Федеральное государственное автономное

образовательное учреждение высшего образования

«Пермский государственный национальный исследовательский университет» (ПГНИУ)

Региональный институт непрерывного образования (РИНО ПГНИУ)

Цифровая кафедра

Выпускная аттестационная (квалификационная) работа

по курсу профессиональной переподготовки

«Анализ данных и машинное обучение»

**Предсказание возраста, классификация Age detection**

Разработчики проекта:

Братчиков Андрей

Богоявленский Всеволод

Пермь, 2024

**ПАСПОРТ ПРОЕКТА**

**Название проекта:**

Предсказание возраста, классификация Age detection

**Сведения об авторах:**

Братчиков Андрей

Богоявленский Всеволод

**Цель и задачи проекта:**

**Цель:**

Разработка и обучение моделей для автоматической классификации возраста человека по фотографии, с целью оптимизации точности и скорости предсказания.

**Задачи:**

1. Проанализировать датасет, провести предварительную обработку: выявление пропусков, обработка выбросов и анализ сбалансированности классов.
2. Выбрать и протестировать три модели: одну из области классического машинного обучения, одну на основе глубоких нейросетей и одну гибридную модель.
3. Определить подходящую функцию потерь для многоклассовой классификации.
4. Подобрать и обосновать метрики для оценки качества работы моделей.
5. Провести обучение моделей, выбрать наилучшую по точности.
6. Подготовить презентацию с подробным анализом данных, выбранных моделей и полученных результатов.

**Краткое описание проекта:**

**Для выполнения задачи классификации возраста использовались следующие подходы:**

1. Глубокая сверточная нейросеть с несколькими слоями (CNN) для извлечения особенностей изображения.
2. Метод случайных лесов (Random Forest), как классический алгоритм машинного обучения для построения предсказаний на основе структурированных данных.
3. XGBoost — метод градиентного бустинга, применяющийся для классификации и известный высокой точностью при работе с разнообразными данными.
4. KNN (k-Nearest Neighbors) — алгоритм классификации, основанный на близости соседей и эффективный для задач с несложными зависимостями.

Оценка качества моделей проводилась по метрике accuracy, что позволяет легко интерпретировать результаты для данной задачи.

**Конкретные ожидаемые результаты:**

По итогам работы было достигнуто среднее значение точности около 70%. Этот результат обусловлен высокой вариативностью данных (например, изображения с различным качеством и освещением), а также несбалансированностью классов в наборе данных. Результаты могут быть улучшены путем дальнейшей предобработки данных и оптимизации моделей.