**Модульное домашняя работа №2**

**по майнору «Прикладной статистический анализ»,**

**дисциплина: «Классификация данных»**

Домашнее задание выполняется индивидуально.

Работа должна быть оформлена в формате MS Word (либо PDF).

Все полученные результаты расчетов должны быть описаны выводами в текстовом формате.

1. **2 балла** – Каждый студент формирует свою базу данных, определяя проблематику исследования. Использовать данные пространственной выборки объемом не менее 80 наблюдений (субъекты РФ, страны, компании и т.п.). В базу данных необходимо включить не менее 6 непрерывных и 2 бинарных переменных, а также бинарный ответ на задачу классификации (целевую переменную). Сформулировать задачу классификации.
2. **3 балла** – По сформированной базе данных провести линейный дискриминантный анализ:
   1. Выделить 1-3 наблюдения (резко выделяющихся, наиболее удаленных от центров кластеров), подлежащих дискриминации;
   2. Провести дискриминантный анализ;
   3. Записать выражение для дискриминантной функции;
   4. Оценить значимость дискриминантной функции (по коэффициенту Уилкса);
   5. Определить относительный вклад каждой переменной в формирование классов;
   6. Определить, чему равны средние значения дискриминантной функции по группам;
   7. Указать, к каким группам были отнесены классифицируемые объекты и вероятности, с которыми объекты входят в эти группы;
   8. На основании анализа таблицы «Wilks’ Lambda» проверьте значимость различий средних значений дискриминантной функции в двух группах;
   9. Оценить качество дискриминантного анализа (на основании результатов таблицы Eigenvalue);
   10. Оценить целесообразность проведения дискриминантного анализа по Вашим данным.
3. **2,5 балла** – Построение дерева классификации (методы CHAID или CRT на выбор):
   1. Выбор зависимой переменной – дискретная шкала;
   2. Построение деревьев с помощью метода CHAID (или CRT) с различными независимыми переменными;
   3. Выбор оптимального дерева с помощью таблицы классификации;
   4. Визуализация результатов;
   5. Интерпретация результатов.
4. **2,5 балла** – Используя данные из файла «задача\_5.xls» провести декомпозицию смеси распределений (количество распределений, входящих в смесь, неизвестно). Описать каждое распределение, входящее в состав смеси, выполнить визуализацию. Определить к какому распределению относится наблюдение, номер которого соответствует Вашему порядковому номеру в списке группы.