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)

Отчет по лабораторной работе №1

по дисциплине

«Глубинное обучение»

**«Задание №1 – Полносвязные сети прямого распространения**

Выполнил

Студент гр. 932209

Пан А.Э.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2023 г

Проверил

Доцент кафедры ТОИ

\_\_\_\_\_\_\_\_\_\_ Аксёнов С.В.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2023 г.

Томск 2023

# Введение

**Цель работы** – получить навыки создания полносвязного бинарного и многоклассового классификаторов, а также нейросетевого регрессора с помощью библиотеки построения нейронных сетей.

**Задачи** – Построить три нейросетевые модели:

a) Бинарный Классификатор

b) Многоклассовый классификатор

c) Регрессор

Варианты (выборки):

Для своего варианта необходимо посмотреть последнюю цифру номера своей зачетной книжки (или студенческого билета). Номер моей зачетной книжки (студенческого билета) 472689. Это значит, что у меня нечетный вариант.

**1 Описание набора данных**

a) Оценка вероятности диагностики диабета у человека <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset> Класс: Diabetes\_012. Класс отрицательный – 0 (no diabetes – нет диабета), класс положительный – 1 & 2 (prediabetes – преддиабетическое состояние & diabetes - диабет). diabetes\_012\_health\_indicators\_BRFSS2015.csv представляет собой чистый набор данных, содержащий 253 680 ответов на опрос CDC BRFSS2015. Целевая переменная Diabetes\_012 имеет 3 класса. 0 — отсутствие диабета или только во время беременности, 1 — преддиабет, 2 — диабет. В этом наборе данных наблюдается классовый дисбаланс. Этот набор данных содержит 21 переменную объекта.

b) Оценка уровня физического развития людей разного возраста: <https://www.kaggle.com/datasets/kukuroo3/body-performance-data> Метка класса: class. Это данные, подтверждающие уровень работоспособности в зависимости от возраста, а также некоторые данные о результатах упражнений. Столбцы в датасете:

* Возраст (age): 20 ~ 64
* Пол (sex): Ж,М
* Рост (height\_cm) : (Если вы хотите перевести в футы, разделите на 30,48)
* Вес (weigth) (кг)
* телесный жир (body fat)\_%
* диастолическое: диастолическое артериальное давление (мин) (diastolic : diastolic blood pressure)
* систолическое: систолическое артериальное давление (мин) (systolic : systolic blood pressure)
* сцеплениесила (gripForce)
* сесть и наклониться вперед\_см (sit and bend forward\_cm)
* кол-во приседаний (sit-ups counts)
* прыжок в длину\_см (broad jump\_cm)
* класс: A,B,C,D (A: лучший) / стратифицированный

c) Оценка расходов домохозяйства на один из видов потребляемых ресурсов. Набор данных DS\_2019\_public.csv. Описание атрибутов – файл recs2009\_public\_codebook.xlsx. Предсказываемый признак – один из атрибутов (использовать только один из): TOTALDOL, TOTALDOLSPH, TOTALDOLCOL, TOTALDOLWTH, TOTALDOLOTH.

**2 Признаки, которые были использованы для анализа**

|  |  |  |
| --- | --- | --- |
| Наборы данных | | |
| a) | b) | c) |
| Все | age | TOTALBTUWTH |
|  | gender | TOTALDOL |
|  | weight\_kg | TOTALBTU |
|  | body fat\_% | TOTALBTU |
|  | diastolic | TOTALDOLOTH |
|  | gripforce | DOLELOTH |
|  | sit and bend forward\_cm | KWH |
|  | sit-ups counts | BTUEL |
|  | broad jump\_cm | BTUELOTH |
|  |  | DOLELWTH |
|  |  | TOTALBTUOTH |

**3 Параметры архитектур и обучения нейронных сетей, использованные для обучения**

|  |  |  |  |
| --- | --- | --- | --- |
| Параметры | a) | b) | c) |
| Optimizer | Adam | Adam | Adam |
| Learning rate | 0.00001 | 0.001 | 0.001 |
| Has dropout? | No | 0.2 | No |
| Loss function | Binary crossentropy | Categorical crossentropy | Huber |
| Batch size | 64 | 64 | 32 |
| Epochs | 50 | 200 (stopped at 77) | 300 |

Архитектуры моделей:

|  |  |  |
| --- | --- | --- |
| A) | B) | C) |
|  |  |  |

**4 Графики обучения для архитектур нейронных сетей с лучшими характеристиками эффективности**

А)

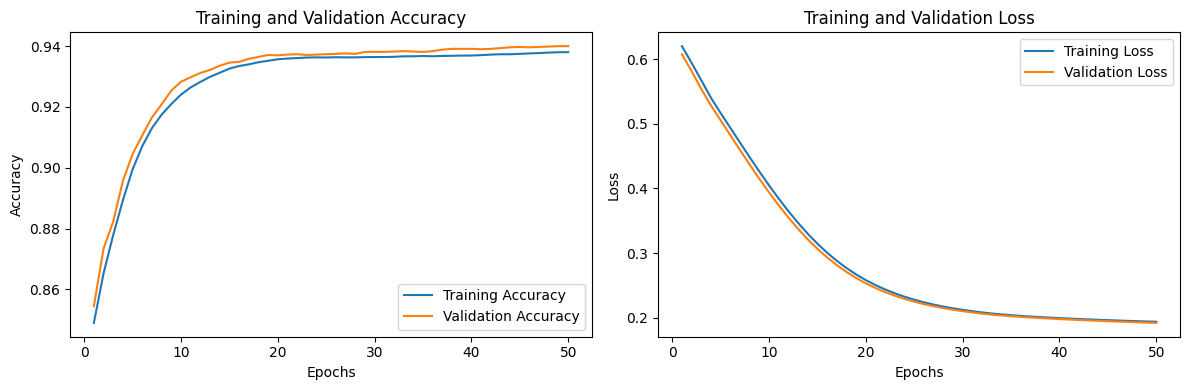
****

Рисунок 1 – График с характеристиками эффективности для задачи а)

B)

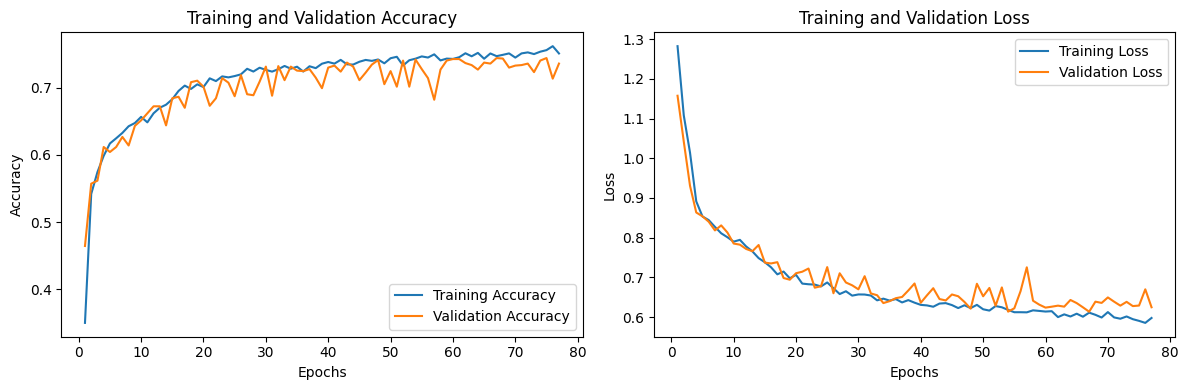


Рисунок 2 – График с характеристиками эффективности для задачи b)

C)

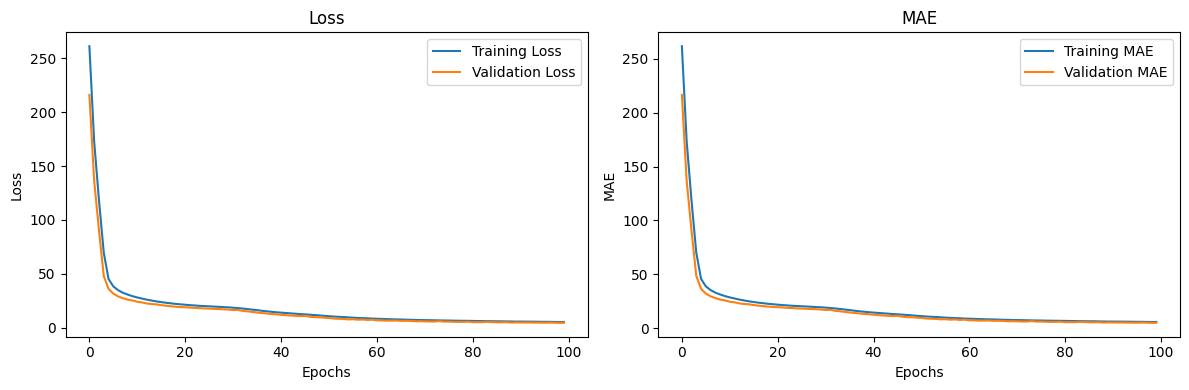


Рисунок 3 – График с характеристиками эффективности для задачи c)

**5 ROC-кривые классов для лучших классификаторов**

a)

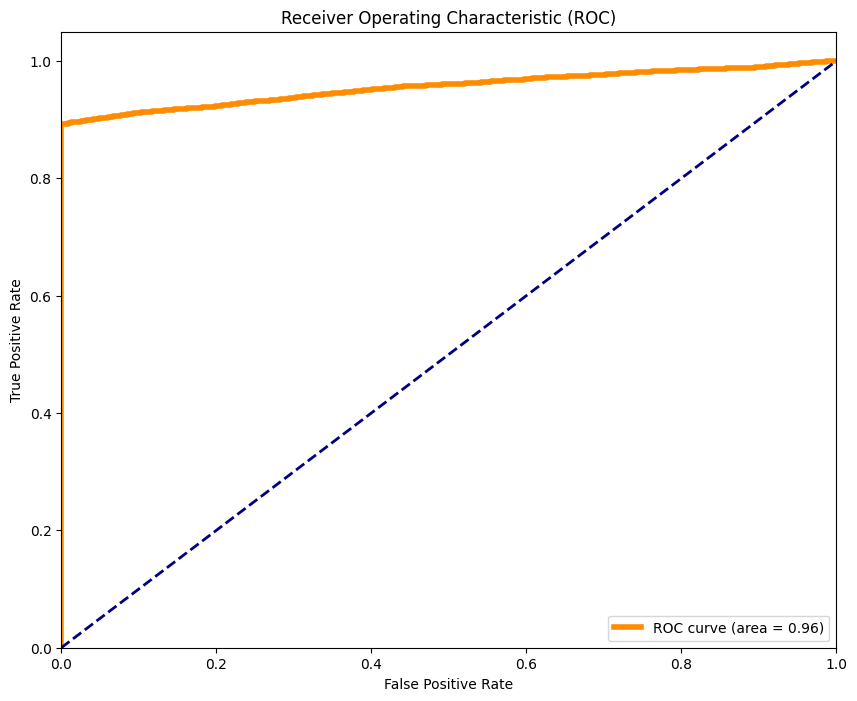


Рисунок 4 – ROC-кривая для задачи a)

b)

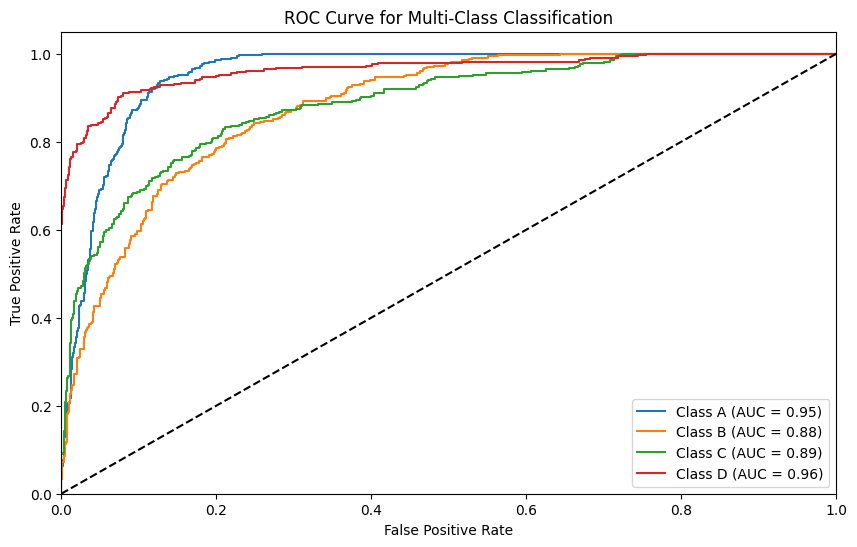
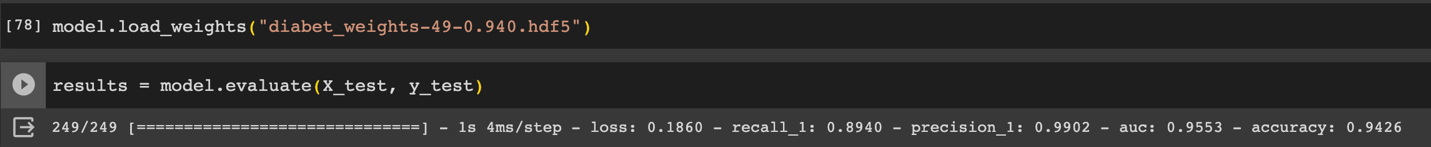
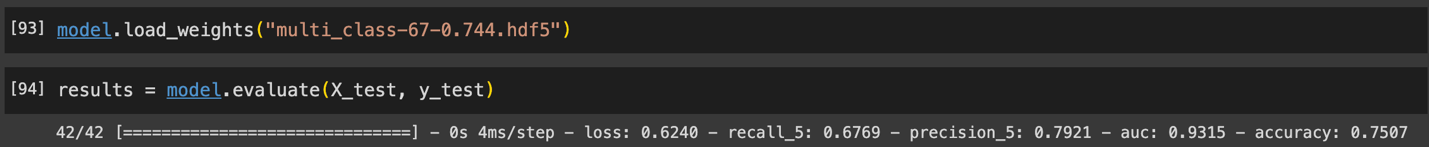


Рисунок 5 – ROC-кривая для задачи b)

**6 Оценки моделей на тестовых выборках в виде таблиц/ диаграмм, отображающих метрики качества**

a) 

b) 

c)

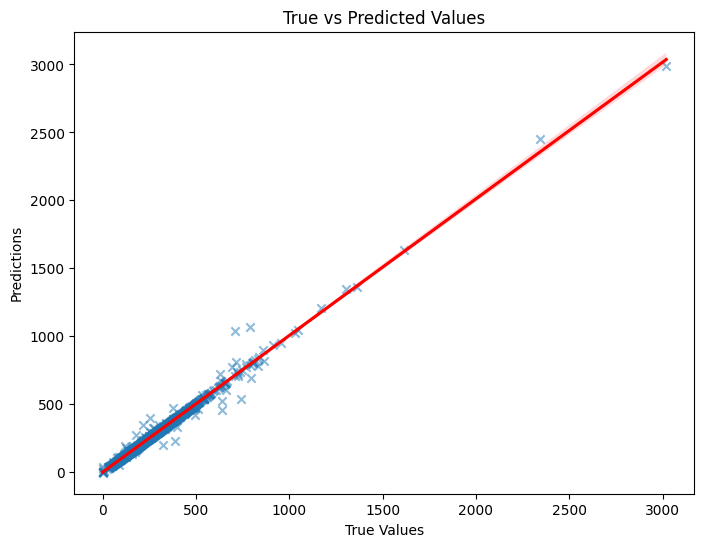
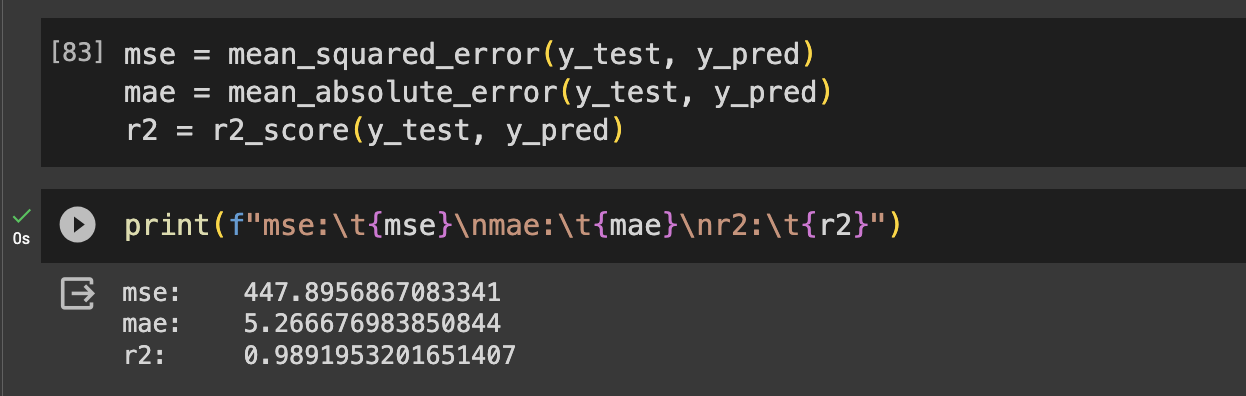


Рисунок 6 – Прямая линия регрессии для задачи c)

**7 Программный код**

Листинг программы представлен в отдельном файле laba\_1.ipynb. Ссылка на файл в среде Google colab:

<https://colab.research.google.com/drive/1JA57aiVYr-xlT8xUGkPzO4CiUTLbQOgd?usp=sharing>

**8 Выводы**

В ходе выполнения работы был получены знания использования полносвязные сети прямого распространения для задач бинарной и многоклассовой классификации, а также нейросетевой регрессии. Для построения нейросетей использовалась библиотека keras.

В задаче бинарной классификации изначально данные были несбалансированы, что приводило к неправильному решению поставленной задачи. Но применив субдискретизацию[[1]](#footnote-1) были получены очень хорошие результаты, которые изложены выше в отчете.

В задаче мультиклассовой классификации также не возникло проблем с распределением целевой переменной, однако ожидаемая точность (> 80%) не была достигнута. Но, с другой стороны, ROC-кривые классов демонстрируют нам очень хорошие результаты.

Задача нейросетевой регрессии решалась путем отбрасывания некоторого количества признаков. Мной было принято решение об использовании признаков, корреляция которых с целевой переменной больше или равна 0.4. Затем был проведен анализ на выбросы и в последствии было принято решение об использовании Huber[[2]](#footnote-2) в качестве функции потерь, потому что использование средней квадратической ошибки влияло на значение потерь (loss) на валидационной выборке: ~150 против ~5.

1. Cубдискретизация (англ. under-sampling) — удаление некоторого количества примеров мажоритарного класса. [↑](#footnote-ref-1)
2. Huber Loss является комбинацией MSE и MAE. Он менее чувствителен к выбросам, чем MSE, и менее чувствителен к малым ошибкам, чем MAE. [↑](#footnote-ref-2)