Министерство науки и высшего образования Российской Федерации Санкт-Петербургский политехнический университет Петра Великого Физико-механический институт

Работа допущена к защите

Руководитель образовательной программы

«Прикладная математика и информатика»

К.Н. Козлов

« » 2023 г.

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

РАБОТА БАКАЛАВРА

**ОПРЕДЕЛЕНИЕ ФЕНОТИПА КОЛОНИЙ**

**ПЛЮРИПОТЕНТНЫХ СТВОЛОВЫХ КЛЕТОК ЧЕЛОВЕКА ПО МОРФОЛОГИЧЕСКИМ ПАРАМЕТРАМ МЕТОДАМИ**

**МАШИННОГО ОБУЧЕНИЯ**

по направлению подготовки 01.03.02 Прикладная математика и информатика по образовательной программе 01.03.02\_4 Биоинформатика

Выполнила

студентка гр. 5030102/90401 Е. Д. Веденеева

Руководитель

доцент ВШПМиВФ, к.ф.-м.н. В. В. Гурский

Консультант

по нормоконтролю Л. А. Арефьева

Санкт-Петербург 2023

**САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО**

## Физико-механический институт

УТВЕРЖДАЮ

Руководитель образовательной программы

«Прикладная математика и информатика»

К.Н. Козлов

« » 202\_ г.

# ЗАДАНИЕ

## на выполнение выпускной квалификационной работы

студенту Веденеевой Екатерине Дмитриевне, гр. 5030102/90401

1. Тема работы: «Определение фенотипа колоний плюрипотентных стволовых клеток человека по морфологическим параметрам методами машинного обуче- ния»
2. Срок сдачи студентом законченной работы: июнь 2023 г.
3. Исходные данные по работе:

Данные с измеренными морфологическими параметрами растущих колоний ин- дуцированных плюрипотентных клеток человека и клеток внутри колоний вме- сте с информацией о качестве колоний («фенотипе»). Качество колоний, характе- ризуемое их потенциальной способностью к дальнейшему росту и поддержанию состояний клональности и плюрипотентности, предварительно оценено экспер- тами.

Инструментальные средства:

Язык программирования Python3.8, среда разработки Jupyter Notebook Ключевые источники литературы:

1. Krasnova, O.A.; Gursky, V.V.; Chabina, A.S.; Kulakova, K.A.; Alekseenko, L.L.; Panova, A.V.; Kiselev, S.L.; Neganova, I.E. Prognostic Analysis of Human Pluripotent Stem Cells Based on Their Morphological Portrait and Expression of Pluripotent Mark- ers. Int. J. Mol. Sci. 2022, 23, 12902. <https://doi.org/10.3390/ijms232112902>
2. Maddah, M.; Shoukat-Mumtaz, U.; Nassirpour, S.; Loewke, K. A system for auto- mated, noninvasive, morphology-based evaluation of induced pluripotent stem cell cul- tures. J Lab Autom 2014, 19, 454–60. <https://doi.org/10.1177/2211068214537258>
3. Kato, R.; Matsumoto, M.; Sasaki, H.; Joto, R.; Okada, M.; Ikeda, Y.; Kanie, K.; Suga, M., Kinehara, M.; Yanagihara, K.; et al. Parametric analysis of colony morpho-

logy of non-labelled live human pluripotent stem cells for cell quality control. Sci Rep 2016, 6, 34009. <https://doi.org/10.1038/srep34009>

1. Содержание работы (перечень подлежащих разработке вопросов):

С помощью методов машинного обучения необходимо разработать эффективную модель классификации колоний по фенотипу, используя измеренные морфологи- ческие параметры в качестве предикторов. Для этого в ходе работы необходимо:

* 1. Определить наиболее эффективный метод классификации, оценить качество полученной классификации, исследовать особенности результатов обучения для разных клеточных линий в данных.
  2. Сравнить методы выделения наиболее значимых предикторов для обучения модели, исследовать, как их применение влияет на качество классификации.
  3. Построить и исследовать эффективность модели, использующей в качестве предикторов комбинацию клеточных и колониальных данных.
  4. На основе полученных результатов сделать выводы о связи морфологии ко- лоний и клеток с фенотипом растущих колоний.

1. Дата выдачи задания: 14 октября 2022 г.

Руководитель ВКР В. В. Гурский

(подпись) Задание принял к исполнению

Студент Е. Д. Веденеева

(подпись)

На 50 с., 10 рисунков, 11 таблиц, 2 приложения

КЛЮЧЕВЫЕ СЛОВА: МАШИННОЕ ОБУЧЕНИЕ, ПЛЮРИПОТЕНТНЫЕ СТВОЛОВЫЕ КЛЕТКИ ЧЕЛОВЕКА, БИНАРНАЯ КЛАССИФИКАЦИЯ, МЕ- ТОДЫ ОТБОРА ПРИЗНАКОВ, МОРФОЛОГИЧЕСКИЕ ПАРАМЕТРЫ

Тема выпускной квалификационной работы: «Определение фенотипа ко- лоний плюрипотентных стволовых клеток человека по морфологическим пара- метрам методами машинного обучения».

Цель данной работы — с помощью методов машинного обучения постро- ить на основе имеющихся данных эффективную модель классификации колоний плюрипотентных стволовых клеток человека, позволяющую определить фено- тип колонии («хороший» или «плохой»), используя морфологические параметры в качестве предикторов.

В ходе работы были рассмотрены различные методы бинарной классифи- кации объектов по числовым признакам и проведено сравнение качества их ра- боты на клеточных и колониальных морфологических данных. После этого был проведен анализ методов отбора признаков и исследовано влияние их примене- ния на качество работы моделей. В итоге были получены оптимальные модели классификации морфологии колоний с кросс-валидационной точностью 68.11% и 76.10% для клеточных и колониальных данных соответственно.

Кроме того, были построены модели, использующие в качестве предикто- ров комбинацию клеточных и колониальных параметров. Оптимальная модель достигает качества классификации 98.28%, что показывает высокую эффектив- ность совместного использования данных о морфологии колонии и клеток внутри нее для определения фенотипа колонии.

Таким образом, была проиллюстрирована взаимосвязь морфологии коло- ний и клеток с фенотипом растущих колоний и предложены классификаторы, позволяющие автоматически определять качество колонии плюрипотентных кле- ток человека.

50 pages, 10 figures, 11 tables, 2 appendices

KEYWORDS: MACHINE LEARNING, HUMAN PLURIPOTENT STEM CELLS, BINARY CLASSIFICATION, FEATURE SELECTION METHODS, MOR- PHOLOGICAL PARAMETERS

The topic of the graduate qualification work is “Determining the phenotype of human pluripotent stem cell colonies by morphological parameters using machine learning methods”.

The aim of this work is to use machine learning methods to construct, based on the available data, an effective model for classification of human pluripotent stem cell colonies, which allows to determine the colony phenotype (“good” or “bad”) using morphological parameters as predictors.

In the course of the work, various methods of binary classification of objects by numerical characteristics were considered and the quality of their work was compared on cellular and colonial morphological data. After that, the analysis of feature selection methods was carried out and the effect of their application on the quality of the models was investigated. As a result, optimal colony morphology classification models were obtained with cross-validation accuracy of 68.11% and 76.10% for cellular and colony data, respectively.

In addition, models were constructed using a combination of cellular and colonial parameters as predictors. The optimal model achieves a classification quality 98.28%, that shows the high efficiency of combining data on the morphology of the colony and cells inside it to determine the phenotype of the colony.

Thus, the relationship between the morphology of colonies and cells with the phenotype of growing colonies was illustrated and classifiers were proposed to automatically determine the quality of a colony of pluripotent human cells.

# СОДЕРЖАНИЕ

[Введение ........................................................................................................... 7](#_TOC_250026)

Глава 1. Описание данных и формальная постановка задачи ............. 9

* 1. [Описание набора данных и способа их получения .............................. 9](#_TOC_250025)
  2. [Постановка задачи машинного обучения ............................................... 11](#_TOC_250024)

[Глава 2. Обзор методов бинарной классификации ................................. 13](#_TOC_250023)

* 1. [Наивный байесовский классификатор .................................................... 13](#_TOC_250022)
  2. [Метод *k* ближайших соседей.................................................................... 14](#_TOC_250021)
  3. [Логистическая регрессия.......................................................................... 15](#_TOC_250020)
  4. [Метод случайного леса............................................................................. 17](#_TOC_250019)
  5. [Метод опорных векторов ......................................................................... 19](#_TOC_250018)
  6. [Искусственная нейронная сеть ................................................................ 22](#_TOC_250017)

Глава 3. Сравнение качества работы методов на исследуемых

данных ................................................................................................................. 27

* 1. [Метрики оценки качества классификации ............................................. 27](#_TOC_250016)
     1. [Матрица ошибок..................................................................................... 27](#_TOC_250015)
     2. [Кросс-валидационная точность ............................................................ 28](#_TOC_250014)
     3. [Площадь под кривой ошибок................................................................ 28](#_TOC_250013)
  2. [Предварительный анализ и предобработка данных .............................. 29](#_TOC_250012)
     1. [Фильтрация данных ............................................................................... 29](#_TOC_250011)
     2. Дисперсионный и корреляционный анализ ........................................ 30
     3. [Стандартизация данных ........................................................................ 33](#_TOC_250010)
  3. [Реализация алгоритмов классификации и подбор гиперпараметров . 34](#_TOC_250009)
  4. [Сравнение качества работы полученных моделей ............................... 35](#_TOC_250008)
  5. [Сравнение моделей, обученных на объединенных данных ................. 37](#_TOC_250007)

[Глава 4. Отбор признаков ............................................................................ 39](#_TOC_250006)

* 1. [Методы отбора признаков......................................................................... 39](#_TOC_250005)
     1. [Метод SHAP............................................................................................ 40](#_TOC_250004)
     2. [Исчерпывающий поиск признаков....................................................... 40](#_TOC_250003)
  2. [Результаты применения методов отбора признаков ............................. 41](#_TOC_250002)

[Заключение.................................................................................................... 45](#_TOC_250001)

[Список использованных источников ...................................................... 46](#_TOC_250000)

Приложение 1 ................................................................................................. 49

Приложение 2 ................................................................................................. 50

# ВВЕДЕНИЕ

Стволовые клетки — это незрелые (недифференцированные) клетки, явля- ющиеся предшественниками всех клеток многоклеточного организма. Ключе- выми особенностями стволовых клеток являются их способность к неограничен- ному самообновлению — образованию новых стволовых клеток без изменения фенотипа, а также способность дифференцироваться в специализированные клетки, то есть превращаться в клетки различных органов и тканей [1, 2].

В соответствии с дифференцирующим потенциалом, или потентностью, стволовые клетки делятся на тотипотентные, плюрипотентные, мультипотент- ные, олигопотентные и унипотентные. Из плюрипотентных стволовых клеток развиваются три зародышевых листка — эктодерма, мезодерма и энтодерма, та- ким образом они могут давать начало всем тканям и органам, за исключением экстраэмбриональных тканей (например, плаценты) [3].

Такие свойства человеческих плюрипотентных стволовых клеток (hPSC) делают их привлекательным источником материала для регенеративной меди- цины, инструментом моделирования заболеваний человека in vitro, создания но- вых лекарств (в том числе от неизлечимых заболеваний) и разработки банков кле- ток и тканей [3–5]. Основными способами получения hPSC является выделение из эмбриона человека на определенных стадиях его развития (ESCs), а также по- лучение в результате перепрограммирования путем экспрессии определенного набора факторов транскрипции (индуцирования) соматических клеток (iPSCs) [3].

Культивирование и размножение колоний hPSC in vitro производится за счет пролиферации и пассирования — переноса клеток в один или несколько со- судов со свежей питательной средой для дальнейшего развития. Колонии могут отличаться по уровню потенциала для поддержания плюрипотентности и кло- нальности, характеризирующему их склонность к появлению нежелательных му- таций, поэтому для их безопасного применения в медицинских исследованиях требуется высокий уровень контроля качества [6].

Для оценки качества колоний hPSC используются различные молекуляр- ные методы, однако применение инвазивных методов оценки не позволяет в дальнейшем использовать эти клетки в клинической практике. Неинвазивная оценка качества клеточной культуры в настоящее время не является количествен- ной и часто может напрямую зависеть от опыта исследователя или тонких мор- фологических различий между клеточными линиями и клонами. Кроме того, от- сутствие единых стандартов затрудняет объективное сравнение результатов раз- личных экспериментов и лабораторий. Существующие коммерческие системы компьютерного анализа не всегда могут быть применимы для многоклеточных колоний hPSC, поскольку они образованы плотно упакованными мелкими клет- ками (10–16 мкм в диаметре), что может не позволить автоматическим анализа- торам выявить различия в морфологических особенностях, отражающие начало процесса дифференцировки [7].

Таким образом, существует необходимость создания специализированных компьютерных методов анализа морфологических параметров колоний hPSC для безопасного отбора лучших линий для дальнейших клинических применений.

**Целью** данной выпускной квалификационной работы является построение модели классификации колоний плюрипотентных стволовых клеток человека по фенотипу (качеству), использующей морфологические параметры в качестве предикторов. В качестве **объекта** исследования будем рассматривать методы ма- шинного обучения, а в качестве **предмета** —применение методов обучения к проблеме классификации колоний hPSC.

# ГЛАВА 1. ОПИСАНИЕ ДАННЫХ И ПОСТАНОВКА ЗАДАЧИ БИНАРНОЙ КЛАССИФИКАЦИИ

В данной главе рассмотрен набор данных, используемый для построения моделей классификации, описан способ получения данных, а также приведена постановка задачи машинного обучения с учетом характера исходных данных.

## Описание набора данных и способа их получения

Для проведения исследований будем использовать данные с измеренными морфологическими параметрами растущих колоний плюрипотентных клеток че- ловека и клеток внутри колоний вместе с информацией о качестве колоний («фе- нотипе»), полученные авторами статьи [7]. Авторы данной статьи сфокусирова- лись на морфологических характеристиках трех генетически различных линий hPSC: наиболее широко используемой в мировых исследованиях линии H9 (WiCell), контрольной линии индуцированных стволовых клеток (hiPSC) AD3 и пациенто-специфичной линии hiPSC HPCASRi002-A (CaSR).

Общепризнано, что здоровые и высококачественные колонии hPSC должны обладать следующими характеристиками: относительно круглая форма, плотно упакованные клетки с ядром, занимающим практически всю клетку; вы- ступающие ядрышки; центры колоний очень плотные и имеют четко очерченный край [8]. В соответствии с этим, авторами было выделено семь параметров, ха- рактеризующих морфологию растущих колоний и клеток в различных пассажах во время их выращивания в идентичных условиях культивирования в течение 120 часов: площадь, периметр, малая ось, диаметр Ферета, минимальный диаметр Ферета, форм-фактор и площадь межклеточного пространства (последнее только для колоний) (таблица 1.1).

Таблица 1.1. Описание исследуемых морфологических параметров колоний и клеток.

|  |  |
| --- | --- |
| **Параметр** | **Описание** |
| Площадь | Площадь колонии или клетки |
| Периметр | Длина границы колонии или клетки |
| Малая ось | Длина малой оси эллипса, вписанного в колонию или клетку |
| Диаметр Ферета | Наибольшее расстояние между двумя точками на границе колонии или клетки (мера вытянутости) |
| Минимальный диаметр Ферета | Наименьшее расстояние между двумя точками на границе колонии или клетки (меря вытянутости) |
| Форм-фактор (изопериметрический коэффициент) | Площадь  4𝜋  Периметр2  (мера округлости и компактности) |
| Площадь межклеточного пространства (только для колоний) | Общая площадь свободного межклеточного пространства в колонии (показатель ком- пактности упаковки клеток) |

Данные были получены из фазово-контрастных изображений с использо- ванием Adobe Photoshop и программы ImageJ. Примеры изображений колоний с

«хорошим» и «плохим» фенотипами представлены на рисунке 1.1.

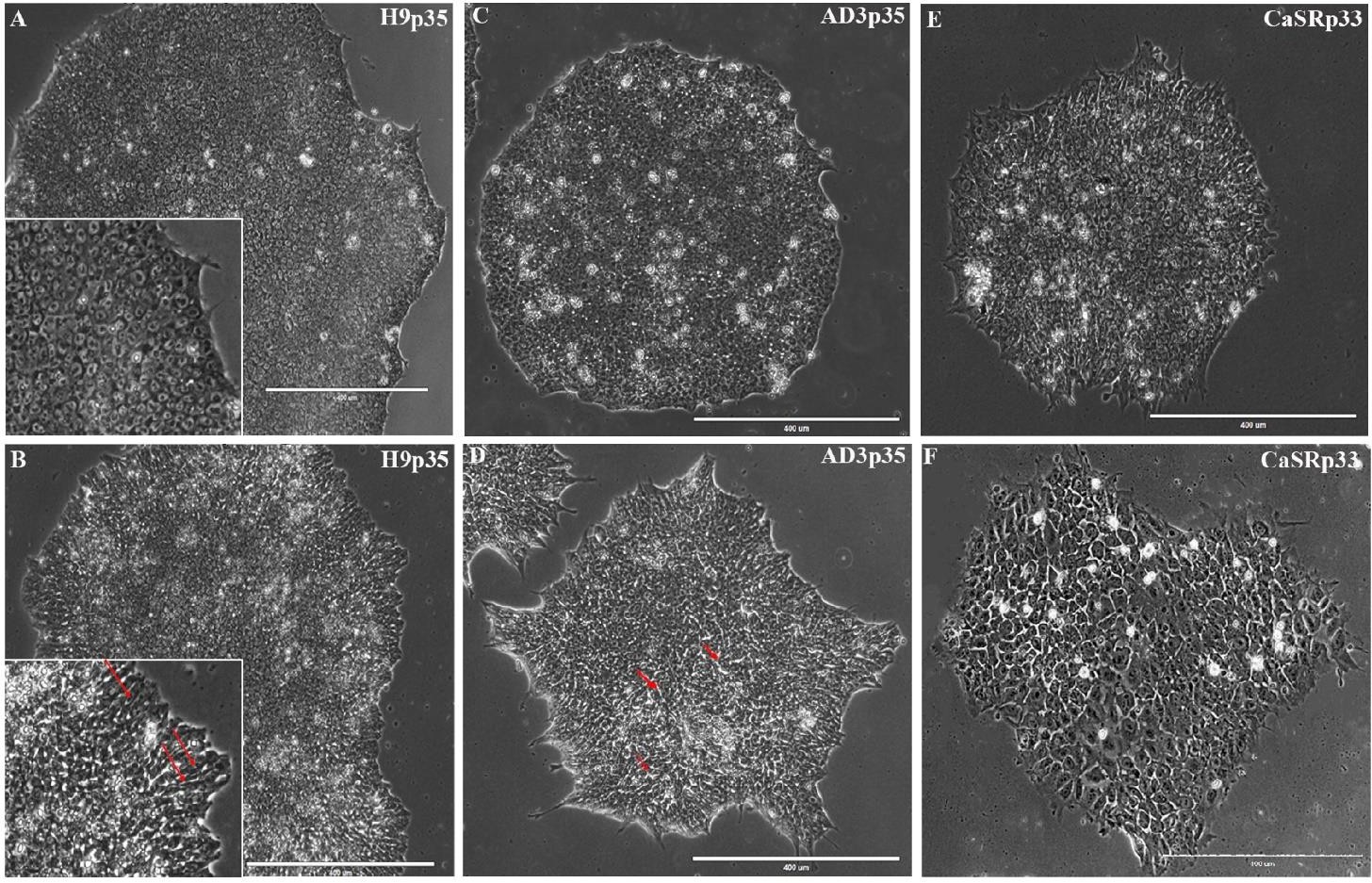


Рисунок 1.1. Примеры изображений колоний hPSC с «хорошим» (a, c, e) и «плохим» (b, d, f) фенотипами [7].

Таким образом значения параметров были получены для 53 колоний и 1602 клеток линии hESC H9, 49 колоний и 569 клеток контрольной линии hiPSC AD3 и 48 колоний и 1315 клеток для специфичных для пациента линий hiPSC CaSR. Фенотип колоний был определен экспертами на основе визуального анализа морфологии колоний и клеток.

## Постановка задачи машинного обучения

Для построения модели, позволяющей определить фенотип колоний на ос- нове морфологических признаков, будем пользоваться методами машинного обу- чения. Сформулируем общую постановку *задачи классификации*, относящейся к *задачам обучения с учителем*.

Пусть задано множество *объектов* 𝑋. Пусть все множество 𝑋 разбито на конечное число непересекающихся подмножеств 𝐾𝑦 = {𝑥 ∈ 𝑋: 𝑓∗(𝑥) = 𝑦}, кото- рые мы будем назвать *классами*, и задано множество *меток классов* 𝑌 =

{0, 1, … , 𝑚}*.* Пусть существует *целевая функция* 𝑓∗: 𝑋 → 𝑌, значения которой из- вестны только на конечном подмножестве объектов {𝑥1, … , 𝑥𝑘} ⊂ 𝑋. Совокуп- ность пар 𝑇 = {(𝑥1, 𝑦1), … , (𝑥𝑘, 𝑦𝑘)} ⊂ 𝑋 × 𝑌 называется *обучающей выборкой*. Задача обучения с учителем состоит в том, чтобы по выборке 𝑇 восстано-

вить зависимость 𝑦∗, то есть получить *решающую функцию* 𝑓: 𝑋 → 𝑌, которая приближала бы целевую функцию 𝑓∗(𝑥), причем не только на обучающей вы- борке, но и на всем множестве 𝑋. Другими словами, решить задачу классифика- ции значит построить алгоритм 𝑓, способный достаточно точно классифициро- вать произвольный объект 𝑥 ∈ 𝑋. Если 𝑚 = 2 и множество 𝑌 = {0, 1}, то такая задача называется *задачей бинарной классификации.*

Под объектом на практике обычно понимают некоторый *вектор признаков*

𝑥 = (𝑥(1), … , 𝑥(𝑛)), где *признак* 𝑥(𝑖) ∈ ℝ — это результат измерения некоторой ха- рактеристики объекта. Таким образом, элементы множества 𝑋 представляют со- бой не реальные объекты, а лишь доступные знания о них. Эти знания могут быть неточными из-за погрешностей измерения, и неполными, так как вектор

признаков обычно не может включать в себя все мыслимые характеристики объ- екта. В результате одному и тому же описанию 𝑥 могут соответствовать различ- ные объекты, а следовательно, и различные метки класса, и 𝑓∗(𝑥), строго говоря, не является функцией. В таком случае имеет смысл перейти к *вероятностной постановке задачи*.

Вместо существования неизвестной целевой функции 𝑓∗(𝑥) предположим существование вероятностного распределения на множестве 𝑋 × 𝑌 с неизвест- ной плотностью 𝑝(𝑥, 𝑦). Пары (𝑥𝑖, 𝑦𝑖) из обучающей выборки 𝑇, представляю- щие собой реализацию случайной величины, будем считать независимыми оди- наково распределенными.

Качество произвольного алгоритма классификации оценивается по *ошибке классификации*, которая определяется как вероятность неправильной классифи- кации

𝑒𝑟𝑟(𝑓) = 𝑃{(𝑥, 𝑦) ∈ 𝑋 × 𝑌 | 𝑓(𝑥) ≠ 𝑦} (1.1)

Основная цель при решении задач классификации — найти такой класси- фикатор 𝑓, при котором ошибка классификации 𝑒𝑟𝑟(𝑓) будет наименьшей [9].

Таким образом, задача построения модели для определения фенотипа ко- лоний hPSC может быть интерпретирована как задача бинарной классификации, где признаками являются измеренные морфологические характеристики, а мет- ками классов — фенотипы колоний «плохой» и «хороший», которые могут быть закодированы как 0 и 1 соответственно.

# ГЛАВА 2. ОБЗОР МЕТОДОВ БИНАРНОЙ КЛАССИФИКАЦИИ

Задачи бинарной классификации встречаются во многих предметных обла- стях, например, в банковском деле при определении кредитоспособности заем- щиков, в медицине при решении задач диагностики заболеваний и т. д. Несмотря на то, что разработано большое число методов, позволяющих решать подобные прикладные задачи с достаточно высокой точностью, не существует универсаль- ных алгоритмов, гарантирующих высокое качество классификации произволь- ных наборов данных [10]. Поэтому для решения поставленной задачи имеет смысл рассмотреть некоторое число наиболее популярных алгоритмов бинарной классификации и сравнить качество их работы на имеющихся данных.

## Наивный байесовский классификатор

Наивный байесовский классификатор — это простой вероятностный клас- сификатор, основанный на использовании теоремы Байеса [11].

В вероятностной постановке задача классификации объекта 𝑥 = (𝑥(1), … , 𝑥(𝑛)) равносильна задаче определения класса, вероятность принадлеж- ности 𝑥 к которому является наибольшей:

𝑦𝑀𝐴𝑃 = argmax 𝑃(𝑦|𝑥(1), … , 𝑥(𝑛)) (2.1)

𝑦∈𝑌

Используя формулу Байеса, это выражение можно переписать следующим образом:

𝑦𝑀𝐴𝑃 = argmax

𝑦∈𝑌

𝑃(𝑥(1), … , 𝑥(𝑛)|𝑦)𝑃(𝑦)

𝑃(𝑥(1), … , 𝑥(𝑛)) =

= argmax 𝑃(𝑦|𝑥(1), … , 𝑥(𝑛))𝑃(𝑦) (2.2)

𝑦∈𝑌

Вероятность 𝑃(𝑦) может быть легко оценена с помощью подсчета частоты встречаемости каждого из классов среди объектов обучающей выборки. В то же время оценка условной вероятности 𝑃(𝑥(1), … , 𝑥(𝑛)|𝑦) представляет намного большую сложность. Для упрощения задачи нахождения данной оценки, при

построении наивного байесовского классификатора принимается гипотеза о том, что каждый признак условно не зависим от других признаков, то есть

𝑃(𝑥(𝑖)|𝑦, 𝑥(𝑗) ) = 𝑃(𝑥(𝑖)|𝑦) при 𝑖 ≠ 𝑗 (2.3)

При данном допущении верна формула:

𝑛

𝑃(𝑥(1), … , 𝑥(𝑛)|𝑦) = 𝖦 𝑃(𝑥(𝑖)|𝑦)

𝑖=1

Таким образом формула (2.1) принимает вид:

𝑛

𝑦𝑀𝐴𝑃 = argmax 𝑃(𝑦) 𝖦 𝑃(𝑥(𝑖)|𝑦)

(2.4)

(2.5)

𝑦∈𝑌

𝑖=1

Апостериорные вероятности 𝑃(𝑥(𝑖)|𝑦) для непрерывных признаков обычно оценивают с помощью нормального распределения, в качестве парамет- ров используя выборочное среднее и среднеквадратическое отклонение, которые могут быть найдены по обучающей выборке.

## Метод k-ближайших соседей

Метод 𝑘-ближайших соседей является одним из наиболее известных при- меров метрических алгоритмов классификации [12]. В основе этой группы мето- дов лежит оценка сходства объектов и предположение о том, что если *мера сход- ства объектов* введена корректно, то схожие объекты чаще лежат в одном классе, чем в разных.

В качестве такой меры сходства часто выбирают евклидово расстояние в пространстве признаков:

𝑛

𝜌(𝑥 , 𝑥 ) = √∑ (𝑥(𝑖) − 𝑥(𝑖)

(2.6)

𝑘 𝑗

2

𝑘

𝑖=1

𝑗 )

При использовании данной метрики важно, чтобы все признаки были нор- мированы, так как в противном случае признаки с наибольшими числовыми зна- чениями будут влиять на значение метрики значительно сильнее, чем остальные.

В методе 𝑘-ближайших соседей класс объекта 𝑥 определяется как класс, наиболее часто встречающийся среди 𝑘 его ближайших соседей — элементов 𝑥𝑗 с наименьшим значением метрики 𝜌(𝑥𝑗, 𝑥). Число 𝑘 является *гиперпараметром* метода, то есть настраиваемым параметр, позволяющим управлять процессом обучения модели.

## Логистическая регрессия

Логистическая регрессия является разновидностью множественной ре- грессии — статистической модели для оценки связей между зависимой перемен- ной и переменными-предикторами [11]. В множественной линейной регрессии предполагается, что непрерывная зависимая переменная является линейной функцией независимых переменных:

𝑛

𝑧 = 𝑏0 + 𝑏1𝑥(1) + ⋯ + 𝑏𝑛𝑥(𝑛) + 𝜀 = 𝑏0 + ∑ 𝑏𝑖𝑥(𝑖) + 𝜀, (2.7)

𝑖=1

где 𝑏𝑖 — коэффициенты регрессии, 𝜀 — ошибка регрессии. Полагая 𝑥(0) = 1, можно переписать эту формулу в виде:

𝑛

𝑧 = ∑ 𝑏𝑖𝑥(𝑖) + 𝜀 = 𝑏𝑇𝑥 + 𝜀 (2.8)

𝑖=0

Для решения задачи бинарной классификации необходимо перейти от пе- ременной 𝑧, которая, вообще говоря, может принимать любые значения, к новой зависимой переменной таким образом, чтобы она лежала в интервале [0, 1]. Для этого в методе логистической регрессии используют следующее преобразование:

1 𝑒𝑧

𝑔(𝑧) = 1 + 𝑒−𝑧 = 𝑒𝑧 + 1 (2.9)

где 𝑔(𝑧) ∈ [0,1] — *логистическая функция*.

Тогда апостериорные вероятности принадлежности произвольного объекта

𝑥 к классу с метками 1 и 0 соответственно могут быть выражены следующими формулами:

𝑃(𝑦 = 1|𝑥; 𝑏) = 𝑔(𝑏𝑇𝑥) (2.10)

𝑃(𝑦 = 0|𝑥; 𝑏) = 1 − 𝑔(𝑏𝑇𝑥) (2.11) Таким образом, найдя коэффициенты регрессии, мы сможем оценить вероят- ность принадлежности объекта к классу с меткой 1 и, если она больше 0.5, отне- сти его к этому классу, а если меньше — к противоположному.

Используя формулы (2.10) и (2.11), можем записать выражение для плот-

ности вероятности принадлежности объекта 𝑥 к классу 𝑦:

𝑝(𝑦|𝑥; 𝑏) = 𝑔(𝑏𝑇𝑥)𝑦(1 − 𝑔(𝑏𝑇𝑥)

1−𝑦

(2.12)

)

Коэффициенты регрессии 𝑏𝑖 можно найти с помощью метода максималь- ного правдоподобия, который заключается в максимизации *функции правдоподо- бия* на обучающей выборке:

𝑘 𝑘

𝐿(𝑏) = 𝖦 𝑃(𝑦 |𝑥 ; 𝑏) = 𝖦 𝑔(𝑏𝑇𝑥 )𝑦𝑖(1 − 𝑔(𝑏𝑇𝑥 )

1−𝑦𝑖

(2.13)

𝑖=1

𝑖 𝑖

𝑖

𝑖=1

𝑖 )

Для этого, пользуясь линейностью логарифма, можем перейти к поиску ми- нимума следующего функционала:

𝑘

𝑙(𝑏) = −𝑙𝑜𝑔𝐿(𝑏) = − ∑[𝑦𝑖 log(𝑔(𝑏𝑇𝑥𝑖)) + (1 − 𝑦𝑖) log(1 − 𝑔(𝑏𝑇𝑥𝑖))] (2.14)

𝑖=1

Для минимизации 𝑙(𝑏) можно воспользоваться, например методом гради- ентного спуска, где коэффициенты пересчитываются по формуле

𝑏(𝑟) = 𝑏(𝑟−1) − 𝛼 𝜕 𝑙(𝑏(𝑟−1)) (2.15)

𝑗 𝑗

𝜕𝑏𝑗

где 𝑟 — номер шага, 0 < 𝛼 < 1 — переменная, задающая скорость спуска, а начальные значения 𝑏(0) задаются произвольным образом.

𝑗

Находя значение частной производной функции 𝑙(𝑏), получаем, что пра-

вило пересчета методе градиентного спуска для поиска коэффициентов логисти- ческой регрессии будет иметь следующий вид:

𝑘

𝑏(𝑟) = 𝑏(𝑟−1) + 𝛼 ∑(𝑦

− 𝑔(𝑏𝑇𝑥 )) 𝑥(𝑗)

(2.16)

𝑗 𝑗

𝑖

𝑖=1

𝑖 𝑖

Одним из допущений, при которых возможно применение логистической

регрессии является независимость переменных-предикторов друг от друга. По- этому *мультиколлинеарность* — высокая корреляция между зависимыми пере- менными, отрицательно влияет на качество модели. Одним из самых популярных методов борьбы с влиянием мультиколлинеарности на точность предсказаний модели является регуляризация, заключающаяся в добавлении штрафного слага- емого, ограничивающего рост коэффициентов регрессии. Например, при 𝐿-2 ре- гуляризации задача оптимизации приобретает следующий вид:

𝑛

𝑙(𝑏) + 𝜆 ∑ 𝑏2 → min

(2.17)

𝑖 𝑏

𝑖=0

где 𝜆 — коэффициент регуляризации, который выступает в качестве гиперпара- метра метода логистической регрессии.

Если обучающая выборка достаточно велика, то для ускорения вычислений имеет смысл использовать стохастический градиентный спуск, в котором для пе- ресчета коэффициентов регрессии на каждом шаге рассматривается не вся обу- чающая выборка, а ее случайные подвыборки заданного размера. Также на прак- тике часто используют другие методы оптимизации, например методы второго порядка, такие как алгоритм Бройдена-Флетчера-Гольдфарба-Шанно [13].

## Метод случайного леса

Случайный лес (англ. Random Forest) — это алгоритм машинного обуче- ния, который заключается в использовании ансамбля деревьев решений [11].

Деревья решений классифицируют объекты путем сортировки их вниз по дереву от корня до одного из листьев, соответствующих меткам класса. Каждый узел в дереве представляет собой предикат — простое решающее правило, опре- деляющее проверку некоторого признака объекта, а каждая ветвь, нисходящая от узла, соответствует одному из возможных диапазонов значений этого признака.

В основе процедуры построения деревьев решений обычно лежат «жад- ные» алгоритмы, обеспечивающие локально-оптимальное разбиение в каждом узле. Дерево строится сверху вниз начиная с корня. При построении каждого

узла-предиката формируется правило вида 𝑥(𝑖) ≤ 𝑐, где 𝑥(𝑖) — один из признаков, а 𝑐 — порог, который часто выбирается как среднее арифметическое некоторых двух соседних значений признака 𝑥(𝑖) среди объектов обучающей выборки. Та- ким образом, на каждом шаге построения дерева алгоритм последовательно срав- нивает все возможные разбиения для всех признаков (или для их заданного числа, которое в таком случае является гиперпараметром метода) и выбирает наилучший признак и наилучший порог для него.

Для выбора оптимального разбиения обычно используют функции оценки качества разбиения, такие как индекс Джини или теоретико-информационный критерий.

Индекс Джини для множества 𝐴 выражается следующей формулой:

𝑔𝑖𝑛𝑖(𝐴) = 1 − 𝑝2 − 𝑝2, (2.18)

0 1

где 𝑝0 и 𝑝1 — частота встречаемости классов с метками 0 и 1 соответственно среди всех объектов множества 𝐴. Если множество 𝐴, состоящее из 𝑙 объектов, разбивается на два подмножества 𝐴1 и 𝐴2 объемами 𝑙1 и 𝑙2 соответственно, то показатель качества разбиения можно определить следующим образом:

𝑙1 𝑙2

𝑔𝑖𝑛𝑖 (𝐴) = 𝑔𝑖𝑛𝑖(𝐴 ) + 𝑔𝑖𝑛𝑖(𝐴 ) (2.19)

𝑠𝑝𝑙𝑖𝑡 𝑙 1 𝑙 2

Наилучшим считают правило, приводящее к разбиению с минимальным значе- нием индекса 𝑔𝑖𝑛𝑖𝑠𝑝𝑙𝑖𝑡 (𝐴).

Теоретико-информационный критерий представляет собой разницу между энтропией родительского узла и взвешенной суммой энтропий его дочерних уз- лов:

𝐺𝑎𝑖𝑛 ( )

𝑙1

( ) + 𝑙2 𝑒𝑛𝑡𝑟𝑜𝑝𝑦(𝐴

)) (2.20)

𝑠𝑝𝑙𝑖𝑡 𝐴 = 𝑒𝑛𝑡𝑟𝑜𝑝𝑦(𝐴) − ( 𝑙 𝑒𝑛𝑡𝑟𝑜𝑝𝑦 𝐴1 𝑙 2

где под энтропией понимают меру информации:

𝑒𝑛𝑡𝑟𝑜𝑝𝑦(𝐴) = −𝑝0log 𝑝0 − 𝑝1log 𝑝1 (2.21) Наилучшим считают правило, приводящее к разбиению с максимальным значе- нием 𝑔𝑖𝑛𝑖𝑠𝑝𝑙𝑖𝑡 (𝐴).

Процесс построения решающего дерева обычно останавливается при

достижении установленного числа разбиений (максимальной глубины) или ми- нимального допустимого числа объектов в узле. Данные значения являются ги- перпараметрами метода. После того, как дерево построено, обычно прибегают к его «стрижке» — проходу по дереву снизу вверх и удалению узлов, которые ока- зывают минимальное влияние на точность модели.

Использование ансамбля решающих деревьев в методе случайного леса позволяет повысить точность в сравнении с одним деревом. Формирование ан- самбля деревьев происходит с помощью бэггинга. Для этого задают желаемое количество деревьев 𝑁 и случайным образом формируют 𝑁 подмножеств обуча- ющей выборки одинакового размера. Каждый элемент обучающей выборки мо- жет как присутствовать в подмножестве несколько раз, так и отсутствовать вовсе. Затем на основе полученных подмножеств строят решающие деревья.

В качестве результата классификации объекта методом случайного леса выбирается тот класс, за который проголосовала большая часть решающих дере- вьев.

## Метод опорных векторов

Метод опорных векторов (англ. Support Vector Machine, SVM) основыва- ется на построении в пространстве признаков гиперплоскости, разделяющей объекты обучающей выборки, относящиеся к разным классам [14]. Для удобства, в рамках данной задачи изменим метки классов с 𝑌 = {0, 1} на 𝑌 = {−1, 1}.

Гиперплоскость задается следующим уравнением:

𝑛

∑ 𝑤̂𝑖𝑥(𝑖) − 𝑤̂0 = 0 ⟺ 𝑤̂𝑇𝑥 − 𝑤̂0 = 0 (2.22)

𝑖=1

где 𝑤̂ — вектор нормали к гиперплоскости.

Если классы объектов *линейно разделимы*, то есть могут быть разделены гиперплоскостью, то объекты одного класса будут удовлетворять неравенству

𝑤̂𝑇𝑥 > 𝑤̂0, (2.23)

а второго — неравенству

𝑤̂𝑇𝑥 < 𝑤̂0 (2.24)

Для линейно разделимых классов существует множество разделяющих ги- перплоскостей. Цель данного метода — найти гиперплоскость, которая наиболее удалена от объектов обоих классов, тем самым увеличив надежность классифи- кации.

Векторы, расстояние от которых до разделяющей гиперплоскости мини- мально, называются *опорными*. Гиперплоскости, параллельные гиперплоскости (2.22) и содержащие опорные вектора и задаются уравнениями:

𝑤̂𝑇𝑥 − 𝑤̂0 = 𝜀, (2.25)

𝑤̂𝑇𝑥 − 𝑤̂0 = −𝜀 (2.26)

(в оптимальном положении гиперплоскость равноудалена от обоих классов). Поделив оба равенства на 𝜀, получим

𝑤𝑇𝑥 − 𝑤0 = 1, (2.27)

𝑤𝑇𝑥 − 𝑤0 = −1, (2.28)

где 𝑤 =

𝑤̂

𝜀 , 𝑤0 =

𝑤̂0

𝜀 .

Вычислим евклидово расстояние между этими гиперплоскостями:

|𝑤0 + 1 − (𝑤0 − 1)| 2

𝜌 = =

(2.29)

||𝑤|| ||𝑤||

где ||𝑤|| — евклидова норма вектора. Таким образом, задача нахождения разде- ляющей гиперплоскости, максимально удаленной от объектов обоих классов, сводится к задаче оптимизации:

||𝑤|| ⟶ min ;

𝑤,𝑤0

𝑤𝑇𝑥𝑖 − 𝑤0 ≥ 1, 𝑦𝑖 = 1, (2.30)

𝑤𝑇𝑥𝑖 − 𝑤0 ≤ 1, 𝑦𝑖 = −1

или в более удобной форме

1 ||𝑤||2 ⟶ min ; 𝑦 (𝑤𝑇𝑥

− 𝑤

) ≥ 1, 𝑖 = 1̅̅̅, 𝑘̅̅

(2.31)

2 𝑤,𝑤0 𝑖

𝑖 0

На практике данные часто не являются линейно разделимыми. В таком слу- чае алгоритму позволяют допускать ошибки на объектах обучающей выборки.

Для этого в исходной задаче смягчают ограничения и добавляют в минимизиру- емый функционал штраф за суммарную ошибку:

𝑛

1 ||𝑤||2 + 𝐶 ∑ 𝜉 ⟶ min ;

2 𝑖

𝑖=1

𝑤,𝑤0,𝜉

𝑦𝑖(𝑤𝑇𝑥𝑖 − 𝑤0) ≥ 1 − 𝜉𝑖, 𝑖 = 1̅̅̅, 𝑘̅̅, (2.32)

𝜉𝑖 ≥ 0, 𝑖 = 1̅̅̅, 𝑘̅̅,

где 𝐶 — гиперпараметр, позволяющий находить компромисс между максимиза- цией расстояний и минимизацией суммарной ошибки.

Полученная задача является задачей выпуклой квадратичной оптимизации с линейными ограничениями. Такие задачи известны как задачи квадратичного программирования [13], и для решения задач данного класса разработано множе- ство готовых прикладных пакетов.

Решив задачу квадратичного программирования и найдя значения парамет- ров 𝑤 и 𝑤0, сможем построить линейный пороговый классификатор:

𝑦𝑝𝑟𝑒𝑑𝑖𝑐𝑡𝑒𝑑 = sign(𝑤𝑇𝑥 − 𝑤0) (2.33)

Еще один подход, позволяющий решить проблему линейной неразделимо- сти, называется *ядерный трюк*. В основе данного подхода лежит теорема Ковера, утверждающая, что при повышении размерности пространства признаков, уве- личивается возможность линейной разделимости объектов.

Для применения ядерного трюка, от задачи (2.32), используя теорему Куна-Таккера, переходят к двойственной задаче поиска седловой точки функции Лагранжа, которая содержит векторы признаков только в виде скалярных произ- ведений (𝑥𝑖, 𝑥𝑗). Переход к новому пространству признаков с большей размерно- стью осуществляется с помощью применения некоторого преобразования

𝜂: 𝑋 ⟶ 𝐻. При этом исходная задача строится так же, за исключением того, что скалярное произведение (𝑥𝑖, 𝑥𝑗) в пространстве 𝑋 заменяется скалярным произ- ведением (𝜂(𝑥𝑖), 𝜂(𝑥𝑗)) в пространстве 𝐻. Вместо того, чтобы напрямую искать преобразование 𝜂, обычно подбирают вид *ядра* 𝐾(𝑥𝑖, 𝑥𝑗) = (𝜂(𝑥𝑖), 𝜂(𝑥𝑗))*,* кото- рым формально заменяется скалярное произведение. Наиболее распространен- ными являются:

1. Линейное ядро

𝐾(𝑥𝑖, 𝑥𝑗) = (𝑥𝑖, 𝑥𝑗) (2.34)

1. Полиномиальное ядро

𝑑

𝐾(𝑥𝑖, 𝑥𝑗) = ((𝑥𝑖, 𝑥𝑗) + 𝑟)

1. Радиальная базисная функция (RBF)

(2.35)

2

𝐾(𝑥𝑖, 𝑥𝑗) = 𝑒−𝛾||𝑥𝑖−𝑥𝑗|| , 𝛾 > 0 (2.36)

1. Сигмоидная функция

𝐾(𝑥𝑖, 𝑥𝑗) = tanh(𝛾(𝑥𝑖, 𝑥𝑗) + 𝑟) , 𝛾 > 0, 𝑐 < 0 (2.37)

Форма ядра и его параметры являются гиперпараметрами метода опорных векторов.

## Искусственная нейронная сеть

Искусственная нейронная сеть (англ. artificial neural network) — это мате- матическая модель, основанная на общих принципах функционирования биоло- гических нейронных сетей [15]. Вычислительной единицей искусственной нейронной сети является *нейрон*. По аналогии с биологическими нейронами ис- кусственный нейрон состоит из «дендритов», через которые входные сигналы по- ступают в ядро, ядра, осуществляющего обработку входных сигналов, и «ак- сона», передающего выходной сигнал на «дендриты» следующих нейронов. Вы- ходной сигнал нейрона определяется как значение функции активации на взве- шенной сумме его входов:

𝑛

𝑦 = 𝜎 (∑ 𝑥(𝑗)𝑤𝑗 − 𝑤0) , (2.38)

𝑗=1

где 𝑥(𝑗) — значения входных сигналов, 𝑤𝑗 — синаптические веса, определяющие степень влияния каждого сигнала на итоговое состояние, 𝑤0 — смещение. Си- наптические веса и смещение каждого нейрона являются параметрами сети, ко- торые настраиваются в ходе ее обучения. В качестве функции активации 𝜎

обычно выступает некоторая нелинейная дифференцируемая функция, напри- мер:

1. ReLU (от англ. Rectified Linear Unit)

ReLU(𝑥) = max(𝑥, 0) (2.39)

1. Логистическая функция

1

sigmoid(𝑥) = 1 + 𝑒−𝑥 (2.40)

1. Гиперболический тангенс

1 − 𝑒−2𝑥

tanh(𝑥) = 1 + 𝑒−2𝑥 (2.41)

В зависимости от количества нейронов и того, как они соединены друг с другом, выделяют различные архитектуры нейронных сетей. В задачах обучения с учителем для классификации объектов с числовыми признаками используются *многослойные полностью связные нейронные сети*. В сетях такого вида нейроны объединены в слои трех типов: входной, скрытые (один или несколько) и выход- ной. Каждый нейрон передает свой выходной сигнал всем нейронам следующего слоя. Общий вид полносвязной сети с одним скрытым слоем представлен на ри- сунке 2.1 [16]. Количество нейронов во входном слое равно количеству призна- ков, а выходной слой в сетях, используемых для решения задач бинарной класси- фикации, обычно содержит всего один нейрон с логистической функцией акти- вации. Количество скрытых слоев и нейронов в них и используемые функции ак- тивации являются гиперпараметрами нейронной сети.

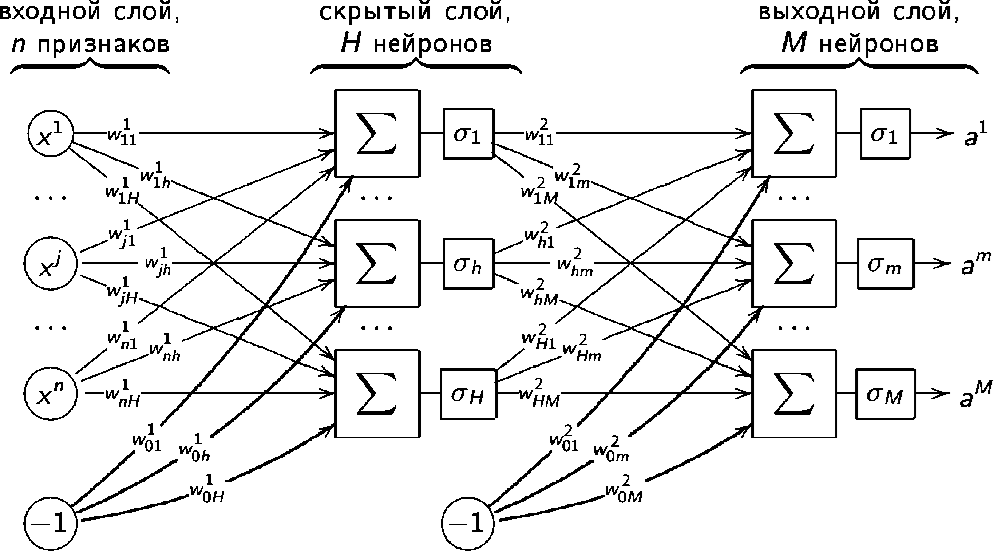


Рисунок 2.1. Общая схема многослойной искусственной нейронной сети.

Согласно теореме Цыбенко [17], с помощью нейронной сети такой архи- тектуры, вообще говоря, можно аппроксимировать любую непрерывную зависи- мость с произвольной точностью.

Обучение нейронной сети представляет собой процесс подбора весовых коэффициентов таким образом, чтобы минимизировать некоторый функционал ошибки на тренировочном наборе данных. Задачу нахождения оптимальной кон- фигурации сети можно рассматривать как задачу многомерной оптимизации, и для ее решения обычно используют вариации метода градиентного спуска. Вы- бор конкретного алгоритма пересчета весов является важным шагом обучения [18]. В данной работе остановимся на рассмотрении **алгоритма Adam** (от англ. Adaptive Moment Estimation), который является одним из наиболее часто исполь- зуемых и показывает хорошие результаты на практике [19].

В качестве минимизируемого функционала (функции потерь) будем ис- пользовать сумму значений бинарной кросс-энтропии на объектах обучающей выборки:

𝑘 𝑘

1 1

𝑄(𝑤) = − 𝑘 ∑ 𝑄𝑖(𝑤) = − 𝑘 ∑[𝑦𝑖 log(𝑝𝑖) + (1 − 𝑦𝑖) log(1 − 𝑝𝑖)] , (2.42)

𝑖=1 𝑖=1

где 𝑝𝑖 — предсказанная вероятность принадлежности к классу с меткой 1 для 𝑖- того объекта, 𝑦𝑖 — истинная метка класса 𝑖-того объекта, 𝑘 — число объектов в выборке.

Формула пересчета весовых коэффициентов в методе стохастического гра- диентного спуска имеет следующий вид:

𝑤(𝑘+1) = 𝑤(𝑘−1) − 𝛼∇𝑄(𝑤(𝑘−1)), (2.43) где 𝛼 — размер шага спуска, а начальные значения весов 𝑤(0) задаются произ- вольным образом. В методе Adam для ускорения сходимости применяются сле- дующие эвристики:

1. Использование информации о предыдущих шагах.

В качестве оценки моментов на 𝑘-ом шаге используется экспоненциальное сколь- зящее среднее между значением момента на 𝑘-ом шаге и значением градиента

(для второго момента — квадрата градиента) в точке 𝑤(𝑘−1):

𝜇(𝑘) = 𝛽 𝜇(𝑘−1) + (1 − 𝛽 )∇𝑄(𝑤(𝑘−1)), 𝜇(0) = 0 (2.44)

1 1 1 1 1

2

𝜇(𝑘) = 𝛽 𝜇(𝑘−1) + (1 − 𝛽 ) (∇𝑄(𝑤(𝑘−1) , 𝜇(0) = 0 (2.45)

2 2 2 2

)) 2

1. Адаптивный подбор размера шага c учетом истории изменения весов

Данная идея заключается в выравнивании скорости адаптации весов с помощью деления шага спуска на квадратный корень из второго момента, который отра- жает частоту обновления каждого веса. Таким образом, формула пересчета весов:

где

𝑤(𝑘) = 𝑤(𝑘−1) − α

√𝜇̂(𝑘) + 𝜀

2

1

𝜇̂(𝑘), (2.46)

𝜇(𝑘) 𝜇(𝑘)

𝜇̂(𝑘) = 1 , 𝜇̂(𝑘) = 2

(2.47)

1 1 − 𝛽𝑘 2 1 − 𝛽𝑘

1

2

* несмещенные моменты, 𝜀 — небольшая константа, позволяющая избежать де- ления на ноль.

Авторы Adam предлагают выбирать значения гиперпараметров 𝛽1, 𝛽2 и 𝜀 равными 0.9, 0.999 и 10−8 соответственно, а размер шага 0 < α < 1 подбирать вручную, отталкиваясь от конкретных исходных данных.

Для вычисления значений градиента функции потерь на каждом шаге бу- дем использовать **метод обратного распространения ошибки** (англ. Backpropa- gation) [16]. Данный метод состоит из следующих шагов:

* 1. Выбрать из обучающей выборки случайным образом подмножество за- данного размера 𝑚, которое еще не участвовало в обучении.
  2. Выбрать некоторый объект данного подмножества.
  3. Совершить прямой проход по сети (слева направо) и вычислить для каж- дого нейрона выходной сигнал

𝐻𝑙−1

𝑥𝑙 = 𝜎𝑙 (𝑠𝑙 ), 𝑠𝑙 = ∑ 𝑥𝑙−1𝑤𝑙

, 𝑙 = 1̅̅̅, 𝐿̅̅, ℎ = 1̅̅,̅𝐻̅̅𝑙̅

(2.48)

ℎ ℎ ℎ

ℎ 𝑘

𝑘=0

𝑘ℎ

и его частные производные

𝜕𝑥𝑙

𝑙 ℎ

𝑧 = |

𝜕𝜎𝑙

= |

ℎ

, 𝑙 = 1̅̅̅, 𝐿̅̅, ℎ = 1̅̅,̅𝐻̅̅̅ (2.49)

ℎ 𝜕𝑠

ℎ

𝑠=𝑠𝑙

𝜕𝑠

𝑙

𝑠=𝑠𝑙

ℎ

где 𝑙 — номер слоя, 𝐿 — число слоев, ℎ — номер нейрона в слое, 𝐻𝑙−1 — размер

𝑙 − 1 слоя, 𝑥𝑙 = −1 для всех слоев. Значения вектора 𝑥0 полагаются равными входному вектору признаков выбранного объекта обучающей выборки (рисунок 2.1).

0

* 1. Совершить обратный проход по сети (справа налево), рекурсивно вычис- ляя ошибки для каждого слоя:
* для выходного слоя (содержит один нейрон с выходным сигналом 𝑥𝐿 = 𝑝):

1

𝜀𝐿 = 𝜕𝑄𝑖(𝑤) = 𝑥𝐿 − 𝑦 = 𝑝 − 𝑦 (2.50)

1

* для скрытых слоев:

𝜕𝑄 (𝑤) 𝜕𝑥𝑙−1

𝜕𝑠𝐿 1

1

𝐻𝑙

𝜀𝑙−1 = 𝑖 ℎ = 𝑧𝑙−1 ∑ 𝜀𝑙 𝑤𝑙 , 𝑙 = 𝐿, … ,2, ℎ = 0̅̅,̅𝐻̅̅̅̅̅̅ (2.51)

ℎ 𝜕𝑥𝑙−1

ℎ

1

𝜕𝑠𝑙−1

ℎ 𝑘

𝑘=0

ℎ𝑘

𝑙−1

Вычислив все значения ошибок, сможем получить значения частных про- изводных минимизируемого функционала, используя формулу дифференцирова- ния суперпозиции функций:

𝜕𝑄𝑖(𝑤) 𝜕𝑄𝑖(𝑤) 𝜕𝑥𝑙 𝜕𝑠𝑙

= ℎ ℎ = 𝜀𝑙 𝑥𝑙−1, 𝑙 = 1̅̅̅, 𝐿̅̅, 𝑘 = 0̅̅,̅𝐻̅̅̅̅̅̅, ℎ = 1̅̅,̅𝐻̅̅̅ (2.52)

𝜕𝑤𝑙

𝑘ℎ

ℎ

ℎ

𝑘ℎ

𝜕𝑥𝑙

𝜕𝑠𝑙

𝜕𝑤𝑙

ℎ 𝑘

𝐿−1 𝑙

* 1. Повторить шаги 2–4 для каждого объекта подвыборки и усреднить зна- чение градиента по всем объектам:

𝑚

𝜕𝑄(𝑤) = 1 ∑ 𝜕𝑄𝑖(𝑤) , 𝑙 = 1̅̅̅, 𝐿̅̅, 𝑘 = 0̅̅,̅𝐻̅̅̅̅̅̅, ℎ = 1̅̅,̅𝐻̅̅̅ (2.53)

𝜕𝑤𝑙

𝑘ℎ

𝑚

𝑘=0

𝜕𝑤𝑙

𝑘ℎ

𝐿−1 𝑙

Теперь с помощью полученных производных можно совершать шаг обнов- ления весов по методу Adam.

Пересчет весов повторяют до тех пор, пока нейронная сеть не обработает всю обучающую выборку заданное число раз. Данное число называется количе- ством эпох и является гиперпараметром.

# ГЛАВА 3. СРАВНЕНИЕ КАЧЕСТBА РАБОТЫ МЕТОДОВ НА ИССЛЕДУЕМЫХ ДАННЫХ

В первом параграфе данной главы описаны метрики качества, используе- мые для оценки качества работы методов классификации. Затем рассмотрены де- тали предобработки данных, реализации методов и способа подбора гиперпара- метров. В заключительной части главы приведены результаты сравнения работы методов на исследуемых данных и сделаны выводы о том, какой метод бинарной классификации наиболее предпочтителен для данной задачи.

## Метрики оценки качества классификации

Рассмотрим метрики, которые будут использоваться в данной работе для оценки точности полученных моделей [12].

### Матрица ошибок

*Матрица ошибок* (англ. confusion matrix) представляет собой таблицу, со- держащую информацию о том, сколько раз классификатор принял верное или не- верное решение. Общий вид матрицы ошибок приведен в таблице 3.1.

Таблица 3.1. Матрица ошибок.

|  |  |  |
| --- | --- | --- |
| **Истинный класс** | **Предсказанный класс** | |
| 0 | 1 |
| 0 | TN | FP |
| 1 | FN | TP |

В ячейках матрицы находятся следующие значения:

1. TN (True Negative) — число объектов класса с меткой 0, которые были опре- делены верно
2. FP (False Positive) — число объектов класса с меткой 0, которые были опре- делены неверно
3. FN (False Negative) — число объектов класса с меткой 1, которые были опре- делены неверно
4. TN (True Positive) — число объектов класса с меткой 1, которые были опре- делены верно

### Кросс-валидационная точность

*Точность* (англ. accuracy) — это отношение числа правильно классифици- рованных объектов к общему числу объектов в наборе данных:

𝑇𝑃 + 𝑇𝑁

𝑎𝑐𝑐𝑢𝑟𝑎𝑐𝑦 =

𝑇𝑃 + 𝐹𝑃 + 𝑇𝑁 + 𝐹𝑁

(3.1)

Точность является одной из наиболее распространенных метрик оценки ка- чества моделей классификации, однако плохо характеризует модель в задачах с несбалансированным числом объектов разных классов в обучающей выборке.

Точность работы модели обычно оценивают с помощью тестовой выборки, которая выделяется из набора данных и не участвует в обучении модели. Если доступный набор данных не обладает большим объемом, часто прибегают к ме- тоду, который называется *кросс-валидацией по* 𝑘 *блокам* (англ. k-fold cross valida- tion). При применении кросс-валидации все данные разделяют на 𝑘 равных ча- стей. После этого 𝑘 раз повторяют следующий процесс: 𝑘 − 1 блок используют для построения модели с выбранными гиперпараметрами, а оставшийся блок ис- пользуется для оценки качества работы модели, таким образом, чтобы каждое подмножество данных использовалось в качестве тестового ровно один раз. Кросс-валидационная оценка точности вычисляется как среднее полученных 𝑘 оценок. Оптимальным числом блоков считается 𝑘 = 10.

### Площадь под кривой ошибок

*ROC-кривая* — графический инструмент оценки точности моделей бинар- ной классификации, позволяющий найти оптимальный баланс между *чувстви- тельностью*:

и *специфичностью* модели:

𝑠𝑒𝑛𝑠𝑖𝑡𝑖𝑣𝑖𝑡𝑦 =

𝑠𝑝𝑒𝑐𝑖𝑓𝑖𝑐𝑖𝑡𝑦 =

𝑇𝑃

𝑇𝑃 + 𝐹𝑁

𝑇𝑁

𝑇𝑁 + 𝐹𝑃

(3.2)

(3.3)

ROC-кривая задается в осях (𝑠𝑒𝑛𝑠𝑖𝑡𝑖𝑣𝑖𝑡𝑦, 1 − 𝑠𝑝𝑒𝑐𝑖𝑓𝑖𝑐𝑖𝑡𝑦) и отображает соотношение этих величин при варьировании порога решающего правила (объ- екты, для которых предсказанная вероятность принадлежности к классу с меткой 1 ниже этого порога, относят к классу с меткой 0, остальные — к классу с меткой 1).

*Площадь под ROC-кривой* (англ. area under the curve, AUС) является ее чис- ленной характеристикой, которой удобно пользоваться для сравнения моделей. Чем выше значение AUC для классификатора, тем лучше его способность разли- чать положительные и отрицательные классы. Идеальный классификатор будет иметь 𝐴𝑈𝐶 = 1, в то время как классификатор, присваивающий метки класса слу- чайным образом, будет иметь 𝐴𝑈𝐶 = 0,5.

## Предварительный анализ и предобработка данных

Перед использованием данных, содержащих измеренные морфологиче- ские параметры растущих колоний hPSC и клеток внутри колоний вместе с ин- формацией о качестве колоний, для построения моделей классификации, прове- дем предварительный анализ и предобработку данных.

### Фильтрация данных

Поскольку некоторые объекты не содержат информацию о качестве коло- нии, необходимо произвести фильтрацию данных, чтобы удалить строки, кото- рые не могут быть использованы при построении классификаторов. Произведя фильтрацию, можем определить соотношение классов в данных (таблица 3.2).

Таблица 3.2. Количество объектов с «хорошим» и «плохим» фенотипом.

|  |  |  |
| --- | --- | --- |
|  | **Клеточные данные** | **Колониальные данные** |
| **«Хороший» фенотип** | 794 | 70 |
| **«Плохой» фенотип** | 1045 | 76 |

Заметим, что данные можно считать сбалансированными, поэтому точ- ность (accuracy) может быть использована в качестве метрики для сравнения ка- чества работы моделей.

### Дисперсионный и корреляционный анализ данных

Дисперсионный анализ данных, проведенный в статье [7], показал, что многие признаки имеют статистически значимую разницу между средними зна- чениями для объектов разных классов, из чего можно сделать вывод, что они ока- зывают значимое влияние на фенотип колонии. Тем не менее, авторы так же по- казали, что разные линии характеризуются различными морфологическими па- раметрами, чувствительными к фенотипическому разделению, следовательно су- ществует необходимость построения комплексной модели классификации.

Для оценки распределения признаков и их взаимосвязи построим точечные диаграммы рассеяния для каждой пары признаков и подсчитаем *коэффициенты корреляции Пирсона*:

∑𝑘 (𝑎𝑖 − 𝑎̅)(𝑏𝑖 − 𝑏̅)

𝑟𝑎𝑏

= 𝑖=1 , (3.4)

𝜎𝑎𝜎𝑏

где 𝑎 = (𝑎1, … 𝑎𝑘), 𝑏 = (𝑏1, … 𝑏𝑘) — некоторые выборки, 𝑎̅ и 𝑏̅ — выборочные средние, 𝜎𝑎 и 𝜎𝑏 — среднеквадратические отклонения.

Результаты корреляционного анализа представлены на рисунках 3.1 (для клеточных данных) и 3.2 (для колониальных данных).

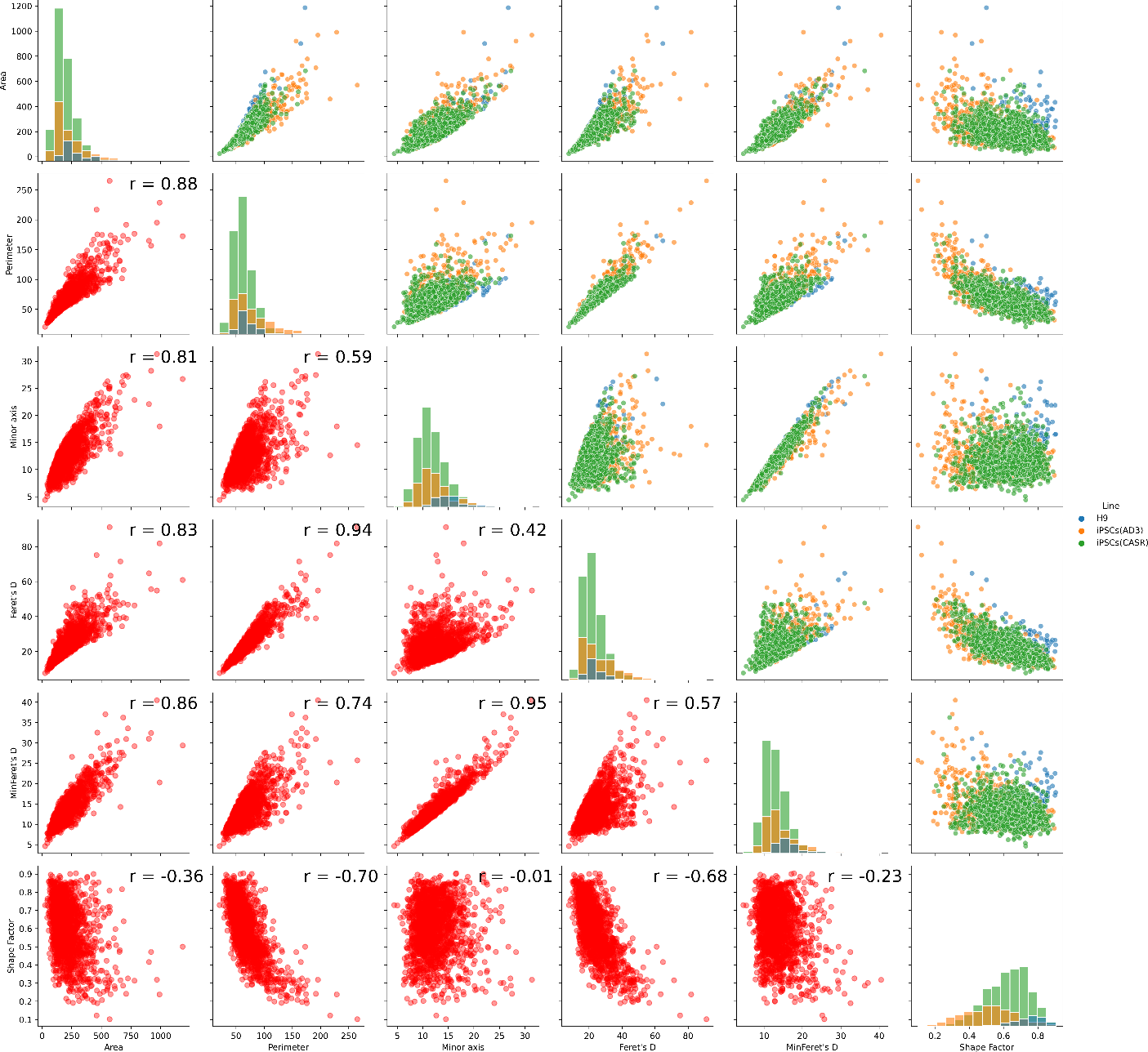


Рисунок 3.1. Точечные диаграммы рассеяния для клеточных данных.

Голубой цвет соответствует линии эмбриональных hPSC H9, желтый — ли- нии индуцированных hPSC AD3, зеленый — линии индуцированных hPSC CASR. На диагоналях расположены диаграммы распределения каждого из при- знаков всех трех линий, над диагональю — диаграммы рассеяния с разделением объектов по линиям, под диагональю — диаграммы рассеяния без разделения. На графиках под диагональю также указано значение коэффициента корреляции

𝑟 для соответствующих пар признаков.

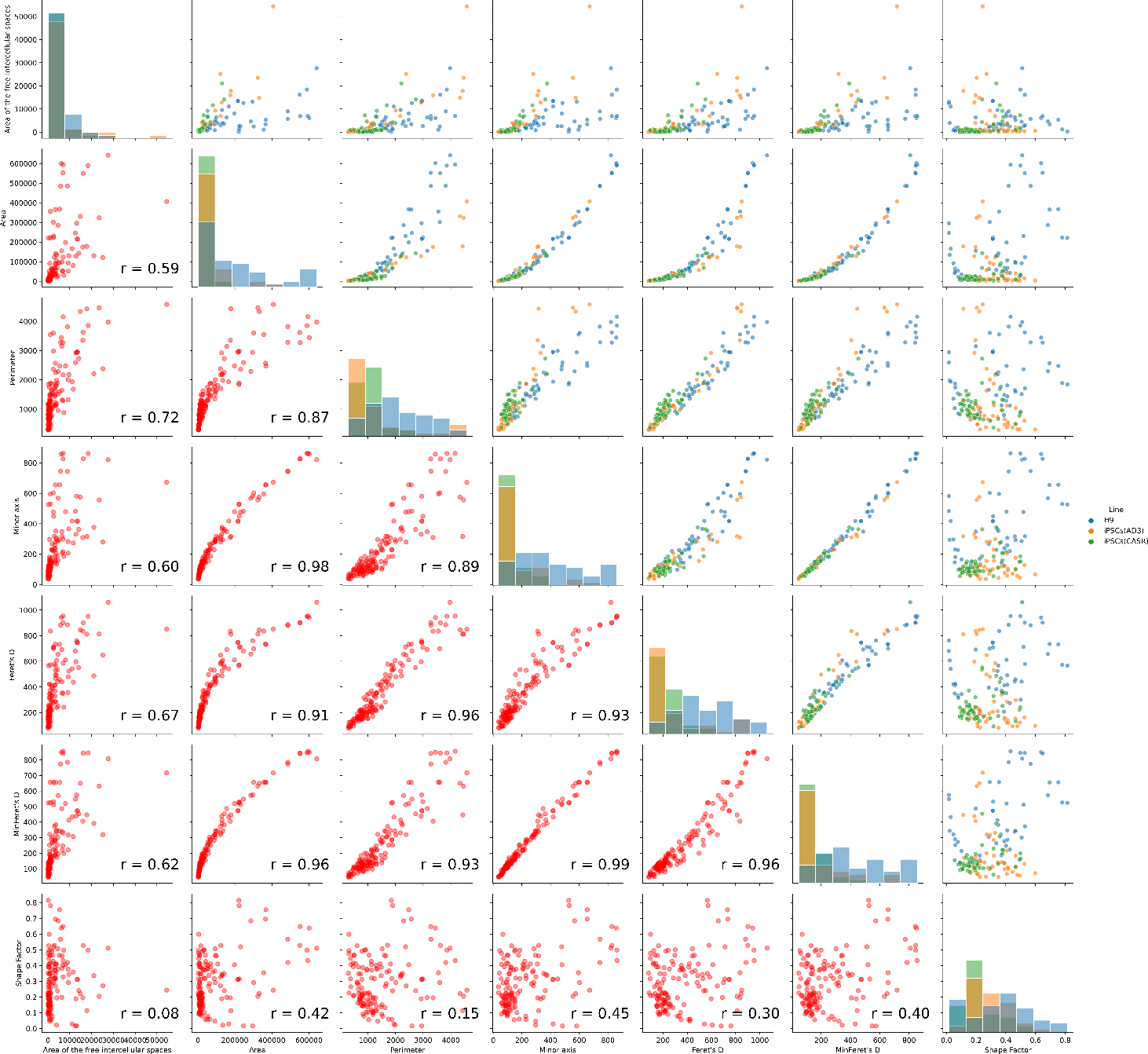


Рисунок 3.2. Точечные диаграммы рассеяния для колониальных данных.

Можно заметить, что как в клеточных, так и в колониальных данных мно- гие признаки достаточно сильно коррелированы (𝑟 по модулю близок к единице). Сильная зависимость признаков может негативно влиять на многие модели клас- сификации, поэтому необходимо прибегать к использованию регуляризации в тех моделях, в которых это возможно. Еще одним способом устранения негативного влияния коррелированности признаков является использование для построения классификатора только части признаков, обладающих наименьшей взаимной корреляцией. Отбор наиболее важных признаков будет рассмотрен в главе 4.

### Стандартизация данных

Рассмотрим среднее значение для каждого признака в клеточных и коло- ниальных данных (таблица 3.3).

Таблица 3.3. Средние значения признаков для колониальных и клеточных данных.

|  |  |  |
| --- | --- | --- |
| **Параметр** | **Клеточные данные** | **Колониальные данные** |
| Площадь, кв. мкм | 212.21 | 96303.19 |
| Периметр, мкм | 67.4 | 1521.65 |
| Малая ось, мкм | 12.3 | 237.13 |
| Диаметр Ферета, мкм | 23.67 | 371.3 |
| Минимальный диаметр  Ферета, мкм | 13.69 | 265.37 |
| Форм-фактор | 0.6 | 0.295 |
| Площадь межклеточного  пространства, кв. мкм | - | 4104.92 |

Из приведённой таблицы видно, что значения различных признаков отли- чаются на порядки. Большая разница в масштабах признаков может негативно сказаться на работе методов классификации, основанных на расстояниях в про- странстве признаков, поэтому перед применением метода 𝑘 ближайших соседей и метода опорных векторов данные необходимо стандартизировать. Кроме того, стандартизацию стоит также использовать при построении логистической ре- грессии и нейронных сетей, поскольку это обычно приводит к ускорению их обу- чения.

Стандартизация изменяет масштаб набора данных, таким образом, чтобы среднее значение было равно нулю, а стандартное отклонение — единице:

𝑥(𝑖)′

𝑥(𝑖) − 𝑥̅(𝑖)

=

𝜎𝑖

, (3.5)

где 𝑥̅(𝑖) — среднее значение признака, 𝜎𝑖 — среднеквадратическое отклонение.

## Реализация алгоритмов классификации и подбор гиперпараметров

Для применения к исследуемым данным методов бинарной классифика- ции, рассмотренных во второй главе, использовался язык программирования Py- thon 3.8 и библиотеки, предназначенные для реализации методов машинного обу- чения, такие как sklearn и keras (для работы с искусственной нейронной сетью).

Таблица 3.4. Гиперпараметры моделей.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Метод** | **Гиперпараметры** | **Диапазон поиска** | **Лучшее значение** | |
| **Клеточные**  **данные** | **Колониаль-**  **ные данные** |
| Наивный байесов-  ский классификатор | - | - | - | - |
| k ближайших  соседей | число соседей | [1, 100] | 29 | 11 |
| Логистическая  регрессия | коэффициент  регуляризации | от 10−7 до 107 | 10−1 | 10−4 |
| Случайный лес | критерий  разбиения | gini, entropy | entropy | gini |
| число деревьев | 10, 20, 50, 100,  150 | 50 | 20 |
| максимальная  глубина дерева | 2, 5, 10, 100 | 5 | 5 |
| число признаков  для разбиения | 2, 3, 5, sqrt, log2 | log2 | log2 |
| Метод опорных векторов | вид ядра | linear, poly, rbf, sigmoid | rbf | rbf |
| коэффициент  регуляризации | от 10−3 до 103 | 1 | 103 |
| Искусственная нейронная сеть | конфигурация скрытых слоев | - | 13:5 | 10:3 |
| функции актива- ции между слоями | - | relu | relu |
| число эпох | - | 150 | 100 |
| размер шага | - | 0.0001 | 0.001 |

Для подбора гиперпараметров всех методов, кроме наивного байесовского классификатора (не имеет гиперпараметров) и нейронной сети (конфигурация была подобрана вручную), был использован метод поиска по сетке (англ. Grid

Search) [20]. Данный метод находит наилучшие гиперпараметры путем полного перебора в заданном диапазоне значений и сравнения полученных моделей с по- мощью кросс-валидации по пяти блокам. Поиск по сетке не всегда применим из- за больших вычислительных затрат, однако он хорошо подходит для исследуе- мых данных, поскольку они не обладают большим объемом.

Конфигурации искусственных сетей для клеточных и колониальных дан- ных были выбраны на основе общих соображений о построении оптимальных сетей для классификации [21] и сравнении топологий с помощью кросс-валида- ционной точности. После скрытых слоев также были добавлены слои, осуществ- ляющие пакетную нормализацию данных.

Гиперпараметры, диапазоны, в которых производился поиск и найденные наилучшие значения гиперпараметров приведены в таблице 3.4.

## Сравнение качества работы полученных моделей

Оценки качества работы моделей с наилучшими гиперпараметрами при- ведены в таблице 3.5 (ROC-кривые приведены в приложении 1).

Таблица 3.5. Сравнение качества работы моделей.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Метод** | **Клеточные данные** | | **Колониальные данные** | |
| **Кросс-валида-**  **ционная точность** | **ROC AUC** | **Кросс-валида-**  **ционная точность** | **ROC AUC** |
| Наивный Байес | 59.54% | 0.685 | 59.52% | 0.708 |
| (±3.59%) | (±13.49%) |
| k ближайших | 64.27% | 0.66 | 66.48% | 0.708 |
| соседей | (±1.88%) | (±11.98%) |
| Логистическая | 59.11% | 0.629 | **74.81%** | **0.895** |
| регрессия | (±3.65%) | **(**±**12.84%)** |
| Случайный лес | 64.06% | 0.665 | 64.19% | 0.786 |
| (±1.56%) | (±12.59%) |
| Метод опорных | 64.55% | 0.676 | 67.81% | 0.857 |
| векторов | (±2.23%) | (±8.21%) |
| Искусственная нейронная сеть | **67.48%**  **(**±**3.79%)** | **0.699** | 71.05%  (±11.54%) | 0.894 |

На рисунке 3.3 приведены боксплоты распределений, полученных в про- цессе кросс-валидационной оценки каждой модели.

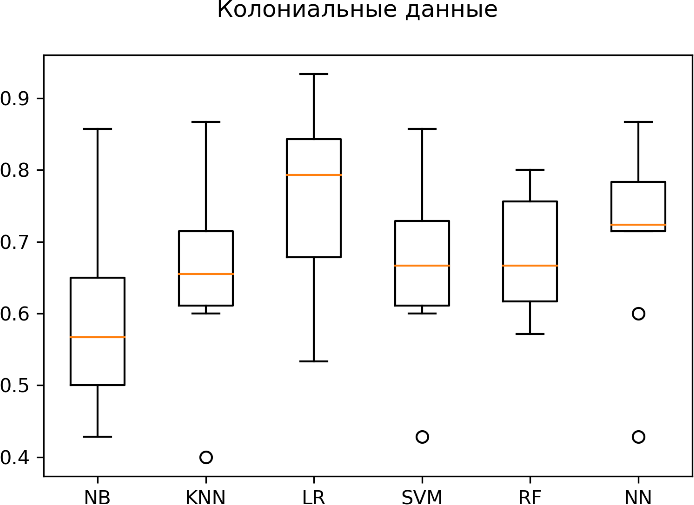
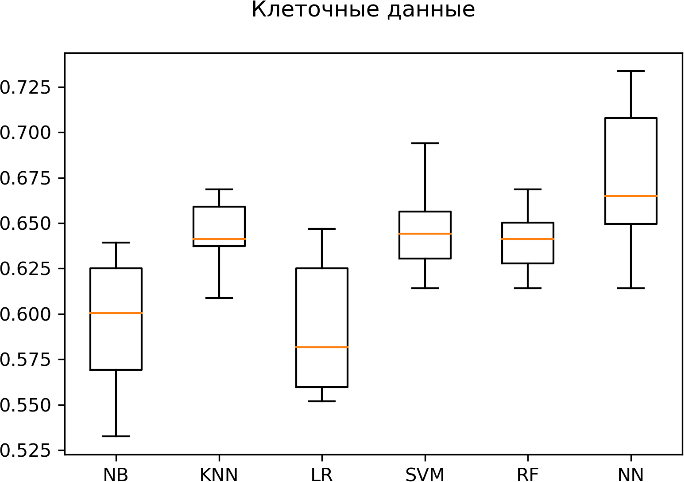


Рисунок 3.3. Боксплоты распределений точности моделей для клеточных (слева) и колониаль- ных данных (справа). NB — наивный байесовский классификатор, KNN — метод 𝑘 ближай- ших соседей, LR — логистическая регрессия, SVC — метод опорных векторов, RF — слу- чайный лес, NN — искусственная нейронная сеть.

В случае клеточных данных среднее значение кросс-валидационной точно- сти *искусственной нейронной сети* (67.48%) превышает точность остальных мо- делей. T-критерий Стьюдента показывает статистически значимое отличие (𝑝 < 0.05) между средними значениями точности для нейронной сети и всех осталь- ных моделей, кроме метода опорных векторов. Учитывая также то, что искус- ственная нейронная сеть обладает наибольшим значением ROC AUC, можно утверждать, что данная модель является наиболее эффективной для классифика- ции на основе морфологии отдельных клеток.

В случае колониальных данных, наиболее высокое среднее значение точ- ности имеет *метод логистической регрессии* (74.81%), однако t-критерий Стью- дента не показывает статистическую значимость отличий. Тем не менее, учиты- вая различия средних значений точности и значений метрики ROC AUC для рас- смотренных моделей, можно считать, что наиболее эффективным для классифи- кации на основе морфологии колоний является метод логистической регрессии.

Обучим модели, показавшие наибольшую точность, на данных каждой из трех клеточных линий по-отдельности, используя найденные гиперпараметры. Кросс-валидационная точность полученных моделей приведена в таблице 3.6.

Таблица 3.6. Сравнение качества работы моделей для отдельных клеточных линий.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Клеточная линия** | | |
| **hESC H9** | **hiPSC AD3** | **hiPSC CaSR** |
| Клеточные данные  (нейронная сеть) | 73.05%  (±7.15%) | 60.32%  (±6.06%) | 63.92%  (±3.06%) |
| Колониальные данные  (логистическая регрессия) | 75.33%  (±17.84%) | 61.50%  (±20.01) | 67.00%  (±17.20%) |

Модели, построенные как на клеточных, так и на колониальных данных клеточной линии hESC H9, в среднем показывают результаты лучше, чем модели, для построения которых использовался полный набор данных. В то же время кросс-валидационная точность моделей, обученных на данных клеточных линий hiPSC AD3 и hiPSC CaSR, в обоих случаях уступает точности полных моделей.

## Сравнение моделей, обученных на объединенных данных

Объединим клеточные и колониальные данные, добавив к морфологиче- ским данным каждой клетки морфологические параметры колонии, в которой эта клетка находится, и найдем оптимальную модель, основанную на комбинации клеточных и колониальных данных (результаты подбора оптимальных гиперпа- раметров приведены в приложении 2).

Таблица 3.7. Сравнение качества работы моделей на объединенных данных.

|  |  |  |
| --- | --- | --- |
| **Метод** | **Кросс-валидационная точность** | **ROC AUC** |
| Наивный Байес | 72.76% (±4.99%) | 0.815 |
| k ближайших соседей | 88.97% (±5.24%) | 0.818 |
| Логистическая регрессия | 79.48% (±5.94%) | 0.826 |
| Случайный лес | **98.28% (**±**1.89%)** | **0.997** |
| Метод опорных векторов | 95.52% (±2.91%) | 0.975 |
| Искусственная нейронная сеть | 97.59% (±2.46%) | 0.956 |

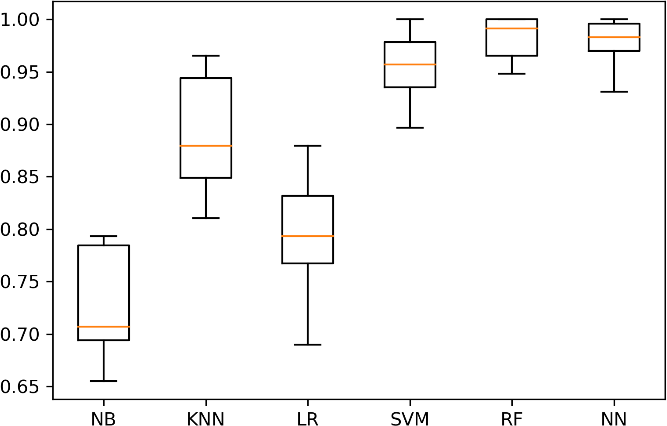
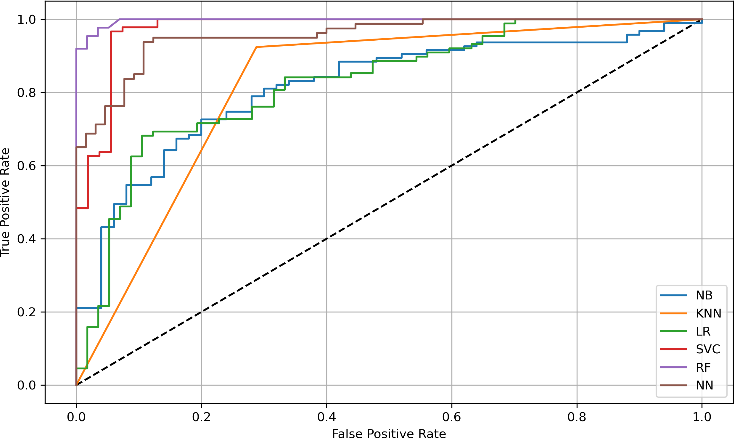
 

Рисунок 3.4. Боксплоты распределений точности моделей (слева) и ROC-кривые (справа) для объединенных данных. Расшифровка легенды аналогична рисунку 3.4.

Результаты сравнения работы методов приведены в таблице 3.7 и на ри- сунке 3.4. Наилучшее качество по совокупности двух метрик показывает метод случайного леса (98.28%), причем t-критерий Стьюдента показывает статистиче- ски различимое отличие средней кросс-валидационной точности для метода слу- чайного леса в сравнении со всеми методами, кроме искусственной нейронной сети. В таблице 3.8 приведена матрица ошибок метода случайного леса на тесто- вых данных (данные разделены на обучающие и тестовые в соотношении 3:1).

Таблица 3.8. Матрица ошибок для наилучшего классификатора, построенного по объединенным данным.

|  |  |  |
| --- | --- | --- |
| **Истинный класс** | **Предсказанный класс** | |
| **bad** | **good** |
| **bad** | 55 | 1 |
| **good** | 3 | 86 |

Качество моделей, обученных на совокупности клеточных и колониальных данных по обоим метрикам значительно превосходит качество моделей, исполь- зующих данные по-отдельности. Такая высокая оценка может быть связана со структурой данных (для клеток из одной колонии все признаки, унаследованные из колониальных данных, полностью совпадают) и не гарантирует столь же вы- сокое качество классификации на новых данных, однако показывает эффектив- ность совместного использования данных о морфологии колонии и клеток внутри нее для определения фенотипа колонии.

# ГЛАВА 4. ОТБОР ПРИЗНАКОВ

Анализ данных о морфологии колоний и клеток hPSC и оценка качества моделей, построенных на этих данных, показали, что полученные модели склонны к переобучению. Одним из способов борьбы с переобучением является отбор наиболее значимых признаков. Кроме того, оценка влияния каждого из признаков на качество построенных классификаторов может помочь лучше по- нять взаимосвязь между морфологическими признаками и фенотипом колоний.

## Методы отбора признаков

Существующие методы отбора признаков делятся на три основных кате- гории [22]:

1. Методы фильтрации (англ. filter methods)

Методы этой группы применяются до обучения моделей и основаны на сравне- нии некоторых статистических показателей. Методы фильтрации обычно рас- сматривают каждый признак независимо, определяя степени корреляции призна- ков с целевой переменной и ранжируя их по полученным значениям.

1. Методы-оболочки (англ. wrapper methods)

В методах-оболочках значимость признаков оценивается с помощью построения моделей классификации на подмножествах признаков и оценки и сравнении ка- чества работы полученных моделей.

1. Встроенные методы (англ. embedded methods)

Встроенные методы позволяют отбирать признаки в процессе построения мо- дели. Данные методы вводят дополнительные ограничения при оптимизации ал- горитма, смещающие модель в сторону меньшей сложности. Наиболее распро- страненными методами этой группы являются методы регуляризации.

В данной работе остановимся на рассмотрении двух алгоритмов, относя- щихся к методам-оболочкам, поскольку методы данной группы показывают до- статочно высокую точность оценки на практике и являются хорошо

интерпретируемыми.

### Метод SHAP

В основе метода SHAP (от англ. Shapley additive explanations, аддитивные объяснения Шепли) лежат понятия из кооперативной теории игр — раздела при- кладной математики, изучающего методы принятия оптимальных решений в кон- фликтных ситуациях [23]. Идея метода состоит в том, чтобы рассматривать от- дельные признаки как игроков, а всю совокупность признаков — как команду. Каждый игрок вносит свой вклад в результат команды, а сумма этих вкладов определяет значение целевой переменной.

Значение SHAP является оценкой вклада отдельного признака в величину предсказания выбранной модели и может быть найдено с помощью формулы:

𝜑(𝑥(𝑖)) = ∑ |𝑆|! (|𝐹| − |𝑆| − 1)! (𝑣(𝑆 ∪ {𝑖}) − 𝑣(𝑆)), (4.1)

|𝐹|!

𝑆⊆𝐹/{𝑖}

где F — полный набор признаков, 𝑆 — произвольный набор признаков, не содер- жащий 𝑖-ый признак, 𝑣(𝑆) — характеристическая функция набора признаков:

𝑣(𝑆) = 𝐸[𝑓(𝑥)|𝑥𝑆], (4.2)

— условное математическое ожидание предсказания 𝑓 на примерах 𝑥′, взятых из распределения данных, таких что 𝑥′ = 𝑥𝑆.

𝑆

Точное вычисление значений SHAP является сложной задачей, однако ав-

торы метода предлагают способы аппроксимации, позволяющие найти оценки значений SHAP [23].

### Исчерпывающий поиск признаков

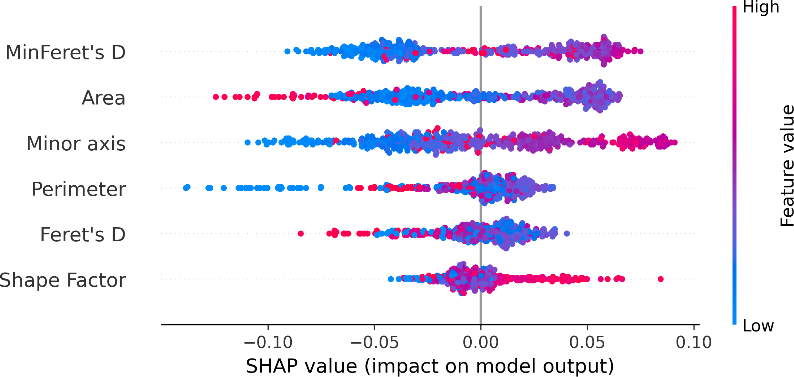
Исчерпывающий поиск — это метод поиска наиболее важных признаков, который гарантирует нахождение наилучшего набора [22]. Суть данного метода заключается в полном переборе всевозможных комбинаций признаков, использу- емых для построения выбранной модели машинного обучения, и сравнения кросс-валидационного качества полученных классификаторов. Метод

исчерпывающего поиска редко применяют на практике из-за его высокой вычис- лительной сложности, тем не менее, он может быть использован на данных, со- держащих небольшое количество признаков.

## Результаты применения методов отбора признаков

Применим рассмотренные методы отбора признаков к тем моделям, кото- рые в главе 3 показали наилучшее качество классификации по совокупности кросс-валидационной точности и ROC AUC. Для реализации методов исполь- зуем Python библиотеку shap и модуль feature\_selection Python библиотеки mlxtend.

В качестве оптимальной модели классификации для клеточных данных в главе 3 была выбрана искусственной нейронная сеть (кросс-валидационная точ- ность на полном наборе признаков равна 67.48%). Результат применения **метода оценки значений SHAP** к данной модели приведен на рисунке 4.1.



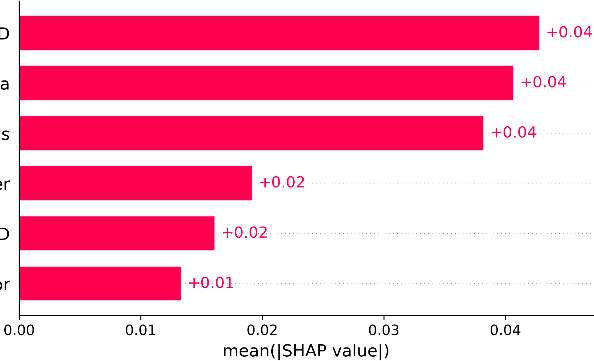


Рисунок 4.1. Значения SHAP для модели, построенной на клеточных данных.

Используя полученное ранжирование признаков, построим модели на уменьшенных множествах признаков, по очереди исключая те признаки, которые имеют наименьшее по модулю среднее значение SHAP. График зависимости кросс-валидационной точности от набора признаков для полученных моделей приведен на рисунке 4.2. Сокращение количества признаков меняет среднюю кросс-валидационную точность модели менее чем на один процент.

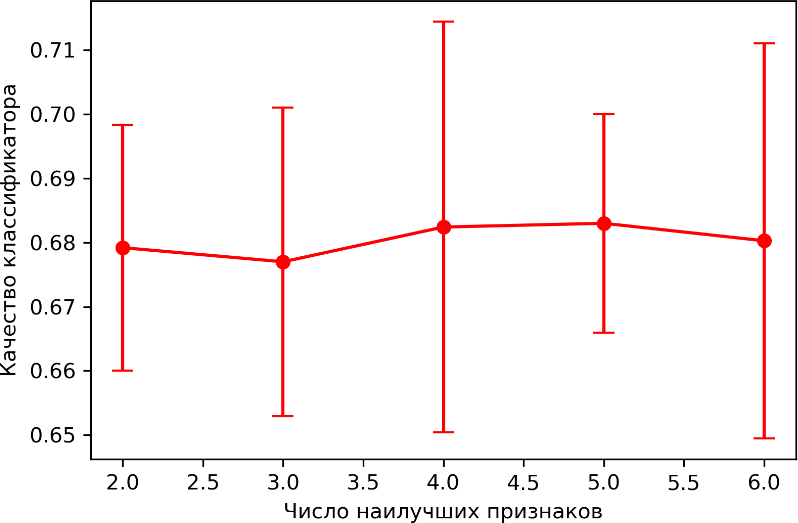


Рисунок 4.2. График зависимости кросс-валидационной точности (± среднеквадратическое отклонение) от набора признаков для моделей, построенных на клеточных данных.

Наибольшее значение средней кросс-валидационной точности достигается на модели, использующей пять параметров (минимальный диаметр Ферета, пло- щадь, малая ось, периметр, диаметр Ферета), и составляет 68.24% (±3.20%).

В то же время **методом исчерпывающего поиска** как оптимальное было найдено подмножество, состоящее из трех признаков: площадь, минимальный диаметр Ферета, форм-фактор. Кросс-валидационная точность модели, постро- енной на данных признаках, составляет 68.11% (±1.54%), что почти не уступает модели, построенной на пяти признаках. В таблице 4.1 приведена матрица оши- бок полученной модели на тестовых данных (данные разделены на обучающие и тестовые в соотношении 3:1).

Таблица 4.1. Матрица ошибок для оптимального классификатора, построенного по клеточным данным.

|  |  |  |
| --- | --- | --- |
| **Истинный класс** | **Предсказанный класс** | |
| **bad** | **good** |
| **bad** | 145 | 128 |
| **good** | 83 | 359 |

Для колониальных данных в качестве оптимальной модели классификации в главе 3 был выбран метод логистической регрессии (кросс-валидационная точ- ность на полном наборе признаков равна 74.81%). Результат применения **метода оценки значений SHAP** к данной модели приведен на рисунке 4.3.

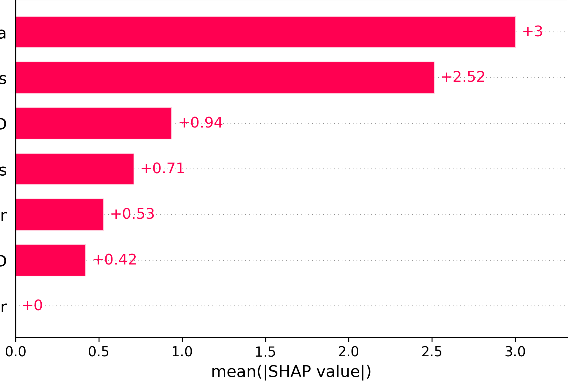
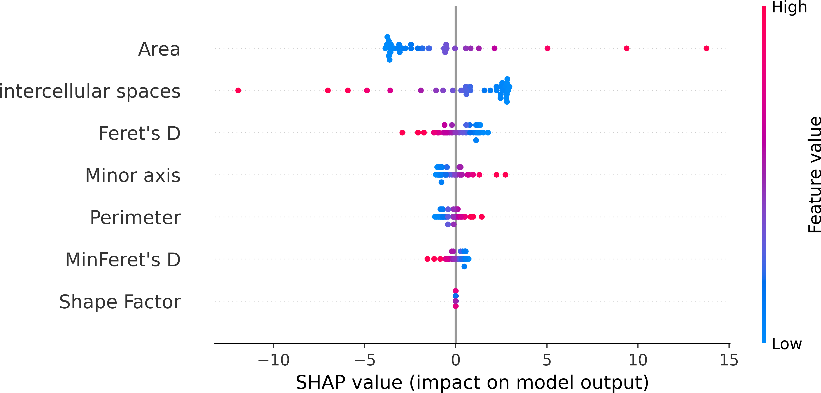


Рисунок 4.3. Значения SHAP для модели, построенной на колониальных данных.

Построим набор моделей, использующих различное количество признаков, так же, как это было сделано для клеточных данных. График зависимости кросс- валидационной точности от набора признаков для полученных моделей приведен на рисунке 4.4.

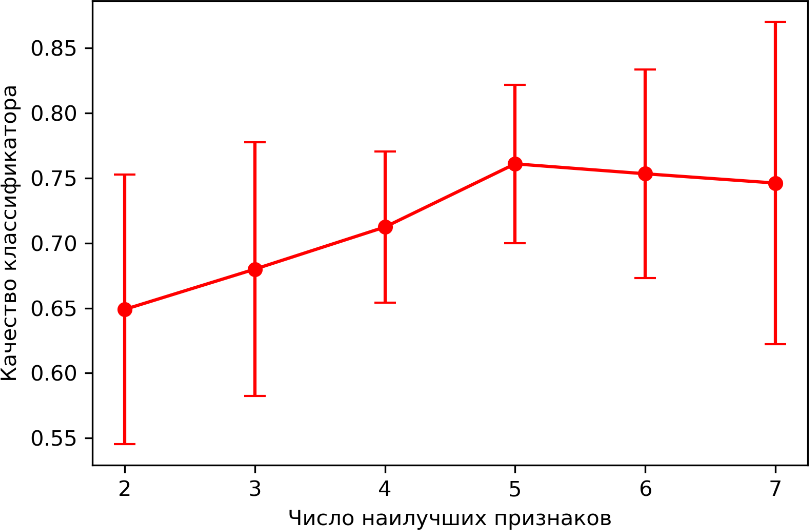


Рисунок 4.4. График зависимости кросс-валидационной точности (± среднеквадратическое отклонение) от набора признаков для моделей, построенных на клеточных данных.

Наибольшее значение средней кросс-валидационной точности достигается на модели, использующей пять параметров (площадь, площадь межклеточного пространства, диаметр Ферета, малая ось, периметр), и составляет 76.10% (±6.09%). Примечательно, что в данном случае **метод исчерпывающего поиска** нашел в качестве оптимального то же самое подмножество признаков. Построим матрицу ошибок полученной модели на тестовых данных (данные разделены на обучающие и тестовые в соотношении 3:1) (таблица 4.2).

Таблица 4.2. Матрица ошибок для оптимального классификатора, построенного по колониальным данным.

|  |  |  |
| --- | --- | --- |
| **Истинный класс** | **Предсказанный класс** | |
| **bad** | **good** |
| **bad** | 18 | 1 |
| **good** | 6 | 12 |

Приведенные исследования показывают, что уменьшение количества признаков, используемых для построения моделей, не снижает качество классификации и может послужить способом борьбы с переобучением.

# ЗАКЛЮЧЕНИЕ

В данной работе была рассмотрена проблема классификации колоний плю- рипотентных стволовых клеток человека по их фенотипу — потенциальной спо- собности к поддержанию плюрипотентности и клональности, на основе морфо- логических признаков колоний и клеток внутри них.

Задача построения классификатора была поставлена как задача машинного обучения с учителем и для ее решения был проведен обзор следующих шести наиболее известных методов бинарной классификации: наивный байесовский классификатор, метод 𝑘 ближайших соседей, логистическая регрессия, метод опорных векторов, метод случайного леса и многослойная полносвязная искус- ственная нейронная сеть. Для этих методов были подобраны оптимальные зна- чения гиперпараметров и проведено сравнение качества их работы на исследуе- мых данных. Полученные оптимальные модели для клеточных и колониальных данных обладают кросс-валидационной точностью, равной 67.48% (искусствен- ная нейронная сеть) и 74.81% (логистическая регрессия), соответственно. Были также рассмотрены особенности обучения моделей на данных отдельных клеточ- ных линий. Кроме того, была построена модель, использующая для предсказа- ния качества колонии комбинацию клеточных и колониальных данных. Опти- мальная модель, построенная на объединенном наборе данных, показала кросс- валидационную точность, равную 98.28%.

Последним шагом работы стало рассмотрение методов отбора признаков и результатов их применения к полученным моделям. Как в случае клеточных, так и в случае колониальных данных удалось, снизив количество признаков до пяти, немного повысить качество моделей — до 68.11% и 76.10% соответственно.

Полученные результаты демонстрируют возможность эффективной оценки качества колоний hPSC на основе морфологических параметров колоний и кле- ток внутри них и могут стать основой для создания систем автоматического не- инвазивного контроля качества колоний плюрипотентных стволовых клеток че- ловека.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Ramalho-Santos M., Willenbring H.* On the origin of the term "stem cell" // Cell Stem Cell — 2007. — № 1(1). — C. 35–38. — DOI 10.1016/j.stem.2007.05.013
2. *Chagastelles P. C., Nardi N. B.* Biology of stem cells: an overview // Kidney Inter- national Supplements — 2011.— № 1(3). — С. 63–67. — DOI 10.1038/kisup.2011.15
3. Stem cells: past, present, and future / *Zakrzewski W., Dobrzyński M., Szymonowicz*

*M. [и др.]* // Stem Cell Research & Therapy — 2019 — Т. 10 № 68. — DOI https://doi.org/10.1186/s13287-019-1165-5

1. *Pir P., Le Novère N.* Mathematical Models of Pluripotent Stem Cells: At the Dawn of Predictive Regenerative Medicine // Methods in Molecular Biology — 2016. —

№ 1386. — С. 331–350. — DOI 1007/978-1-4939-3283-2\_15

1. *Engle S. J., Puppala D.* Integrating human pluripotent stem cells into drug devel- opment // Cell Stem Cell — 2013. — № 12(6). — С. 669–677. — DOI 10.1016/j.stem.2013.05.011
2. *Maddah M., Shoukat-Mumtaz U., Nassirpour S., Loewke K.* A system for auto- mated, noninvasive, morphology-based evaluation of induced pluripotent stem cell cultures // Journal of Laboratory Automation — 2014. — T. 19 № 5. — С. 454–

460. — DOI 10.1177/2211068214537258

1. Prognostic Analysis of Human Pluripotent Stem Cells Based on Their Morpholog- ical Portrait and Expression of Pluripotent Markers / *Krasnova O. A., Gursky V. V., Chabina A. S. [и др.]* // International Journal of Molecular Sciences — 2022. — № 23(21):12902. — DOI https://doi.org/10.3390/ijms232112902
2. Assessing Morphology of hPSCs: STEMCELL Technologies [Электронный ре- сурс] — URL: https://[www.stemcell.com/technical-resources/methods-li-](http://www.stemcell.com/technical-resources/methods-li-) brary/cell-culture/pluripotent-stem-cells/maintenance-of-pluripotent-stem-

cells/assessing-morphology-of-hpscs.html (дата обращения 19.04.2023)

1. *Вьюгин В. В.* Элементы математической теории машинного обучения — Москва: МФТИ — ИППИ РАН, 2010. — 252 c. — ISBN 978-5-4439-1691-0
2. *Wolpert D.H.* The lack of a priori distinctions between learning algorithms // Neural computation — 1996. — Т. 8 № 7. — С. 1341–1390. — DOI 10.1162/neco.1996.8.7.1341
3. *T. M. Mitchell* Machine Learning — USA: McGraw-Hill, 1997. — 432 c. — ISBN 0070428077
4. *Kelleher J. D., Namee B. M., D’Arcy A.* Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies — USA: The MIT Press, 2015. — 624 c. — ISBN 978-0-262-02944-5
5. *Nocedal J., Wright S. J.* Numerical Optimization — USA: Springer New York, 1990. — 651 c. — ISBN 0-387-98793-2
6. *Воронцов К. В.* Лекции по методу опорных векторов // Курс лекций МФТИ — 2007.
7. *Rojas R.* Neural Networks — Germany: Springer-Verlag, Berlin, 1996. — 502 c.

— ISBN 978-3-642-61068-4

1. *Воронцов К. В.* Нейронные сети: градиентные методы оптимизации // Курс лекций МФТИ — 2022.
2. *Cybenko G.* Approximation by superpositions of a sigmoidal function // Mathemat- ics of Control, Signals and Systems — 1989. — № 2. — С. 303–314. — DOI https://doi.org/10.1007/BF02551274
3. *Soydaner D.* A comparison of optimization algorithms for deep learning // Interna- tional Journal of Pattern Recognition and Artificial Intelligence — 2020. — Т. 34

№ 13, 2052013. — DOI https://doi.org/10.1142/S0218001420520138

1. *Kingma D., Ba J*. Adam: A Method for Stochastic Optimization // 3rd International Conference for Learning Representations, San Diego, CA — 2015. — 12 c. — DOI https://doi.org/10.48550/arXiv.1412.6980
2. Search Algorithms for Automated Hyper-Parameter Tuning / *Zahedi L., Moham- madi F., Rezapour S. [и др.]* // arXiv: 2104.14677 — 2021. — 10 c. — DOI https://doi.org/10.48550/arXiv.2104.14677
3. *Stathakis D.* How many hidden layers and nodes? // International Journal of Re- mote Sensing — 2009. — Т. 30 № 8. — С. 2133–2147, DOI

10.1080/01431160802549278

1. *V. Kumar, S. Minz* Feature Selection: A literature Review // Smart Computing Re- view — 2014. — № 4. — С. 211–229. — DOI: 10.1145/2740070.2626320
2. Lundberg S. M., S.-I. Lee A unified approach to interpreting model predictions // 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA — 2017. — С. 4768–4777 — DOI:

https://doi.org/10.48550/arXiv.1705.07874

Приложение 1.

## ROC-кривые для моделей, построенных на полных наборах данных.

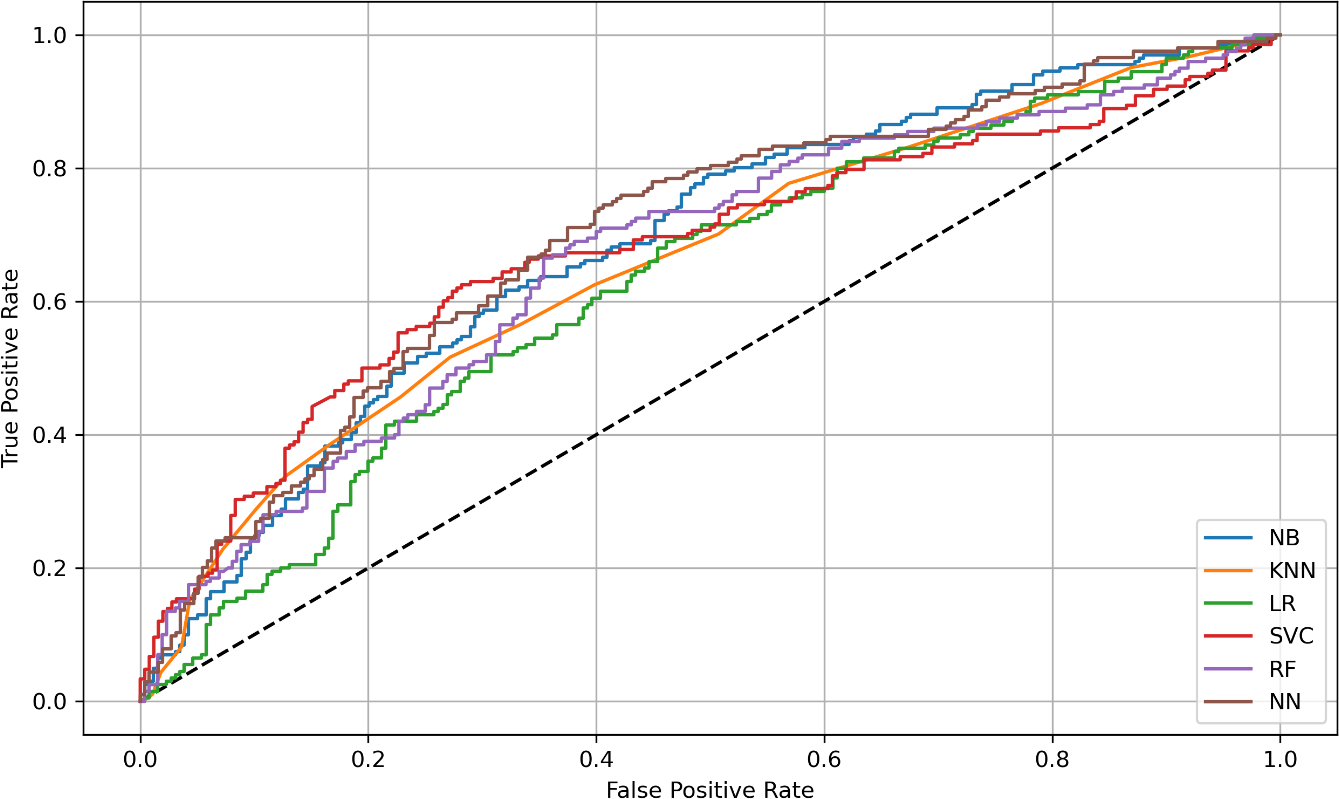
****

Рисунок П1.1. ROC-кривые для моделей, построенных на полных клеточных данных. NB со- ответствует наивному байесовскому классификатору, KNN — методу 𝑘 ближайших соседей, LR — логистической регрессии, SVC — методу опорных векторов, RF — случайному лесу, NN — искусственной нейронной сети.

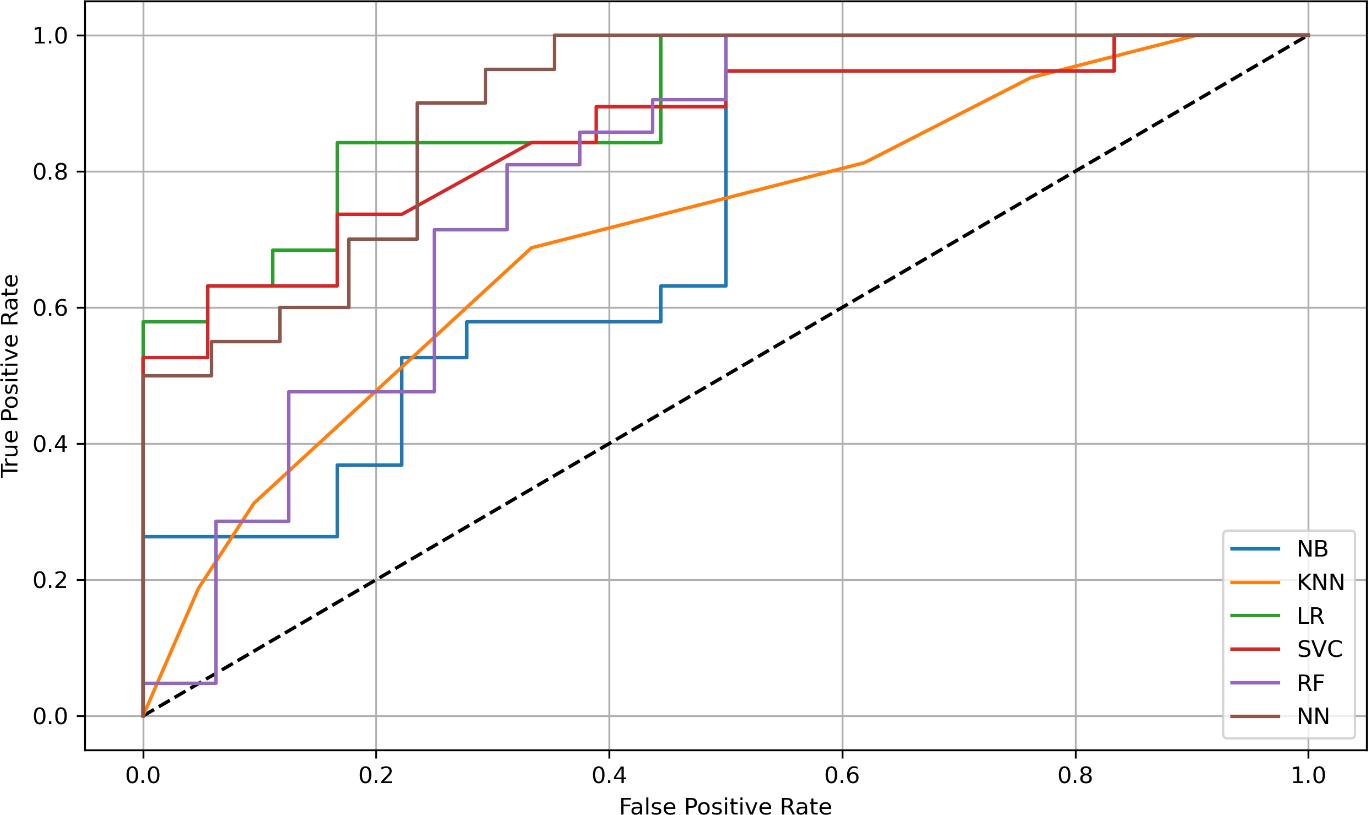


Рисунок П1.1. ROC-кривые для моделей, построенных на полных колониальных данных.

Расшифровка легенды аналогична рисунку П.1.2.

Приложение 2.

## Подбор гиперпараметров для моделей, основанных на объединенных кле- точных и колониальных данных.

Таблица П2.1. Гиперпараметры моделей, построенных на объединенных данных.

|  |  |  |  |
| --- | --- | --- | --- |
| **Метод** | **Гиперпараметры** | **Диапазон**  **поиска** | **Лучшее значение** |
| Наивный байесов-  ский классификатор | - | - | - |
| k ближайших  соседей | число соседей | [1, 100] | 1 |
| Логистическая  регрессия | коэффициент  регуляризации | от 10−7 до 107 | 1 |
| Случайный лес | критерий  разбиения | gini, entropy | entropy |
| число деревьев | 10, 20, 50, 100, 150 | 10 |
| максимальная  глубина дерева | 2, 5, 10, 100 | 20 |
| число признаков для  разбиения | 2, 3, 5, sqrt, log2 | 5 |
| Метод опорных векторов | вид ядра | linear, poly, rbf, sigmoid | rbf |
| коэффициент  регуляризации | от 10−3 до 103 | 102 |
| Искусственная нейронная сеть | конфигурация скрытых слоев | - | 17:6 |
| функции активации между слоями | - | relu |
| число эпох | - | 100 |
| размер шага | - | 0.01 |