УДК 004.891.2 URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1035.html

# КЛАССИФИКАЦИЯ ЛИЗИНГОВОЙ ДОКУМЕНТАЦИИ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

**Д.И. Насибуллин** nasibullindi@student.bmstu.ru

*МГТУ им. Н.Э. Баумана, Москва, Российская Федерация*

Работа посвящена актуальной задаче использования автоматизированной классификации документов, связанной с необходимостью сокращения времени и количества ошибок при обработке большого числа документов. Различные лизинговые документы собраны и предварительно разделены на типы. Определены основные алгоритмы машинного обучения, предназначенные для классификации данных. Построены графики тестовых и обучающих выборок, необходимые для определения наиболее приемлемых гиперпараметров моделей, что позволяет достичь лучшего результата предсказания обученных моделей. Проведен анализ и составлена сравнительная характеристика обученных моделей на исследуемых данных. Выявлено, что наиболее подходящей моделью машинного обучения для классификации лизинговой документации является наивный байесовский классификатор. Подчеркнуто, что его преимущество перед другими моделями связано с высокой скоростью обучения и предсказания, а также прогнозированием типа документа с точностью более 90 %.

**Ключевые слова:** машинное обучение, классификация текстовой документации, матрица ошибок, автоматизация документооборота, лизинговая документация, метод деревьев принятия решений, метод ближайших соседей, метод опорных векторов, байесовский классификатор

**Введение.** Современные информационные технологии являются неотъемлемой частью жизни каждого человека. Компании стремятся внедрять различные автоматизированные системы, улучшая удобство работы сотрудников и предоставляя клиентам более качественные услуги. В направлениях с высокой конкурентностью это может оказаться одним из важнейших факторов успеха организации в борьбе за привлечение потребителей товаров.

Большую долю рынка России занимает банковский сектор, который предоставляет большое количество различных услуг. Одной из популярных для населения услуг является кредитование, среди которых можно выделить кредиты наличными, потребительские кредиты, ипотеки или лизинг. Особенностью данной сферы является анализ рисков кредитования, оформление различных договоров, исследование подозрительных активностей клиента, поиск актуальной информации и многое другое. Для полного предоставления услуги требуется задействовать разные отделы компании, от менеджеров до отдела

 ISSN 2541-8009

рисков, а весь процесс оформления может занимать несколько дней, что негативно сказывается на клиентском опыте.

Внедрение автоматизированных систем проверки документации в данной области позволит упростить процессы оформления кредитования, что предоставит возможность снизить нагрузку на внутренние отделы банка, а также увеличить поток клиентов. В качестве рассматриваемой области автоматизации было выбрано кредитование по лизинговому договору.

Анализ лизинговой документации может состоять из большого количества этапов, таких как определение типов документов, извлечение финансовой информации и персональных данных, проверка печатей и подписей и другое. В данной работе рассмотрено применение методов машинного обучения для классификации документов по их типу.

Целью работы является определение наиболее подходящей модели машинного обучения для анализа лизинговой документации с точки зрения достижения приемлемой точности классификации. Применение машинного обучения для решения поставленной задачи включает в себя такие этапы, как предварительная обработка данных, разработка векторизации текста и обучение модели машинного обучения [1].

Эксперименты с подготовкой обучающей выборки и обучения моделей проводили с помощью языка программирования Python и специализированных библиотек.

В исследовании решены следующие задачи:

– подготовка документов различных типов;

– предварительная обработка текста для обучения моделей;

– обучение моделей;

– анализ метрик моделей;

– определение модели, наиболее качественно справляющейся с поставленной задачей.

Результатом работы служит готовая модель машинного обучения, показавшая наибольшую точность классификации лизинговой документации.

**Подготовка обучающей выборки.** Подготовка обучающей выборки — важный этап решения задачи классификации текстовых документов. От ее качества и размера зависит результат предсказания полученной модели.

В качестве набора данных были собраны документы различных типов [2]: учредительные документы юридического лица; выписка из ЕГРЮЛ или ЕГРИП; бухгалтерская отчетность за прошедшие 3–12 месяцев; справка из кредитного учреждения об обороте финансовых средств лизингополучателя; информация о надлежащем исполнении налоговых обязательств за последние 3–12 месяцев. По причине наличия в документах персональных данных информация

о собранном датасете не может быть предоставлена. В исходной выборке содержится 2 573 документа, которые подразделены на пять типов:

1) учредительные документы юридического лица — 282;

2) выписка из ЕГРЮЛ или ЕГРИП — 36;

3) бухгалтерская отчетность за прошедшие 3–12 месяцев — 317;

4) справка из кредитного учреждения об обороте финансовых средств лизингополучателя — 702;

5) информация о надлежащем исполнении налоговых обязательств за последние 3–12 месяцев — 1236.

После определения набора данных требуется выполнить его предобработку, которая включает в себя следующие этапы [3]:

1) нормализация — приведение текста к одному регистру, удаление знаков пунктуации;

2) токенизация — разделение строк на более короткие (обычно разделение происходит по словам);

3) удаление стоп-слов — очистка текста от предлогов, союзов и других конструкций языка, которые не несут смысловой нагрузки;

4) лемматизация — приведение слов к основной форме.

Следующий этап обработки текста — его векторизация. Необходимость представления текстов в виде векторов признаков обусловлено тем, что для использования рассматриваемых методов классификации требуется, чтобы классифицируемые объекты были представлены в виде последовательностей чисел одинакового размера и одинакового формата [4].

Среди самых популярных методов векторизации можно выделить следующие [5]:

1) BinaryBOW — градация признаков по наличию или отсутствию в документе;

2) Bag of words — векторное представление всего документа и индексация каждого токена в порядке следования слов в словаре;

3) TF-IDF — показателем данного метода является вес векторного представления документа, таким образом, чем меньше вес, тем чаще слово встречается в документе и не несет отличительного признака;

4) Word2Vec — это метод векторного представления слов с помощью неглубоких нейронных сетей.

На основе представленных методов было решено использовать векторизацию с помощью TF-IDF, поскольку он учитывает вес каждого признака в документе, а использование нейронных сетей для решения поставленной задачи является избыточным. На практике был использован инструмент TfidfVectorizer из пакета sklearn [6].

 ISSN 2541-8009

**Обучение моделей.** После завершения подготовки обучающей выборки необходимо определить наиболее подходящую модель для решения поставленной задачи. Для этого существует два основных подхода [7]:

1) классификация — предсказание класса нового экземпляра данных;

2) кластеризация — выделение группы схожих объектов без предварительно определенных классов.

В данном случае наиболее подходящими являются алгоритмы классификации, поскольку на предыдущем этапе уже был собран и обработан массив данных для дальнейшей работы. Среди существующих методов машинного обучения были применены следующие [6, 8]:

– метод деревьев принятия решений — непараметрический метод, используемый в машинном обучении, анализе данных и статистике;

– метод ближайших соседей (KNN) — отнесение объекта к классу, к которому относится большинство из $k$ ближайших к нему объектов;

– метод опорных векторов (SVM) — его идея состоит в построении оптимальной гиперплоскости, выступающей в роли поверхности решений, максимально разделяющей объекты разных классов.

Кроме того, можно использовать наивный байесовский классификатор, который является одним из самых простых методов классификации, но показывает на практике хорошие результаты [7, 9].

Для каждого из методов была разработана функция, позволяющая определить гиперпараметры, свойственные каждой модели. Практическая часть была реализована с помощью инструментов из библиотеки sklearn. Качество обученной модели оценивали по сбалансированной точности, поскольку исходный набор данных содержит неравномерное распределение среди всех типов.

Решение задачи методом KNN основано на определении отношения к классу путем выявления типов ближайших соседей к исследуемому элементу. Точность прогнозирования результата данного алгоритма зависит от гиперпараметра «количество соседей», который обозначается как “n_neighbors” в библиотеке sklearn. На рис. 1 видно, что графики обучающей и тестовой выборок коррелируются, а наилучшая точность прогнозирования модели достигается при значении n_neighbors = 42.

Одним из используемых инструментов для анализа полученных моделей служит матрица ошибок, характерная для многоклассовой классификации. С ее помощью удобно рассматривать качество предсказаний принадлежности документов к конкретному классу. Главная диагональ содержит количество корректно спрогнозированных образцов, а остальные значения — количество ошибок [7, 9].

В соответствии с определенным выше гиперпараметром n_neighbors для модели KNN была составлена матрица ошибок на тестовой выборке (рис. 2).

Полученные результаты позволяют убедиться в высокой точности прогнозирования большинства классов, поскольку элементы сконцентрированы в основном на главной диагонали. Единственным исключением является низкое качество определения элементов 3-го типа.

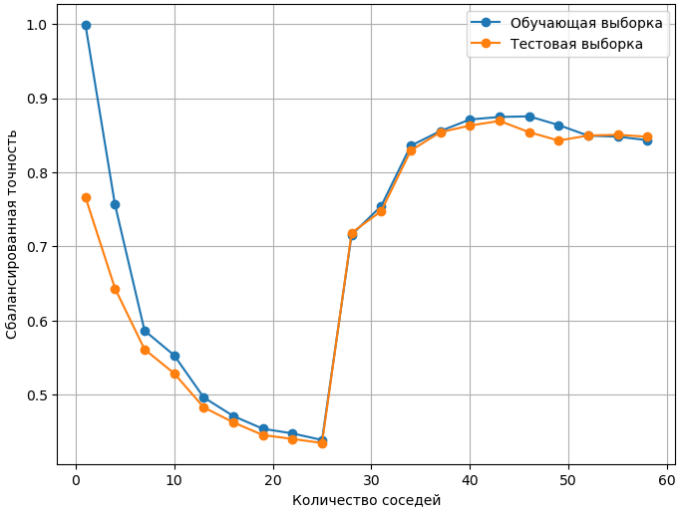


**Рис. 1.** Зависимость сбалансированной KNN точности от n_ neighbors

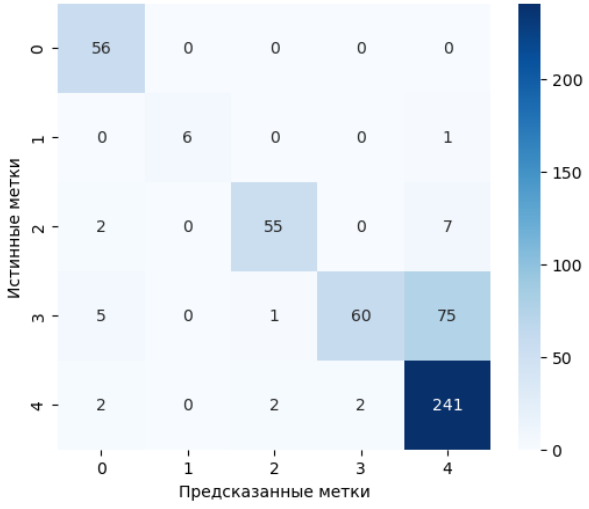


**Рис. 2.** Матрица ошибок KNN

 ISSN 2541-8009

Решение задачи методом дерева принятия решений основано на определении отношения к классу путем поэтапной проверки заранее прописанного набора правил или простейших условий. Точность прогнозирования результатов зависит от гиперпараметра минимального количества выборок, необходимых для разделения узла — “min_samples_split”. На рис. 3 видно, что графики обучающей и тестовой выборок не зависят друг от друга, а наиболее точное предсказание модель выдает при значении гиперпараметра min_samples_split = 3.

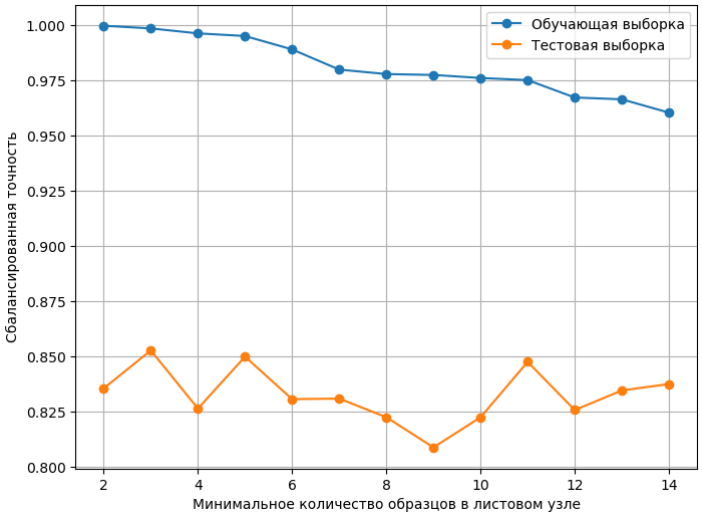


**Рис. 3.** Зависимость сбалансированной точности дерева принятия решения от минимального количества образцов в листовом узле

В соответствии с определенным выше гиперпараметром min_samples_split для модели дерева принятия решений была составлена матрица ошибок на тестовой выборке (рис. 4).

По сравнению с методом KNN матрица ошибок дерева принятия решений более точно определяет класс 3, при этом можно заметить ухудшение качества предсказания класса 4.

Решение задачи методом опорных векторов основывается на определении отношения к классу путем поиска разделяющей гиперплоскости классов с наибольшим зазором в пространстве. Точность прогнозирования результатов зависит от гиперпараметра «коэффициент регуляризации». На рис. 5 видно, что графики обучающей и тестовой выборок коррелируются, а наиболее точное предсказание модель выдает при значении гиперпараметра «коэффициент регуляризации» = 26,8.

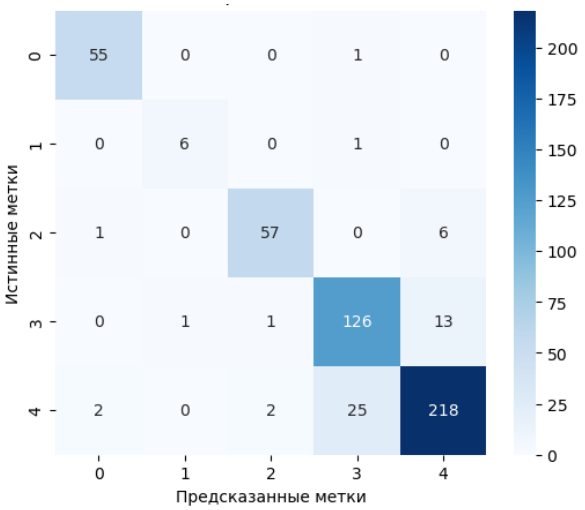


**Рис. 4.** Матрица ошибок дерева принятия решений

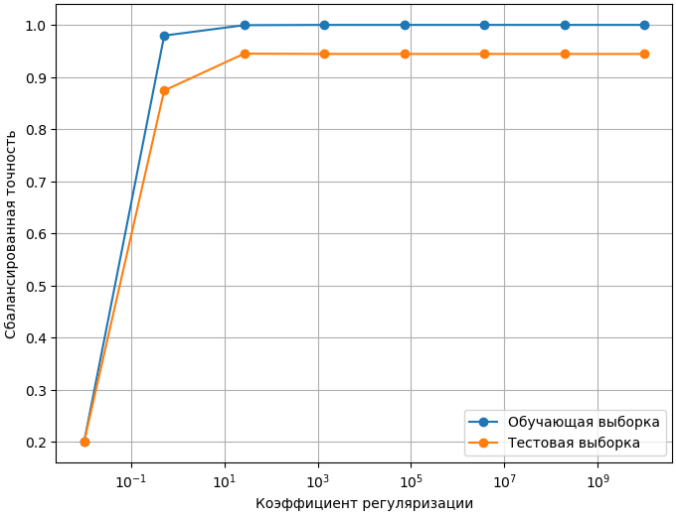


**Рис. 5.** Зависимость сбалансированной точности SVM от коэффициента регуляризации

В соответствии с определенным выше гиперпараметром коэффициента регуляризации для модели SVM была составлена матрица ошибок на тестовой выборке (рис. 6).

В полученной матрице ошибок можно заметить, что метод SVM более точно справляется с задачей прогнозирования различных типов документов по сравнению с методами KNN и деревьев принятия решений.

 ISSN 2541-8009

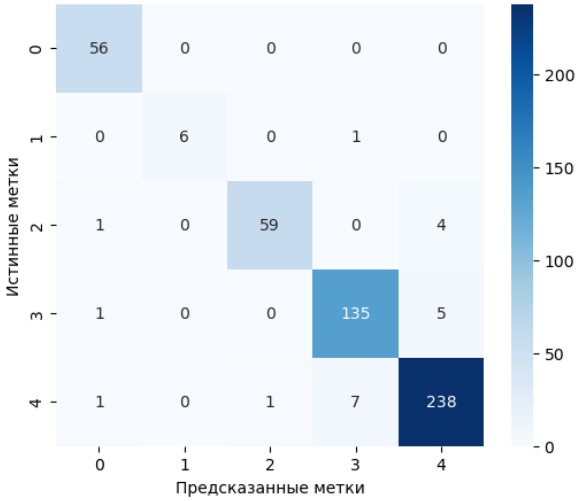


**Рис. 6.** Матрица ошибок метода SVM

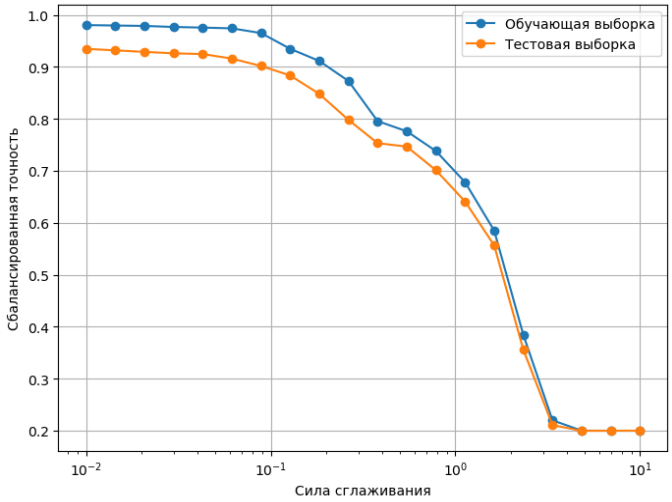


**Рис. 7.** Зависимость сбалансированной точности наивного байесовского классификатора от силы сглаживания

Решение задачи с помощью метода наивного байесовского классификатора основывается на определении вероятности объекта, он основан на теореме Байеса и в основном используется для классификации текста. Точность прогнозирования результатов зависит от гиперпараметра силы сглаживания — “alpha”. На рис. 7 видно, что графики обучающей и тестовой выборок коррелируются, а наиболее точное предсказание модель выдает при значении гиперпараметра “alpha” = 0,01.

В соответствии с определенным выше гиперпараметром коэффициента регуляризации для модели наивного байесовского классификатора была составлена матрица ошибок на тестовой выборке (рис. 8).

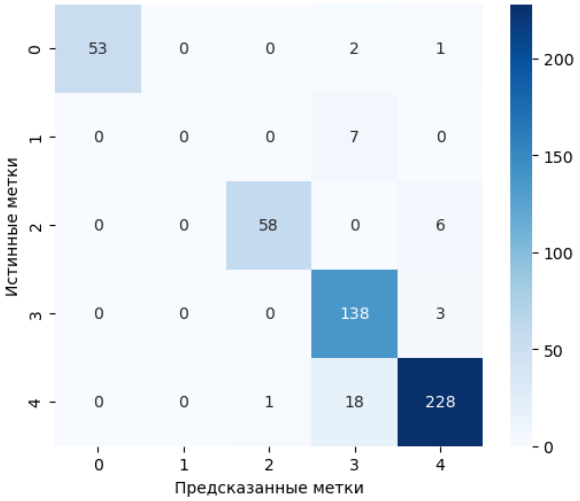


**Рис. 8.** Матрица ошибок метода опорных векторов

Данная матрица ошибок отражает высокое качество предсказания модели. Несмотря на это метод наивного байесовского классификатора показывает результаты немного хуже по сравнению с методом SVM.

**Сравнительный анализ.** Сравнительный анализ ранее обученных моделей позволит определить наиболее подходящую модель для решения задачи классификации текстовых документов. Для данного типа задач применимы следующие метрики [8, 10]:

– скорость обучения (предсказания) — позволяет определить затрачиваемые ресурсы на обучение (предсказание);

– сбалансированная точность — позволяет проводить более справедливое сравнение между различными моделями, обученными на несбалансированных данных, обеспечивая более надежную оценку их производительности;

– F1-метрика — позволяет оценивать модели в случаях, когда важен баланс между точностью и полнотой предсказания.

В процессе обучения моделей были рассчитаны значения с помощью всех перечисленных методов машинного обучения. Все измерения проводили с помощью программы Jupyter Notebook и библиотеки для работы с машинным обучением на языке python sklearn, установленной на операционной си-

 ISSN 2541-8009

стеме семейства UNIX. Технические характеристики аппаратного обеспечения представлены ниже:

– процессор Inter Core i7;
– количество ядер процессора 12;
– частота процессора 2,1 ГГц;
– ОЗУ 16 Гб;
– частота ОЗУ 3,2 ГГц;
– ПЗУ 256 Гб.

Результаты вычислений представлены в таблице.

**Сравнение обученных моделей**

| Метод | Время обучения | Время предсказания | Сбалансированная точность | F1 |
|---|---|---|---|---|
| KNN | 18,637 | 0,062 | 0,82 | 0,79 |
| Дерево принятия решений | 22,152 | 0,001 | 0,84 | 0,90 |
| Метод опорных векторов | 80,384 | 0,330 | 0,94 | 0,96 |
| Наивный байесовский классификатор | 2,667 | 0,005 | 0,91 | 0,90 |

Представленные выше результаты обучения моделей с помощью используемых моделей машинного обучения позволяют определить модель, наиболее подходящую для решения поставленной задачи. Согласно данным, приведенным в таблице, наиболее приемлемой моделью является наивный байесовский классификатор. По критериям сбалансированной точности и метрики F1 он показывает результаты, приближенные к результатам метода SVM, но скорость обучения и предсказания данной модели выше в десятки раз.

**Заключение.** Таким образом, в данной работе были исследованы возможности применения методов машинного обучения к анализу лизинговой документации на примере задачи классификации различных документов. Для этого были собраны и размечены данные таких документов, как:

– учредительные документы юридического лица;
– выписка из ЕГРЮЛ или ЕГРИП;
– бухгалтерская отчетность за прошедшие 3–12 месяцев;
– справка из кредитного учреждения об обороте финансовых средств лизингополучателя;
– информация о надлежащем исполнении налоговых обязательств за последние 3–12 месяцев.

В процессе подготовки текстовое содержание документов было обработано для последующего применения нормализации, токенизации, удаления

 

стоп-слов и лемматизации, после данные были преобразованы в векторный формат с помощью метода TF-IDF.

Было проведено обучение различных моделей для выявления наиболее подходящей для решения поставленной задачи. Среди моделей KNN, дерева принятия решений, SVM и наивного байесовского классификатора была выделена последняя, поскольку она обладает высокой скоростью обучения и предсказания, а точность результатов составила более 90 %. Данный результат обусловлен хорошими показателями модели при работе с текстовыми данными.

В перспективе планируется обобщить полученные результаты на случай работы с различной документацией, а также рассмотреть эффективность применения других методов машинного обучения к классификации произвольных текстовых документов из различных предметных областей.

Работа может быть полезна специалистам, занимающимся анализом текстовой документации с использованием методов машинного обучения.

**Литература**

[1] Гусев П.Ю. Обработка текстов и подготовка моделей векторизации для программного комплекса классификации научных текстов. *Моделирование, оптимизация и информационные технологии*, 2021, № 9 (1). https://doi.org/10.26102/2310-6018/2021.32.1.010

[2] *Что нужно для оформления лизинга?* URL: https://www.ileasing.ru/about/clients/on-leasing/detail/chto-nuzhno-dlya-oformleniya-lizinga/ (дата обращения 20.02.2025).

[3] Чижик А.В., Жеребцова Ю.А. Создание чат-бота: обзор архитектур и векторных представлений текста. *International Journal of Open Information Technologies*, 2020, № 7 (8), с. 50–56.

[4] Бурлаева Е.И. Обзор методов классификации текстовых документов на основе подхода машинного обучения. *Программная инженерия*, 2017, № 7 (8), с. 328–336. https://doi.org/10.17587/prin.8.328-336

[5] Попова О.А. Анализ методов векторизации текстовых документов. *Вестник РГРТУ*, 2023, № 85, с. 96–102. https://doi.org/10.21667/1995-4565-2023-85-96-102

[6] *Scikit-learn: machine learning in Python — scikit-learn 1.6.1 documentation*. URL: https://scikit-learn.org/stable/index.html (дата обращения 15.02.2025).

[7] Боженко В.В., Клюканов В.К. Применение алгоритмов машинного обучения в задачах классификации и кластеризации. *Обработка, передача и защита информации в компьютерных системах. Вторая Междунар. науч. конф.: сб. ст.* Санкт-Петербург, ГУАП, 2022, с. 28–33. https://doi.org/10.31799/978-5-8088-1701-2-2022-2-28-33

 ISSN 2541-8009

[8] Бабаев А.М., Шемякина М.А. Обзор классических методов машинного обучения в контексте решения задач классификации. *Форум молодых ученых*, 2018, № 11 (27), с. 137–142.

[9] Золина Е.В., Гамова Н.А. Наивный классификатор Байеса для решения задачи сентимент-анализа текстов. *Шаг в науку*, 2019, № 4, с. 140–142.

[10] Михайличенко А.А. Аналитический обзор методов оценки качества алгоритмов классификации в задачах машинного обучения. *Вестник АГУ*, 2022, № 4 (311), с. 52–59. https://doi.org/10.53598/2410-3225-2022-4-311-52-59

***Поступила в редакцию 26.02.2025***

**Насибуллин Данил Ильнурович** — студент магистратуры кафедры «Компьютерные системы и сети», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

**Научный руководитель** — Гуренко Владимир Викторович, кандидат технических наук, доцент, первый заместитель декана факультета ИУ МГТУ им. Н.Э. Баумана, доцент кафедры «Компьютерные системы и сети» МГТУ им. Н.Э. Баумана. E-mail: wgurenko@bmstu.ru; SPIN-код: 2675-4796.

**Ссылку на эту статью просим оформлять следующим образом:**
Насибуллин Д.И. Классификация лизинговой документации на основе методов машинного обучения. *Политехнический молодежный журнал*, 2025, № 02 (97). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1035.html

 

# CLASSIFICATION OF LEASING DOCUMENTATION USING THE MACHINE LEARNING METHODS

**D.I. Nasibullin** nasibullindi@student.bmstu.ru

*Bauman Moscow State Technical University, Moscow, Russian Federation*

The paper is devoted to a relevant problem of using the documents automated classification, which is associated with the need to reduce the time and number of errors in processing a large number of documents. Various leasing documents are collected and pre-divided into several types. The paper defines the main machine learning algorithms designed for data classification. It provides constructed graphs of the test and learning samples required in determining the most acceptable hyperparameters of the models, which allows achieving the best result in forecasting a learned model. The paper analyzes and compiles a comparative characteristic of the learned models using the studied data. It indicates that the naïve Bayesian classifier appears to be the most suitable machine learning model for classifying the leasing documentation. The paper emphasizes that its advantage over the other models is associated with the high speed in learning and forecasting, as well as in predicting the document type with more than 90% accuracy.

**Keywords:** machine learning, text documentation classification, error matrix, document flow automation, leasing documentation, tree decision making method, nearest neighbor method, support vector method, Bayesian classifier

***Received 26.02.2025***

**Nasibullin D.I.** — Master's Program Student, Department of Computer Systems and Networks, Bauman Moscow State Technical University, Moscow, Russian Federation.

**Scientific advisor** — Gurenko V.V., Ph. D. (Eng.), Associate Professor, First Deputy Dean, Faculty of Information Technologies, Bauman Moscow State Technical University; Associate Professor, Department of Computer Systems and Networks, Bauman Moscow State Technical University. E-mail: wgurenko@bmstu.ru; SPIN-code: 2675-4796.

**Please cite this article in English as:**
Nasibullin D.I. Classification of leasing documentation using the machine learning methods. *Politekhnicheskiy molodezhnyy zhurnal*, 2025, no. 02 (97). (In Russ.). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/1035.html

 ISSN 2541-8009