Московский Авиационный Институт

(Национальный Исследовательский Университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

**Лабораторная работа №2 по курсу**

**«Машинное обучение»**

Студент: Кузьмичев Александр Николаевич

Группа: М80 – 306Б-18

Преподаватель: Ахмед Самир Халид

Оценка: \_\_\_\_\_\_\_\_\_\_\_

Дата: \_\_\_\_\_\_\_\_\_\_\_

Подпись: \_\_\_\_\_\_\_\_\_\_\_

Москва, 2021

**Задание**

Необходимо реализовать алгоритмы машинного обучения. Применить данные алгоритмы на наборы данных, подготовленных в первой лабораторной работе. Провести анализ полученных моделей, вычислить метрики классификатора. Произвести тюнинг параметров в случае необходимости. Сравнить полученные результаты с моделями реализованными в scikit-learn. Аналогично построить метрики классификации. Показать, что полученные модели не переобучились. Также необходимо сделать выводы о применимости данных моделей к вашей задаче. Задачи со звездочкой бьются по вариантам: N по списку % 2 + 1.

1. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ
2. \*SVM-ПЕРВЫЙ ВАРИАНТ
3. ДЕРЕВО РЕШЕНИЙ
4. \*RANDOMFOREST- ВТОРОЙ ВАРИАНТ

**Описание датасета**

Для данной лабораторной работы я использовал датасет в формате csv, содержащий информацию о 5172 электронных письмах. Отличительной особенностью датасета является большое число столбцов в таблице (3002). Датасет взят с сайта kaggle.com, он доступен по ссылке[Email Spam Classification Dataset CSV | Kaggle](https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv).

**Метод решения**

Самой тяжелой задачей для меня была реализация собственных модели логистической регрессии и дерева решений.

**Выводы**

Выполнив вторую лабораторную работу по машинному обучению, я сравнил собственную реализацию дерева решений с библиотечной. Библиотечная оказалась эффективнее. Данные обрабатывались очень долго. Также стоит уделить внимание подбору макропараметров.