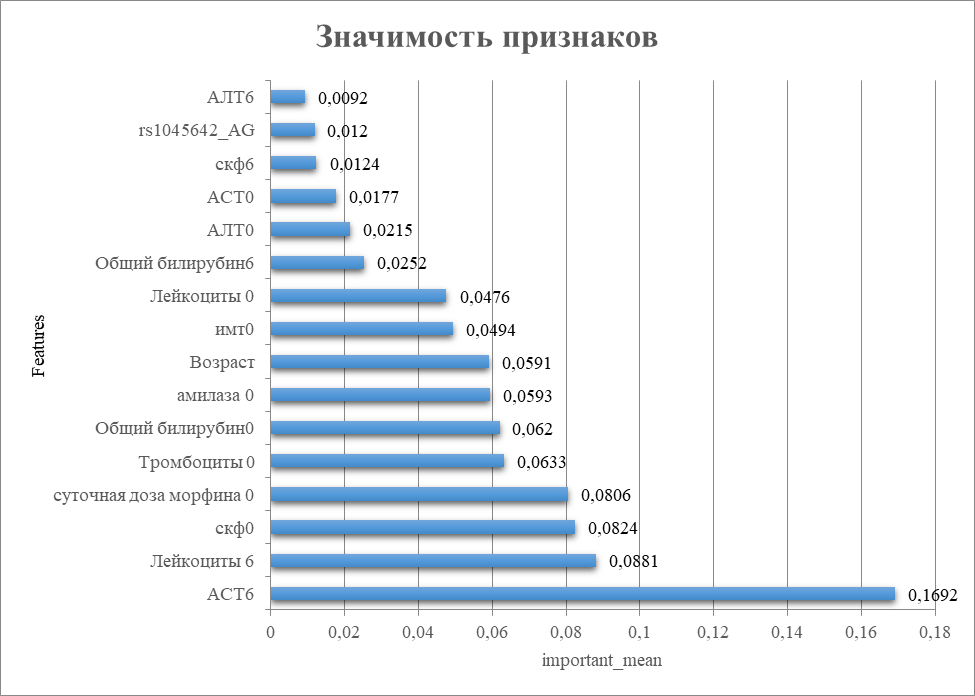
****

Рис. 10 Значимость признаков, отобранных LR при решении задачи №4

Табл. 13. Значимость признаков, отобранных LR при решении задачи №4   
(показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **features** | **important\_mean** |
| LR | **АСТ6** | **0,1692** |
| LR | Лейкоциты 6 | 0,0881 |
| LR | скф0 | 0,0824 |
| LR | суточная доза морфина 0 | 0,0806 |
| LR | Тромбоциты 0 | 0,0633 |
| LR | Общий билирубин0 | 0,062 |
| LR | амилаза 0 | 0,0593 |
| LR | Возраст | 0,0591 |
| LR | имт0 | 0,0494 |
| LR | Лейкоциты 0 | 0,0476 |
| LR | Общий билирубин6 | 0,0252 |
| LR | АЛТ0 | 0,0215 |
| LR | АСТ0 | 0,0177 |
| LR | скф6 | 0,0124 |
| LR | rs1045642\_AG | 0,012 |
| LR | АЛТ6 | 0,0092 |
| LR | Лимфоциты 0 | 0,0089 |
| LR | коморбидность | 0,0083 |
| LR | Общий белок 0 | 0,0066 |
| LR | rs1128503\_AG | 0,0063 |
| LR | Пол\_bin | 0,0057 |
| LR | Глюкоза 0 | 0,0044 |
| LR | rs1143627\_AA | 0,0029 |
| LR | rs1800795\_CC | 0,0026 |
| LR | амилаза 6 | 0,0025 |
| LR | Общий белок 6 | 0,0019 |
| LR | rs1128503\_GG | 0,0017 |
| LR | ХБСстарт | 0,0012 |
| LR | СТАДИЯ\_2а | 0,0011 |
| LR | rs1045642\_AA | 0,0009 |
| LR | ммса0 | 0,0009 |
| LR | rs1143627\_GG | 0,0008 |
| LR | Гемоглобин 0 | 0,0008 |
| LR | rs12233719\_bin | 0,0007 |
| LR | диспепсия | 0,0006 |
| LR | эког 6\_bin | 0,0006 |
| LR | Эритроциты 0 | 0,0006 |
| LR | патогенетический вариант боли\_bin | 0,0005 |
| LR | локализация\_4 | 0,0005 |
| LR | желтуха | 0,0005 |
| LR | rs7438135\_AG | 0,0003 |
| LR | rs1045642\_GG | 0,0001 |
| LR | Лимфоциты 6 | 0,0001 |
| LR | СТАДИЯ\_3 | 0 |
| LR | rs1799971\_AG | 0 |

Наибольшую значимость в точность модели LR вкладывает признак АСТ6. Кроме того, существенное влияние оказывают те же признаки, что и в модели GBC: Лейкоциты 6, амилаза 0, Возраст, но есть и существенные расхождения. Это связано с разными принципами работы моделей LR и GBC. LR работает на агрегирование признаков. При этом признаки могут компенсировать друг друга в итоговой сумме. GBC работает на декомпозицию признаков и разделение пространства на подпространства. Этим объясняется в том числе больший разброс признаков 45 признаков с ненулевой значимостью на 50 прогонах.

Табл. 14 Пример значений коэффициентов LR при решении задачи №4

|  |  |
| --- | --- |
| **LR** | **coef** |
| АСТ6 | 0,341981 |
| скф6 | 0,080842 |
| ХБСстарт | 0,0701 |
| эког 6\_bin | 0,020119 |
| Эритроциты 0 | 0,015856 |
| диспепсия | 0,012643 |
| rs2032582\_TT | 0,005786 |
| патогенетический вариант боли\_bin | 0,00284 |
| Лейкоциты 6 | 0,00211 |
| rs1799971\_GG | 0,001001 |
| название операции\_bin | -0,00327 |
| rs1143627\_AA | -0,01712 |
| Тромбоциты 0 | -0,01963 |
| суточная доза морфина 0 | -0,03071 |
| скф0 | -0,04146 |
| Общий билирубин6 | -0,0903 |
| имт0 | -0,1329 |
| Общий билирубин0 | -0,15041 |
| амилаза 0 | -0,1611 |

В таблице 14 представлен пример коэффициентов логистической регрессии на одном из 50 прогонов для задачи №4.

Рассмотрим результаты решения задачи №5, в которой терапия производилась Фентонилом. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось бинарное значение резистентности (сопротивляемость) пациентов с данным генотипом к терапии.

Табл. 15 Описательные статистики (задача №5)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | n\_feature\_mean | n\_feature\_std | Ptrain\_mean | Ptrain\_std | Ptest\_mean | Ptest\_std | P\_Cross\_mean | P\_Cross\_std |
| DTC | 20,94 | 5,74 | 1 | 0 | 0,56 | 0,12 | 0,74 | 0,07 |
| GBC | 11,76 | 4,5 | 1 | 0 | 0,71 | 0,15 | 0,83 | 0,04 |
| KNC | 2,02 | 0,25 | 0,75 | 0,06 | 0,43 | 0,04 | 0,58 | 0,05 |
| LR | 13,88 | 4,06 | 0,98 | 0,04 | 0,56 | 0,13 | 0,86 | 0,04 |
| LSVC | 35,48 | 3,97 | 0,75 | 0,17 | 0,55 | 0,12 | 0,62 | 0,09 |
| MLPC | 31,7 | 5,33 | 0,75 | 0,27 | 0,57 | 0,18 | 0,63 | 0,1 |
| RFC | 13,44 | 4,28 | 0,96 | 0,05 | 0,71 | 0,19 | 0,67 | 0,11 |
| SVC | 5,34 | 1,1 | 0,9 | 0,06 | 0,88 | 0,2 | 0,8 | 0,04 |

Табл. 16 Проверка статистической значимости результатов между моделями (задача №5)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *U-тест Манна-Уитни (р)* | | | | *n\_feature\_mean* | | | | | | | | | | | | | | | |
| **13,44** | | **20,94** | | **5,34** | | **35,48** | | **11,76** | | **13,88** | | **31,7** | | **2,02** | |
| **RFC** | | **DTC** | | **SVC** | | **LSVC** | | **GBC** | | **LR** | | **MLPC** | | **KNC** | |
| *n\_feature\_mean* | **13,44** | **RFC** | 1,0 | | 0,0 | | 0,0 | | 0,0 | | 0,4 | | 0,0 | | 0,0 | | 0,0 | |
| **20,94** | **DTC** | 0,0 | | 1,0 | | 0,0 | | 0,22 | | 0,0 | | 0,31 | | 0,07 | | 0,0 | |
| **5,34** | **SVC** | 0,0 | | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | |
| **35,48** | **LSVC** | 0,0 | | 0,22 | | 0,0 | | 1,0 | | 0,0 | | 0,39 | | 0,45 | | 0,0 | |
| **11,76** | **GBC** | 0,4 | | 0,0 | | 0,0 | | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | |
| **13,88** | **LR** | 0,0 | | 0,31 | | 0,0 | | 0,39 | | 0,0 | | 1,0 | | 0,15 | | 0,0 | |
| **31,7** | **MLPC** | 0,0 | | 0,07 | | 0,0 | | 0,45 | | 0,0 | | 0,15 | | 1,0 | | 0,0 | |
| **2,02** | **KNC** | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 1,0 | |

На этой задаче наилучшим образом показывает себя модель SVC. Данная модель дает достаточно высокую точность, хотя и имеет самый большой разброс по точности. При этом использует в среднем всего около пяти признаков. Несмотря на высокий разброс данная модель значимо отличается от остальных моделей и принимается в качестве результирующей. Несколько хуже справляются ансамблевые методы RFC, GBC. При этом имеют близкий к SVC разброс по точности. Наихудшим образом справляется KNC, хотя использует в среднем всего около 2 признаков.

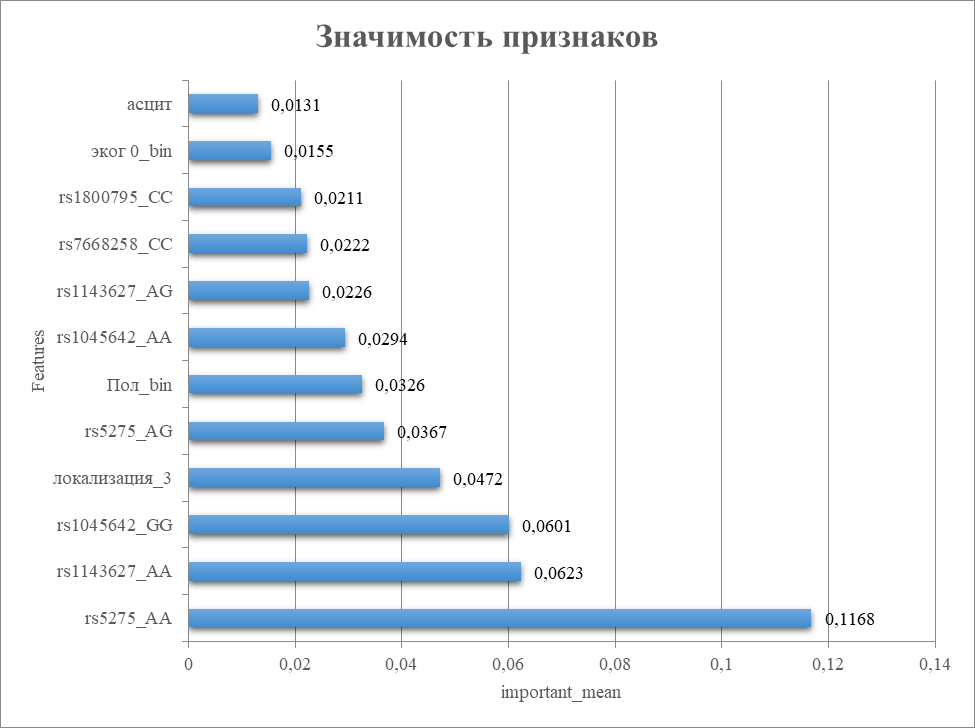


Рис. 11 Значимость признаков, отобранных SVC при решении задачи №5

Табл. 17. Значимость признаков, отобранных SVC при решении задачи №5   
(показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **features** | **important\_mean** |
| SVC | **rs5275\_AA** | **0,1168** |
| SVC | rs1143627\_AA | 0,0623 |
| SVC | rs1045642\_GG | 0,0601 |
| SVC | локализация\_3 | 0,0472 |
| SVC | rs5275\_AG | 0,0367 |
| SVC | Пол\_bin | 0,0326 |
| SVC | rs1045642\_AA | 0,0294 |
| SVC | rs1143627\_AG | 0,0226 |
| SVC | rs7668258\_CC | 0,0222 |
| SVC | rs1800795\_CC | 0,0211 |
| SVC | эког 0\_bin | 0,0155 |
| SVC | асцит | 0,0131 |
| SVC | диспепсия | 0,0089 |
| SVC | rs7668258\_CT | 0,0078 |
| SVC | rs1045642\_AG | 0,004 |
| SVC | СТАДИЯ\_3 | 0,004 |
| SVC | rs2032582\_GG | 0,0038 |
| SVC | rs1128503\_GG | 0,0037 |
| SVC | название операции\_bin | 0,0036 |
| SVC | rs1128503\_AG | 0,0034 |
| SVC | патогенетический вариант боли\_bin | 0,0032 |
| SVC | Общий белок 6 | 0,0026 |
| SVC | rs1128503\_AA | 0,0025 |
| SVC | rs2032582\_TT | 0,0023 |
| SVC | rs7668258\_TT | 0,0019 |
| SVC | локализация\_2 | 0,0016 |
| SVC | rs2032582\_GT | 0,0014 |
| SVC | rs1799971\_bin | 0,0008 |

Наибольший вклад в точность модели обеспечивает признак rs5275\_AA. В отличие от предыдущих задач в этой задачи существенное влияние на точность оказывает влияние генотип человека. При этом в среднем модель использует 5 признаков в различном сочетании.

Рассмотрим результаты решения задачи №6, в которой терапия производилась Фентонилом. В качестве входа использовался весь набор из 57 признаков. В качестве выхода использовалось *бинарное значение показателя появления НПР в виде седации.*

Табл. 18 Описательные статистики (задача №6)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | n\_feature\_mean | n\_feature\_std | Ptrain\_mean | Ptrain\_std | Ptest\_mean | Ptest\_std | P\_Cross\_mean | P\_Cross\_std |
| DTC | 14,08 | 3,91 | 1 | 0 | 0,67 | 0,11 | 0,84 | 0,03 |
| GBC | 9,2 | 2,92 | 1 | 0 | 0,57 | 0,14 | 0,92 | 0,03 |
| KNC | 3,6 | 2,07 | 0,75 | 0,07 | 0,68 | 0,08 | 0,68 | 0,11 |
| LR | 13,48 | 4,76 | 0,91 | 0,05 | 0,73 | 0,15 | 0,86 | 0,02 |
| LSVC | 30,82 | 3,77 | 0,77 | 0,15 | 0,56 | 0,14 | 0,66 | 0,08 |
| MLPC | 35,1 | 4,56 | 0,83 | 0,24 | 0,63 | 0,16 | 0,6 | 0,09 |
| RFC | 27,24 | 4,87 | 0,97 | 0,03 | 0,53 | 0,1 | 0,63 | 0,07 |
| SVC | 7,82 | 2,58 | 0,98 | 0,05 | 0,78 | 0,13 | 0,83 | 0,08 |

Табл. 19 Проверка статистической значимости результатов между моделями (задача №6)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *U-тест Манна-Уитни (р)* | | | | *n\_feature\_mean* | | | | | | | | | | | | | | | |
| **27,24** | | **14,08** | | **7,82** | | **30,82** | | **9,2** | | **13,48** | | **35,1** | | **3,6** | |
| **RFC** | | **DTC** | | **SVC** | | **LSVC** | | **GBC** | | **LR** | | **MLPC** | | **KNC** | |
| *n\_feature\_mean* | **27,24** | **RFC** | 1,0 | | 0,0 | | 0,0 | | 0,1 | | 0,07 | | 0,0 | | 0,0 | | 0,0 | |
| **14,08** | **DTC** | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | | 0,02 | | 0,05 | | 0,07 | |
| **7,82** | **SVC** | 0,0 | | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,1 | | 0,0 | | 0,0 | |
| **30,82** | **LSVC** | 0,1 | | 0,0 | | 0,0 | | 1,0 | | 0,43 | | 0,0 | | 0,02 | | 0,0 | |
| **9,2** | **GBC** | 0,07 | | 0,0 | | 0,0 | | 0,43 | | 1,0 | | 0,0 | | 0,13 | | 0,0 | |
| **13,48** | **LR** | 0,0 | | 0,02 | | 0,1 | | 0,0 | | 0,0 | | 1,0 | | 0,0 | | 0,0 | |
| **35,1** | **MLPC** | 0,0 | | 0,05 | | 0,0 | | 0,02 | | 0,13 | | 0,0 | | 1,0 | | 0,06 | |
| **3,6** | **KNC** | 0,0 | | 0,07 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,06 | | 1,0 | |

Наилучший результат на этой задаче дают SVC и LR. Однако SVC использует в среднем всего около 8 признаков, а LR в среднем около 13. Поэтому SVC является предпочтительнее и принимается в качестве результирующей. Однако LR может быть использована в качестве вспомогательной модели. Наихудшую точность дают ансамблевые модели GBC, RFC, и модель LSVC.

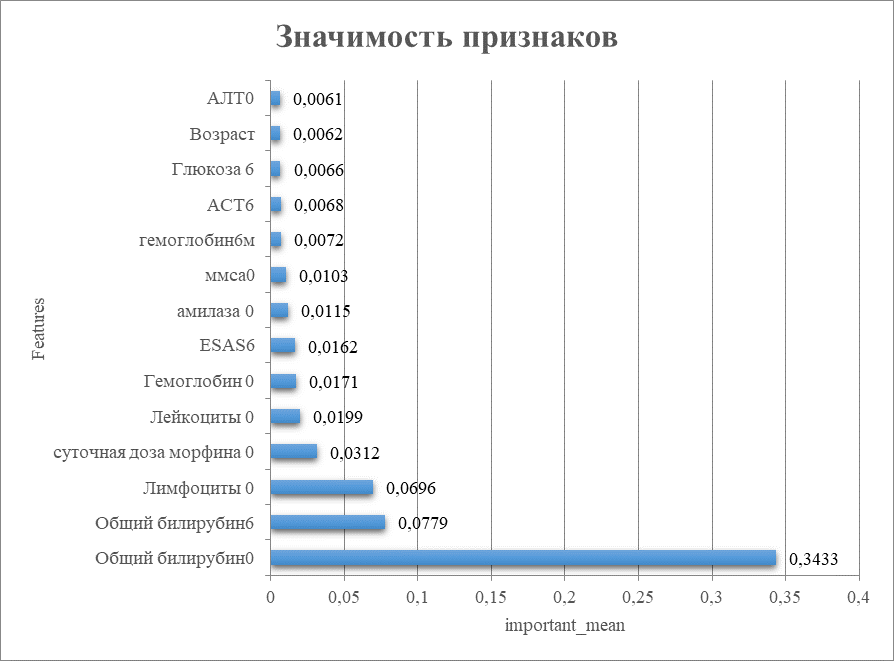


Рис. 12 Значимость признаков, отобранных LR при решении задачи №6

Табл. 20. Значимость признаков, отобранных LR при решении задачи №6   
(показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **features** | **important\_mean** |
| LR | **Общий билирубин0** | **0,3433** |
| LR | Общий билирубин6 | 0,0779 |
| LR | Лимфоциты 0 | 0,0696 |
| LR | суточная доза морфина 0 | 0,0312 |
| LR | Лейкоциты 0 | 0,0199 |
| LR | Гемоглобин 0 | 0,0171 |
| LR | ESAS6 | 0,0162 |
| LR | амилаза 0 | 0,0115 |
| LR | ммса0 | 0,0103 |
| LR | гемоглобин6м | 0,0072 |
| LR | АСТ6 | 0,0068 |
| LR | Глюкоза 6 | 0,0066 |
| LR | Возраст | 0,0062 |
| LR | АЛТ0 | 0,0061 |
| LR | ESAS0 | 0,006 |
| LR | Тромбоциты 6 | 0,0057 |
| LR | АСТ0 | 0,0055 |
| LR | rs1045642\_AG | 0,0055 |
| LR | диспепсия | 0,0045 |
| LR | rs1128503\_AG | 0,0041 |
| LR | амилаза 6 | 0,0039 |
| LR | rs1800795\_CC | 0,0036 |
| LR | rs1128503\_GG | 0,0027 |
| LR | Общий белок 6 | 0,0027 |
| LR | rs5275\_AG | 0,0023 |
| LR | Лимфоциты 6 | 0,0022 |
| LR | скф6 | 0,0021 |
| LR | rs1799971\_bin | 0,0019 |
| LR | Лейкоциты 6 | 0,0016 |
| LR | rs2032582\_GT | 0,0014 |
| LR | эког 0\_bin | 0,0012 |
| LR | локализация\_1 | 0,001 |
| LR | rs7668258\_CC | 0,0009 |
| LR | локализация\_5 | 0,0008 |
| LR | rs7438135\_AA | 0,0008 |
| LR | rs1143627\_GG | 0,0007 |
| LR | Пол\_bin | 0,0007 |
| LR | патогенетический вариант боли\_bin | 0,0006 |
| LR | коморбидность | 0,0006 |
| LR | rs1128503\_AA | 0,0005 |
| LR | имт0 | 0,0005 |
| LR | rs7668258\_CT | 0,0004 |
| LR | ХБСстарт | 0,0004 |
| LR | СТАДИЯ\_2б | 0,0004 |
| LR | rs1045642\_AA | 0,0003 |
| LR | рак-обусловл.слабость | 0,0003 |
| LR | rs7438135\_GG | 0,0003 |
| LR | скф0 | 0,0003 |
| LR | асцит | 0,0002 |
| LR | rs776746\_bin | 0,0002 |
| LR | rs5275\_AA | 0,0001 |
| LR | локализация\_2 | 0,0001 |

В отличие от предыдущих задач, в данной задачи есть один доминирующий признак обеспечивающий точность модели: Общий билирубин0.

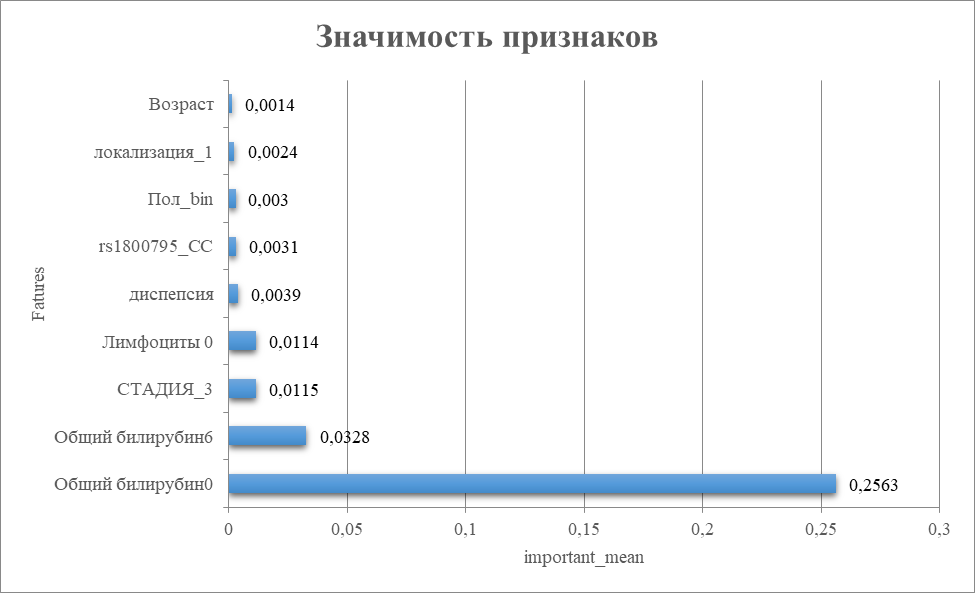


Рис. 13 Значимость признаков, отобранных SVC при решении задачи №6

Табл. 21 Пример значений коэффициентов LR при решении задачи №6

|  |  |
| --- | --- |
| **LR** | **Coef** |
| Глюкоза 6 | 0,942761 |
| патогенетический вариант боли\_bin | 0,643628 |
| rs1128503\_AG | 0,58308 |
| rs1143627\_AA | 0,256475 |
| Общий билирубин0 | 0,106755 |
| АСТ6 | 0,029479 |
| гемоглобин6м | 0,020732 |
| амилаза 0 | -0,00524 |
| Тромбоциты 6 | -0,00692 |
| рак-обусловл.слабость | -0,22771 |
| rs12233719\_bin | -0,3028 |
| ммса6 | -0,33829 |
| Пол\_bin | -0,50277 |
| rs1045642\_AA | -0,51667 |
| rs1800795\_CC | -0,62224 |
| ESAS0 | -0,85555 |

В таблице 21 представлен пример коэффициентов логистической регрессии на одном из 50 прогонов для задачи №6.

Табл. 22. Значимость признаков, отобранных SVC при решении задачи №6   
(показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **features** | **important\_mean** |
| SVC | **Общий билирубин0** | **0,2563** |
| SVC | Общий билирубин6 | 0,0328 |
| SVC | СТАДИЯ\_3 | 0,0115 |
| SVC | Лимфоциты 0 | 0,0114 |
| SVC | диспепсия | 0,0039 |
| SVC | rs1800795\_CC | 0,0031 |
| SVC | Пол\_bin | 0,003 |
| SVC | локализация\_1 | 0,0024 |
| SVC | Возраст | 0,0014 |
| SVC | эког 0\_bin | 0,001 |
| SVC | rs776746\_bin | 0,0009 |
| SVC | локализация\_3 | 0,0008 |
| SVC | rs1128503\_GG | 0,0007 |
| SVC | rs1143627\_AA | 0,0004 |
| SVC | rs12233719\_bin | 0,0003 |
| SVC | rs1045642\_GG | 0,0001 |

В этой модели так же подавляющее влияние на точность модели оказывает признак: Общий билирубин0. При этом модель SVC использует меньше признаков и с меньшим разбросом (7 признаков против 13).

Рассмотрим результаты решения задачи №7, в которой терапия производилась Фентонилом. В качестве входа использовался весь набор из 57 признаков. В качестве выхода *использовалось объединение выходов всех показателей НПР для каждого пациента в бинарный признак*, имеющий значение 0 – в случае отсутствия любой НПР, 1 – присутствует хотя бы одна НПР.

Табл. 23 Описательные статистики (задача №7)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | n\_feature\_mean | n\_feature\_std | Ptrain\_mean | Ptrain\_std | Ptest\_mean | Ptest\_std | P\_Cross\_mean | P\_Cross\_std |
| DTC | 9,28 | 6,44 | 1 | 0 | 0,81 | 0,1 | 0,87 | 0,04 |
| GBC | 4,6 | 1,7 | 1 | 0 | 0,98 | 0,05 | 0,94 | 0,01 |
| KNC | 6,7 | 1,89 | 0,64 | 0,05 | 0,61 | 0,14 | 0,65 | 0,04 |
| LR | 15,02 | 3,42 | 0,89 | 0,04 | 0,86 | 0,1 | 0,89 | 0,03 |
| LSVC | 29,72 | 5,08 | 0,77 | 0,17 | 0,7 | 0,15 | 0,69 | 0,14 |
| MLPC | 32,9 | 5,09 | 0,93 | 0,17 | 0,66 | 0,17 | 0,64 | 0,08 |
| RFC | 28,5 | 5,88 | 0,99 | 0,02 | 0,73 | 0,12 | 0,71 | 0,08 |
| SVC | 9,06 | 3,35 | 0,97 | 0,05 | 0,75 | 0,07 | 0,8 | 0,06 |

Наилучшим образом на данной задаче отработала модель GBC. Точность работы модели близка к 1 при этом разброс по точности меньше чем у других моделей и для выполнения прогноза модель GBC использует в среднем всего около 5 признаков. Именно эта модель принимается результирующей для данной задачи.

Табл. 24 Проверка статистической значимости результатов между моделями (задача №7)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *U-тест Манна-Уитни (р)* | | | | *n\_feature\_mean* | | | | | | | | | | | | | | | |
|  | |  | |  | |  | |  | |  | |  | |  | |
| **RFC** | | **DTC** | | **SVC** | | **LSVC** | | **GBC** | | **LR** | | **MLPC** | | **KNC** | |
| *n\_feature\_mean* |  | **RFC** | 1,0 | | 0,0 | | 0,15 | | 0,26 | | 0,0 | | 0,0 | | 0,02 | | 0,0 | |
|  | **DTC** | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | | 0,02 | | 0,0 | | 0,0 | |
|  | **SVC** | 0,15 | | 0,0 | | 1,0 | | 0,09 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | |
|  | **LSVC** | 0,26 | | 0,0 | | 0,09 | | 1,0 | | 0,0 | | 0,0 | | 0,06 | | 0,0 | |
|  | **GBC** | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 1,0 | | 0,0 | | 0,0 | | 0,0 | |
|  | **LR** | 0,0 | | 0,02 | | 0,0 | | 0,0 | | 0,0 | | 1,0 | | 0,0 | | 0,0 | |
|  | **MLPC** | 0,02 | | 0,0 | | 0,0 | | 0,06 | | 0,0 | | 0,0 | | 1,0 | | 0,07 | |
|  | **KNC** | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,0 | | 0,07 | | 1,0 | |

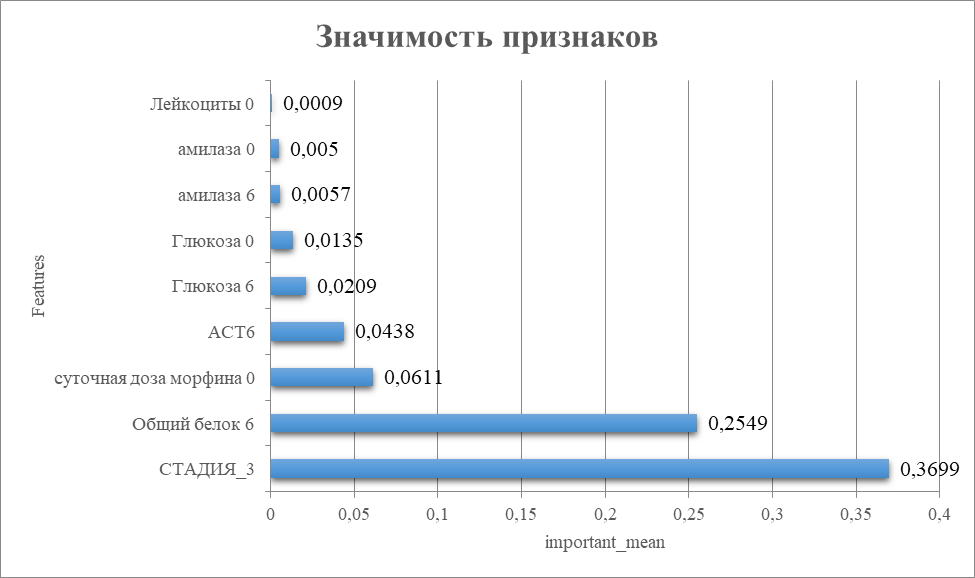


Рис. 14 Значимость признаков, отобранных GBC при решении задачи №7

Табл. 25. Значимость признаков, отобранных GBC при решении задачи №7 (показатель - среднее значение значимости по 50 прогонам)

|  |  |  |
| --- | --- | --- |
| **Метод** | **Features** | **important\_mean** |
| GBC | СТАДИЯ\_3 | 0,3699 |
| GBC | Общий белок 6 | 0,2549 |
| GBC | суточная доза морфина 0 | 0,0611 |
| GBC | АСТ6 | 0,0438 |
| GBC | Глюкоза 6 | 0,0209 |
| GBC | Глюкоза 0 | 0,0135 |
| GBC | амилаза 6 | 0,0057 |
| GBC | амилаза 0 | 0,005 |
| GBC | Лейкоциты 0 | 0,0009 |

В данной задаче наибольшее влияние оказывают признаки: СТАДИЯ\_3, общий белок 6. В среднем модель используют 5 признаков в разных сочетаниях. При этом модель использует всего 9 признаков на 50 прогонах.