**Задание 7**

*Решение задачи классификации методом kNN*

Выполните следующее:

1. Найдите реальные данные для задачи классификации.
2. Перечислите факторы (и укажите единицы их измерения) и поясните смысл переменной отклика.
3. Если число факторов m <= 2, то визуализуйте данные (постройте облако точек).
4. Реализуйте метод kNN для найденных данных.
5. Оцените точность полученного решения с помощью метода кросс-валидации. (Скриншот кросс-валидационнной таблицы включите в отчёт. Поясните, как Вы обеспечили репрезентативность обучающей выборки.)
6. Задайте нескольких новых данных, если m <= 2, то покажите соответствующие точки на графике (выделите их другим цветом).
7. Определите значение переменной отклика (номер класса) для новых данных.
8. Оформите отчёт в виде Word-файла и прикрепите его здесь.

Были использованы данные обследования 789 женщин индейского народа Пима (см. [статью](https://ru.wikipedia.org/wiki/%D0%9F%D0%B8%D0%BC%D0%B0) в Википедии для дополнительной информации) старше 21 года на предмет наличия сахарного диабета. На самом деле датасет *“Pima Indians Diabet Database”* достаточно популярен.

Источник данных: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.

Данные содержат следующие характеристики:

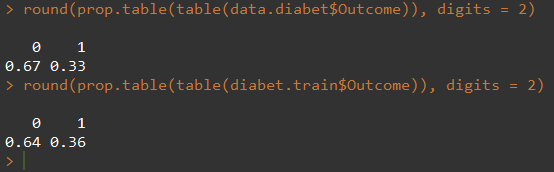
1. Pregnancies — число беременностей обследуемой;
2. Glucose — концентрация глюкозы в плазме через 2 часа после проведения преорального глюкозотолерантного теста, ;
3. BloodPressure[[1]](#footnote-1) — диастолическое кровяное давление, ;
4. SkinThickness — толщина кожной складки трицепса, ;
5. Insulin[[2]](#footnote-2) — 2-Hour serum insulin, ;
6. BMI— индекс массы тела, определяется как (;
7. DiabetesPedigreeFunction — вероятность заболевания сахарным диабетом, основанная на истории болезни предков;
8. Age — возраст, в годах;
9. Outcome — обнаружен ли диабет у обследуемой: 0 — не обнаружен, 1 — обнаружен.

В качестве факторов будем использовать характеристики 1-8, характеристика 9 — переменная отклика.

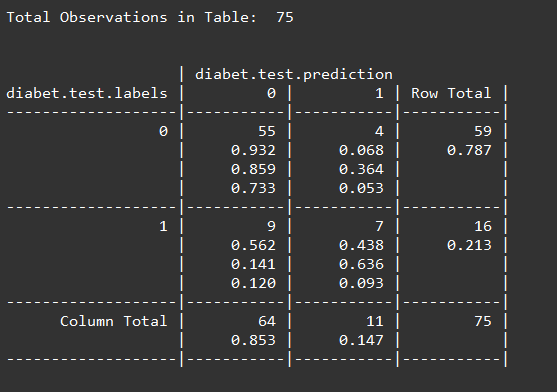
После изучения существующей информации о данных, оказалось, что присутствуют некорректные наблюдения. К примеру, уровень кровяного давления составляет или же индекс массы тела равен . После удаления записей с некорректными значениями, объем выборки составил **392** наблюдения.

Исходная выборка была разделена обучающую и тестовую, объемами **317** и **75** наблюдений соответственно. Для сохранения репрезентативности обучающей выборки с помощью функции sample () был сгенерирован вектор индексов длиной в 317 (объем обучающей выборки) элементов, состоящий из целых чисел от 1 до 392 (объем исходной выборки).

Соотношения классов в исходной и обучающей выборках почти одинаковы:



Приведем кросс-валидационную таблицу для анализа эффективности метода **kNN**:



Заметим, что метод ошибочно **определил** наличие диабета у **4** обследуемых и ошибочно **не определил** наличие диабета у **9** обследуемых. от общего числа прогнозов — ошибочны, следовательно, метод точен на .

Результаты прогноза для пользовательских значение приведены [здесь](#_Результаты_прогноза_пользовательски).

Скрипт программы на R:

data.diabet <- read.csv("data/diabetes.csv",

header = TRUE,

sep = ",")

na.fail(data.diabet)

# отберем только те наблюдения, которые не содержат логических ошибок

data.diabet <- data.diabet[data.diabet$Glucose > 0 &

data.diabet$BloodPressure > 0 &

data.diabet$SkinThickness > 0 &

data.diabet$Insulin > 0 &

data.diabet$BMI > 0,]

# определим объемы тестовой и обучающей выборок

data.size <- length(data.diabet[, 1])

train.size <- 317

test.size <- data.size - train.size

# определим индексы наблюдений тестовой и обучающей выборок

train.indices <- sample(1:data.size, train.size)

test.indices <- setdiff(1:data.size, train.indices)

diabet.train <- data.diabet[train.indices,]

# проверим соотношения классов в исходной и обучающей выборках

round(prop.table(table(data.diabet$Outcome)), digits = 2)

round(prop.table(table(diabet.train$Outcome)), digits = 2)

# запомним отклики для обучающей и тестовой выборок

diabet.train.labels <- diabet.train[, 9]

diabet.train <- diabet.train[, -9]

diabet.test <- data.diabet[test.indices,]

diabet.test.labels <- diabet.test[, 9]

diabet.test <- diabet.test[, -9]

# предскажем значение отклика для тестовой выборки

library("class")

diabet.test.prediction <- knn(train = diabet.train,

test = diabet.test,

cl = diabet.train.labels,

k = round(sqrt(train.size)))

# проанализируем полученные результаты

library("gmodels")

CrossTable(diabet.test.labels, diabet.test.prediction, prop.chisq = FALSE)

# спрогнозируем нзначение переменной-отклика для произвольных значений

my.diabet.test <- **matrix**(nrow = 2, ncol = 8)

my.diabet.test[1,] <- c(2, 76, 70, 23, 67, 34.7, 0.23, 22)

my.diabet.test[2,] <- c(0, 90, 88, 26, 80, 31.4, 0.63, 25)

my.diabet.test <- as.data.frame(my.diabet.test)

colnames(my.diabet.test) <- colnames(diabet.train)

my.diabet.test.prediction <- knn(train = diabet.train,

test = my.diabet.test,

cl = diabet.train.labels,

k = round(sqrt(train.size)))

my.diabet.test.prediction

Приложение

# Результаты прогноза пользовательских значений

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Diabetes Pedigree Function | Age | Outcome |
| 2 | 76 | 70 | 23 | 67 | 34.7 | 0.23 | 22 | 0 |
| 0 | 90 | 88 | 26 | 80 | 31.4 | 0.63 | 25 | 0 |

1. Не совсем ясно, в какой момент времени измерялось давление обследуемой. Скорее всего, на момент проведения тестов. [↑](#footnote-ref-1)
2. К сожалению, не удалось найти точной информации об этом показателе. Возможно, он показывает содержание инсулиновой сыворотки в крови по прошествии двух часов после проведения теста на инсулинорезсистентность. Однако, это лишь мои домыслы. [↑](#footnote-ref-2)