Задание «Пакеты для анализа данных, рецензирование».
Срок сдачи первого этапа – 14 апреля 2013 г., 23:59
Максимальный балл – 5 баллов

Данное задание направлено на приобретение студентами навыков использования готовых пакетов для анализа данных. Примерами таких пакетов являются Weka, RapidMiner, scikit-learn.

Данное задание состоит из нескольких этапов:

1) При помощи одного из пакетов анализа данных (возможно использование каких-либо альтернатив по согласованию с преподавателем) необходимо провести анализ данных задачи своего варианта из репозитория UCI. В рамках данного анализа необходимо затронуть следующие аспекты: предобработка данных, оценка информативности признаков, отбор признаков, применение различных алгоритмов, оценка точности решения, адекватность полученного решения поставленной задаче, и т.д. По итогам проведенного исследования необходимо написать отчет в формате PDF с описанием задачи, всех проведенных исследований, графиками, иллюстрациями. В данном отчете фамилия автора должна быть вынесена на отдельный титульный лист. Больше нигде в отчете не должно быть ничего, раскрывающего авторство.
2) Студенты, сдавшие отчет по итогам первого этапа задания, должны отрецензировать отчет одного из своих коллег. Назначение рецензентов проводится по системе double blind (как на ведущих конференциях): рецензент не знает автора, автор не знает рецензента. Автор обязуется не указывать в отчете информацию, позволяющую установить его личность (титульный лист будет отдельно удален преподавателем), рецензент обязуется не искать личность автора и проводить рецензирование беспристрастно и объективно. Рецензия должна содержать в себе: оценку адекватности текста отчета (с точки зрения понятности описания задачи и того, что же было сделано), оценку адекватности алгоритма решения задачи, качества ее решения, рекомендации автору по улучшению отчета: алгоритмические и технические замечания, грамматические и стилистические ошибки, опечатки.
3) Студенты, прошедшие первые два этапа, исправляют отчет, согласно замечаниям, высказанным в решении, а также пишут небольшой отзыв о рецензии, в котором они могут согласиться, либо не согласиться с претензиями рецензента, а также оценить качество рецензии ☺

Распределение задач по вариантам:
Вариант 1 – Adult, вариант 2 – Abalone; вариант 3 – Thyroid; вариант 4 – Covertype
Сроки сдачи задания:

1) Этап 1 (исследование): 4-14 апреля.
2) Этап 2 (рецензирование) – 22-28 апреля.
3) Этап 3 (исправление отчета) – 6-12 мая.

Первый этап оценивается из 2-х баллов, второй – их 2-х, третий из одного.
Штраф за опоздание составляет 0.1 балл в день за каждый этап.
Студент, просрочивший какой-то из этапов более чем на неделю (пропустивший начало следующего этапа), к дальнейшим этапам не допускается.
Полезные ссылки:

1) Weka: http://www.cs.waikato.ac.nz/ml/weka/
2) RapidMiner: http://www.rapidminer.com/
3) scikit-learn: http://scikit-learn.org/stable/
4) Задача Adult: http://archive.ics.uci.edu/ml/datasets/Adult
5) Задача Abalone: http://archive.ics.uci.edu/ml/datasets/Abalone
6) Задача Covertype: http://archive.ics.uci.edu/ml/datasets/Covertype
7) Задача Thyroid: http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease