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 обучения»

2. Срок сдачи студентом законченной работы: июнь 2023 г.

3. Исходные данные по работе:

Данные с измеренными морфологическими параметрами растущих колоний индуцированных плюрипотентных клеток человека и клеток внутри колоний вместе с информацией о качестве колоний («фенотипе»). Качество колоний, характеризуемое их потенциальной способностью к дальнейшему росту и поддержанию состояний клональности и плюрипотентности, предварительно оценено экспертами.

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 Markers. 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 automated, noninvasive, morphology-based evaluation of induced pluripotent stem cell cultures. 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

4. Содержание работы (перечень подлежащих разработке вопросов):

С помощью методов машинного обучения необходимо разработать эффективную модель классификации колоний по фенотипу, используя измеренные морфологические 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 колоний и клеток с фенотипом растущих колоний.

5. Дата выдачи задания: 14 октября 2022 г.

Руководитель ВКР _____ В. В. Гурский
(подпись)

Задание принял к исполнению

Студент _____ Е. Д. Веденеева
(подпись)

РЕФЕРАТ

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

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

Тема выпускной квалификационной работы: «Определение фенотипа колоний плюрипотентных стволовых клеток человека по морфологическим параметрам методами машинного обучения».

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

В ходе работы были рассмотрены различные методы бинарной классификации объектов по числовым признакам и проведено сравнение качества их работы на клеточных и колониальных морфологических данных. После этого был проведен анализ методов отбора признаков и исследовано влияние их применения 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 предикторов комбинацию клеточных и колониальных параметров. Оптимальная модель достигает качества классификации 98.28%, что показывает высокую эффективность совместного использования данных о морфологии колонии и клеток внутри нее для определения фенотипа колонии.

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

ABSTRACT

50 pages, 10 figures, 11 tables, 2 appendices

KEYWORDS: MACHINE LEARNING, HUMAN PLURIPOTENT STEM CELLS, BINARY CLASSIFICATION, FEATURE SELECTION METHODS, MORPHOLOGICAL 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

Стволовые клетки — это незрелые (недифференцированные) клетки, являющиеся предшественниками всех клеток многоклеточного организма. Ключевыми особенностями стволовых клеток являются их способность к неограниченному 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, мультипотентные, олигопотентные и унипотентные. Из плюрипотентных стволовых клеток развиваются три зародышевых листка — эктодерма, мезодерма и энтодерма, таким 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 медицины, инструментом моделирования заболеваний человека in vitro, создания новых лекарств (в том числе от неизлечимых заболеваний) и разработки банков клеток и тканей [3–5]. Основными способами получения hPSC является выделение из эмбриона человека на определенных стадиях его развития (ESCs), а также получение 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 сосудов со свежей питательной средой для дальнейшего развития. Колонии могут отличаться по уровню потенциала для поддержания плюрипотентности и клональности, характеризирующему их склонность к появлению нежелательных мутаций, поэтому для их безопасного применения в медицинских исследованиях требуется высокий уровень контроля качества [6].

Для оценки качества колоний hPSC используются различные молекулярные 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 количественной и часто может напрямую зависеть от опыта исследователя или тонких морфологических различий между клеточными линиями и клонами. Кроме того, отсутствие единых стандартов затрудняет объективное сравнение результатов различных экспериментов и лабораторий. Существующие коммерческие системы компьютерного анализа не всегда могут быть применимы для многоклеточных колоний hPSC, поскольку они образованы плотно упакованными мелкими клетками (10–16 мкм в диаметре), что может не позволить автоматическим анализаторам 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 машинного 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.

1.1. Описание набора данных и способа их получения

Для проведения исследований будем использовать данные с измеренными морфологическими параметрами растущих колоний плюрипотентных клеток человека и клеток внутри колоний вместе с информацией о качестве колоний («фенотипе»), полученные авторами статьи [7]. Авторы данной статьи сфокусировались 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; выступающие ядрышки; центры колоний очень плотные и имеют четко очерченный край [8]. В соответствии с этим, авторами было выделено семь параметров, характеризующих 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)	$4\pi \frac{\text{Площадь}}{\text{Периметр}^2}$ (мера округлости и компактности)
Площадь межклеточного пространства (только для колоний)	Общая площадь свободного межклеточного пространства в колонии (показатель компактности упаковки клеток)

Данные были получены из фазово-контрастных изображений с использованием Adobe Photoshop и программы ImageJ. Примеры изображений колоний с «хорошим» и «плохим» фенотипами представлены на рисунке 1.1.

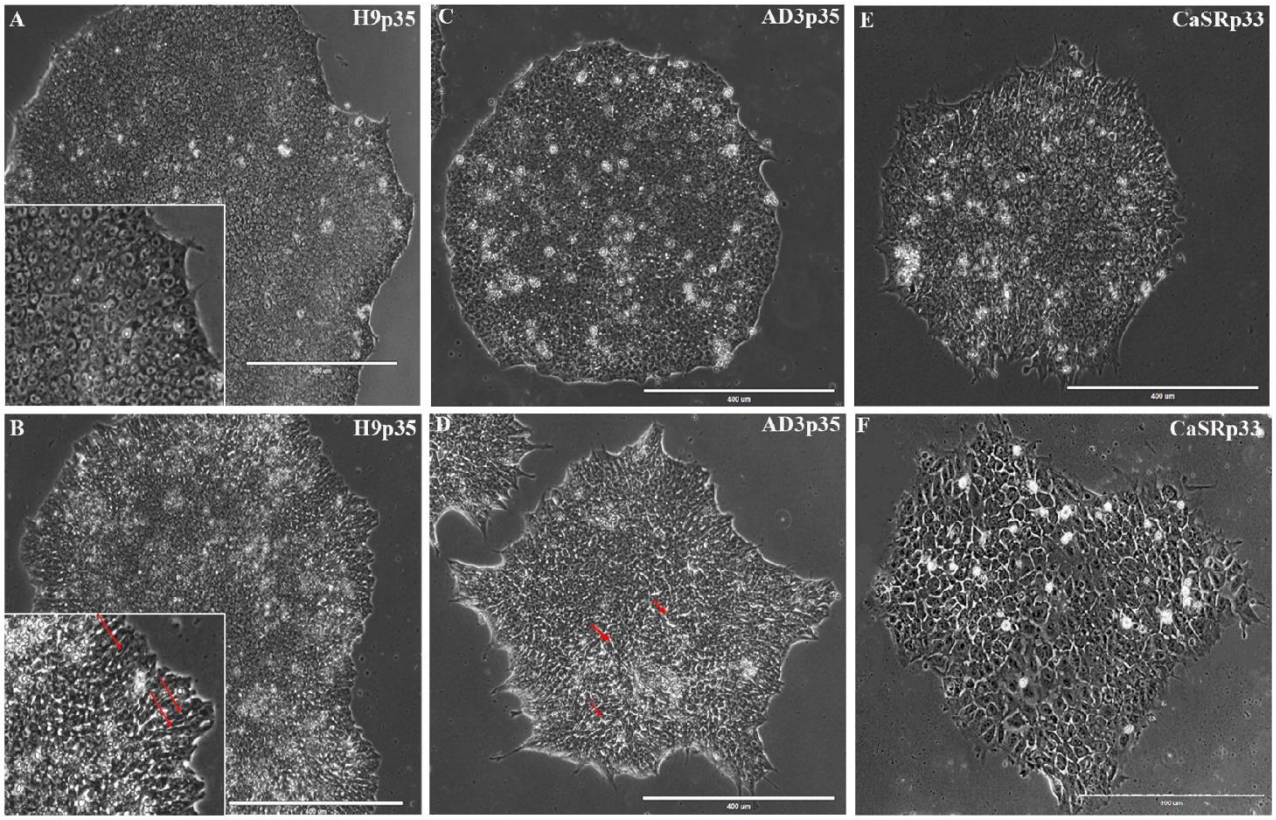

Рисунок 1.1. Примеры изображений колоний hPSC с «хорошим» (a, c, e) и «плохим» (b, d, f) фенотипами [7].

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.

1.2. Постановка задачи машинного обучения

Для построения модели, позволяющей определить фенотип колоний на основе морфологических признаков, будем пользоваться методами машинного обучения. 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 $K_y = \{x \in X: f^*(x) = y\}$, которые мы будем назвать *классами*, и задано множество *меток классов* $Y = \{0, 1, \dots, m\}$. Пусть существует *целевая функция* $f^*: X \to Y$, значения которой известны только на конечном подмножестве объектов $\{x_1, \dots, x_k\} \subset X$. Совокупность пар $T = \{(x_1,\ y_1), \dots, (x_k,\ y_k)\} \subset X \times Y$ называется *обучающей выборкой*.

Задача обучения с учителем состоит в том, чтобы по выборке T восстановить зависимость y^*, то есть получить *решающую функцию* $f: X \to Y$, которая приближала бы целевую функцию $f^*(x)$, причем не только на обучающей выборке, но и на всем множестве X. Другими словами, решить задачу классификации значит построить алгоритм f, способный достаточно точно классифицировать произвольный объект $x \in X$. Если $m = 2$ и множество $Y = \{0, 1\}$, то такая задача называется *задачей бинарной классификации.*

Под объектом на практике обычно понимают некоторый *вектор признаков* $x = \left(x^{(1)}, \dots, x^{(n)}\right)$, где *признак* $x^{(i)} \in \mathbb{R}$ — это результат измерения некоторой характеристики объекта. Таким образом, элементы множества X представляют собой 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 объекта. В результате одному и тому же описанию x могут соответствовать различные 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 $X \times Y$ с неизвестной плотностью $p(x, y)$. Пары $(x_i,\ y_i)$ из обучающей выборки T, представляющие собой реализацию случайной величины, будем считать независимыми одинаково распределенными.

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

$$err(f) = P\{(x, y) \in X \times Y \mid f(x) \neq y\} \tag{1.1}$$

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

Таким образом, задача построения модели для определения фенотипа колоний hPSC может быть интерпретирована как задача бинарной классификации, где признаками являются измеренные морфологические характеристики, а метками 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 областях, например, в банковском деле при определении кредитоспособности заемщиков, 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 универсальных алгоритмов, гарантирующих высокое качество классификации произвольных 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.

2.1. Наивный байесовский классификатор

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

В вероятностной постановке задача классификации объекта $x = \left(x^{(1)}, \dots, x^{(n)}\right)$ равносильна задаче определения класса, вероятность принадлежности x к которому является наибольшей:

$$y_{MAP} = \underset{y \in Y}{\operatorname{argmax}} P\left(y \middle| x^{(1)}, \dots, x^{(n)}\right) \tag{2.1}$$

Используя формулу Байеса, это выражение можно переписать следующим образом:

$$y_{MAP} = \underset{y \in Y}{\operatorname{argmax}} \frac{P(x^{(1)}, \dots, x^{(n)}|y)P(y)}{P(x^{(1)}, \dots, x^{(n)})} =$$

$$= \underset{y \in Y}{\operatorname{argmax}} P\left(y \middle| x^{(1)}, \dots, x^{(n)}\right)P(y) \tag{2.2}$$

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 $P(x^{(1)}, \dots, x^{(n)}|y)$ 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

$$P\left(x^{(i)}\middle|y, x^{(j)}\right) = P\left(x^{(i)}\middle|y\right) \text{ при } i \neq j \tag{2.3}$$

При данном допущении верна формула:

$$P\left(x^{(1)}, \dots, x^{(n)}\middle|y\right) = \prod_{i=1}^{n} P\left(x^{(i)}\middle|y\right) \tag{2.4}$$

Таким образом формула (2.1) принимает вид:

$$y_{MAP} = \underset{y \in Y}{\operatorname{argmax}} P(y) \prod_{i=1}^{n} P\left(x^{(i)}\middle|y\right) \tag{2.5}$$

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

2.2. Метод k-ближайших соседей

Метод k-ближайших соседей является одним из наиболее известных примеров метрических алгоритмов классификации [12]. В основе этой группы методов лежит оценка сходства объектов и предположение о том, что если *мера сходства 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:

$$\rho\left(x_k, x_j\right) = \sqrt{\sum_{i=1}^{n} \left(x_k^{(i)} - x_j^{(i)}\right)^2} \tag{2.6}$$

При использовании данной метрики важно, чтобы все признаки были нормированы, так как в противном случае признаки с наибольшими числовыми значениями 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 x_j с наименьшим значением метрики $\rho(x_j, x)$. Число k является *гиперпараметром* метода, то есть настраиваемым параметр, позволяющим управлять процессом обучения модели.

2.3. Логистическая регрессия

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

$$z = b_0 + b_1 x^{(1)} + \cdots + b_n x^{(n)} + \varepsilon = b_0 + \sum_{i=1}^{n} b_i x^{(i)} + \varepsilon, \qquad (2.7)$$

где b_i — коэффициенты регрессии, ε — ошибка регрессии. Полагая $x^{(0)} = 1$, можно переписать эту формулу в виде:

$$z = \sum_{i=0}^{n} b_i x^{(i)} + \varepsilon = b^T x + \varepsilon \qquad (2.8)$$

Для решения задачи бинарной классификации необходимо перейти от переменной 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:

$$g(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{e^z + 1} \qquad (2.9)$$

где $g(z) \in [0,1]$ — *логистическая функция*.

Тогда апостериорные вероятности принадлежности произвольного объекта x к классу с метками 1 и 0 соответственно могут быть выражены следующими формулами:

$$P(y = 1|x;\ b) = g(b^T x) \tag{2.10}$$

$$P(y = 0|x;\ b) = 1 - g(b^T x) \tag{2.11}$$

Таким образом, найдя коэффициенты регрессии, мы сможем оценить вероятность принадлежности объекта к классу с меткой 1 и, если она больше 0.5, отнести его к этому классу, а если меньше — к противоположному.

Используя формулы (2.10) и (2.11), можем записать выражение для плотности вероятности принадлежности объекта x к классу y:

$$p(y|x; b) = g(b^T x)^y \left(1 - g(b^T x)\right)^{1-y} \tag{2.12}$$

Коэффициенты регрессии b_i можно найти с помощью метода максимального правдоподобия, который заключается в максимизации *функции правдоподобия* на обучающей выборке:

$$L(b) = \prod_{i=1}^{k} P(y_i|x_i;\ b) = \prod_{i=1}^{k} g(b^T x_i)^{y_i} \left(1 - g(b^T x_i)\right)^{1-y_i} \tag{2.13}$$

Для этого, пользуясь линейностью логарифма, можем перейти к поиску минимума следующего функционала:

$$l(b) = -logL(b) = -\sum_{i=1}^{k} \left[y_i \log\left(g(b^T x_i)\right) + (1 - y_i) \log\left(1 - g(b^T x_i)\right)\right] \tag{2.14}$$

Для минимизации $l(b)$ можно воспользоваться, например методом градиентного спуска, где коэффициенты пересчитываются по формуле

$$b_j^{(r)} = b_j^{(r-1)} - \alpha \frac{\partial}{\partial b_j} l\left(b^{(r-1)}\right) \tag{2.15}$$

где r — номер шага, $0 < \alpha < 1$ — переменная, задающая скорость спуска, а начальные значения $b_j^{(0)}$ задаются произвольным образом.

Находя значение частной производной функции $l(b)$, получаем, что правило пересчета методе градиентного спуска для поиска коэффициентов логистической регрессии будет иметь следующий вид:

$$b_j^{(r)} = b_j^{(r-1)} + \alpha \sum_{i=1}^{k} \left(y_i - g(b^T x_i)\right) x_i^{(j)} \tag{2.16}$$

Одним из допущений, при которых возможно применение логистической

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

$$l(b) + \lambda \sum_{i=0}^{n} b_i^2 \to \min_b \tag{2.17}$$

где λ — коэффициент регуляризации, который выступает в качестве гиперпараметра метода логистической регрессии.

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

2.4. Метод случайного леса

Случайный лес (англ. Random Forest) — это алгоритм машинного обучения, 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, определяющее 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 «жадные» алгоритмы, обеспечивающие локально-оптимальное разбиение в каждом узле. Дерево строится сверху вниз начиная с корня. При построении каждого

узла-предиката формируется правило вида $x^{(i)} \leq c$, 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. Таким образом, на каждом шаге построения дерева алгоритм последовательно сравнивает 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:

$$gini(A) = 1 - p_0^2 - p_1^2, \tag{2.18}$$

где p_0 и p_1 — 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 A_1 и A_2 объемами l_1 и l_2 соответственно, то показатель качества разбиения можно определить следующим образом:

$$gini_{split}(A) = \frac{l_1}{l} gini(A_1) + \frac{l_2}{l} gini(A_2) \tag{2.19}$$

Наилучшим считают правило, приводящее к разбиению с минимальным значением индекса $gini_{split}(A)$.

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

$$Gain_{split}(A) = entropy(A) - \left(\frac{l_1}{l} entropy(A_1) + \frac{l_2}{l} entropy(A_2) \right) \tag{2.20}$$

где под энтропией понимают меру информации:

$$entropy(A) = -p_0 \log p_0 - p_1 \log p_1 \tag{2.21}$$

Наилучшим считают правило, приводящее к разбиению с максимальным значением $gini_{split}(A)$.

Процесс построения решающего дерева обычно останавливается при

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

Использование ансамбля решающих деревьев в методе случайного леса позволяет повысить точность в сравнении с одним деревом. Формирование ансамбля деревьев происходит с помощью бэггинга. Для этого задают желаемое количество деревьев N и случайным образом формируют N подмножеств обучающей выборки одинакового размера. Каждый элемент обучающей выборки может 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 деревьев.

2.5. Метод опорных векторов

Метод опорных векторов (англ. Support Vector Machine, SVM) основывается 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:

$$\sum_{i=1}^{n} \widehat{w}_i x^{(i)} - \widehat{w}_0 = 0 \iff \widehat{w}^T x - \widehat{w}_0 = 0 \tag{2.22}$$

где $\widehat{w}$ — вектор нормали к гиперплоскости.

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

$$\widehat{w}^T x > \widehat{w}_0, \tag{2.23}$$

а второго — неравенству

$$\widehat{w}^T x < \widehat{w}_0 \tag{2.24}$$

Для линейно разделимых классов существует множество разделяющих гиперплоскостей. Цель данного метода — найти гиперплоскость, которая наиболее удалена от объектов обоих классов, тем самым увеличив надежность классификации.

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

$$\widehat{w}^T x - \widehat{w}_0 = \varepsilon, \tag{2.25}$$

$$\widehat{w}^T x - \widehat{w}_0 = -\varepsilon \tag{2.26}$$

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

$$w^T x - w_0 = 1, \tag{2.27}$$

$$w^T x - w_0 = -1, \tag{2.28}$$

где $w = \dfrac{\widehat{w}}{\varepsilon}$, $w_0 = \dfrac{\widehat{w}_0}{\varepsilon}$.

Вычислим евклидово расстояние между этими гиперплоскостями:

$$\rho = \frac{|w_0 + 1 - (w_0 - 1)|}{||w||} = \frac{2}{||w||} \tag{2.29}$$

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

$$\begin{aligned} &||w|| \longrightarrow \min_{w,w_0}; \\ &w^T x_i - w_0 \geq 1, \qquad y_i = 1, \\ &w^T x_i - w_0 \leq 1, \qquad y_i = -1 \end{aligned} \tag{2.30}$$

или в более удобной форме

$$\frac{1}{2}||w||^2 \longrightarrow \min_{w,w_0};\ \ y_i(w^T x_i - w_0) \geq 1, \qquad i = \overline{1,k} \tag{2.31}$$

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

Для этого в исходной задаче смягчают ограничения и добавляют в минимизируемый функционал штраф за суммарную ошибку:

$$\begin{aligned} &\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \longrightarrow \min_{w,w_0,\xi}; \\ &y_i(w^T x_i - w_0) \geq 1 - \xi_i, \quad i = \overline{1,k}, \\ &\xi_i \geq 0, \quad i = \overline{1,k}, \end{aligned} \tag{2.32}$$

где C — гиперпараметр, позволяющий находить компромисс между максимизацией 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 множество готовых прикладных пакетов.

Решив задачу квадратичного программирования и найдя значения параметров w и w_0, сможем построить линейный пороговый классификатор:

$$y_{predicted} = \text{sign}(w^T x - w_0) \tag{2.33}$$

Еще один подход, позволяющий решить проблему линейной неразделимости, называется *ядерный трюк*. В основе данного подхода лежит теорема Ковера, утверждающая, что при повышении размерности пространства признаков, увеличивается 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 произведений (x_i, x_j). Переход к новому пространству признаков с большей размерностью осуществляется с помощью применения некоторого преобразования $\eta: X \longrightarrow H$. При этом исходная задача строится так же, за исключением того, что скалярное произведение (x_i, x_j) в пространстве X заменяется скалярным произведением $(\eta(x_i),\ \eta(x_j))$ в пространстве H. Вместо того, чтобы напрямую искать преобразование η, обычно подбирают вид *ядра* $K(x_i, x_j) = (\eta(x_i),\ \eta(x_j))$, которым формально заменяется скалярное произведение. Наиболее распространенными являются:

1. Линейное ядро

$$K(x_i, x_j) = (x_i, x_j) \tag{2.34}$$

2. Полиномиальное ядро

$$K(x_i, x_j) = \left((x_i, x_j) + r\right)^d \tag{2.35}$$

3. Радиальная базисная функция (RBF)

$$K(x_i, x_j) = e^{-\gamma\left||x_i - x_j|\right|^2}, \qquad \gamma > 0 \tag{2.36}$$

4. Сигмоидная функция

$$K(x_i, x_j) = \tanh\left(\gamma(x_i, x_j) + r\right), \qquad \gamma > 0, c < 0 \tag{2.37}$$

Форма ядра и его параметры являются гиперпараметрами метода опорных векторов.

2.6. Искусственная нейронная сеть

Искусственная нейронная сеть (англ. artificial neural network) — это математическая модель, основанная на общих принципах функционирования биологических нейронных сетей [15]. Вычислительной единицей искусственной нейронной сети является *нейрон*. По аналогии с биологическими нейронами искусственный нейрон состоит из «дендритов», через которые входные сигналы поступают в ядро, ядра, осуществляющего обработку входных сигналов, и «аксона», передающего выходной сигнал на «дендриты» следующих нейронов. Выходной сигнал нейрона определяется как значение функции активации на взвешенной сумме его входов:

$$y = \sigma\left(\sum_{j=1}^{n} x^{(j)} w_j - w_0\right), \tag{2.38}$$

где $x^{(j)}$ — значения входных сигналов, w_j — синаптические веса, определяющие степень влияния каждого сигнала на итоговое состояние, w_0 — смещение. Синаптические веса и смещение каждого нейрона являются параметрами сети, которые настраиваются в ходе ее обучения. В качестве функции активации σ

обычно выступает некоторая нелинейная дифференцируемая функция, например:

1. ReLU (от англ. Rectified Linear Unit)

$$\mathrm{ReLU}(x) = \max(x, 0) \tag{2.39}$$

2. Логистическая функция

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.40}$$

3. Гиперболический тангенс

$$\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{2.41}$$

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 выходной. Каждый нейрон передает свой выходной сигнал всем нейронам следующего слоя. Общий вид полносвязной сети с одним скрытым слоем представлен на рисунке 2.1 [16]. Количество нейронов во входном слое равно количеству признаков, а выходной слой в сетях, используемых для решения задач бинарной классификации, обычно содержит всего один нейрон с логистической функцией активации. Количество скрытых слоев и нейронов в них и используемые функции активации являются гиперпараметрами нейронной сети.

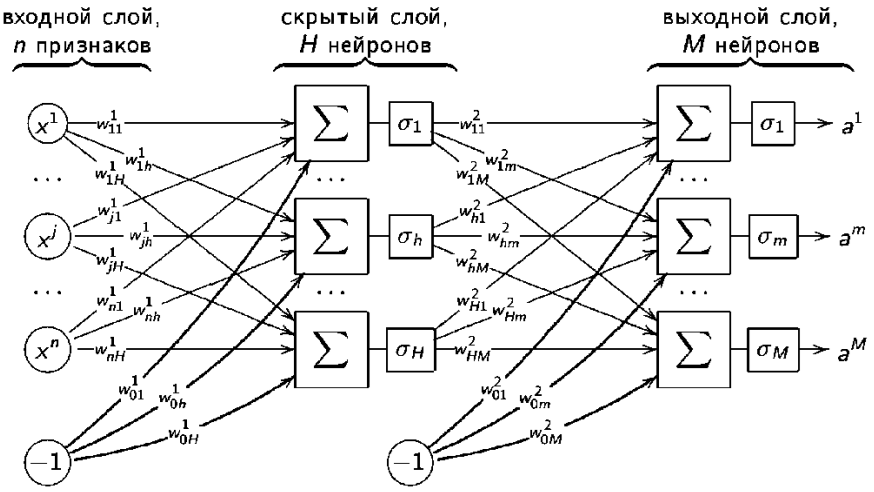

Рисунок 2.1. Общая схема многослойной искусственной нейронной сети.

Согласно теореме Цыбенко [17], с помощью нейронной сети такой архитектуры, вообще говоря, можно аппроксимировать любую непрерывную зависимость 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 конфигурации сети можно рассматривать как задачу многомерной оптимизации, и для ее решения обычно используют вариации метода градиентного спуска. Выбор 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 используемых и показывает хорошие результаты на практике [19].

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

$$Q(w) = -\frac{1}{k}\sum_{i=1}^{k} Q_i(w) = -\frac{1}{k}\sum_{i=1}^{k}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)], \qquad (2.42)$$

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

Формула пересчета весовых коэффициентов в методе стохастического градиентного спуска имеет следующий вид:

$$w^{(k+1)} = w^{(k-1)} - \alpha\nabla Q\left(w^{(k-1)}\right), \qquad (2.43)$$

где α — размер шага спуска, а начальные значения весов $w^{(0)}$ задаются произвольным образом. В методе Adam для ускорения сходимости применяются следующие эвристики:

1. Использование информации о предыдущих шагах.

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

(для второго момента — квадрата градиента) в точке $w^{(k-1)}$:

$$\mu_1^{(k)} = \beta_1 \mu_1^{(k-1)} + (1-\beta_1)\nabla Q\left(w^{(k-1)}\right), \qquad \mu_1^{(0)} = 0 \tag{2.44}$$

$$\mu_2^{(k)} = \beta_2 \mu_2^{(k-1)} + (1-\beta_2)\left(\nabla Q\left(w^{(k-1)}\right)\right)^2, \qquad \mu_2^{(0)} = 0 \tag{2.45}$$

2. Адаптивный подбор размера шага с учетом истории изменения весов

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

$$w^{(k)} = w^{(k-1)} - \frac{\alpha}{\sqrt{\hat{\mu}_2^{(k)} + \varepsilon}}\hat{\mu}_1^{(k)}, \tag{2.46}$$

где

$$\hat{\mu}_1^{(k)} = \frac{\mu_1^{(k)}}{1-\beta_1^k}, \qquad \hat{\mu}_2^{(k)} = \frac{\mu_2^{(k)}}{1-\beta_2^k} \tag{2.47}$$

— несмещенные моменты, ε — небольшая константа, позволяющая избежать деления на ноль.

Авторы Adam предлагают выбирать значения гиперпараметров β_1, β_2 и ε равными 0.9, 0.999 и 10^{-8} соответственно, а размер шага $0 < \alpha < 1$ подбирать вручную, отталкиваясь от конкретных исходных данных.

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

1. Выбрать из обучающей выборки случайным образом подмножество заданного 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 каждого нейрона выходной сигнал

$$x_h^l = \sigma_h^l\left(s_h^l\right), \qquad s_h^l = \sum_{k=0}^{H_{l-1}} x_k^{l-1} w_{kh}^l, \qquad l = \overline{1,L}, \qquad h = \overline{1,H_l} \tag{2.48}$$

и его частные производные

$$z_h^l = \left.\frac{\partial x_h^l}{\partial s}\right|_{s=s_h^l} = \left.\frac{\partial \sigma_h^l}{\partial s}\right|_{s=s_h^l}, \qquad l = \overline{1,L}, \qquad h = \overline{1,H_l} \tag{2.49}$$

где l — номер слоя, L — число слоев, h — номер нейрона в слое, H_{l-1} — размер $l-1$ слоя, $x_0^l = -1$ для всех слоев. Значения вектора x^0 полагаются равными входному вектору признаков выбранного объекта обучающей выборки (рисунок 2.1).

4. Совершить обратный проход по сети (справа налево), рекурсивно вычисляя ошибки для каждого слоя:

- для выходного слоя (содержит один нейрон с выходным сигналом $x_1^L = p$):

$$\varepsilon_1^L = \frac{\partial Q_i(w)}{\partial s_1^L} = x_1^L - y = p - y \tag{2.50}$$

- для скрытых слоев:

$$\varepsilon_h^{l-1} = \frac{\partial Q_i(w)}{\partial x_h^{l-1}} \frac{\partial x_h^{l-1}}{\partial s_1^{l-1}} = z_h^{l-1} \sum_{k=0}^{H_l} \varepsilon_k^l w_{hk}^l, \qquad l = L, \dots, 2, \qquad h = \overline{0, H_{l-1}} \tag{2.51}$$

Вычислив все значения ошибок, сможем получить значения частных производных минимизируемого функционала, используя формулу дифференцирования суперпозиции функций:

$$\frac{\partial Q_i(w)}{\partial w_{kh}^l} = \frac{\partial Q_i(w)}{\partial x_h^l} \frac{\partial x_h^l}{\partial s_h^l} \frac{\partial s_h^l}{\partial w_{kh}^l} = \varepsilon_h^l x_k^{l-1}, \qquad l = \overline{1,L}, k = \overline{0, H_{L-1}}, h = \overline{1,H_l} \tag{2.52}$$

5. Повторить шаги 2–4 для каждого объекта подвыборки и усреднить значение градиента по всем объектам:

$$\frac{\partial Q(w)}{\partial w_{kh}^l} = \frac{1}{m} \sum_{k=0}^{m} \frac{\partial Q_i(w)}{\partial w_{kh}^l}, \qquad l = \overline{1,L}, k = \overline{0, H_{L-1}}, h = \overline{1,H_l} \tag{2.53}$$

Теперь с помощью полученных производных можно совершать шаг обновления весов по методу Adam.

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

ГЛАВА 3. СРАВНЕНИЕ КАЧЕСТВА РАБОТЫ МЕТОДОВ НА ИССЛЕДУЕМЫХ ДАННЫХ

В первом параграфе данной главы описаны метрики качества, используемые для оценки качества работы методов классификации. Затем рассмотрены детали предобработки данных, реализации методов и способа подбора гиперпараметров. 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.

3.1. Метрики оценки качества классификации

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

3.1.1. Матрица ошибок

Матрица ошибок (англ. confusion matrix) представляет собой таблицу, содержащую информацию о том, сколько раз классификатор принял верное или неверное 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 определены верно
2. FP (False Positive) — число объектов класса с меткой 0, которые были определены неверно

3. FN (False Negative) — число объектов класса с меткой 1, которые были определены неверно
4. TN (True Positive) — число объектов класса с меткой 1, которые были определены верно

3.1.2. Кросс-валидационная точность

Точность (англ. accuracy) — это отношение числа правильно классифицированных объектов к общему числу объектов в наборе данных:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.1}$$

Точность является одной из наиболее распространенных метрик оценки качества 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 методу, который называется *кросс-валидацией по k блокам* (англ. k-fold cross validation). При применении кросс-валидации все данные разделяют на k равных частей. После этого k раз повторяют следующий процесс: $k - 1$ блок используют для построения модели с выбранными гиперпараметрами, а оставшийся блок используется 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 $k = 10$.

3.1.3. Площадь под кривой ошибок

ROC-кривая — графический инструмент оценки точности моделей бинарной классификации, позволяющий найти оптимальный баланс между *чувствительностью*:

$$sensitivity = \frac{TP}{TP + FN} \tag{3.2}$$

и *специфичностью* модели:

$$specificity = \frac{TN}{TN + FP} \tag{3.3}$$

ROC-кривая задается в осях $(sensitivity, 1 - specificity)$ и отображает соотношение этих величин при варьировании порога решающего правила (объекты, 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, AUC) является ее численной характеристикой, которой удобно пользоваться для сравнения моделей. Чем выше значение AUC для классификатора, тем лучше его способность различать положительные и отрицательные классы. Идеальный классификатор будет иметь $AUC = 1$, в то время как классификатор, присваивающий метки класса случайным образом, будет иметь $AUC = 0{,}5$.

3.2. Предварительный анализ и предобработка данных

Перед использованием данных, содержащих измеренные морфологические параметры растущих колоний hPSC и клеток внутри колоний вместе с информацией о качестве колоний, для построения моделей классификации, проведем предварительный анализ и предобработку данных.

3.2.1. Фильтрация данных

Поскольку некоторые объекты не содержат информацию о качестве колонии, необходимо произвести фильтрацию данных, чтобы удалить строки, которые 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 точность (accuracy) может быть использована в качестве метрики для сравнения качества работы моделей.

3.2.2. Дисперсионный и корреляционный анализ данных

Дисперсионный анализ данных, проведенный в статье [7], показал, что многие признаки имеют статистически значимую разницу между средними значениями для объектов разных классов, из чего можно сделать вывод, что они оказывают значимое влияние на фенотип колонии. Тем не менее, авторы так же показали, что разные линии характеризуются различными морфологическими параметрами, чувствительными к фенотипическому разделению, следовательно существует 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*:

$$r_{ab} = \frac{\sum_{i=1}^{k}(a_i - \bar{a})(b_i - \bar{b})}{\sigma_a \sigma_b}, \tag{3.4}$$

где $a = (a_1, \dots a_k), b = (b_1, \dots b_k)$ — некоторые выборки, $\bar{a}$ и $\bar{b}$ — выборочные средние, σ_a и σ_b — среднеквадратические отклонения.

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

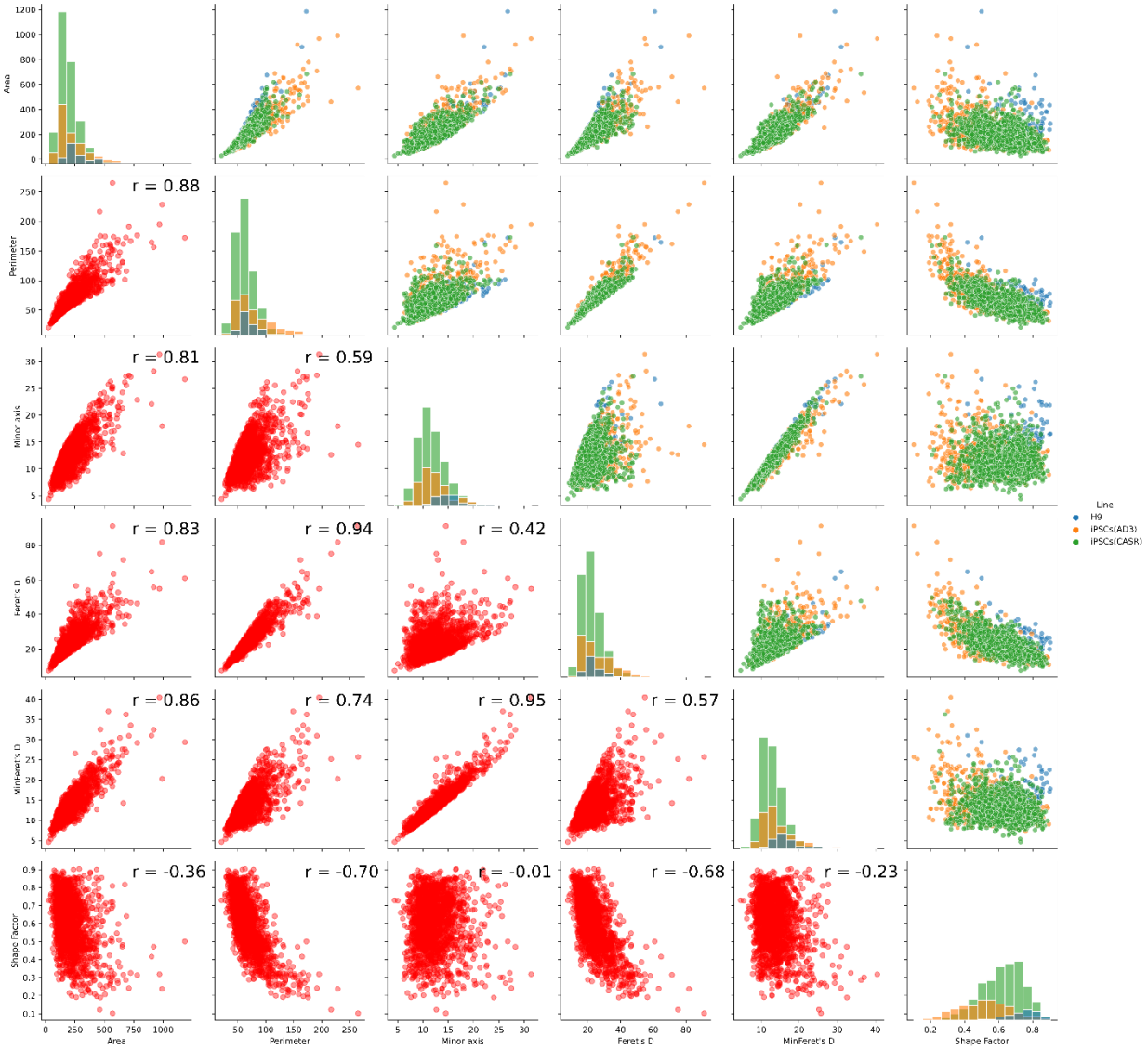

Рисунок 3.1. Точечные диаграммы рассеяния для клеточных данных.

Голубой цвет соответствует линии эмбриональных hPSC H9, желтый — линии индуцированных hPSC AD3, зеленый — линии индуцированных hPSC CASR. На диагоналях расположены диаграммы распределения каждого из признаков 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.

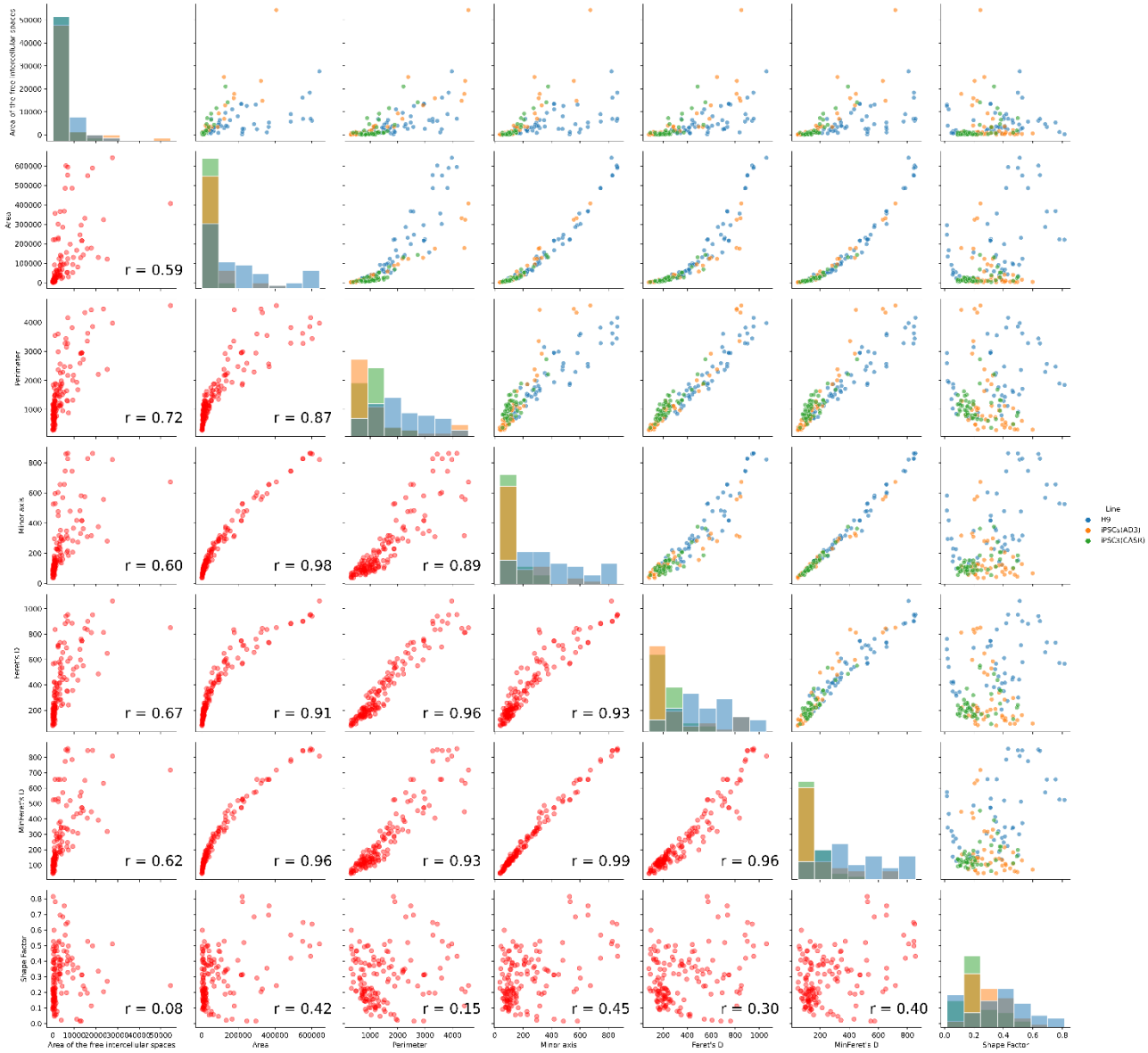

Рисунок 3.2. Точечные диаграммы рассеяния для колониальных данных.

Можно заметить, что как в клеточных, так и в колониальных данных многие признаки достаточно сильно коррелированы (r по модулю близок к единице). Сильная зависимость признаков может негативно влиять на многие модели классификации, 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.

3.2.3. Стандартизация данных

Рассмотрим среднее значение для каждого признака в клеточных и колониальных 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 отличаются на порядки. Большая разница в масштабах признаков может негативно сказаться на работе методов классификации, основанных на расстояниях в пространстве признаков, поэтому перед применением метода k ближайших соседей и метода опорных векторов данные необходимо стандартизировать. Кроме того, стандартизацию стоит также использовать при построении логистической регрессии и нейронных сетей, поскольку это обычно приводит к ускорению их обучения.

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

$$x^{(i)\prime} = \frac{x^{(i)} - \bar{x}^{(i)}}{\sigma_i}, \tag{3.5}$$

где $\bar{x}^{(i)}$ — среднее значение признака, σ_i — среднеквадратическое отклонение.

3.3. Реализация алгоритмов классификации и подбор гиперпараметров

Для применения к исследуемым данным методов бинарной классификации, рассмотренных во второй главе, использовался язык программирования Python 3.8 и библиотеки, предназначенные для реализации методов машинного обучения, такие как sklearn и keras (для работы с искусственной нейронной сетью).

Таблица 3.4. Гиперпараметры моделей.

Метод	**Гиперпараметры**	**Диапазон поиска**	**Лучшее значение**	
			Клеточные данные	**Колониальные данные**
Наивный байесовский классификатор	-	-	-	-
k ближайших соседей	число соседей	$[1, 100]$	29	11
Логистическая регрессия	коэффициент регуляризации	от 10^{-7} до 10^{7}	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 10^{3}	1	10^{3}
Искусственная нейронная сеть	конфигурация скрытых слоев	-	13:5	10:3
	функции активации 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 помощью кросс-валидации по пяти блокам. Поиск по сетке не всегда применим из-за больших вычислительных затрат, однако он хорошо подходит для исследуемых данных, поскольку они не обладают большим объемом.

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

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

3.4. Сравнение качества работы полученных моделей

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

Таблица 3.5. Сравнение качества работы моделей.

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

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

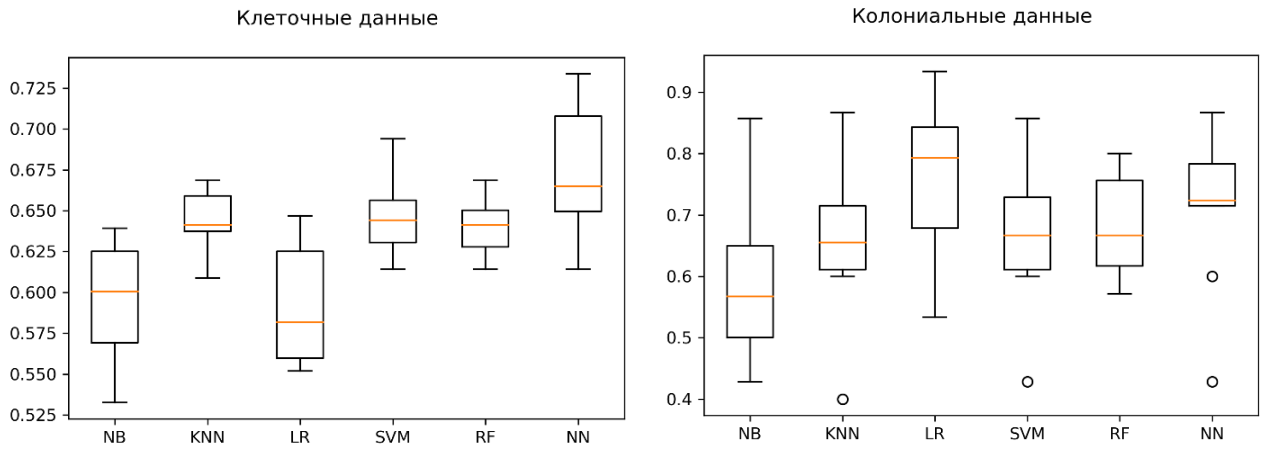

Рисунок 3.3. Боксплоты распределений точности моделей для клеточных (слева) и колониальных данных (справа). NB — наивный байесовский классификатор, KNN — метод k ближайших соседей, LR — логистическая регрессия, SVC — метод опорных векторов, RF — случайный лес, NN — искусственная нейронная сеть.

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

В случае колониальных данных, наиболее высокое среднее значение точности имеет *метод логистической регрессии* (74.81%), однако t-критерий Стьюдента не показывает статистическую значимость отличий. Тем не менее, учитывая различия средних значений точности и значений метрики ROC AUC для рассмотренных моделей, можно считать, что наиболее эффективным для классификации 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.

3.5. Сравнение моделей, обученных на объединенных данных

Объединим клеточные и колониальные данные, добавив к морфологическим 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 гиперпараметров 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

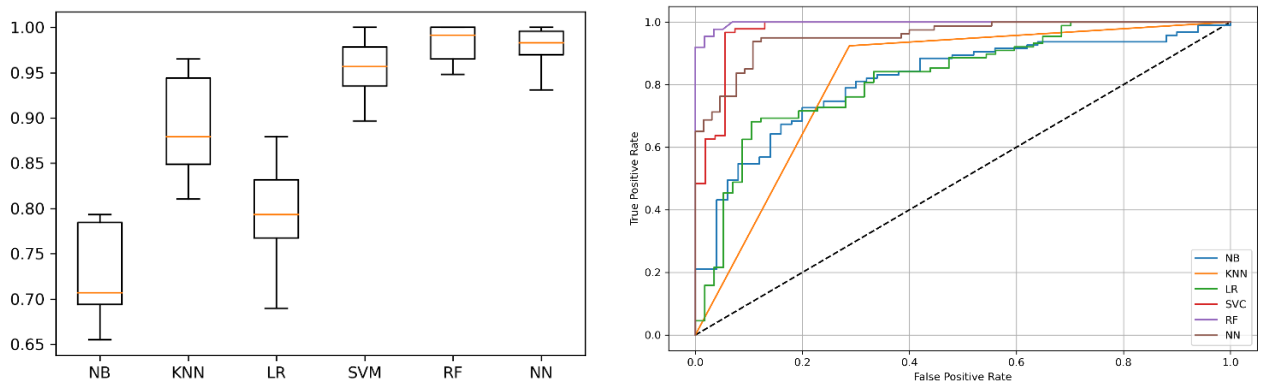

Рисунок 3.4. Боксплоты распределений точности моделей (слева) и ROC-кривые (справа) для объединенных данных. Расшифровка легенды аналогична рисунку 3.4.

Результаты сравнения работы методов приведены в таблице 3.7 и на рисунке 3.4. Наилучшее качество по совокупности двух метрик показывает метод случайного леса (98.28%), причем t-критерий Стьюдента показывает статистически различимое отличие средней кросс-валидационной точности для метода случайного леса в сравнении со всеми методами, кроме искусственной нейронной сети. В таблице 3.8 приведена матрица ошибок метода случайного леса на тестовых 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, использующих 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 высокое качество классификации на новых данных, однако показывает эффективность 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 понять взаимосвязь между морфологическими признаками и фенотипом колоний.

4.1. Методы отбора признаков

Существующие методы отбора признаков делятся на три основных категории [22]:

1. Методы фильтрации (англ. filter methods)

Методы этой группы применяются до обучения моделей и основаны на сравнении некоторых статистических показателей. Методы фильтрации обычно рассматривают каждый признак независимо, определяя степени корреляции признаков с целевой переменной и ранжируя их по полученным значениям.

2. Методы-оболочки (англ. wrapper methods)

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

3. Встроенные методы (англ. embedded methods)

Встроенные методы позволяют отбирать признаки в процессе построения модели. Данные методы вводят дополнительные ограничения при оптимизации алгоритма, смещающие модель в сторону меньшей сложности. Наиболее распространенными методами этой группы являются методы регуляризации.

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

интерпретируемыми.

4.1.1. Метод SHAP

В основе метода SHAP (от англ. Shapley additive explanations, аддитивные объяснения Шепли) лежат понятия из кооперативной теории игр — раздела прикладной математики, изучающего методы принятия оптимальных решений в конфликтных ситуациях [23]. Идея метода состоит в том, чтобы рассматривать отдельные 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:

$$\varphi\left(x^{(i)}\right) = \sum_{S \subseteq F/\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left(v(S \cup \{i\}) - v(S)\right), \qquad (4.1)$$

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

$$v(S) = E[f(x)|x_S], \qquad (4.2)$$

— условное математическое ожидание предсказания f на примерах x', взятых из распределения данных, таких что $x'_S = x_S$.

Точное вычисление значений SHAP является сложной задачей, однако авторы метода предлагают способы аппроксимации, позволяющие найти оценки значений SHAP [23].

4.1.2. 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, используемых для построения выбранной модели машинного обучения, и сравнения кросс-валидационного качества полученных классификаторов. Метод

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

4.2. Результаты применения методов отбора признаков

Применим рассмотренные методы отбора признаков к тем моделям, которые в главе 3 показали наилучшее качество классификации по совокупности кросс-валидационной точности и ROC AUC. Для реализации методов используем 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 точность на полном наборе признаков равна 67.48%). Результат применения **метода оценки значений SHAP** к данной модели приведен на рисунке 4.1.

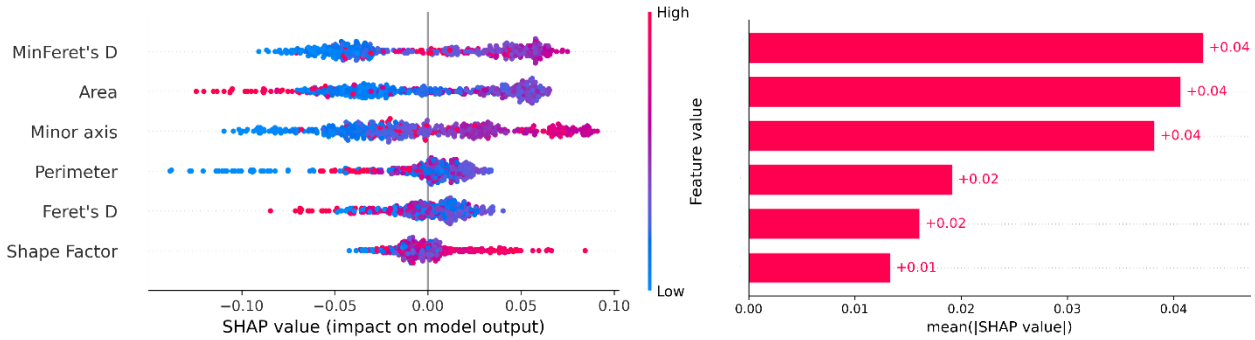

Рисунок 4.1. Значения SHAP для модели, построенной на клеточных данных.

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.

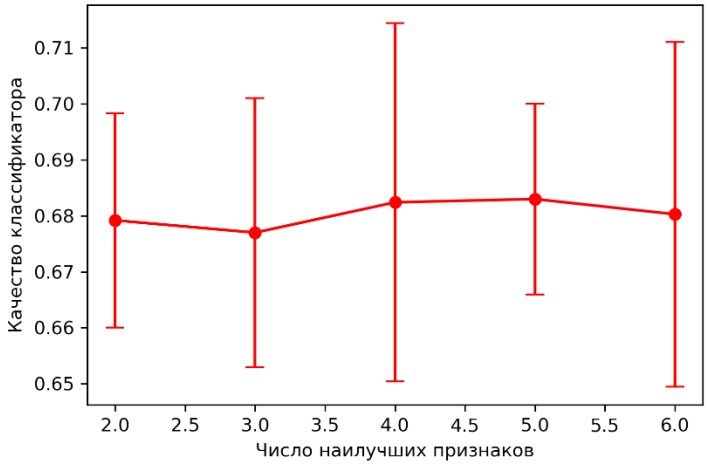

Рисунок 4.2. График зависимости кросс-валидационной точности (± среднеквадратическое отклонение) от набора признаков для моделей, построенных на клеточных данных.

Наибольшее значение средней кросс-валидационной точности достигается на модели, использующей пять параметров (минимальный диаметр Ферета, площадь, 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, построенной на данных признаках, составляет 68.11% (±1.54%), что почти не уступает модели, построенной на пяти признаках. В таблице 4.1 приведена матрица ошибок 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 точность на полном наборе признаков равна 74.81%). Результат применения **метода оценки значений SHAP** к данной модели приведен на рисунке 4.3.

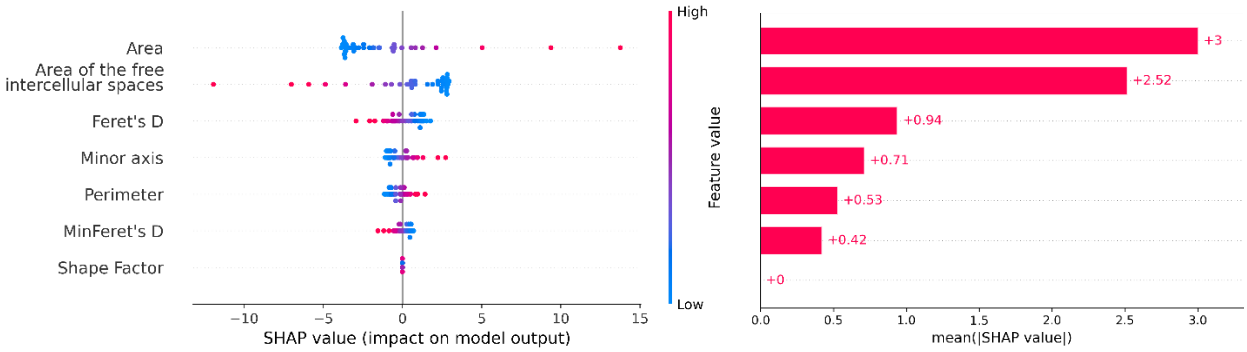

Рисунок 4.3. Значения SHAP для модели, построенной на колониальных данных.

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.

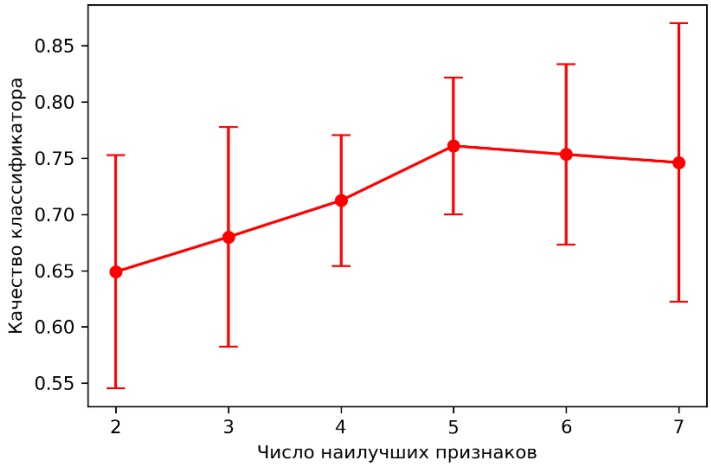

Рисунок 4.4. График зависимости кросс-валидационной точности (± среднеквадратическое отклонение) от набора признаков для моделей, построенных на клеточных данных.

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 плюрипотентных стволовых клеток человека по их фенотипу — потенциальной способности к поддержанию плюрипотентности и клональности, на основе морфологических 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 искусственная нейронная сеть. Для этих методов были подобраны оптимальные значения гиперпараметров и проведено сравнение качества их работы на исследуемых данных. Полученные оптимальные модели для клеточных и колониальных данных обладают кросс-валидационной точностью, равной 67.48% (искусственная нейронная сеть) и 74.81% (логистическая регрессия), соответственно. Были также рассмотрены особенности обучения моделей на данных отдельных клеточных линий. Кроме того, была построена модель, использующая для предсказания качества колонии комбинацию клеточных и колониальных данных. Оптимальная 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 клеток внутри них и могут стать основой для создания систем автоматического неинвазивного контроля качества колоний плюрипотентных стволовых клеток человека.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Ramalho-Santos M., Willenbring H.* On the origin of the term "stem cell" // Cell Stem Cell — 2007. — № 1(1). — С. 35–38. — DOI 10.1016/j.stem.2007.05.013
2. *Chagastelles P. C., Nardi N. B.* Biology of stem cells: an overview // Kidney International 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
4. *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
5. *Engle S. J., Puppala D.* Integrating human pluripotent stem cells into drug development // Cell Stem Cell — 2013. — № 12(6). — С. 669–677. — DOI 10.1016/j.stem.2013.05.011
6. *Maddah M., Shoukat-Mumtaz U., Nassirpour S., Loewke K.* A system for automated, noninvasive, morphology-based evaluation of induced pluripotent stem cell cultures // Journal of Laboratory Automation — 2014. — Т. 19 № 5. — С. 454–460. — DOI 10.1177/2211068214537258
7. Prognostic Analysis of Human Pluripotent Stem Cells Based on Their Morphological 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
8. Assessing Morphology of hPSCs: STEMCELL Technologies [Электронный ресурс] — URL: https://www.stemcell.com/technical-resources/methods-library/cell-culture/pluripotent-stem-cells/maintenance-of-pluripotent-stem-cells/assessing-morphology-of-hpscs.html (дата обращения 19.04.2023)
9. *Вьюгин В. В.* Элементы математической теории машинного обучения — Москва: МФТИ — ИППИ РАН, 2010. — 252 с. — ISBN 978-5-4439-1691-0

10. *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
11. *T. M. Mitchell* Machine Learning — USA: McGraw-Hill, 1997. — 432 с. — ISBN 0070428077
12. *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 с. — ISBN 978-0-262-02944-5
13. *Nocedal J., Wright S. J.* Numerical Optimization — USA: Springer New York, 1990. — 651 с. — ISBN 0-387-98793-2
14. *Воронцов К. В.* Лекции по методу опорных векторов // Курс лекций МФТИ — 2007.
15. *Rojas R.* Neural Networks — Germany: Springer-Verlag, Berlin, 1996. — 502 с. — ISBN 978-3-642-61068-4
16. *Воронцов К. В.* Нейронные сети: градиентные методы оптимизации // Курс лекций МФТИ — 2022.
17. *Cybenko G.* Approximation by superpositions of a sigmoidal function // Mathematics of Control, Signals and Systems — 1989. — № 2. — С. 303–314. — DOI https://doi.org/10.1007/BF02551274
18. *Soydaner D.* A comparison of optimization algorithms for deep learning // International Journal of Pattern Recognition and Artificial Intelligence — 2020. — Т. 34 № 13, 2052013. — DOI https://doi.org/10.1142/S0218001420520138
19. *Kingma D., Ba J.* Adam: A Method for Stochastic Optimization // 3rd International Conference for Learning Representations, San Diego, CA — 2015. — 12 с. — DOI https://doi.org/10.48550/arXiv.1412.6980
20. Search Algorithms for Automated Hyper-Parameter Tuning / *Zahedi L., Mohammadi F., Rezapour S. [и др.]* // arXiv: 2104.14677 — 2021. — 10 с. — DOI https://doi.org/10.48550/arXiv.2104.14677
21. *Stathakis D.* How many hidden layers and nodes? // International Journal of Remote Sensing — 2009. — Т. 30 № 8. — С. 2133–2147, DOI

10.1080/01431160802549278

22. *V. Kumar, S. Minz* Feature Selection: A literature Review // Smart Computing Review — 2014. — № 4. — С. 211–229. — DOI: 10.1145/2740070.2626320
23. 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.

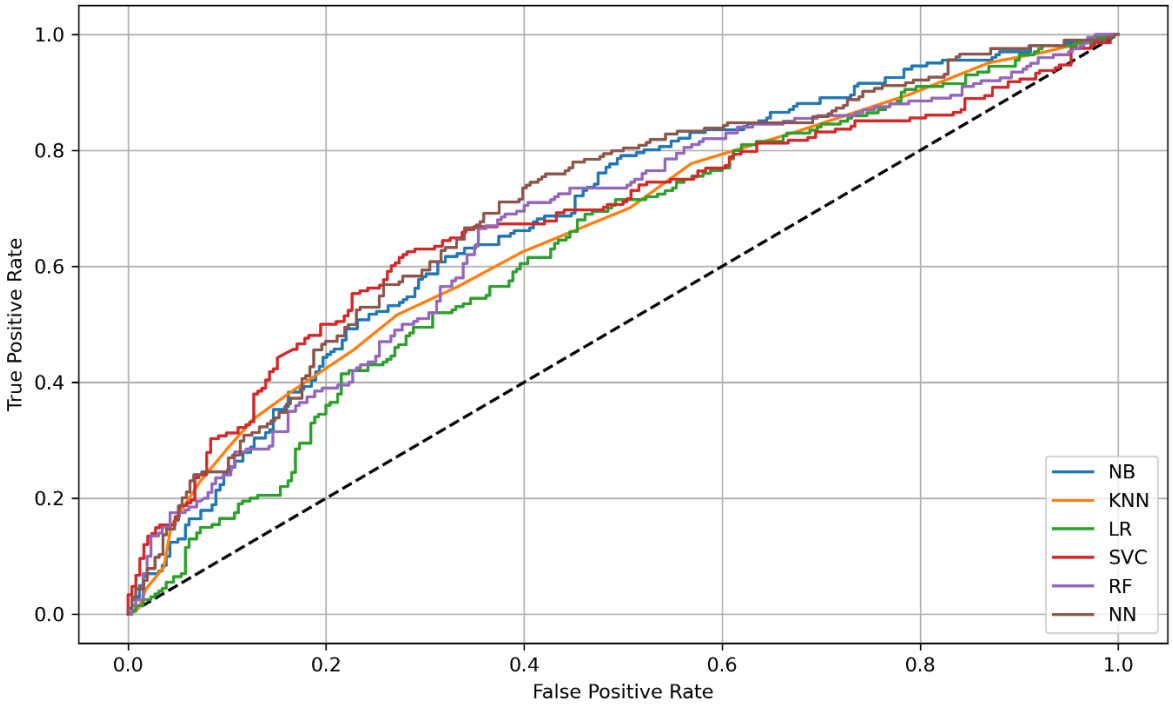

Рисунок П1.1. ROC-кривые для моделей, построенных на полных клеточных данных. NB соответствует наивному байесовскому классификатору, KNN — методу *k* ближайших соседей, LR — логистической регрессии, SVC — методу опорных векторов, RF — случайному лесу, NN — искусственной нейронной сети.

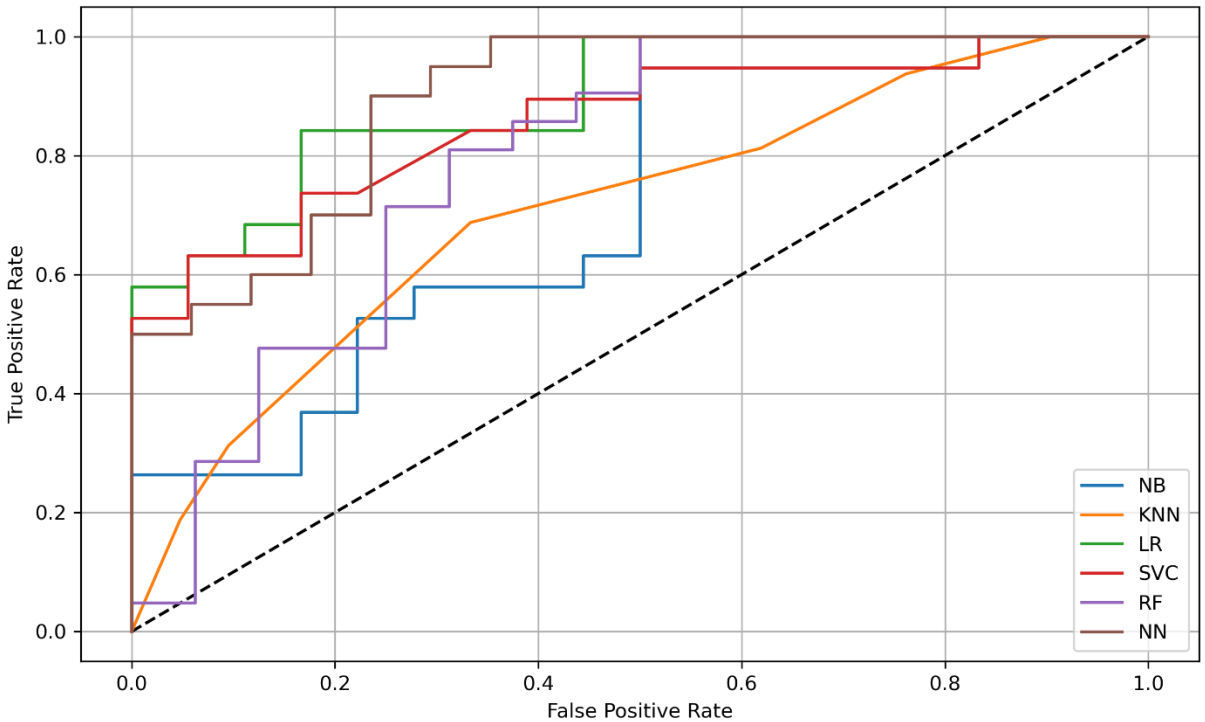

Рисунок П1.1. ROC-кривые для моделей, построенных на полных колониальных данных. Расшифровка легенды аналогична рисунку П.1.2.

Приложение 2.

Подбор гиперпараметров для моделей, основанных на объединенных клеточных и колониальных данных.

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

Метод	**Гиперпараметры**	**Диапазон поиска**	**Лучшее значение**
Наивный байесовский классификатор	-	-	-
k ближайших соседей	число соседей	$[1, 100]$	1
Логистическая регрессия	коэффициент регуляризации	от 10^{-7} до 10^{7}	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 10^{3}	10^{2}
Искусственная нейронная сеть	конфигурация скрытых слоев	-	17:6
	функции активации между слоями	-	relu
	число эпох	-	100
	размер шага	-	0.01