**НИТУ «МИСиС»**

Квалификация (степень): **магистр**

Курс: **1**

Семестр: **1**

Дисциплина: Нейронные сети

**ОТЧЕТ**

**ПО ЛАБОРАТОРНОЙ РАБОТЕ**

на тему:

**«Линейный классификатор и однослойный перцептрон»**

**Выполнил:**

Группа: МПИ-20-4-2

Студент: Новицкий Д. А.

**Проверил:** доц., к.т.н. Курочкин И. И.

Оценка:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Москва 2020**

**Оглавление**

[Постановка задачи 3](#_Toc60764191)

[Основное условие 3](#_Toc60764192)

[Требования к результатам 3](#_Toc60764193)

[Требования к визуализации (для 2-мерных данных) 3](#_Toc60764194)

[Требования к демонстрации работоспособности лабораторной работы 4](#_Toc60764195)

[Требования к входным/выходным данным 4](#_Toc60764196)

[Требования к загрузке файлов 4](#_Toc60764197)

[Описание работы программы 5](#_Toc60764198)

[Используемые датасеты 5](#_Toc60764199)

[Students performance in exams dataset 5](#_Toc60764200)

[Chess game dataset 6](#_Toc60764201)

[Red wine quality dataset 7](#_Toc60764202)

[Iris flower dataset 8](#_Toc60764203)

[Points dataset (самостоятельно сгенерированный датасет) 8](#_Toc60764204)

[Описание работы программы 10](#_Toc60764205)

[Результаты работы программы 11](#_Toc60764206)

[Список используемой литературы 19](#_Toc60764207)

# Постановка задачи

## Основное условие

1. Выбрать 2 датасета с линейно неразделимыми данными с количеством классов более 4 (2 признака и более 4). Для каждого класса количество точек не менее 50. Хотя бы 1 датасет должен быть уникальным во всей группе.

2. Реализовать линейный классификатор.

3. Реализовать однослойный перцептрон

4. Сравнить результаты работы двух методов классификации на одних и тех же данных (возможность работы для данных с различным количеством признаков)..

5. Реализовать визуализацию работы методов для 2-мерных данных. (Различными цветами на плоскости обозначаются точки истинных классов и заливкой различных цветов обозначаются области классов, определённых с помощью линейного классификатора и однослойного перцептрона).

## Требования к результатам

1. Параметры методов (к примеру, количество нейронов в скрытом слое, функции активации, начальные значения).

2. Метрики качества Accuracy, Precision, Recall, F1-measure для каждой пары {датасет, метод}.

3. Сравнительная таблица результатов по каждому методу + характеристики используемого датасета (количество точек, размерность, соотношение обучающего и тестового множеств и др.).

## Требования к визуализации (для 2-мерных данных)

1. Точки на плоскости с исходной классификацией (различные классы разными цветами) (отдельно обучающее, отдельно тестовое множество).

2. Результат классификации точек на плоскости (разноцветные области) по каждому методу.

## Требования к демонстрации работоспособности лабораторной работы

1. Продемонстрировать и объяснить отличия одного метода от другого (на результатах).

2. Определить лучший метод для каждого датасета.

## Требования к входным/выходным данным

1. Входные данные (датасеты) в виде текстового файла (к примеру, экспорт таблицы из Excel в формате TXT или CSV).

2. Выходные данные (результаты) сохраняются в виде HTML или DOCX отчёта (текст и графики).

## Требования к загрузке файлов

Необходимо загрузить следующие файлы:

1. Работающее приложение

2. Датасеты

3. Текстовый файл со сценариями запуска (к примеру: myapp.exe dataset1 results)

4. Отчет по ЛР с приведенными результатами и пояснениями (к примеру, если Вы получили какой-то график, приведите ниже сценарий и параметры запуска вычислительного приложения).

# Описание работы программы

## Используемые датасеты

В данной лабораторной работе используются различные наборы датасетов:

* Students performance in exams dataset [1];
* Chess game dataset [2]
* Red wine quality dataset [3];
* Iris flower dataset [4];
* Points dataset (самостоятельно сгенерированный датасет).

Рассмотрим подробнее каждый из представленных датасетов.

### Students performance in exams dataset

Этот набор данных состоит из оценок, полученных студентами по различным предметам.

Всего в данном датасете присутствует:

* 8 столбцов
* 1000 строк
* 5 первых столбцов – исходные данные (признаки)
* 3 последних столбца – выходные данные (классы)

Общая информация по столбцам представлена в таблице 1.

|  |  |  |
| --- | --- | --- |
| № | Имя столбца | Содержащиеся значения |
| 1 | Пол | female (52%) / male (48%) |
| 2 | Раса/этническое происхождение | group C (32%) / group D (26%) / Other (42%) |
| 3 | Уровень образования родителей | some college (23%) / associate’s degree (22%) /  other(55%) |
| 4 | Ланч | standard (65%) / free/reduced (36%) |
| 5 | Курс подготовки к экзаменам | none (64%) / completed (36%) |
| 6 | Оценка по математике | 0-100 |
| 7 | Оценка за чтение | 17-100 |
| 8 | Оценка за письменную работу | 10-100 |

Таблица 1. Информация по датасету «Students performance in exams»

### Chess game dataset

Это набор из чуть более 20 000 игр, собранных от избранных пользователей на сайте Lichess.org.

В одной шахматной партии содержится много информации, не говоря уже о полном наборе данных из нескольких партий. Это в первую очередь игра в шаблоны, а наука о данных - это все, чтобы выявлять закономерности в данных, поэтому в прошлом шахматы были одной из самых популярных областей ИИ. Этот набор данных собирает всю информацию, доступную по 20 000 игр, и представляет ее в формате, который легко обрабатывать для анализа, например, что позволяет игроку выиграть черным или белым цветом, сколько мета (вне игры) факторы влияют на игру, соотношение дебютов и победы черных и белых и многое другое.

Всего в данном датасете присутствует:

* 16 столбцов
* 20058 строк
* 7-ой (по очереди, начиная с первого) столбец – выходные данные (классы)
* Остальные столбцы – входные данные (признаки)

Общая информация по столбцам представлена в таблице 2.

|  |  |  |
| --- | --- | --- |
| № | Имя столбца | Содержащиеся значения |
| 1 | Номер игры | Уникальные значения |
| 2 | Оценённость | true (81%) / false (19%) |
| 3 | Время начала партии | Уникальные значения |
| 4 | Время конца партии | Уникальные значения |
| 5 | Количество ходов в партии | 1-349 |
| 6 | Статус игры | resign (56%) / mate (32%) / other (13%) |
| 7 | Победитель | white (50%) / black (45%) / other (5%) |
| 8 | Приращение времени | 10+0 (38%) / 15+0 (7%) / other(55%) |
| 9 | ID игрока белыми фигурами | Уникальные значения |
| 10 | Рейтинг игрока белыми фигурами | 784-2700 |
| 11 | ID игрока чёрными фигурами | Уникальные значения |
| 12 | Рейтинг игрока чёрными фигурами | 789-2723 |
| 13 | Все ходы в стандартной шахматной записи | Уникальные значения |
| 14 | Эко открытия | А00 (5%) / С00 (4%) /  other (91%) |
| 15 | Открытое имя/имя позиции | Van't Kruijs Opening(2%) / Sicilian Defense (2%) /  Other (96%) |
| 16 | Начальный слой (количество ходов в начальной фазе) | 1-28 |

Таблица 2. Информация по датасету «Chess game»

### Red wine quality dataset

Этот набор данных относятся к красному варианту португальского вина "Винью Верде". Из-за проблем с конфиденциальностью и логистики доступны только физико-химические (исходные) и сенсорные (выходные) переменные (например, нет данных о сортах винограда, марке вина, цене продажи вина и т. д.).

Эти наборы данных можно рассматривать как задачи классификации или регрессии. Классы упорядочены и не сбалансированы (например, нормальных вин гораздо больше, чем отличных или плохих).

Всего в данном датасете присутствует:

* 12 столбцов
* 1599 строк
* 11 первых столбцов – исходные данные (признаки)
* 1 последний столбец – выходные данные (классы)

Общая информация по столбцам представлена в таблице 3.

|  |  |  |
| --- | --- | --- |
| № | Имя столбца | Содержащиеся значения |
| 1 | Фиксированная кислотность | 4,6-15,9 |
| 2 | Летучая кислотность | 0,12-1,58 |
| 3 | Лимонная кислота | 0-1 |
| 4 | Остаточный сахар | 0,9-15,5 |
| 5 | Хлориды | 0,01-0,61 |
| 6 | Свободный диоксид серы | 1-72 |
| 7 | Диоксид серы общий | 6-289 |
| 8 | Плотность | 0,99-1 |
| 9 | PH | 2,74-4,01 |
| 10 | Сульфаты | 0,33-2 |
| 11 | Алкоголь | 8,4-14,9 |
| 12 | Качество (оценка от 0 до 10) | 3-8 |

Таблица 3. Информация по датасету «Red wine quality»

### Iris flower dataset

Набор данных о цветках ириса – это многомерный набор данных, представленный британским статистиком и биологом Рональдом Фишером в его статье 1936 года «Использование множественных измерений в таксономических задачах». Его иногда называют набором данных ириса Андерсона, потому что Эдгар Андерсон собрал данные для количественной оценки морфологической изменчивости цветков ириса трех родственных видов. Набор данных состоит из 50 образцов каждого из трех видов ириса (Iris Setosa, Iris virginica и Iris versicolor). Для каждого образца были измерены четыре характеристики: длина и ширина чашелистиков и лепестков в сантиметрах.

Этот набор данных стал типичным тестовым примером для многих методов статистической классификации в машинном обучении, таких как машины опорных векторов.

Набор данных содержит набор из 150 записей по 5 атрибутам – длина лепестка, ширина лепестка, длина чашелистиков, ширина чашелистиков и класс (виды).

### Points dataset (самостоятельно сгенерированный датасет)

Данный датасет содержит самостоятельно сгенерированный набор точек, принадлежащими к определённому классу. Данный датасет содержит 3 файла:

* «points\_linear.csv»
* «points\_10\_20.csv»
* «points\_50\_70.csv»

В первом случае представлен набор линейно разделимых классов, во втором случае представлен набор классов с пересечением 10-20%, а в третьем случае представлен набор классов с пересечением 50-70%.

Алгоритм генерации точек представлен в методе «Generate\_points» файла «Lab\_1.py».

Визуализация сгенерированных данных представлена на рисунках 1-3.

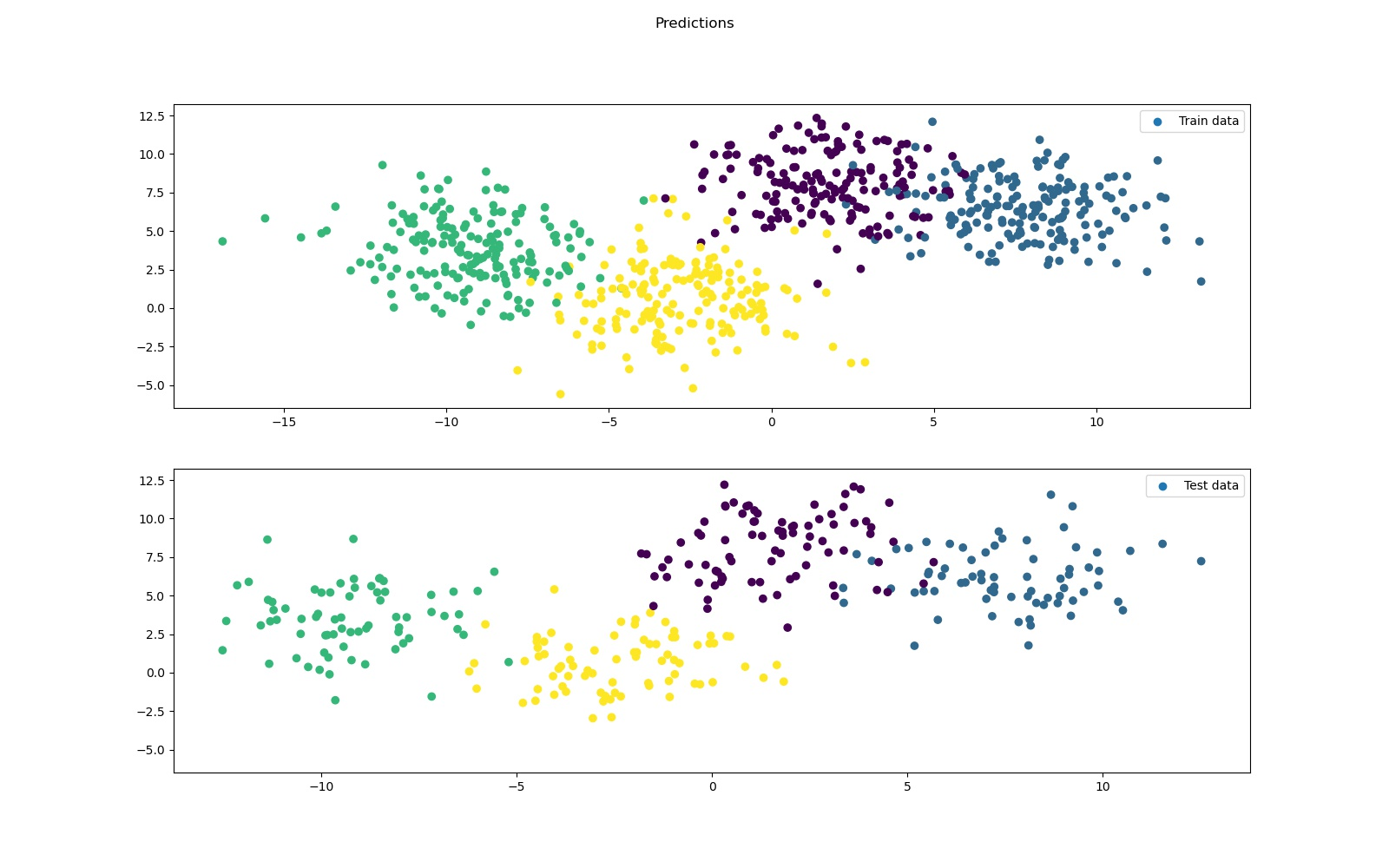


Рис. 1. Визуализация сгенерированного датасета «points\_10\_20»\*

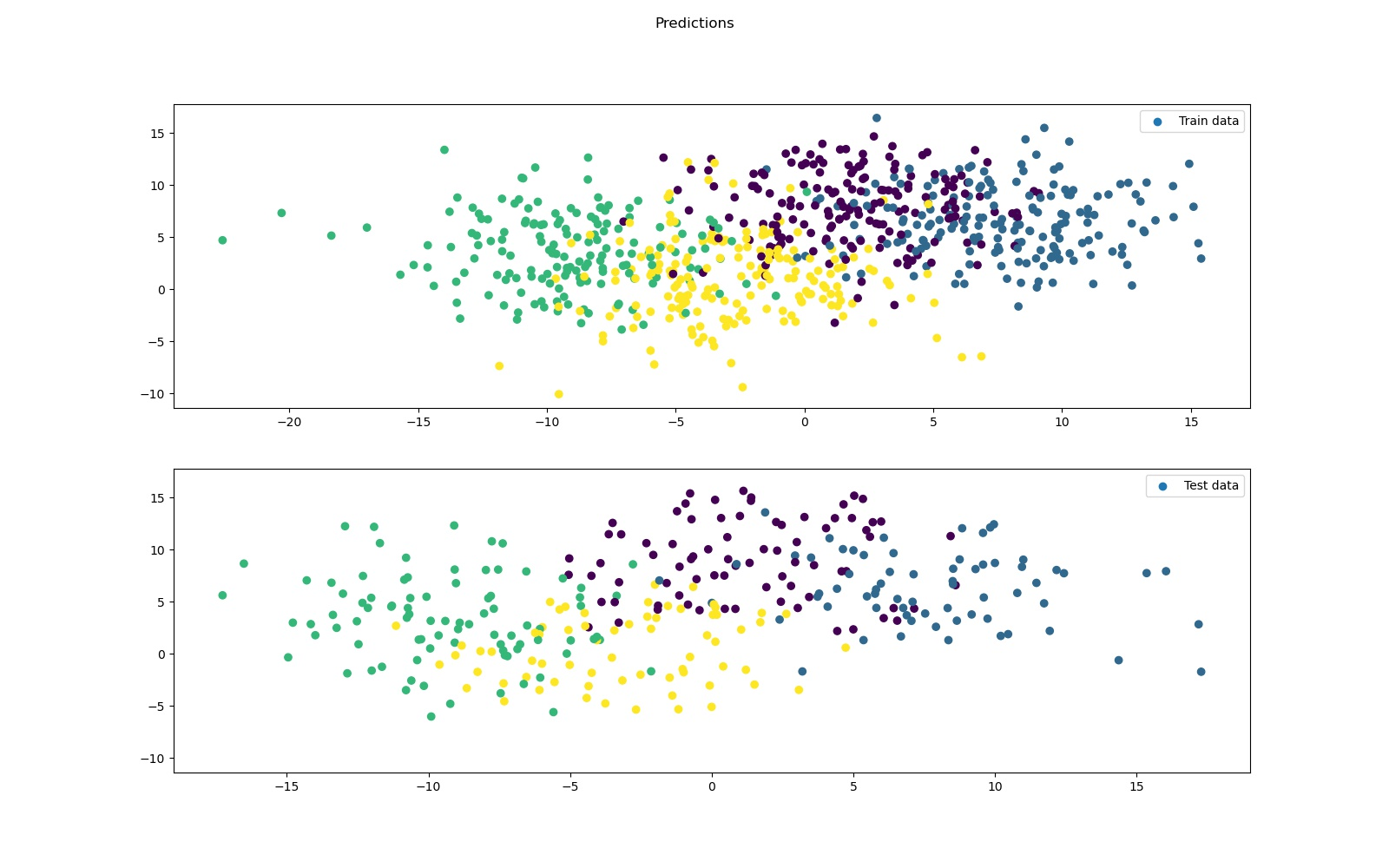


Рис. 2. Визуализация сгенерированного датасета «points\_50\_70»

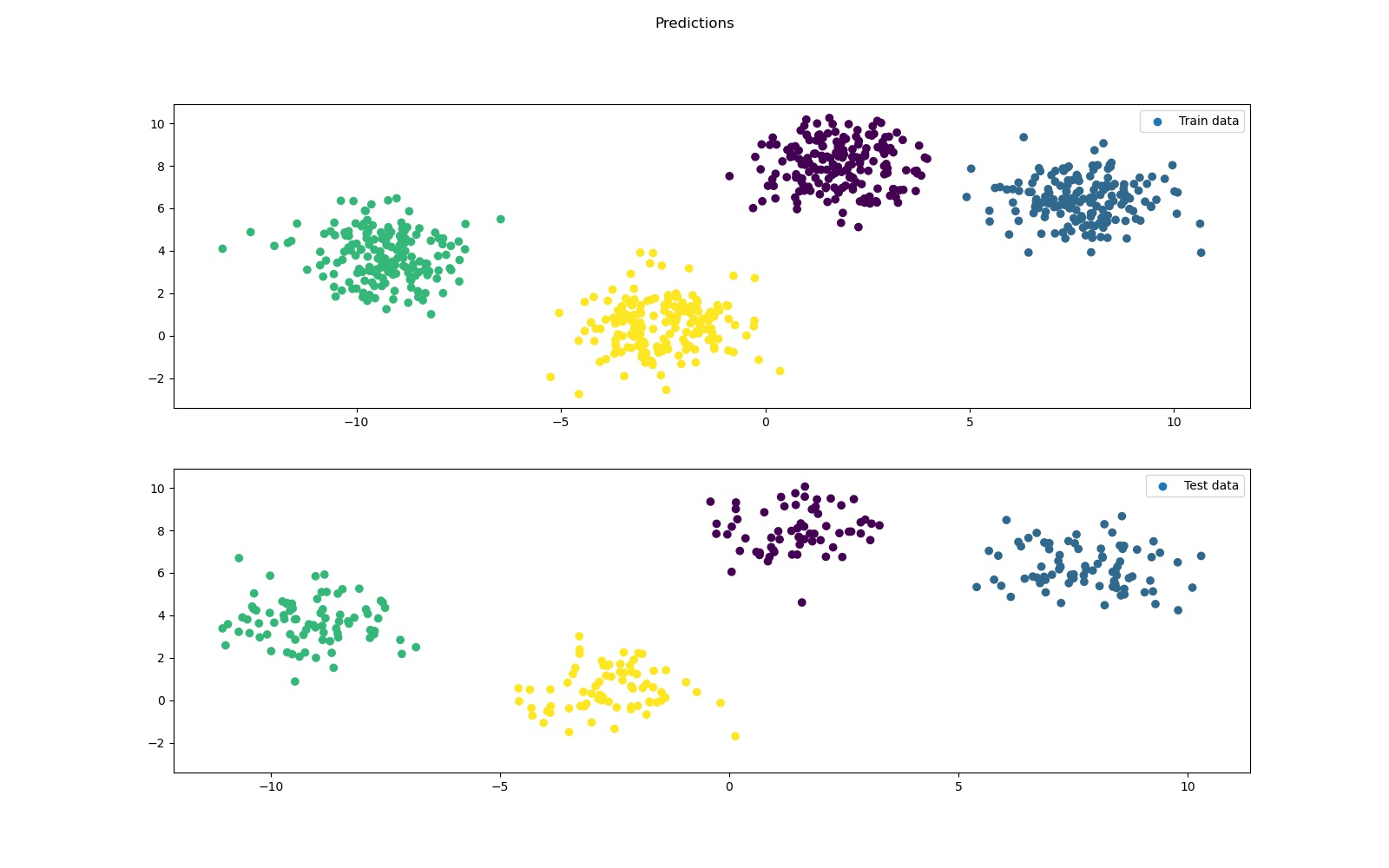


Рис. 3. Визуализация сгенерированного датасета «points\_linear»

\* Сверху представлены данные для обучения (train data)

Снизу представлены данные для тестирования (test data)

Коэффициент разделения данных – 0,7 (70% данных предназначены для обучения, 30% данных для тестирования).

# Описание работы программы

Работа программы состоит из следующих последовательных этапов:

1. Генерация датасета «Points\_dataset» (если датасет уже был ранее сгенерирован, то используется уже сгенерированный датасет).
2. Загрузка и предобработка датасетов, а именно:

2.1) Разделение входных данных на признаки и классы

2.2) Замена строковых значений на категориальные

2.3) Перемешивание данных

3) Процесс обучения для каждого датасета, а именно:

3.1) Разделение датасета на обучающую и тестовую выборки

3.2) Обучение линейного классификатора на заданных данных

3.3) Обучение однослойного персептрона на заданных данных

3.4) Обучение многослойного персептрона на заданных данных

3.5) Отрисовка и запись результатов классификации в jpeg файлы (только для двумерных данных)

3.6) Получение метрик по всем видам обучения и их запись в отдельный массив

3.6) Отрисовка и запись результатов классификации по всем видам обучения в jpeg файлы (только для двумерных данных)

4) Запись полученного массива метрик по всем датасетам в xlsx файл.

# Результаты работы программы

Результаты работы программы представлены на рисунках 4-15 и в таблице 4. Параметры обучения по каждому датасету для каждого из типов обучения одинаковы.

Для логистической регрессии «LogisticRegression» (линейного классификатора) входящие параметры устанавливаются по умолчанию.

Для однослойного персептрона «Perceptron» входящие параметры следующие:

* tol = 1e-3;
* random\_state = 0;

остальные параметры устанавливаются по умолчанию.

Для многослойного персептрона «MLP» входящие параметры следующие:

* hidden\_layer\_sizes = (4);
* max\_iter = 1000;

остальные параметры устанавливаются по умолчанию.



Рис. 4. Результаты обучения для датасета «points\_10\_20»\*

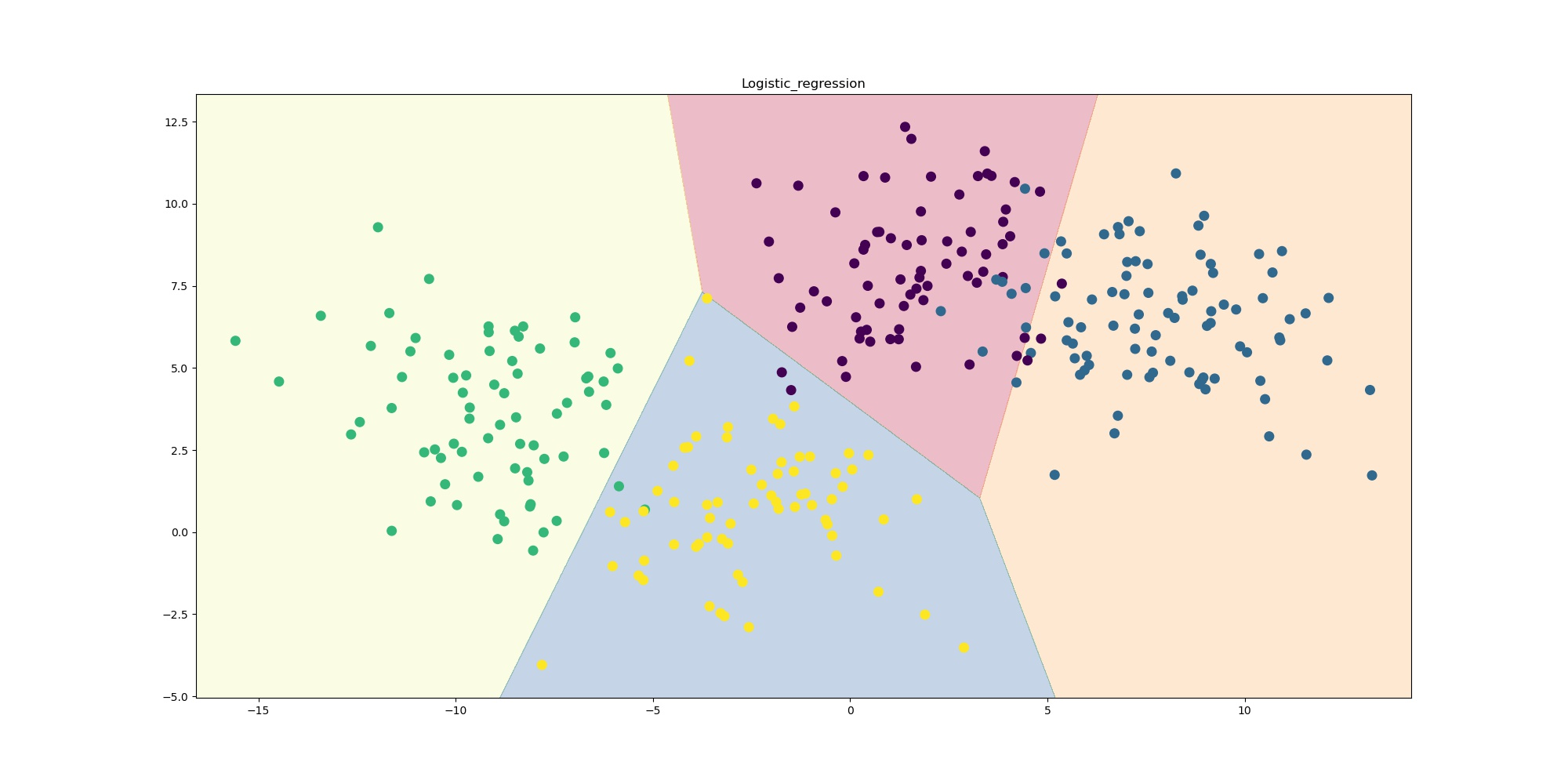


Рис. 5. Результаты работы линейного классификатора на датасете «points\_10\_20»

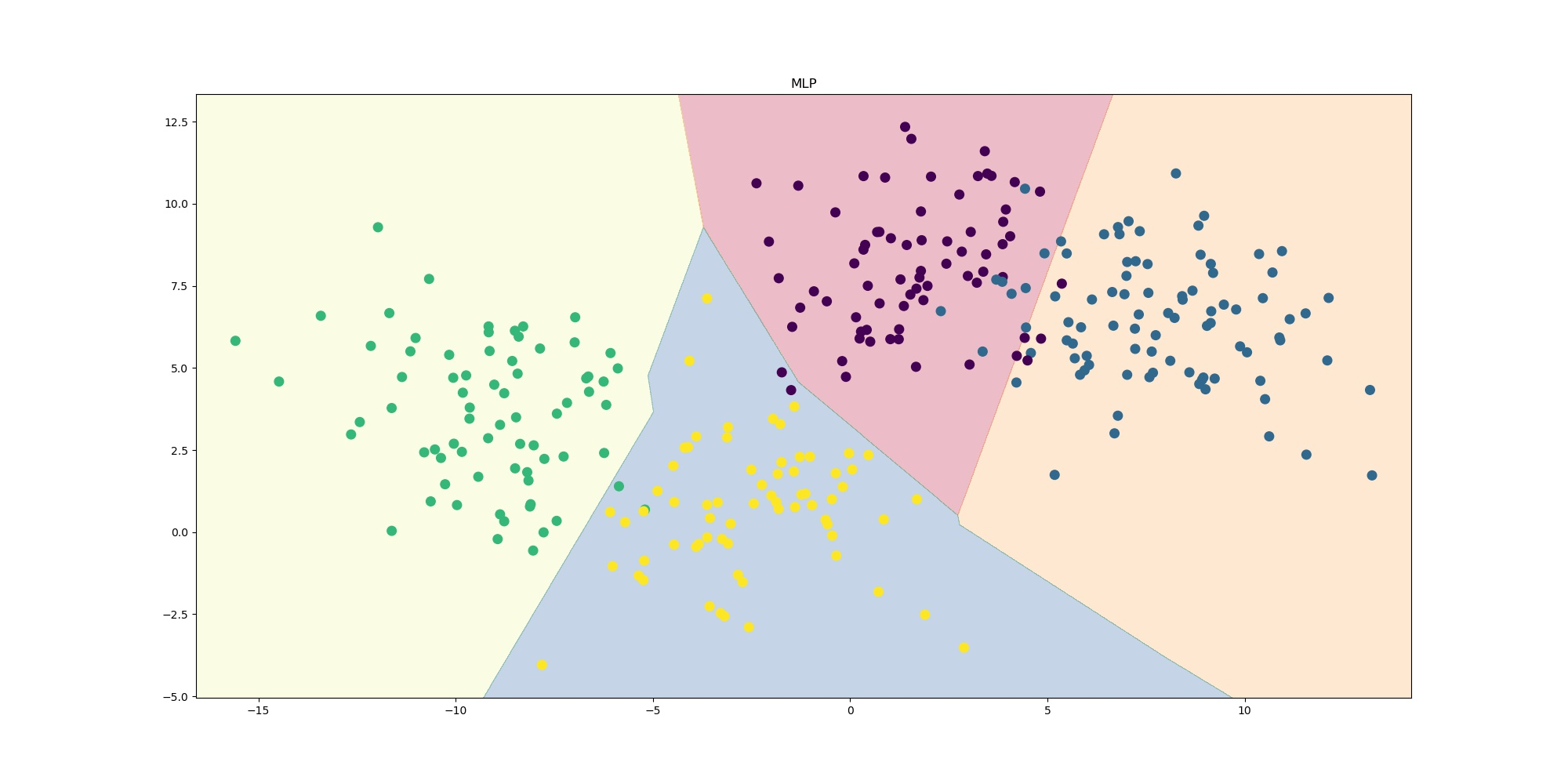


Рис. 6. Результаты работы многослойного персептрона на датасете «points\_10\_20»

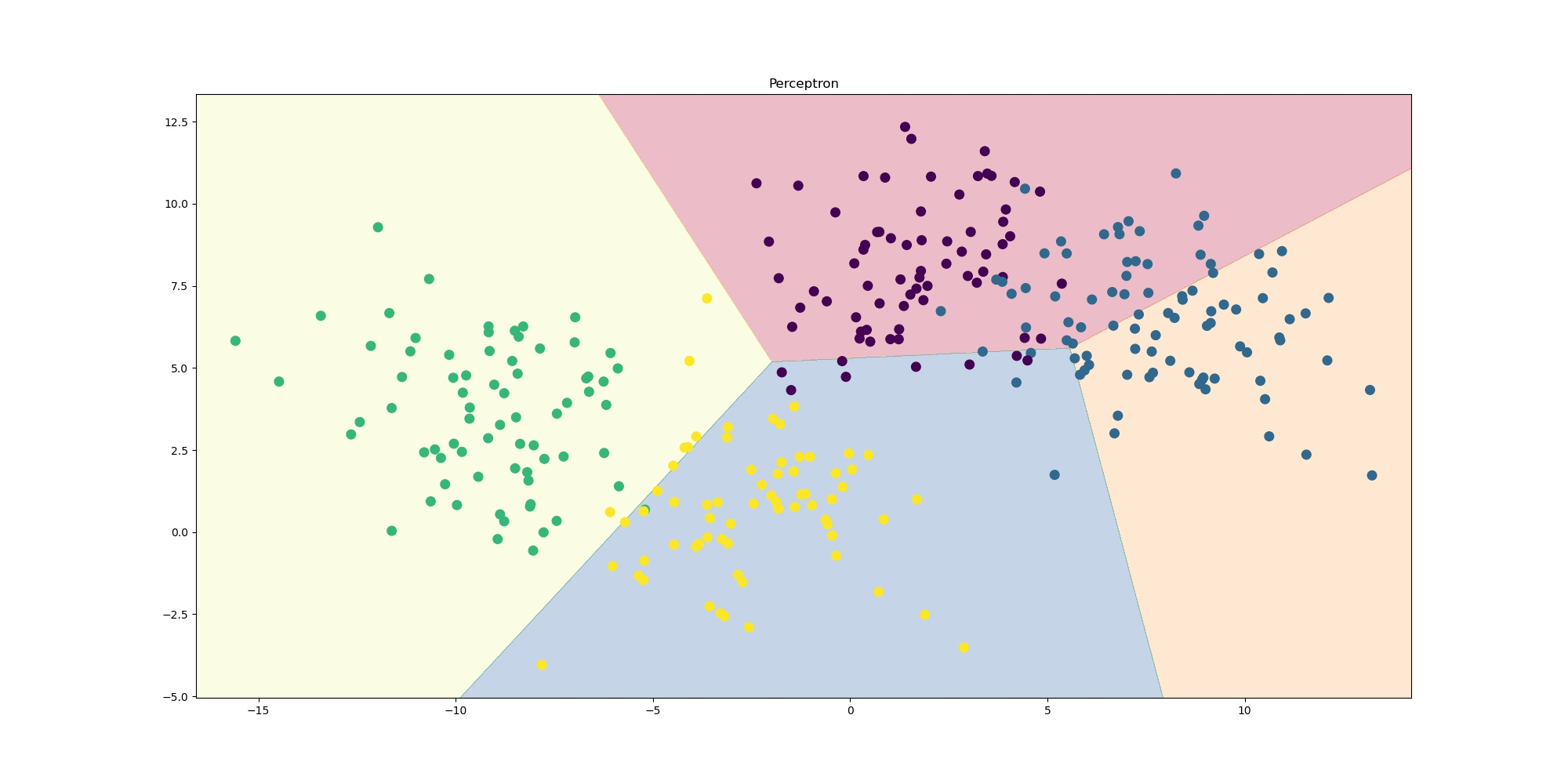


Рис. 7. Результаты работы однослойного персептрона на датасете «points\_10\_20»

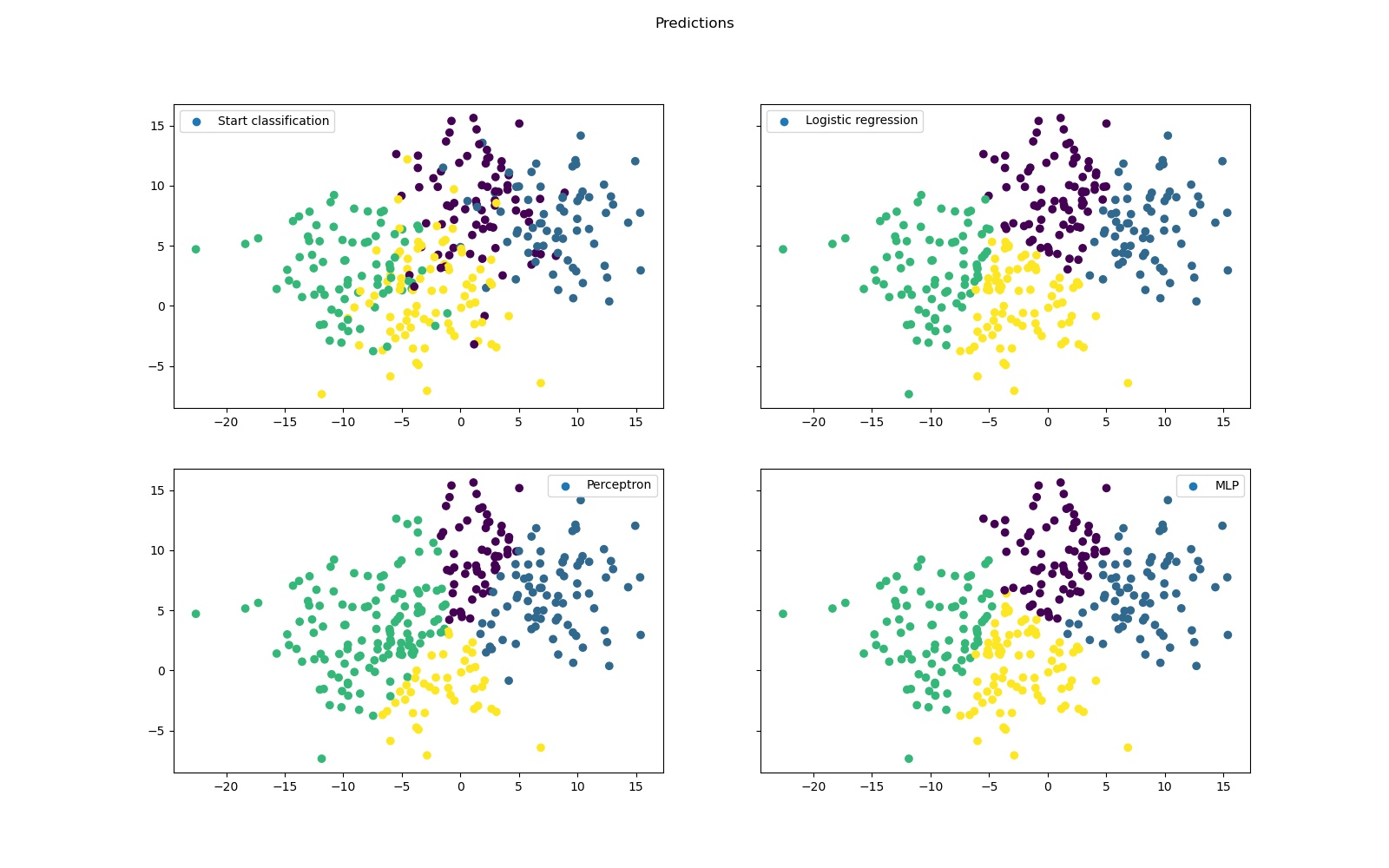


Рис. 8. Результаты обучения для датасета «points\_50\_70»



Рис. 9. Результаты работы линейного классификатора на датасете «points\_50\_70»

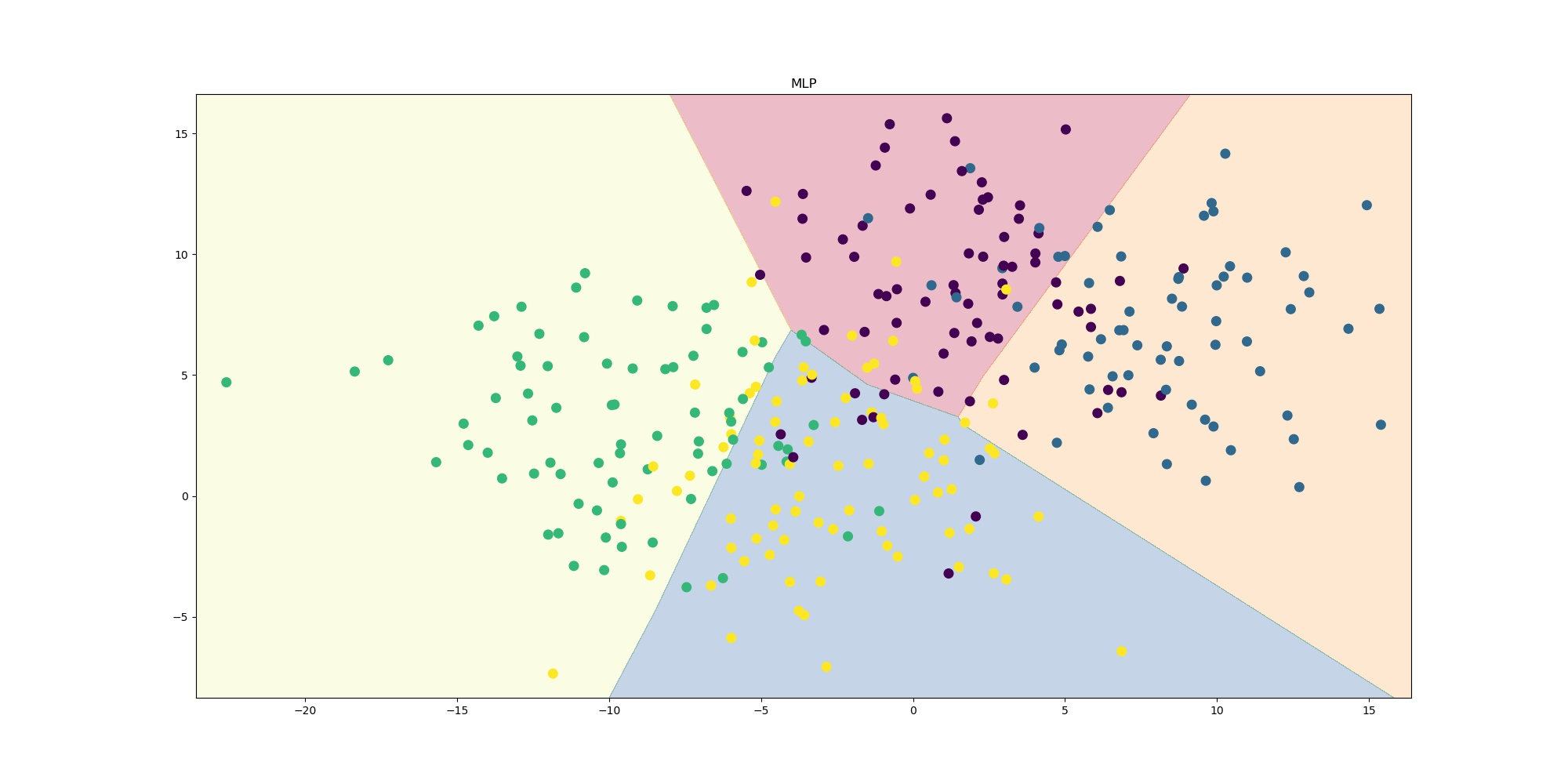


Рис. 10. Результаты работы многослойного персептрона на датасете «points\_50\_70»

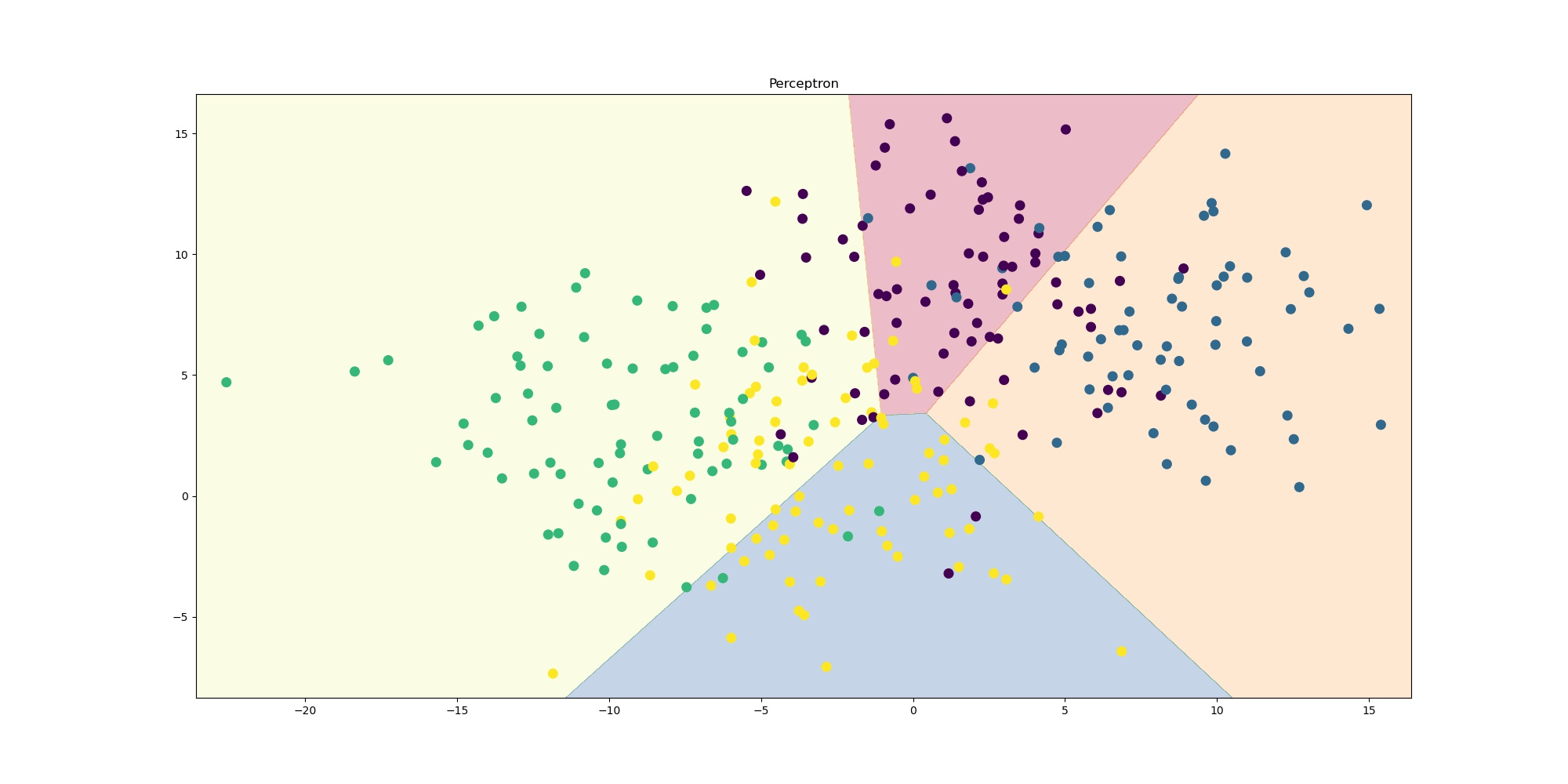


Рис. 11. Результаты работы однослойного персептрона на датасете «points\_50\_70»

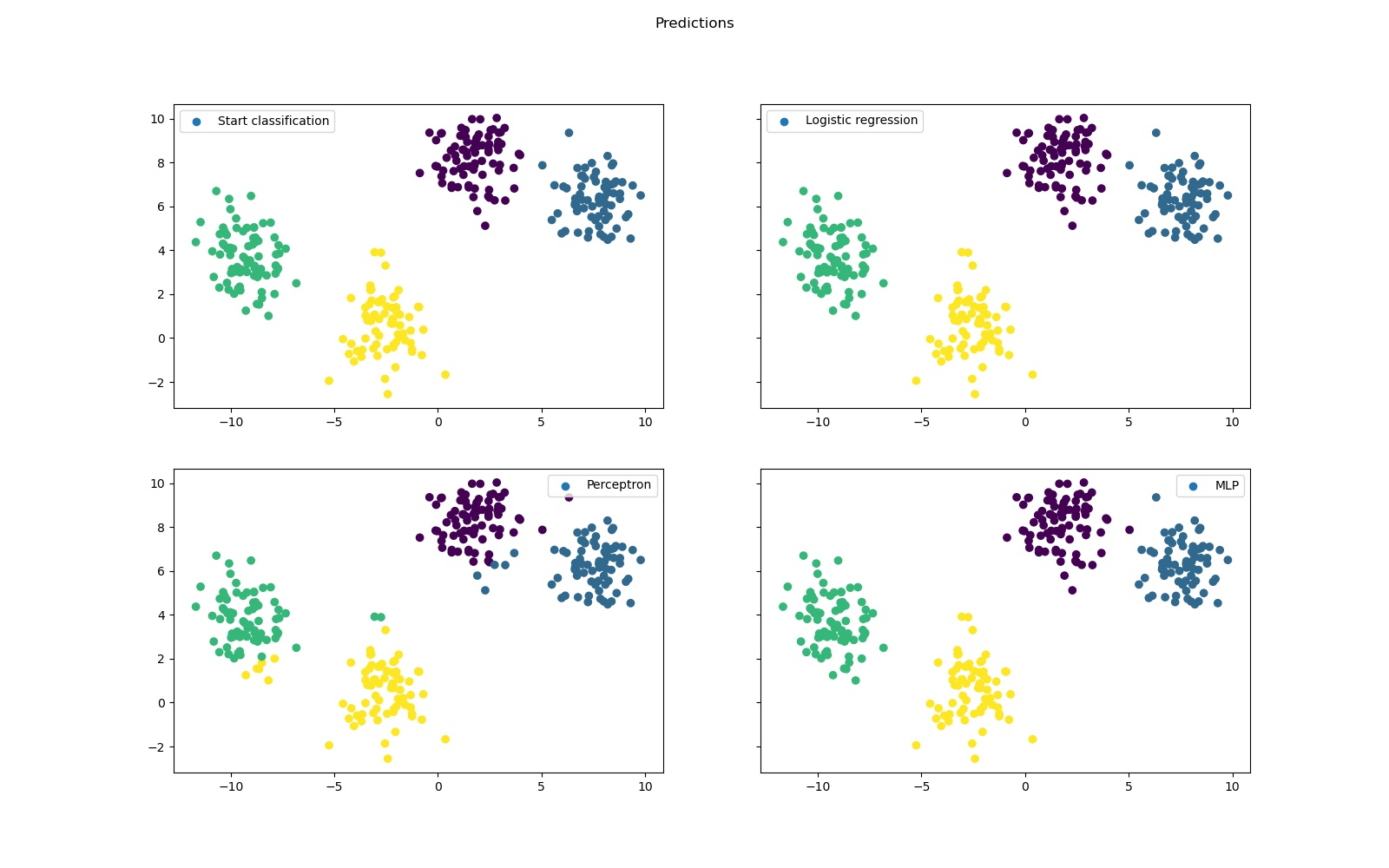


Рис. 12. Результаты обучения для датасета «points\_linear»

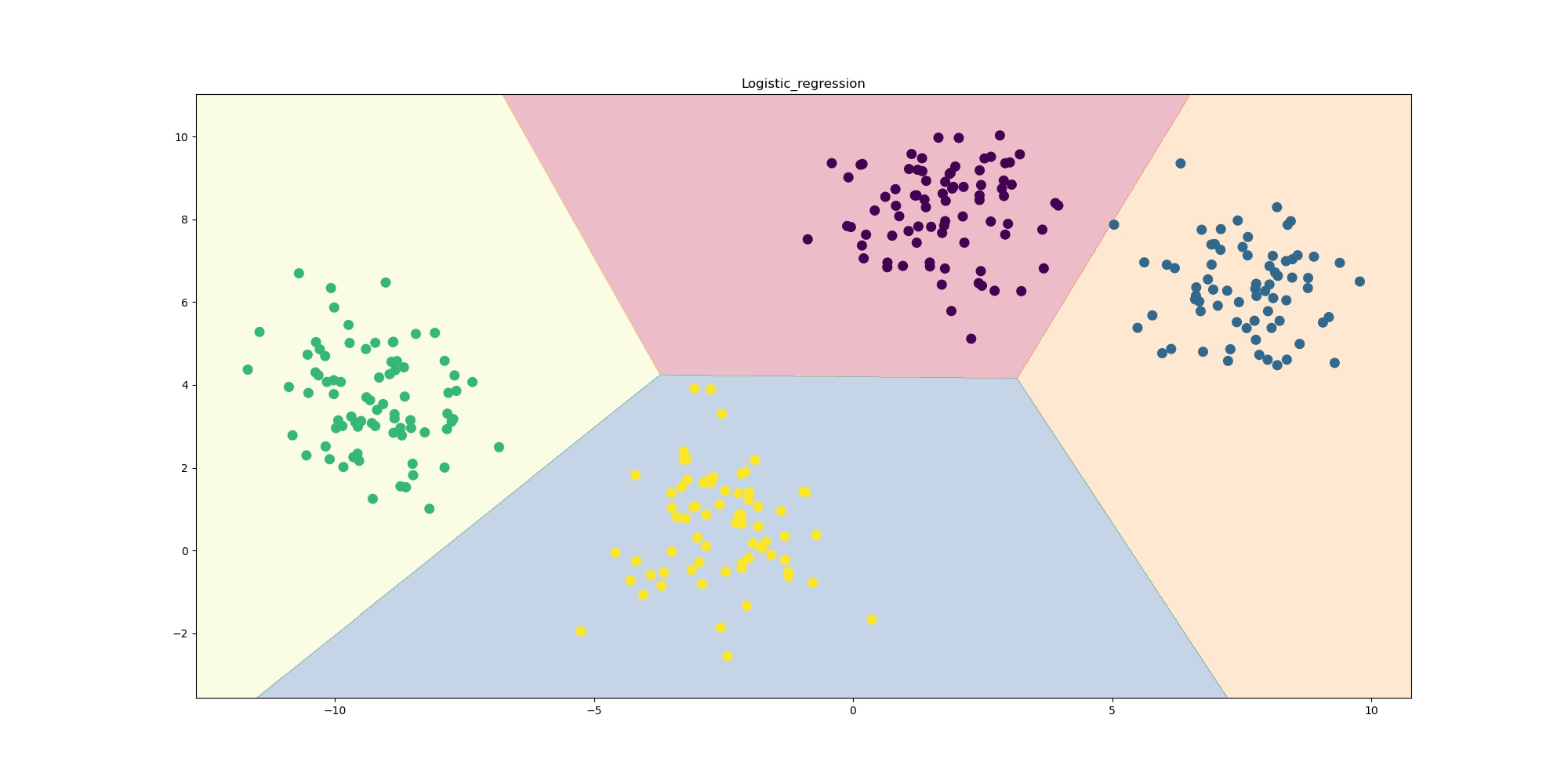


Рис. 13. Результаты работы линейного классификатора на датасете «points\_linear»

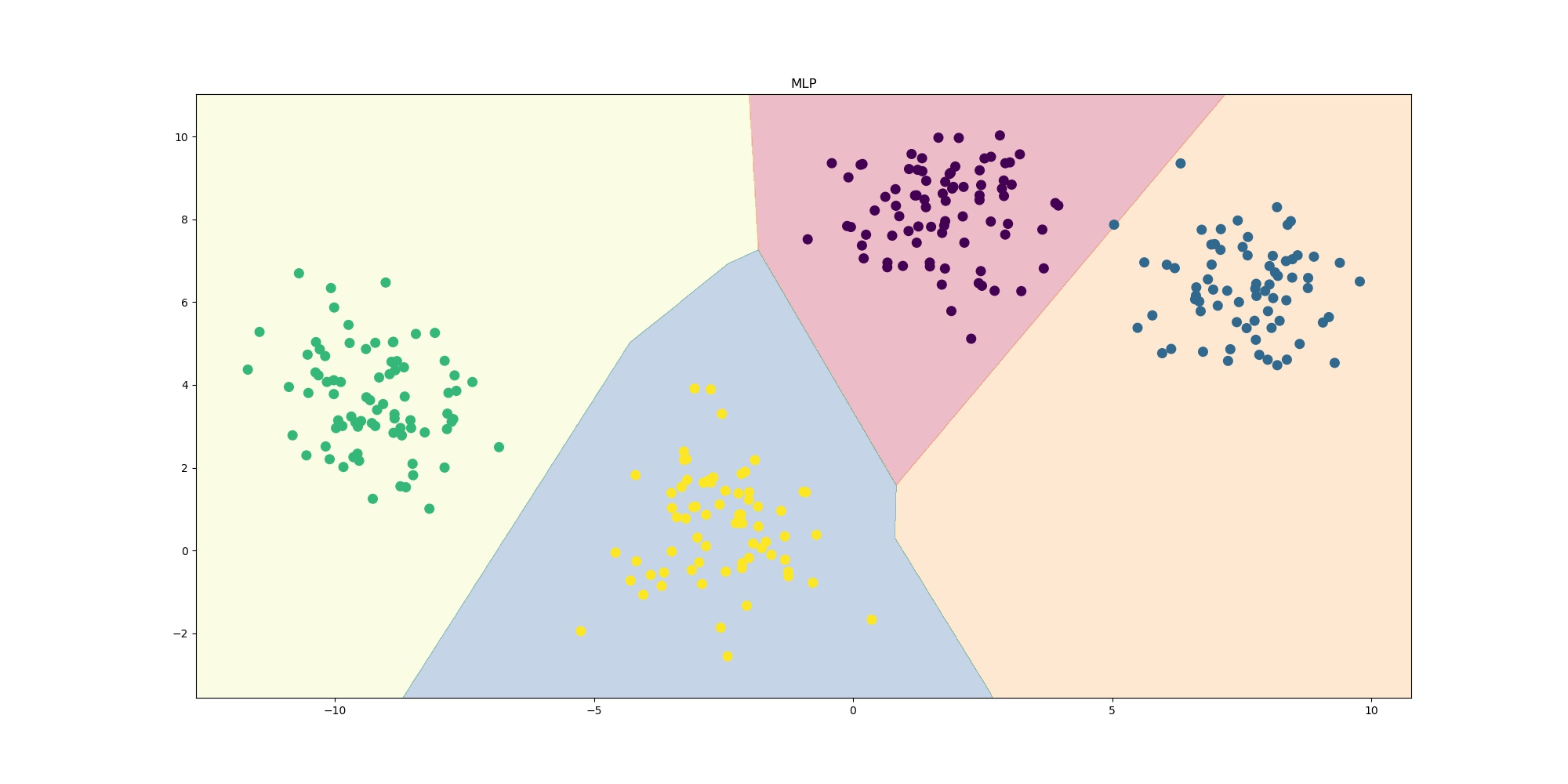


Рис. 14. Результаты работы многослойного персептрона на датасете «points\_linear»

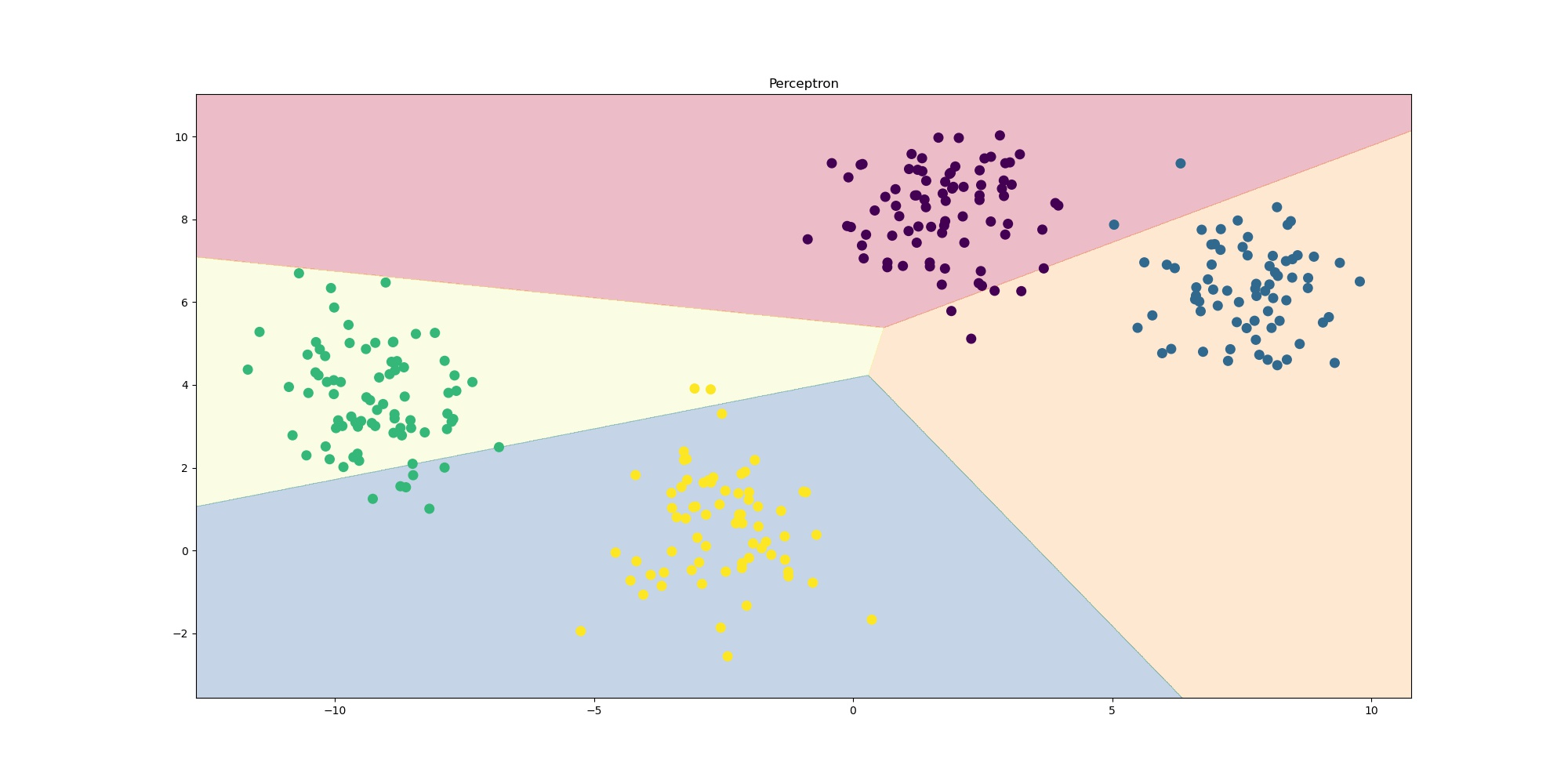


Рис. 15. Результаты работы однослойного персептрона на датасете «points\_linear»

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **learning\_type** | **Accuracy** | **Precision** | **Recall** | **F\_score** |
| **points\_linear** |  |  |  |  |
| Linear\_classifier | 1 | 1 | 1 | 1 |
| Perceptron | 0,95 | 0,94953 | 0,951295 | 0,94988 |
| MLP | 0,996667 | 0,996951 | 0,996429 | 0,996668 |
| **points\_10\_20** |  |  |  |  |
| Linear\_classifier | 0,946667 | 0,947274 | 0,950414 | 0,948142 |
| Perceptron | 0,823333 | 0,849163 | 0,838698 | 0,825126 |
| MLP | 0,94 | 0,940804 | 0,943469 | 0,94172 |
| **points\_50\_70** |  |  |  |  |
| Linear\_classifier | 0,766667 | 0,76863 | 0,769912 | 0,768992 |
| Perceptron | 0,706667 | 0,74705 | 0,715106 | 0,698119 |
| MLP | 0,756667 | 0,755451 | 0,759645 | 0,756847 |
| **SPE\_dataset** |  |  |  |  |
| Linear\_classifier | 0,03 | 0,007281 | 0,018159 | 0,008508 |
| Perceptron | 0,016667 | 0,00089 | 0,014706 | 0,001669 |
| MLP | 0,016667 | 0,008807 | 0,008872 | 0,007444 |
| **CG\_dataset** |  |  |  |  |
| Linear\_classifier | 0,619578 | 0,412393 | 0,431239 | 0,420999 |
| Perceptron | 0,595978 | 0,407025 | 0,421148 | 0,404727 |
| MLP | 0,501911 | 0,500554 | 0,333701 | 0,223449 |
| **RWQ\_dataset** |  |  |  |  |
| Linear\_classifier | 0,570833 | 0,266474 | 0,252498 | 0,249494 |
| Perceptron | 0,504167 | 0,169361 | 0,204456 | 0,184815 |
| MLP | 0,564583 | 0,27213 | 0,26671 | 0,266 |
| **iris\_dataset** |  |  |  |  |
| Linear\_classifier | 0,933333 | 0,935143 | 0,931548 | 0,93266 |
| Perceptron | 0,688889 | 0,492297 | 0,666667 | 0,55579 |
| MLP | 0,888889 | 0,920635 | 0,880952 | 0,882491 |

Табл. 4. Полученные метрики по результатам обучения\*\*

\* Сверху слева – исходная классификация точек.

Сверху справа – классификация точек по результатам работы линейного классификатора.

Снизу слева – классификация точек по результатам работы однослойного персептрона.

Снизу справа – классификация точек по результатам работы многослойного персептрона.

\*\* Зелёным показан наилучший тип обучения для обозначенного датасета

Жёлтым показан второй по уровню классификации тип обучения для обозначенного датасета.

Красным показан наихудший тип обучения для обозначенного датасета.

Из полученных данных можно сделать вывод, что линейный классификатор показал наилучшие результаты на каждом из представленных датасетов. Наихудшие результаты на большинстве представленных датасетов показал однослойный персептрон. Многослойный персептрон занимает 2 позицию по уровню классификации на рассмотренных датасетах.

# Список используемой литературы

1. Kaggle – Students performance in exams dataset. URL: <https://www.kaggle.com/spscientist/students-performance-in-exams?select=StudentsPerformance.csv> (дата обращения: 01.01.21).
2. Kaggle – Chess game dataset. URL: <https://www.kaggle.com/datasnaek/chess?select=games.csv> (дата обращения: 01.01.21).
3. Kaggle – Red wine quality dataset. URL: <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009> (дата обращения: 01.01.21).
4. Kaggle – Iris flower dataset. URL: <https://www.kaggle.com/arshid/iris-flower-dataset> (дата обращения: 01.01.21).