**Задание на курсовой проект**

Курсовой проект заключается в разработке классификаторов для реальной базы данных, визуализации данных, исследовании и настройки классификаторов.

Программное обеспечение: R-project или R-Studio (по выбору)

Список методов классификации:

1. Наивный байесовский классификатор.
2. Метод ближайших соседей.
3. Деревья решений.
4. Метод опорных векторов.
5. Нейронные сети.
6. Беггинг.
7. Бустинг.

Список методов кластеризации:

1. Метод k средних.
2. Метод медоидов.
3. Иерархическая кластеризация.

Базы данных (по вариантам, в скобках имена соответствующих файлов):

1. Protein Localization Sites (Ecoli)
2. Heart disease (heart)
3. Hepatitis Domain (Hepatitis)
4. Liver patient records (Indian Liver Patient Dataset)
5. Mammographic Mass Data (mammographic\_masses)
6. Parkinsons Disease Data Set (parkinsons)
7. Pima Indians Diabetes Database (pima-indians-diabetes)
8. Diagnosing of cardiac Single Proton Emission Computed Tomography images (SPECT)
9. Wisconsin Diagnostic Breast Cancer (wdbc)

Для каждой базы данных необходимо:

1. Разработать 3 классификатора и осуществить настройку их параметров для минимизации ошибки классификации на тестовых данных. Выполнить визуализацию данных при помощи метода t-SNE.
2. Сравнить классификаторы (по критерию вероятность ошибки классификации для тестовых данных) и обосновать выбор наилучшего из них.
3. Удалить их базы метки классов и осуществить кластеризацию данных. Построить дендограмму. Сравнить полученные результаты с реальными метками данных. Определить долю ошибочно кластеризованных данных.
4. Используя логистическую регрессию в рамках метода Лассо, определить наиболее значимые признаки, влияющие на отнесение объектов к определенному классу.
5. Использовать автокодер для сокращения размерности или для реализации разреженного скрытого слоя нейронной сети. Преобразовать обучающую выборку при помощи автокодера и осуществить классификацию новых данных с оценкой ошибки классификации. Выполнить визуализацию новых обучающих данных при помощи метода t-SNE. Определить, когда качество классификации лучше, если использовать сокращение размерности или разреженность скрытого слоя. Выполнить классификацию с использованием зашумленного автокодера (denoising autoencoder). Сравнить полученные результаты с пп.1 и 2.
6. Подготовить пояснительную записку по курсовому проекту и листинги программ.