Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Информационных технологий и управления

Кафедра Интеллектуальных информационных технологий

Индивидуальная практическая работа №1

По дисциплине статические основы индукционного вывода

на тему

# Построение бинарного классификатора средствами MS EXCEL

Выполнил Мулярчик Д.С.

Проверил Ефремов А.А

Минск 2023

**Задание:**

1. **Подобрать в открытых источниках data set, состоящий из результативного признака (заданного бинарной переменной) и нескольких факторных признаков (не менее 3).**
2. **Построить бинарный классификатор, пользуясь методическими указаниями из примера ниже.**
3. **В отчёте представить: постановку задачи с описанием переменных (А), фрагмент таблицы с исходными данными (Б), уравнение логистической регрессии (В), значение Zгр (Г), оценку надёжности классификатора через расчёт процента ошибок (Д).**

**Оценка наличия ухудшения ментального здоровья:**

В качестве dataseta использовался https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data/

1. Увеличение риска
2. Нормальная частота пульса в состоянии покоя в ударах в минуту. HeartRate - x1
3. Верхнее значение артериального давления в мм рт. ст. SystolicBP - x2
4. Нижнее значение артериального давления в мм рт.ст. DiastolicBP - x3
5. Уровень глюкозы в крови выражен в молярной концентрации, ммоль/л. BS - x4

Коэффициенты потенциального x1,x2,x3,x4 человека являются следующими 102.5, 78.7, 10.6, 70.9.

Требуется:

1) построить линейную регрессионную модель для оценки кредитного риска и с ее помощью оценить вероятность дефолта для потенциального заемщика;
2) построить регрессионную дискриминантную модель, найти граничное значение $z_{гр}$ и отнести потенциального заемщика к группе с высоким либо низким кредитным риском.

Информация по кредитам приведена в следующей таблице.

| Age | SystolicBP | DiastolicBl | BS | HeartRate | RiskLevel |
|---|---|---|---|---|---|
| 35 | 85 | 60 | 11 | 86 | 1 |
| 42 | 130 | 80 | 18 | 70 | 1 |
| 23 | 90 | 60 | 7 | 76 | 0 |
| 25 | 110 | 89 | 7 | 77 | 0 |
| 15 | 120 | 80 | 7 | 70 | 0 |
| 50 | 140 | 90 | 15 | 90 | 1 |
| 25 | 140 | 100 | 7 | 80 | 1 |
| 10 | 70 | 50 | 6 | 70 | 0 |
| 40 | 140 | 100 | 18 | 90 | 1 |

**Пункт 1**

Введем таблицу с данными в Excel. Линейная регрессионная модель для оценки кредитного риска в данном случае имеет вид: Z = b0 +b1x1+b2x2+b3x3+b4x4+e.

Для оценки коэффициентов bк, к=0.4, будем использовать модуль «Анализ данных», который вызывается из «Сервиса» в главном меню. В «Анализе данных» найдем инструмент «Регрессия» и вызовем его. В появившемся окне укажем входные интервалы Y и X.

Входной интервал Y – это массив ячеек (в таблице исходных данных), содержащих значения объясняемой переменной Z. Входной интервал X – это массив ячеек, содержащих зна чения объясняющих переменных x1, x2, x3,x4.

После ввода входных интервалов, нажмем на кнопку «ОК». В результате появится новый лист с параметрами регрессионной модели. Оценка b0 коэффициента b0 равна значению коэффициента для «Y-пересечения», а оценки b1, b2, b3, b4 коэффициентов b1, b2, b3, b4 равны значениям коэффициентов для переменных X1, X2, X3, X4.

Вероятность ухудшения ментального здоровья оценивается по формуле $\widehat{\beta_2}x_2 + \widehat{\beta_3}x_3 + \widehat{\beta_4}x_4$, где x1,x2,x3,x4 заданные значения для потенциального пациента.

В результате вычислений мы получили следующие значения:

1. $\widehat{\beta_0}$ = 14.0374285.
2. $\widehat{\beta_1}$ = 0.4658391.
3. $\widehat{\beta_2}$ = -2.0573952.
4. $\widehat{\beta_3}$ = -0.7652419.
5. $\widehat{\beta_4}$ = 0.9274294.

**Пункт 2**

В качестве регрессионной дискриминантной модели можно взять модель из п.1.

Для каждого наблюдения вычисляются прогнозные значения показателя z по формуле: $\widehat{Z^i} = \widehat{\beta_0} + \widehat{\beta_1}x_1^i + \widehat{\beta_2}x_2^i + \widehat{\beta_3}x_3^i i = \overline{1,N}$

Затем с помощью функций СРЗНАЧ и СТАНДОТКЛОН нужно найти средние значения z1 и z2, и стандартные отклонения o1 и 02 для наблюдений с дефолтом (1-й массив) и для наблюдений без дефолта. (Для этого предварительно следует упорядочить таблицу соответствующим образом.)

1. z1 = 3.5723852
2. z2 = 0.2473618
3. o1 = 2.5063951
4. o2 = 1.8294629

$$Z_{\text{гр}} = \frac{\sigma_1\overline{Z_2} + \sigma_2\overline{Z_1}}{\sigma_1 + \sigma_2}$$

Граничное значение zгр вычисляется по формуле: равное 1.6498662053056.

Поскольку z1 > z2, вероятность заболевания оценивается как низкая, если z < zгр, и высокая если z > zгр.

Для потенциального пациента вычислим Zпот = 1.442.

Следовательно, у данного пациента низкая вероятность ухудшения ментального здоровья.