**1. Постановка задачи**

<https://archive.ics.uci.edu/ml/datasets/QSAR+aquatic+toxicity>  
Информация о структуре молекулы вещества и его токсичности для определенного вида рыб.

Структура:  
tpsa - Площадь поверхности молекулы, с атомами, которые могут соединяться с фосфором и серой  
saacc - Площадь поверхности молекулы, с атомами, которые могут соединяться с водородом  
h-050 - Количество атомов, которые могут соединяться с водородом  
mlogp - Коэффициент липофильноси (стремления модекулы соединяться с ограническими веществами)  
rdchi - Топологический коэффициент (определяет форму молекулы)  
gats1p - Коэффициент поляризуемости молекулы  
nn - Количество атомов азота  
c-040 - Количество атомов углерода в определенных связях  
lc50 - Концентрация, при которой погибает более 50% рыб

Задача: вычисление значения параметра lc50 на основании остальных параметров.

**2. Подход к обработке данных.**

К столбцам был применен квадратный корень так как он давал лучшее распределение, чем логарифм.

В столбцах «tpsa», «saacc», «h-050б», «nn», «c-040» нули были заменены на среднее значение.

К столбцам «mlogp», «rdchi», «gats1p», «lc50» применена функция StandardScaler(), так как их графики ближе к нормальному распределению.

Во всех столбцах наблюдались выбросы значений, поэтому значения в каждом из столбцов были срезаны до определённого интервала.

В конце обработки данных была применена функция dataset[i] = dataset[i].fillna(dataset[i].mean()), для того чтобы исправить ошибку «not a number» при обучении нейронной сети.

**3.Базовые параметры нейронной сети.**

функция активации - ReLU

Количество нейронов входного слоя – 8

Первый скрытый слой – 6

Второй скрытый слой – 6

Выходной слой – 1

Количество эпох – 50

Батчей – 8

Выборка 20%

**4. Описание проведенных экспериментов.**

Для всех моделей были одинаковые параметры, изменялось только соотношении размеров обучающей и тестовой выборок.

* Выборка 20%

В первой модели графики «среднего абсолютного отклонения от номера эпохи обучения» и «динамика среднеквадратического отклонения, т.е. значения функции потерь, от номера эпохи обучения» почти совпали

* Выборка 30%

В второй модели графики «среднего абсолютного отклонения от номера эпохи обучения» и «динамика среднеквадратического отклонения, т.е. значения функции потерь, от номера эпохи обучения» стали отличатся больше.

Остальные графики стали чуть более точнее.

* Выборка 10%

В 3 модели графики «среднего абсолютного отклонения от номера эпохи обучения» и «динамика среднеквадратического отклонения, т.е. значения функции потерь, от номера эпохи обучения» лучше, чем во 2-ой.

Остальные графики стали чуть более точнее.

**5. Итоговый вывод.**

Выборка данных нужна для обучения модели на части данных. Выборку обычно делают 20 тестовых на 80 обучающих или 30 тестовых на 70 обучающих. Обычно выборка происходит случайным образом из этого следует что, если изначальных данных слишком много, то нейронная сеть будет переобучена, то есть она будет слишком узконаправленна, а если данных будет слишком мало, то конечная сеть будет слишком свободна и может не соответствовать поставленным требованиям.