МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»

Кафедра телекоммуникационных систем и вычислительных средств (ТС и ВС)

Отчет
по дисциплине
«Системы искусственного интеллекта»

по теме:
ЛАБОРАТОРНАЯ РАБОТА 1. ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ

Студент:
Группа ИА331 *Р.К. Рубцов*

Предподаватель:
. *К.И. Брагин*

Новосибирск 2026 г.

СОДЕРЖАНИЕ

ЛАБОРАТОРНАЯ РАБОТА 1. ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ

0.1 Цель работы

Изучение программных средств для организации рабочего места специалиста по анализу данных и машинному обучению; выполнение первичного анализа набора данных об именах; получение обобщённых характеристик и визуализация структуры данных.

0.2 Описание набора данных

В качестве обучающего набора данных использован датасет `name_gender_dataset.csv`, содержащий 147 269 записей о популярных именах с указанием пола, количества носителей и вероятности встречаемости.

0.2.1 Назначение и возможные задачи

Набор предназначен для обучения методам анализа данных, в частности:

- анализа распределения имён по полу;
- изучения статистики популярности имён;
- визуализации демографических данных;
- выявления закономерностей в частоте использования имён.

0.2.2 Описание признаков

Таблица 1 — Характеристики признаков датасета

№	Признак	Описание и тип
1	`name`	Имя (строковый, до 20 символов)
2	`gender`	Пол: `M` (мужской) или `F` (женский) (категориальный)
3	`count`	Количество носителей имени (целочисленный)
4	`probability`	Вероятность встречаемости имени (вещественный)

0.3 Форма и статистика набора данных

- Количество объектов: 147 269
- Количество признаков: 4 (3 входных + 1 целевой)
- Распределение по полу: мужские имена — 57 520 (39.0%), женские имена — 89 749 (61.0%)

0.3.1 Статистические показатели

Таблица 2 — Основные статистики по числовым признакам

Признак	Среднее	Мин.	Макс.
`count`	$\approx$1 200	1	5 304 407
`probability`	$\approx$0.003	$<$0.0001	0.0145

0.3.2 Топ-5 самых популярных имён

Таблица 3 — Наиболее распространённые имена

№	Имя	Пол	Количество
1	James	M	5 304 407
2	John	M	5 260 831
3	Robert	M	4 970 386
4	Michael	M	4 579 950
5	William	M	4 226 608

0.3.3 Предположения на основе анализа

- Мужские имена в топ-15 доминируют по количеству носителей.
- В датасете представлено больше женских имён (89 749 против 57 520), но их общая популярность ниже.

0.4 Исходный код программы

Для выполнения первичного анализа и визуализации данных был написан следующий скрипт на языке Python:

```
import numpy as np
from matplotlib import pyplot as plt

dt = np.dtype([
    ('name', 'U20'),
    ('gender', 'U1'),
    ('count', 'i4'),
    ('probability', 'f8')
])

data = np.genfromtxt("name_gender_dataset.csv", delimiter=",",
                     dtype=dt, skip_header=1, encoding='utf-8')

print(f'Тип␣данных:␣{type(data)}')
print(f'Тип␣записи:␣{type(data[0])}')
print(f'Тип␣имени:␣{type(data[0][0])}')
```

```
print(f'Форма␣данных:␣{data.shape}')
print(f'\Первыеп␣5␣записей:')
print(data[:5])

male_mask = data['gender'] == 'M'
female_mask = data['gender'] == 'F'

male_data = data[male_mask]
female_data = data[female_mask]

print(f'\Мужскихп␣имён:␣{len(male_data)}')
print(f'Женских␣имён:␣{len(female_data)}')

sorted_by_count = np.sort(data, order='count')[::-1]

# График 1: Топ-15 имён
plt.figure(1)
top_15 = sorted_by_count[:15]
colors = ['#1f77b4' if g == 'M' else '#ff7f0e' for g in top_15['
   gender']]
plt.barh(range(15), top_15['count'], color=colors)
plt.yticks(range(15), top_15['name'])
plt.xlabel('Количество␣носителей')
plt.title('Топ-15␣самых␣популярных␣имён')
plt.gca().invert_yaxis()
plt.tight_layout()

# График 2: Распределениепополу
plt.figure(2)
male_total = male_data['count'].sum()
female_total = female_data['count'].sum()
plt.pie([male_total, female_total],
        labels=[f'Мужские␣({male_total:,})', f'Женские␣({
           female_total:,})'],
        colors=['#1f77b4', '#ff7f0e'])
plt.title('Доля␣имён␣по␣полу␣общее(␣количество␣носителей)')
plt.axis('equal')

# График 3: Зависимость Probability от Count
plt.figure(3)
plt.scatter(male_data['count'], male_data['probability'],
           c='#1f77b4', label='Мужские')
```

```
plt.scatter(female_data['count'], female_data['probability'],
            c='#ff7f0e', label='Женские')
plt.xlabel('Количество␣носителей␣(Count)')
plt.ylabel('Вероятность␣(Probability)')
plt.title('Зависимость␣Probability␣от␣Count')
plt.legend()
plt.grid()

# График 4: Топ-10 пополу
plt.figure(4)
male_sorted = np.sort(male_data, order='count')[::-1][:10]
plt.subplot(1, 2, 1)
plt.barh(range(10), male_sorted['count'], color='#1f77b4')
plt.yticks(range(10), male_sorted['name'])
plt.xlabel('Количество')
plt.title('Топ-10␣мужских␣имён')
plt.gca().invert_yaxis()
plt.grid()

female_sorted = np.sort(female_data, order='count')[::-1][:10]
plt.subplot(1, 2, 2)
plt.barh(range(10), female_sorted['count'], color='#ff7f0e')
plt.yticks(range(10), female_sorted['name'])
plt.xlabel('Количество')
plt.title('Топ-10␣женских␣имён')
plt.gca().invert_yaxis()
plt.grid()

plt.tight_layout()
plt.show()
```

Листинг 1 — Исходный код программы

0.5 Консольный вывод

При выполнении скрипта в терминале были получены следующие данные:

```
Типданных
 : <class 'numpy.ndarray'>Типзаписи
 : <class 'numpy.void'>Типимени
 : <class 'numpy.str_'>Формаданных
```

```
 : (147269,)Первые

 5 записей:
[('James', 'M', 5304407, 0.01451679) ('John', 'M', 5260831,
   0.01439753)
 ('Robert', 'M', 4970386, 0.01360266)
 ('Michael', 'M', 4579950, 0.01253414)
 ('William', 'M', 4226608, 0.01156713)]Мужскихимён

 : 57520Женскихимён
 : 89749
```

Листинг 2 — Результаты в терминале

А также графики (рис. 1–4).

0.6 Графические представления

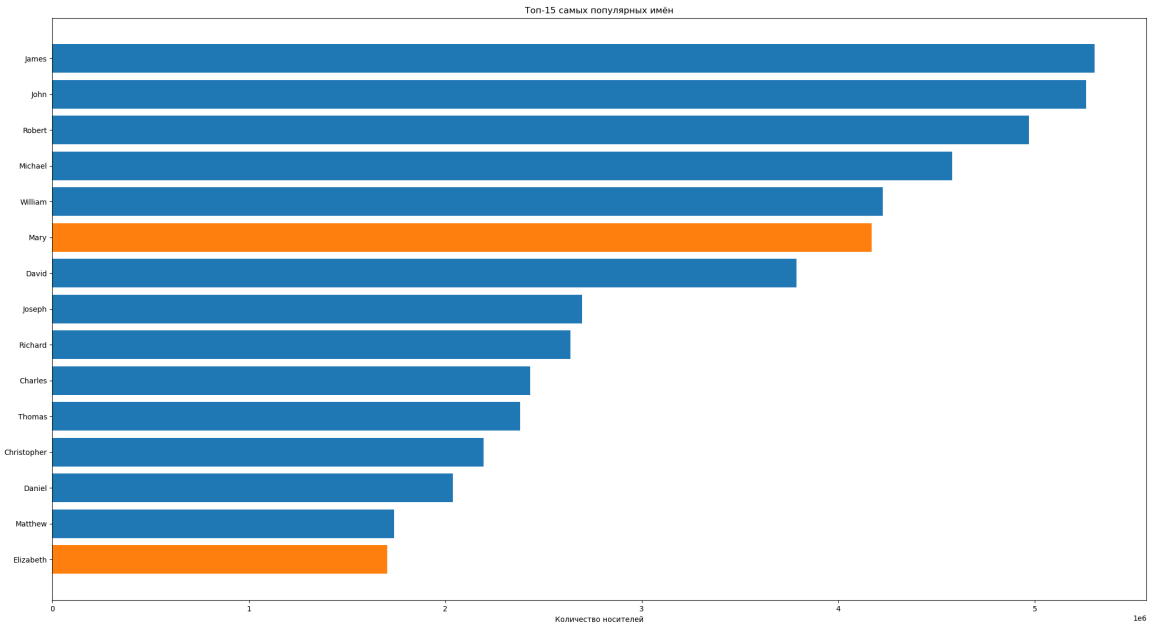

Рисунок 1 — Топ-15 самых популярных имён. Видно доминирование мужских имён (синие) в верхней части рейтинга. Женские имена (оранжевые) появляются реже, но также входят в топ (Mary, Elizabeth).

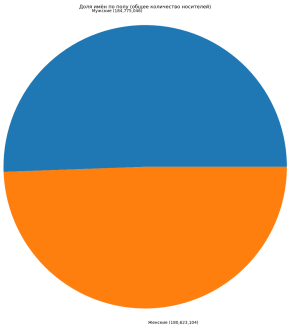

Рисунок 2 — Распределение общего количества носителей по полу. Несмотря на то, что женских имён в датасете больше (89 749 против 57 520), общее количество носителей мужских имён превышает количество носителей женских имён.

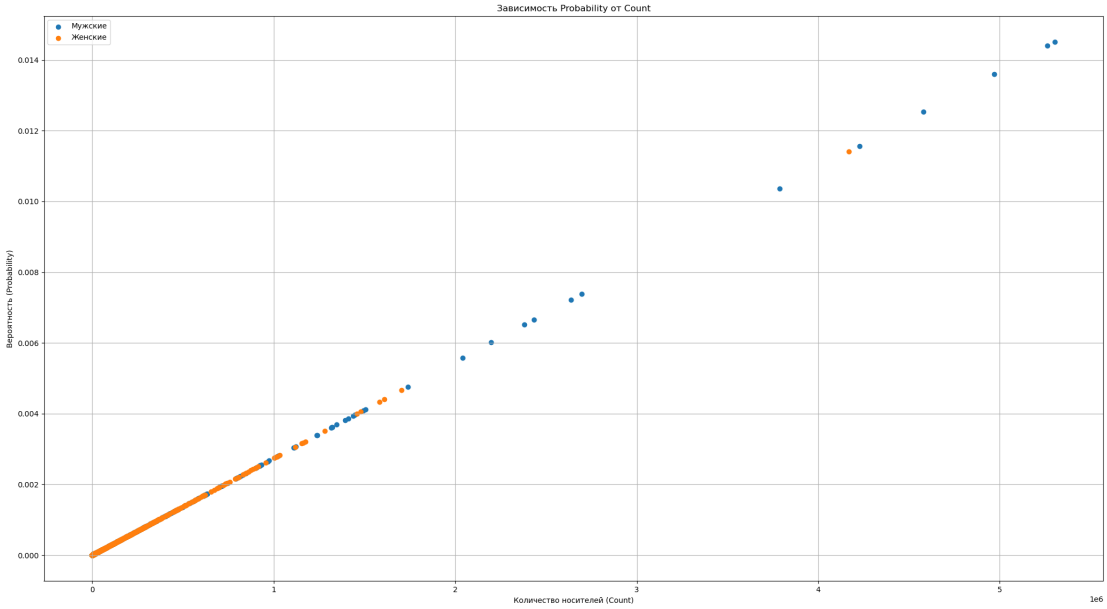

Рисунок 3 — Зависимость Probability от Count. Мужские имена (синие) занимают верхнюю часть графика, что подтверждает их большую популярность.

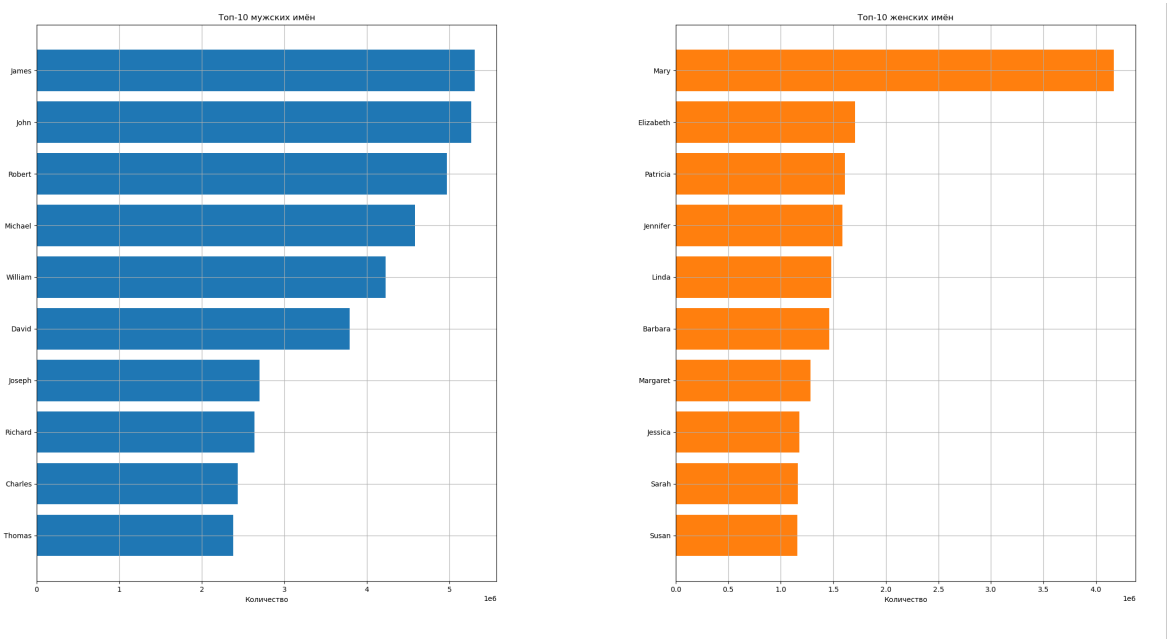

Рисунок 4 — Топ-10 мужских и женских имён. James, John и Robert лидируют среди мужских имён с показателями свыше 4-5 миллионов носителей. Среди женских имён Mary значительно опережает остальные с показателем около 4 миллионов.

0.7 Ответы на контрольные вопросы

1. **Инструментальные средства для Data Science:** Jupyter Notebook, Google Colab, VS Code с расширениями, PyCharm, Anaconda, Git, терминал Linux/WSL.
2. **Библиотеки Python для ML:**
   - `NumPy` — работа с многомерными массивами и математическими операциями.
   - `Pandas` — загрузка, обработка и анализ табличных данных.
   - `Matplotlib/Seaborn` — визуализация данных.
   - `Scikit-learn` — реализация алгоритмов ML (классификация, регрессия, кластеризация).
   - `SciPy` — научные вычисления и статистика.
3. **Почему Python популярен в ML?** Простой синтаксис, огромное сообщество, богатая экосистема библиотек, поддержка научных вычислений, удобство для прототипирования и обучения.

0.8 Выводы

Проведённый первичный анализ показал, что датасет `name_gender_dataset.csv` обладает следующими свойствами:

- большой объём данных (147 269 записей),
- неравномерное распределение по полу (больше женских имён, но мужские более популярны),
- явное доминирование нескольких имён (James, John, Robert, Mary),
- распределение подчиняется степенному закону (немного очень популярных имён, много редких).

Датасет подходит для задач:

- анализа популярности имён,
- демографических исследований,
- прогнозирования трендов именования,
- классификации имён по полу.

Студент группы ИА331
Рубцов Р.К.