Министерство образования Республики Беларусь

Учреждение образования

Брестский государственный технический университет

Кафедра ИИТ

Лабораторная работа №5

За 7 семестр

По дисциплине «Методы обработки данных»

Тема: «Машинное обучение. Метод k-ближайших соседей»

Выполнил:

Студент 4 курса, группы ПО-7(1)

Вощук А.Я.

Проверил:

Савицкий Ю.В.

Брест, 2023

**ЛАБОРАТОРНАЯ РАБОТА №5**

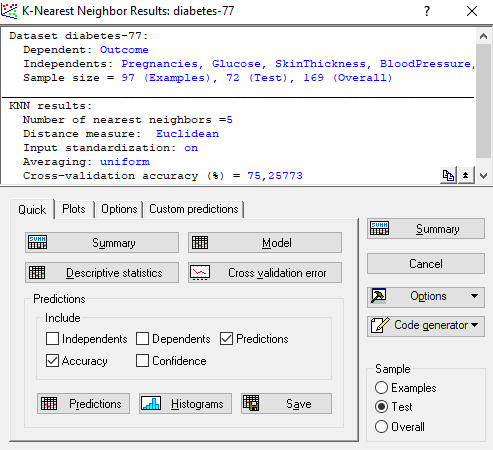
**Машинное обучение. Метод k-ближайших соседей**

**Цель:** изучить основы методов Machine Learning в контексте задачи классификации методом k-ближайших соседей (k-nearest neighbors, k-NN), приобрести навыки работы с методами Machine Learning в системе STATISTICA StatSoft, осуществить обработку методами Machine Learning индивидуального набора данных и интерпретацию результатов.

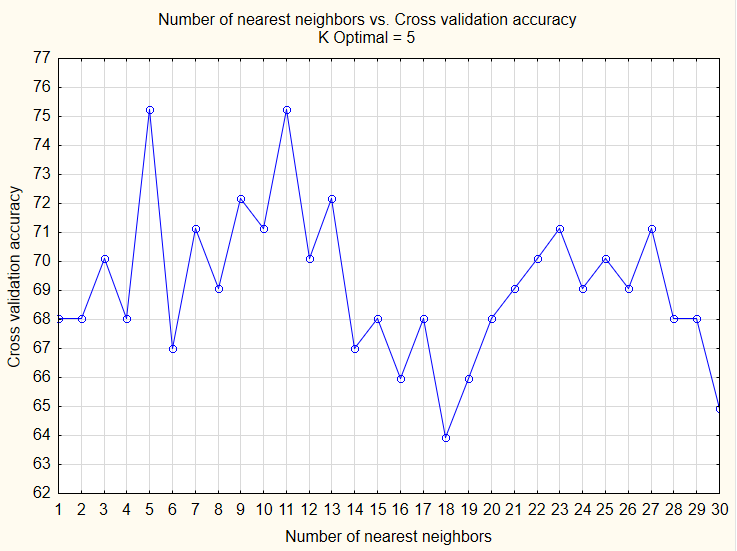
**Вариант 1**

**Ход работы:**

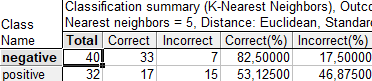
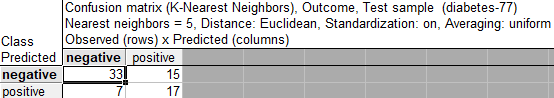
Проведем анализ наших данных с помощью методом K-ближайшего соседа, указав столбец Outcome, в качестве зависимой категории, а столбцы Pregnancies-Age, в качестве независимых переменных. Также установим “Стандартизация растояний”, “Apply v-fold cross-validation”, и использование nnset в качестве sample variable. В результате получим:



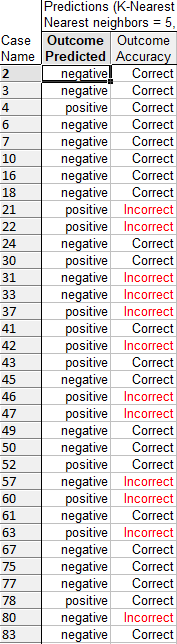
Создадим график ошибки перекрёстной проверки для каждого значения К, опробованного алгоритмом перекрёстной проверки:



Получим данные классификации матрица ошибок:

Также была получена следующая таблица прогнозов:

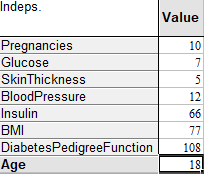


Далее рассмотрим разницу между результатами в виде графиков для параметра Insulin в зависимости от настроек оптимизации:

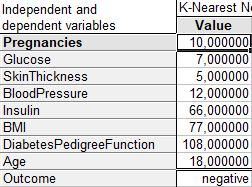
|  |  |
| --- | --- |
| С использование параметров перекрестной проверки (K=30) |  |
| Без использование параметров перекрестной проверки (K=40) |  |
| Без использования параметров перекрестной проверки (K=97, то есть размеру выборки) |  |
| Без использования параметров перекрестной проверки (K=97, то есть размеру выборки) с установленным флажком “Взвешенное расстояние” |  |

Как видим исходя из результатов, с повышением максимального количества в выборки снижается точность анализа. Также, как можно заметить, включение “Взвешенного расстояния” может реабилитировать в некоторой степени наш анализ.

Благодаря нашему анализу, теперь мы можешь производить наши собственные прогнозы. Для примера заполним следующие данные:



В результате мы получили предсказание об отрицательном тесте о сахарном диабете:



**Вывод**: изучил основы методов Machine Learning в контексте задачи классификации методом k-ближайших соседей (k-nearest neighbors, k-NN), приобрёл навыки работы с методами Machine Learning в системе STATISTICA StatSoft, осуществил обработку методами Machine Learning индивидуального набора данных и интерпретировал результаты.