**1.**

Здравствуйте, тема моей научной работы – применение классификаторов в машинном обучении, выполнил Бут А группа ПРИб-201.

**2.**

Целью являлось проведение сравнительного анализа классификаторов для задач анализа данных.

Основные задачи: изучить литературу, рассмотреть модели классификации, программно их реализовать и провести сравнительный анализ результатов.

**3.**

Сами методы машинного обучения делятся на алгоритмы обучения с учителем и без. Основные методы обучения без учителя – Кластеризация (основные методы DBSCAN и метод k-средних) и регрессия (линейная и логистическая). В данной работе были применен метод классификации с помощью алгоритмов классификатор K-ближайших соседей;  
- наивный байесовский классификатор;  
- алгоритм деревьев решений;  
- алгоритм случайного леса;

- метод опорных векторов;

- алгоритм логистической регрессии;

**4.**

Для обучения необходимо предпринять следующие шаги

Собрать и подготовить данные, выбрать и обучить модель, настроить гиперпараметры и совершить предсказание, а затем оценить результаты.

**5.**

Научным руководителем была предоставлена база студентов, поступивших в прошлом году и их ответов на определенный перечень вопросов. База обрабатывалась следующим образом:

Расставлены идентификаторы специальностей 0 – физическая, 1 – математическая (колонка flag)

Удалены студенты, с баллами ниже 160 и те студенты, которые обучаются на мат специальности, но при этом в вопросе “Нравится ли вам физика?” проставили высокие баллы и такие же студенты физ специальностей, которые в вопросе “Нравится ли вам математика” проставили высокие баллы.

**6.**

На основании входных данных были построены диаграммы размаха для каждого вопроса, на экране часть из них

Из них видно, что разбираться в электронных устройствах больше нравится студентам физических специальностей, а решать математические задачи и головоломки нравится математикам, однако существует область, где студенты обоих классов проставили высокие баллы. Такие вопросы повлияют на обучение системы негативным образом. А в вопросе “Нравится ли вам паять схемы?” ставили высокие баллы только физики, этот вопрос повлияет положительно. Однако существует один объект-артефакт, отстоящий от остальной массы, который повлияет негативно.

**7.**

Матрица корреляций – особый объект, который характеризует предсказание. Содержит информацию о верных и неверных предсказаниях каждого класса. На основании нее вычисляются характеристические величины: чувствительность – отношение верных предсказаний первого класса к общему числу его предсказаний, специфичность - отношение верных предсказаний второго класса к общему числу его предсказаний и эффективность – корень из произведения этих величин.

На экране одна из матриц, построенных в ходе работы.

**8.**

На выходе получилась сравнительная таблица значений этих характеристик для каждого метода. Лучше всего себя показал метод логистической регрессии, его эффективность 0,6252. Видно, что произошло недообучение, объекты класса 0 каждым классификатором определяются хуже просто потому что в базе слишком мало экземпляров этого класса.

**9.**

Загрузив базу объектов, содержащую ненадежные элементы, получаем другие значения, где лучше остальных справились метод деревьев решений (эффективность 0,73) и наивный байесовский классификатор (эффективность 0,73). Тем не менее, поскольку существовали ненадежные элементы – обучение велось на не полностью корректных данных, а следовательно и точность предсказания снижена.

**10.**

В ходе работы была собрана и изучена литература по теме исследования, было рассмотрено несколько линейных моделей классификации данных, они были программно реализованы и был проведен их сравнительный анализ.