|  |  |
| --- | --- |
|  | |
| МИНистерство Науки и высшего образования российской федерации | |
| Федеральное государственное бюджетное образовательное учреждениевысшего образования"МИРЭА - Российский технологический университет"РТУ МИРЭА | |
| Институт информационных технологий (ИТ) | |
| Кафедра прикладной математики (ПМ) | |

|  |  |
| --- | --- |
| **ОТЧЕТ**  **ПО ПРАКТИЧЕСКИМ РАБОТАМ** | |
| **по дисциплине** | |
| «Технологии и инструментарий анализа больших данных» | |
|  | |
| Выполнили студенты группы ИВБО-07-19 | Гридасов Е.А. |
| Принял ассистент кафедры ПМ | Горячев А.А. |

|  |  |  |
| --- | --- | --- |
| Практическая работа выполнена | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |
|  |  |  |
| «Зачтено» | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |

Москва 2022

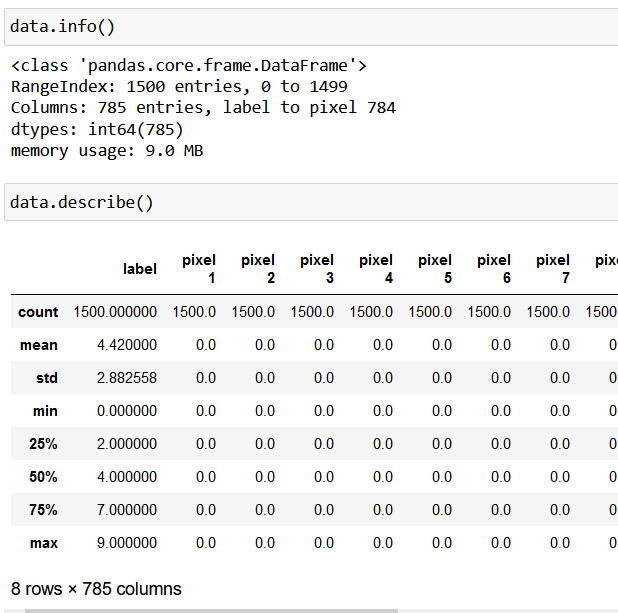
# ВЫПОЛНЕНИЕ ЗАДАЧ

Целью работы было изучение классификации на примере трёх различных моделей из библиотеки sklearn: LogisticRegression, SVC, KNeigborsClassifier.

Для выполнения задания был скачан модернизированный датасет MNIST.



**Рисунок 1 – Загрузка датасета**



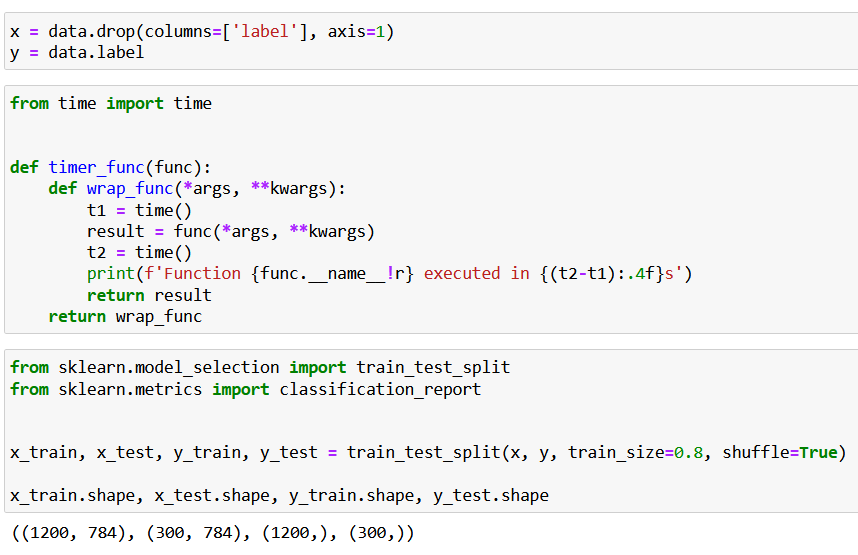
**Рисунок 2 — Проверка значений**



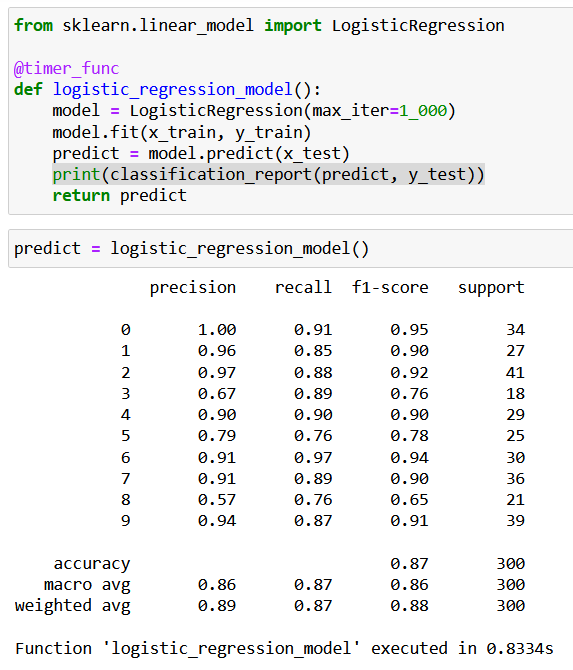
**Рисунок 3 — Сравнение количества данных по классам**

Видно, что количество данных разнятся, но не сильно. Максимум — на 44 процента.

После исследования датасета было произведено разбитие на обучающие и валидационные данные, а также обучение моделей тремя необходимыми способами.

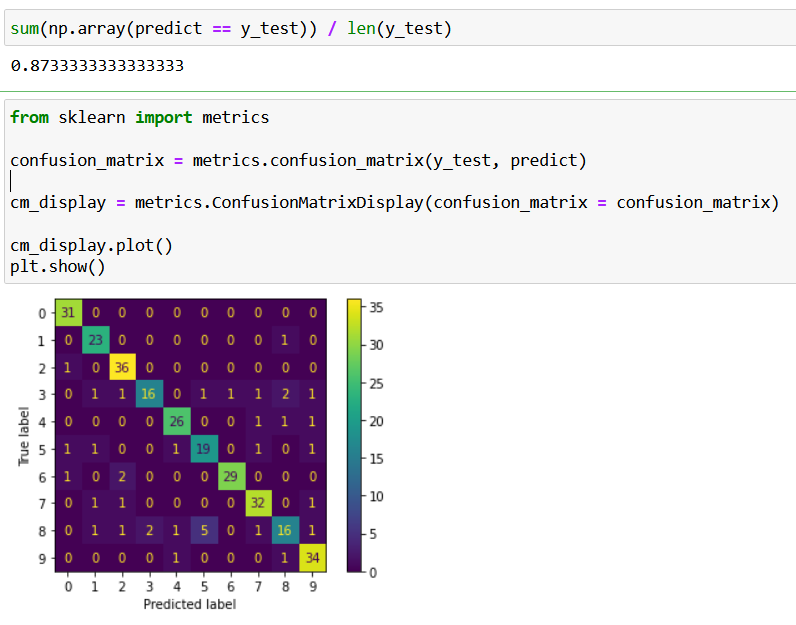


**Рисунок 4 — Разделение на данные, создание декоратора для сравнения времени**

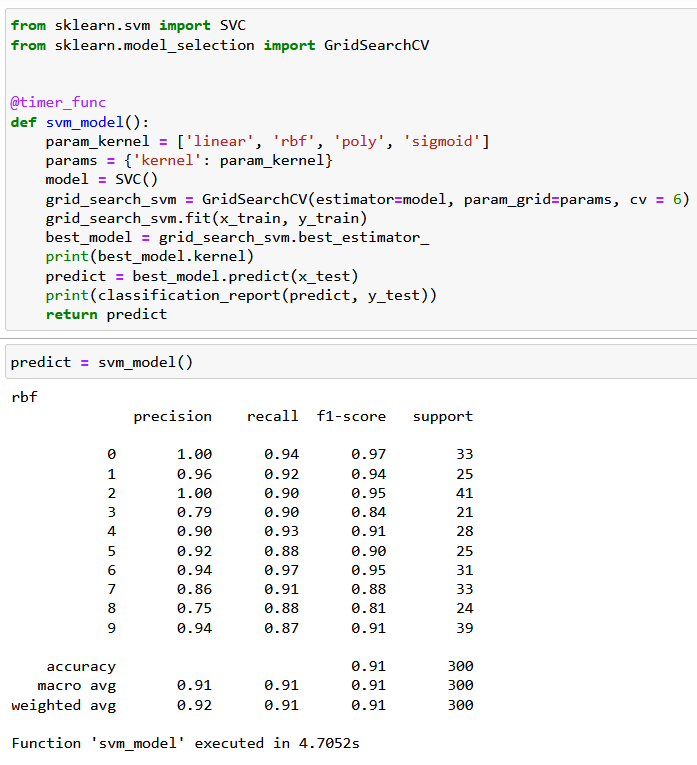


**Рисунок 5 — Обучение модели линейной регрессии**

Время обучения равно 0.83 секунды, точность составляет 87 процентов, полнота держится выше 75 процентов.



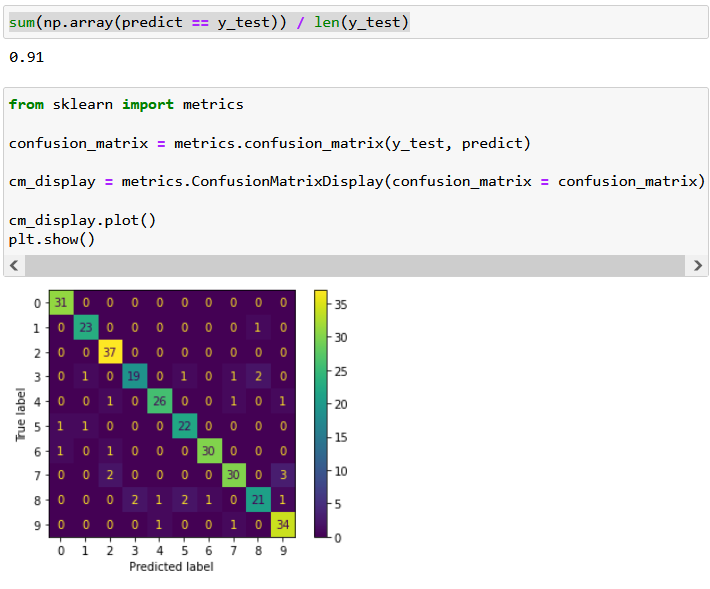
**Рисунок 6 — Вручную подсчитанная точность и матрица ошибок**



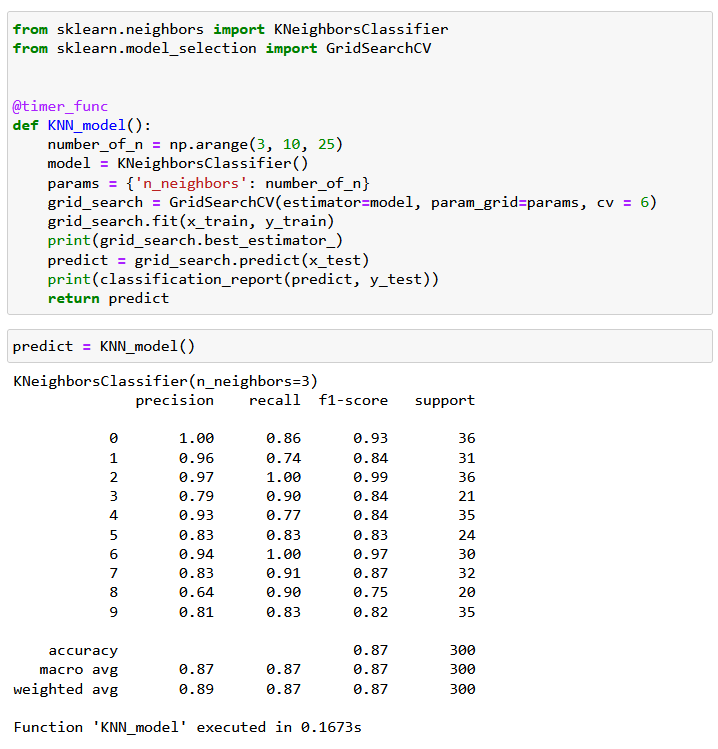
**Рисунок 7 — Модель с использованием поддерживающих векторов**

Время обучения равно 4.7 секунды, точность составляет 91 процент, а полнота держится на уровне выше 79 процентов.

SVM-модель обучается почти в 6 раз дольше, но при этом дает лучшие показатели по сравнению с линейной регрессией.



**Рисунок 8 — Вручную посчитанная точность и матрица ошибок**



**Рисунок 9 — KNN-модель**

Время 0.17 секунд, точность 87 процентов, полнота держится выше 64 процента.

Эта модель обучается быстрее, чем линейная регрессия, почти в 5 раз, при этом не уступая в точности и полноте. Подобный исход можно объяснить отсутствием вариативности в данных.

# ВЫВОДЫ

В ходе выполнения работы были изучены библиотеки matplotlib, seaborn, sklearn, stats. Освоена тема классификации данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Python Documentation [Электронный ресурс] - <https://www.python.org/>
2. Jupyter Documentation [Электронный ресурс] - <https://docs.jupyter.org/en/latest/index.html>
3. Pandas Documentation [Электронный ресурс] - <https://pandas.pydata.org/docs/getting_started/index.html>
4. Matplotlib Documentation [Электронный ресурс] - <https://matplotlib.org/>
5. Plotly Documentation [Электронный ресурс] - <https://plotly.com/python/>
6. Numpy Documentation [Электронный ресурс] - <https://numpy.org/doc/>
7. Stats Documentation [Электронный ресурс] - <https://docs.scipy.org/doc/scipy/reference/stats.html>