Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра автоматизации обработки информации (АОИ)

Вариант 3

Отчёт к Лабораторной работе №4

по дисциплине «Параллельные вычисления и системы»

Выполнил:

Студент группы 422–M1

\_\_\_\_Белоус Г.В.

Принял:

К.т.н., Доцент кафедры АОИ

\_\_\_\_Аксёнов С.В.

Оглавление

[1 Введение 3](#_Toc116226139)

[1.1 Цель лабораторной работы 3](#_Toc116226140)

[1.2 Задание на лабораторную работу 3](#_Toc116226141)

[2 Ход выполнения работы 5](#_Toc116226142)

[3 Вывод 8](#_Toc116226143)

# Введение

## 1.1 Цель лабораторной работы

Изучить инструменты PySpark и написать программу с помощью трёх алгоритмов, из которых нужно взять логистическую регрессию, случайный лес и дерево решений.

## **1.2 Задание на лабораторную работу**

Напишите программу на PySpark согласно вашему варианту для набора данных “DS\_2019\_public.csv” с помощью трёх алгоритмов, из которых нужно взять и логистическую регрессию, случайный лес и дерево решений. Напишите небольшой отчёт, который должен содержать титульный лист, вариант задания, код программы и полученные вами результаты. Для помощи в выполнении задания следует пользоваться книгой «Machine Learning with PySpark -2019.pdf» и «PySpark SQL Recipes».

Описание столбцов набора данных (расширенное) находится в файле recs2009\_public\_codebook.xlsx. В отчёте обязательно должно быть описание выбранных признаков, предсказываемых признаков/категорий, почему вы сделали такой выбор и выведена точность работы полученной модели. Если точность модели получилась невысокой, внесите изменения в выбор признаков, чтобы повысить точность, и отобразите это в отчёте.

Задача по полученному варианту.

Для этого необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета) и выполнить следующие корректировки:

Метка класса – Climate\_Region\_Pub. Так как необходимо создать бинарные классификаторы а возможных классов – 5, то сначала необходимо изменить значение метки Climate\_Region\_Pub.

Содержание отчета.

1. Титульный лист
2. Задача по полученному варианту.

(если последняя цифра 2 или 7: вариант 3)

1. Листинг программы
2. Примеры входных и выходных файлов
3. Описание результатов тестирования
4. Выводы

Выбрать признаки, использующиеся при обучении, и, если необходимо, выполнить их предобработку. Разделить выборку на обучающую и тестовую. Написать короткий отчет по работе, включив в него программу с комментариями, значения качества моделей TP, TN, FP, FN.

# Ход выполнения работы

Программа для обработки данных была написана на Python с использованием платформа Google Colaboratory для блокнотов Jupyter. Кроме среды запуска блокнотов Python и R Colab позволяет совместно использовать свободный доступ к ограниченному количеству GPU и TPU.

Для работы с данными были подключены необходимые фрейморки PySpark. Так же, в соответствии с вариантом, данные были изменены с помощью метода withColomn и вложенной в него функцией, которая делит колонку “Climate\_Region\_Pub”, в которой пять меток, на два класса A и B, метка под номером три входит в класс A (0), а остальные четыре метки в класс B (1).

В список выбранных признаков были включены все имеющиеся в датасете, кроме тех в работе с которыми метод transform для VectorAssembler выдал сообщение о невозможности использовать данную категорию.

Листинг программы можно увидеть в прикрепленном файле “PC\_Lab4.ipynb”.

Исходя из задания, для классификации по признаку “Climate\_Region\_Pub” было создано три модели работающих на трех алгоритмах, а именно: линейной регрессии, случайного леса и дерева решения. Созданные модели были обучены на обучающей выборке (75% от всего набора данных). А для оценки точности модели были созданы предсказания на основе тестовой выборки (25% от всего набора данных) и сравнены с оригинальными значениями, результаты работы моделей трех алгоритмов показаны на рисунках 2.1 – 2.3:

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 2.1 – Результаты модели на основе случайного леса

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 2.2 – Результаты модели на основе дерева решения

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 2.3 – Результаты модели на основе логистической регрессии

Точность предсказаний занесена в таблицу 2.1:

Таблица 2.1 – Точность предсказаний различных моделей

|  |  |  |
| --- | --- | --- |
| Алгоритм модели | Точность для класса A(0), % | Точность для класса B(1), % |
| Случайный лес | 70.4 | 99 |
| Дерево решений | 86.1 | 97.6 |
| Линейная регрессия | 85.8 | 97.4 |

# Вывод

В ходе выполнения лабораторной работы изучены инструменты PySpark и написана программа с помощью трёх алгоритмов, из которых взяты логистическая регрессия, случайный лес и дерево решений.

Полученный набор данных был разделен на обучающую и тестовую выборки, на них были обучены модели. Наибольшую точность для предсказания класса B показал алгоритм случайного леса, но для класса A точность случайного леса оказалась наименьшей и серьезно уступает двум другим алгоритмам, чья точность примерна на одном уровне. При этом случайный лес показал высокую точность в определении класса значения в случае с наибольшим классом, возможно случайному лесу недостаточно 521 записей класса A для обучения точной модели. В то же время двум остальным алгоритмам данного количества записей класса A в тестовую выборку хватило для показания точности большей на 16%.

На основе полученных результатов точности моделей классификации можно сделать вывод о том, что алгоритм случайного леса показывает высокую точность для больших объемов данных, но в то же время недостаток данных сильно влияет на точность предсказания.