МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ   
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ   
«ВЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт математики и информационных систем

ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ И ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК

КАФЕДРА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Допущен к защите

Заведующий кафедрой ПМИ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е.В. Разова

**Тематическая классификация коротких текстовых сообщений**

Курсовой проект по дисциплине  
«Проектная и научно-исследовательская деятельность»

Выполнил студент группы ФИб-3301-51-00     / В.Р.Кочкин /

Руководитель К. п. н., зав. кафедрой ПМИ     / Е.В.Разова /

Работа защищена с оценкой     \_\_\_\_.\_\_\_\_.2021 г.

Члены комиссии:     /     /

    /     /

Киров 2021

**Содержание**

[Введение 4](#_Toc102929859)

[1. Машинное обучение и его применение 6](#_Toc102929860)

[2. Обзор размеченных корпусов для обучения 9](#_Toc102929861)

[2.1 Articles from russian site habr.com 9](#_Toc102929862)

[2.2 News dataset from Lenta.Ru 12](#_Toc102929863)

[2.3 Выводы по разделу 13](#_Toc102929864)

[3. Обзор классификаторов 13](#_Toc102929865)

[3.1 Метод K-ближайших соседей (k-nearest neighbors algorithm, KNN) 14](#_Toc102929866)

[3.1.1 Алгоритм. 14](#_Toc102929867)

[3.1.2 Определение класса нового объекта. 16](#_Toc102929868)

[3.1.3 Выбор значения параметра k. 17](#_Toc102929869)

[3.1.4 Особенности работы алгоритма. 18](#_Toc102929870)

[3.2 Модель RandomForest 19](#_Toc102929871)

[3.3 Модель LogisticRegression 19](#_Toc102929872)

[3.4 Выводы по разделу 20](#_Toc102929873)

[4. Подготовка размеченного корпуса 20](#_Toc102929874)

[4.1 Анализ корпуса 20](#_Toc102929875)

[4.2 Предварительная обработка данных 20](#_Toc102929876)

[4.3 Выводы по разделу 20](#_Toc102929877)

[5. Обучение моделей 20](#_Toc102929878)

[5.1 Обучение модели K-ближайших соседей 20](#_Toc102929879)

[5.2 Обучение модели RandomForest 20](#_Toc102929880)

[5.3 Обучение модели LogisticRegression 20](#_Toc102929881)

[5.4 Выводы по разделу 20](#_Toc102929882)

[6. Сравнение моделей 20](#_Toc102929883)

[6.1 Результаты оценщика 20](#_Toc102929884)

[6.2 Выводы по разделу 20](#_Toc102929885)

[Заключение 21](#_Toc102929886)

[Библиографический список 23](#_Toc102929887)

[Приложения 24](#_Toc102929888)

[Приложение А. Листинг программы 24](#_Toc102929889)

# 

# Введение

С каждым днем в интернете появляется огромное количество текстовой информации, и ориентироваться в ней становится практически невозможно. Миллионы новостей, статей, блогов и других данных ежедневно загружается во всемирную паутину. Появляется потребность в их тематической классификации для упрощения поиска нужной информации в интернете. Поэтому задача классификации текстов стала одной из актуальных тем нашего времени.

На помощь в классификации информации приходит популярное на сегодняшний день направление в анализе данных — машинное обучение. Это наука о том, как заставить искусственный интеллект учиться и действовать как человек, а также сделать так, чтобы он сам постоянно улучшал свое обучение и способности на основе предоставленных нами данных о реальном мире.

Исходя из сформулированной актуальности работы, её целью является обучение качественной модели, способной классифицировать по темам небольшие тексты на русском языке.

Для достижения поставленной цели должны быть достигнуты следующие задачи:

* провести анализ литературы по данной теме;
* определить предметную область текстов, подобрать готовый набор данных или написать парсер для его формирования;
* провести обзор методов тематической классификации текстов;
* подготовить выбранные данные;
* построить нейронную модель;
* дать оценку качества построенной модели.

Курсовой проект состоит из 6 разделов, заключения, библиографического списка и приложения. Порядок и содержание разделов соответствует поставленным задачам.

# Машинное обучение и его применение

Машинное обучение – это класс методов [искусственного интеллекта](https://ru.wikipedia.org/wiki/Искусственный_интеллект), характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства [математической статистики](https://ru.wikipedia.org/wiki/Математическая_статистика), [численных методов](https://ru.wikipedia.org/wiki/Численные_методы), [математического анализа](https://ru.wikipedia.org/wiki/Математический_анализ), [методов оптимизации](https://ru.wikipedia.org/wiki/Оптимизация_(математика)), [теории вероятностей](https://ru.wikipedia.org/wiki/Теория_вероятностей), [теории графов](https://ru.wikipedia.org/wiki/Теория_графов), различные техники работы с [данными в цифровой форме](https://ru.wikipedia.org/wiki/Данные_(вычислительная_техника)). Отчасти это похоже на процесс обучения младенца, который учится самостоятельно классифицировать объекты и события, определять взаимосвязи между ними.

В основе машинного обучения лежат три одинаково важных компонента:

* **Данные.** Собираются всевозможными способами. Чем больше данных, тем эффективней машинное обучение и точнее будущий результат.
* **Признаки.** Определяют, на каких параметрах строится машинное обучение.
* **Алгоритм.** Выбор метода машинного обучения (при условии наличия хороших данных) будет влиять на точность, скорость работы и размер готовой модели.

Задачи, которые способно решить машинное обучение, напрямую определяют выгоды для бизнеса и возможности решения социальных проблем государствами разных стран. К основным задачам относятся:

* **Регрессия.** Предоставляет прогноз на основе выборки объектов с различными признаками. По итогам анализа данных на выходе получается число или числовой вектор. Например, таким образом работает кредитный скоринг — оценка кредитоспособности потенциального заёмщика.
* **Классификация.** Выявляет категории объектов на основе имеющихся параметров. Продолжает традиции машинного зрения, поэтому часто можно встретить термин «распознавание образов»: например, идентификация разыскиваемых людей по фото или на основании словесного описания внешности.
* **Кластеризация.** Разделяет данные на схожие категории по объединяющему признаку. Например, космические объекты кластеризируют по удаленности, размерам, типам и другим признакам.
* **Идентификация.** Отделяет данные с заданными параметрами от остального массива данных. К примеру, участвует в постановке медицинского диагноза по набору симптомов.
* **Прогнозирование.** Работает с объемами данных за определенный период и предсказывает на основе анализа их значение через заданный период времени. Примером может служить прогноз погоды.
* **Извлечение знаний.** Исследует зависимости между рядом показателей одного и того же явления или события. Например, находит закономерности во взаимодействии биржевых показателей.

В таблице 1.1 перечислены некоторые популярные области применения машинного обучения.

*Таблица 1.1. Популярные области применения машинного обучения*

|  |  |
| --- | --- |
| Автономные машины  Анализ эмоциональной окраски (например, классификация рецензий на фильмы на отрицательные, положительные и нейтральные)  Выявление аномалий  Выявление закономерностей в данных  Выявление попыток мошенничества с кредитными картами  Выявление попыток страхового мошенничества  Глубокий анализ данных в социальных сетях (Facebook, Twitter, LinkedIn)  Диагностическая медицина  Исследование данных  Классификация новостей: спорт, финансы, политика и т. д.  Классификация электронной почты и выделение спама  Маркетинг: деление клиентов на группы  Обнаружение вторжений в компьютерные системы | Обнаружение объектов в сценах  Перевод естественных языков (с английского на испанский, с французского на японский и т. д.)  Прогнозирование временных рядов — например, предсказание будущих котировок акций и прогнозы погоды  Прогнозирование нарушений выплат ипотечных кредитов  Прогнозирование оттока клиентов  Распознавание голоса  Распознавание лиц  Распознавание образов и классификация изображений  Распознавание рукописного текста  Рекомендательные системы («тем, кто купил этот продукт, также понравились…»)  Сжатие данных  Фильтрация спама  Чат-боты |

# Обзор размеченных корпусов для обучения

Для построения модели с высокой точностью, в первую очередь, необходимы хорошо организованные данные. Поэтому было решено использовать один из размеченных корпусов данных, представленных в большом количестве на сайте kaggle.com. Так как ресурс иностранный, то в большей степени подходящие под тематическую классификацию корпуса были на английском языке, что не удовлетворяет цели работы.

После анализа сайта были выделены следующие корпусы данных на русском языке, которые подходят для обучения на определение тематики текстов. У каждого из них были выделены положительные и отрицательные аспекты, исходя из которых был выбран более подходящий для данной работы корпус.

## Articles from russian site habr.com

Корпус данных, содержащий статьи с русского информационного ресурса habr.com.

Характеристики:

* размер 2,7 Гб;
* состоит из трех файлов:
  + habs.csv

Соотносит статьи к хабам.

Атрибуты: post\_id, hab.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1.1 – Содержание файла habs.csv

* + posts.csv

Статьи с метаданными.

Атрибуты: post\_id, title, text, date, views\_count, comments\_count, bookmarks\_count, rating, author\_nickname.

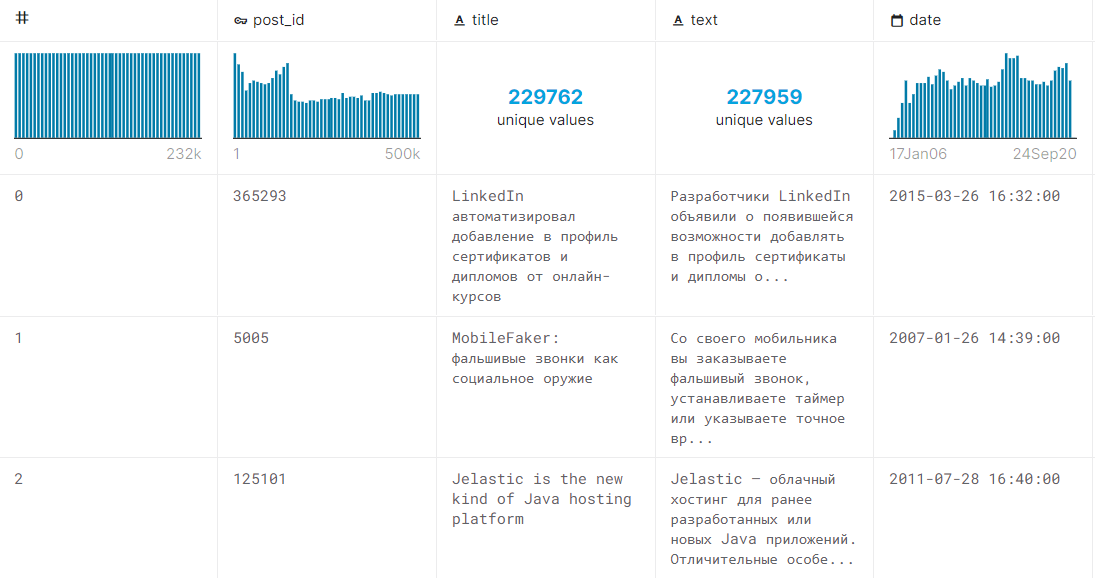


Рисунок 1.2 – Содержание файла posts.csv

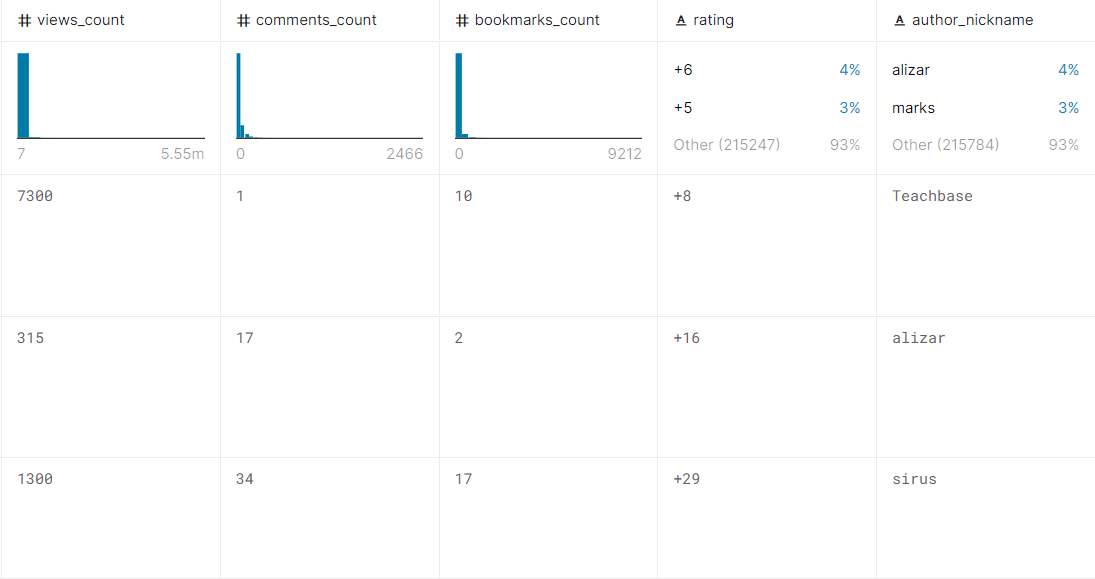


Рисунок 1.3 – Содержание файла posts.csv(продолжение)

* + tags.csv

Соотносит статьи к тегам.

Атрибуты: post\_id, tag.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1.4 – Содержание файла tags.csv

* количество уникальных текстов 227959.

В качестве атрибута можно взять текст статьи, а в качестве класса – её тег.

Набор данных среднего размера, но с довольно большими текстами (средний размер 7000 символов), что будет сказываться на точности модели. Неудобство использования из-за отделения текстов и тегов.

Решающий недостаток, из-за которого данный набор данных не подходит под задачи проекта, - большинство тегов уникально. То есть сколько текстов, столько и классов. При таком раскладе не получится обучить модель.

## News dataset from Lenta.Ru

Размеченный корпус статей с русского новостного портала Lenta.ru.

Характеристики:

* размер 2,08 Гб;
* состоит из одного файла:
  + lenta-ru-news.csv

Атрибуты: url, title, text, topic, tags, date.

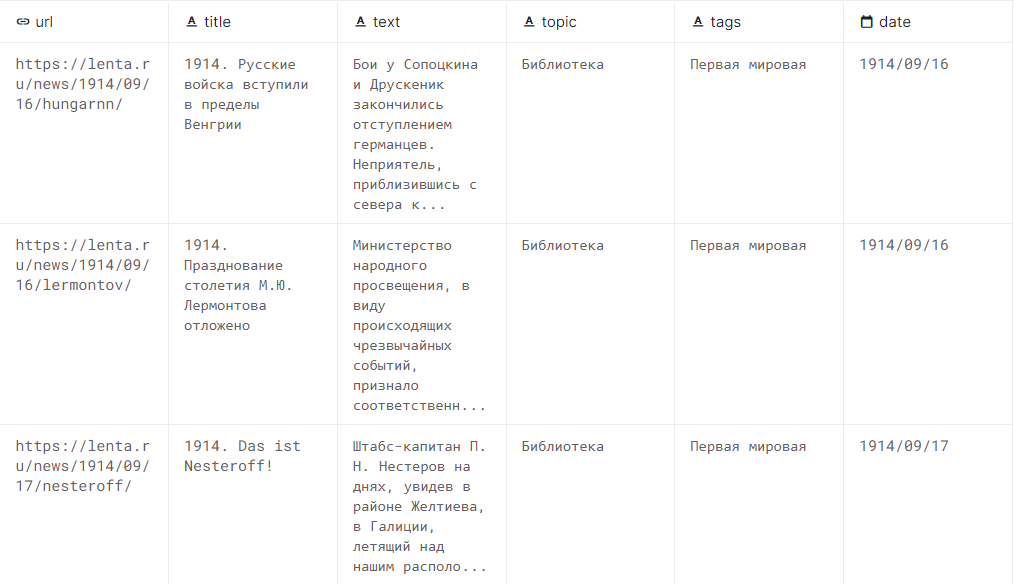


Рисунок 1.5 – Содержание файла lenta-ru-news.csv

* количество уникальных текстов 800038

В качестве атрибута можно использовать текст статьи, а в качестве класса – её топик или тег.

Набор данных довольно большого размера и с небольшими текстами (средний размер N символов). Недостатком является то, что 57% всех тегов являются обобщенными и называются “Все”. Их следует исключить и набора данных. Рассказать, почему следует взять теги, а не топики.

## Выводы по разделу

В этом разделе были рассмотрены готовые размеченные корпуса на русском языке, которые подходят для обучения тематической классификации. В обзор попали два корпуса: статей из habr.com и новостей из lenta.ru. В ходе их анализа было выявлены плюсы и минусы, исходя из которых сделан выбор в пользу корпуса с новостями.

# Обзор классификаторов

В настоящее время разработано большое количество различных видов классификаторов, для построения которых используются как статистические методы (логистическая регрессия, дискриминантный анализ), так и методы машинного обучения (нейронные сети, деревья решений).

Необходимость использования в анализе данных большого числа разнообразных методов классификации, обусловлена тем, что решаемые с её помощью задачи могут иметь свои особенности, связанные, например, с представлением исходных данных, их количеством и качеством, что требует выбора адекватного классификатора. Поэтому выбор классификатора, соответствующего особенностям решаемой задачи анализа, является важным фактором получения правильного решения.

В работе будут рассмотрены и в последующем сравнены три самых популярных метода классификации.

## Метод K-ближайших соседей (k-nearest neighbors algorithm, KNN)

Метод был впервые разработан Эвелином Фиксом и Джозефом Лоусоном Ходжесом в 1951 году, и позднее развит Томасом Ковером.

Метод относится к классу непараметрических, т. е. не требует предположений о том, из какого статистического распределения была сформирована обучающее множество. Следовательно, классификационные модели, построенные с помощью метода KNN, также будут непараметрическими. Это означает, что структура модели не задаётся жёстко изначально, а определяется данными.

Поскольку признаки, на основе которых производится классификация, могут иметь различную физическую природу и, соответственно, диапазоны значений, для улучшения результатов классификации будет полезно выполнить нормализацию обучающих данных.

### **Алгоритм.**

Пусть имеется набор данных, состоящий из  наблюдений , для каждого из которых задан . Тогда на его основе может быть сформировано обучающее множество, все [примеры](https://wiki.loginom.ru/articles/training-sample.html) которого представляют собой пары .

Алгоритм KNN можно разделить на две простые фазы: обучения и классификации. При обучении алгоритм просто запоминает векторы признаков наблюдений и их метки классов (т. е. примеры). Также задаётся параметр алгоритма , который задаёт число “соседей”, которые будут использоваться при классификации.

На фазе классификации предъявляется новый объект, для которого метка класса не задана. Для него определяются ближайших (в смысле некоторой метрики) предварительно классифицированных наблюдений. Затем выбирается класс, которому принадлежит большинство из ближайших примеров-соседей, и к этому же классу относится классифицируемый объект.

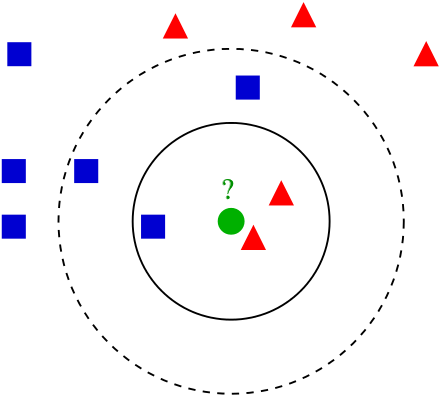


Рисунок 3.1 – Работа алгоритма K-ближайших соседей

Есть тестовый образец в виде зеленого круга. Синие квадраты относятся к классу *1*, красные треугольники – к классу *2*. Зеленый круг должен быть классифицирован как класс *1* или класс *2*.

Если рассматриваемая область является малым кругом (), то объект классифицируется как класс *2*, потому что внутри данного круга два треугольника и только один квадрат.

Если рассматривается большой круг (), то круг будет классифицирован как класс *1*, так как внутри круга три квадрата и два треугольника.

### **Определение класса нового объекта.**

В простейшем случае класс нового объекта может быть определён простым выбором наиболее часто встречающегося класса среди  примеров. Однако на практике это не всегда удачное решение, например, в случае, когда частота появления для двух или более классов оказывается одинаковой. Кроме этого, разумно предположить, что не все обучающие имеют одинаковую значимость для определения класса. В этом случае используют некоторую функцию, с помощью которой определяется класс, называемую функцией сочетания (*combination function*).

В обычном случае используют так называемое простое невзвешенное голосование (*simple unweighted voting*). При это предполагается, что все  примеров имеют одинаковое право «голоса» независимо от расстояния до классифицируемого объекта.

Однако, логично предположить, что чем дальше пример расположен от классифицируемого объекта в пространстве признаков, тем ниже его значимость для определения класса. Поэтому для улучшения результатов классификации вводят взвешивание примеров в зависимости от их удалённости. В этом случае используют взвешенное голосование (*weighted voting*).

В основе идеи взвешенного голосования лежит введение «штрафа» для класса, в зависимости от того, насколько относящиеся к нему примеры удалены от классифицируемого объекта. Такой “штраф” представляется как сумма величин, обратных квадрату расстояний от примера -го класса до классифицируемого объекта (часто данное значение называют показателем близости):

где  — оператор вычисления расстояния,  – вектор признаков классифицируемого объекта, ​ – -ый пример -го класса. Таким образом, “побеждает” тот класс, для которого величина ​ окажется наибольшей. При этом также снижается вероятность того, что классы получат одинаковое число голосов.

### **Выбор значения параметра k.**

Выбор параметра является важным для получения корректных результатов классификации. Если значение параметра мало, то возникает эффект переобучения, когда решение по классификации принимается на основе малого числа примеров и имеет низкую значимость. Это похоже на переобучение в деревьях решений, когда в них много правил, относящихся к небольшому числу примеров. Если установить , то алгоритм будет просто присваивать любому новому наблюдению метку класса ближайшего объекта.

Кроме этого, следует учитывать, что использование небольших значений увеличивает влияние шумов на результаты классификации, когда небольшие изменения в данных приводят к большим изменениям в результатах классификации. Но при этом границы классов оказываются более выраженными (класс при голосовании побеждает с большим счётом).

Напротив, если значение параметра слишком велико, то в процессе классификации принимает участие много объектов, относящихся к разным классам. Такая классификация оказывается слишком грубой и плохо отражает локальные особенности набора данных. Таким образом, выбор параметра является компромиссом между точностью и обобщающей способностью модели.

При больших значениях параметра уменьшается зашумленность результатов классификации, но снижается выраженность границ классов.

В задачах бинарной классификации бывает целесообразно выбрать как нечетное число, так как это позволяет избежать равенства “голосов” при определении класса для нового наблюдения.

### **Особенности работы**

Если значения признаков непрерывные, то в качестве меры расстояния между объектами обычно используется расстояние Евклида, а если категориальные, то может использоваться расстояние Хэмминга.

Алгоритма KNN является чувствительным к дисбалансу классов в обучающих данных: алгоритм “склонен” к смещению решения в сторону доминирующего класса, поскольку относящиеся к нему объекты просто чаще попадают в число ближайших соседей. Одним из способов решения данной проблемы является применение различных способов взвешивания при “голосовании”.

Следует отметить, что отношение соседства не является коммутативным, т. е. если для записи ближайшем соседом является запись , то это не означает, что – ближайший сосед .

Ещё одной проблемой алгоритма KNN, характерной, впрочем, и для большинства методов классификации, является различная значимость признаков с точки зрения определения класса объектов. Учет фактора значимости признаков в алгоритме может позволить повысить точность классификации.

Для этого аналитик или эксперт на основе субъективной, либо некоторой формальной оценки может задать уровень значимости признака, выразив его с помощью числового коэффициента (обозначим его от англ *significance* — значимость), который учитывается при вычислении расстояния между примерами и классифицируемым объектом:

где  – коэффициент значимости для -го признака,  –количество признаков исходного набора данных. Такой приём называется растяжением осей**,** и он позволяет увеличить или уменьшить вклад признака в вычисление расстояния от примера до классифицируемого объекта. Если , то благодаря соответствующему признаку расстояние между примером классифицируемым объектом растёт и вклад в определение класса падает, а если , то наоборот.

### **Области применения**

Алгоритм KNN может применяться практически во всех задачах классификации, особенно в тех случаях, когда оценить параметры вероятностного распределения данных сложно или невозможно. Наиболее типичными приложениями алгоритма KNN являются:

* классификация клиентов (например, по уровню лояльности);
* медицина – классификация пациентов по медицинским показателям;
* маркетинг – классификация товаров по уровню популярности и т. д.

### **Достоинства и недостатки**

К достоинствам алгоритма можно отнести:

* устойчивость к выбросам и аномальным значениям, поскольку вероятность попадания содержащих их записей в число k-ближайших соседей мала. Если же это произошло, то влияние на голосование (особенно взвешенное) также, скорее всего, будет незначительным, и, следовательно, малым будет и влияние на результаты классификации;
* программная реализация алгоритма относительно проста;
* результаты работы алгоритма легко поддаются интерпретации. Логика работы алгоритма понятна экспертам в различных областях.

К недостаткам алгоритм KNN можно отнести:

* данный метод не создает каких-либо моделей, обобщающих предыдущий опыт, а интерес могут представлять и сами правила классификации;
* при классификации объекта используются все доступные данные, поэтому метод KNN является достаточно затратным в вычислительном плане, особенно в случае больших объёмов данных;
* высокая трудоёмкость из-за необходимости вычисления расстояний до всех примеров;
* повышенные требования к репрезентативности исходных данных.

## Модель RandomForest

## Модель LogisticRegression

## Выводы по разделу

# Подготовка размеченного корпуса

## Анализ корпуса

## Предварительная обработка данных

## Выводы по разделу

# Обучение моделей

## Обучение модели K-ближайших соседей

## Обучение модели RandomForest

## Обучение модели LogisticRegression

## Выводы по разделу

# Сравнение моделей

## Результаты оценщика

## Выводы по разделу

# Заключение

В данном курсовом проекте были решены все задачи и достигнута поставленная цель.

Во-первых, был произведен обзор аналогов игры Судоку и выявлен недостающий функционал, а именно отсутствие возможности ввода собственного Судоку и его решения.

Во-вторых, были рассмотрены алгоритм генерации и алгоритмы решения головоломки, а именно Алгоритм перебора с возвратом и его модификация. Они представлены в виде псевдокода и для наглядности продублированы примерами. Данные алгоритмы были протестированы на различных входных данных, благодаря чему сделаны выводы и выбран оптимальный алгоритм для реализации игрового приложения.

В-третьих, было спроектировано и разработано приложение Судоку, которое включает в себя следующие функции:

* Выбор режима игры: генерируемая игра или решение собственного Судоку.
* Выбор размера и сложности игры. Для генерируемой игры доступны размеры 4×4, 9×9, 16×16, 25×25 и сложности легкая, средняя, высокая и эксперт. Для собственной игры доступны размеры 4×4 и 9×9.
* Подсказки для обоих режимов.
* Возможность показать всё решение для обоих режимов.
* Пауза.
* Фоновая музыка, громкостью которой можно управлять из приложения.

В-четвертых, были обработаны все непредвиденные ситуации поведения пользователя при использовании приложения, а также написаны Unit-тесты, которые показывают корректность основных методов программы.

Таким образом, результатом курсового проекта стало игровое приложение Судоку, основанное на рассмотренных алгоритмах, которые были описаны благодаря анализу научной литературы.

# Библиографический список

1. Дейтел Пол, Дейтел Харви. Python: Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.: ил. — (Серия «Для профессионалов»).
2. Рашид, Тарик. Создаем нейронную сеть.: Пер. с англ. — СПб.: ООО “Альфа-книга”, 2017. — 272 с.: ил. — Парал. тит. англ.
3. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина. - М.: ДМАК Пресс, 2015. - 400 с.: ил.
4. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. Москва 2016–2017.
5. Домингос, Педро. Верховный алгоритм: как машинное обучение изменит наш мир / Педро Домингос; пер. с англ. В. Горохова; [науч. ред. А. Сбоев, А. Серенко]. — М.: Манн, Иванов и Фербер, 2016. — 336 с.
6. Шолле Франсуа. Глубокое обучение на Python. — СПб.: Питер, 2018. — 400 с.: ил. — (Серия «Библиотека программиста»).
7. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.
8. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.
9. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001.

# Приложения

## Приложение А. Листинг программы