УДК 004.62

ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТОВ DATA SCIENCE В ПРОЦЕССЕ ПРОЕКТНОЙ УЧЕБНОЙ ДЕЯТЕЛЬНОСТИ

И. А. БЕККЕР, Р. И. КОЗЫРЕВ, В. С. ДАШКО
Белорусско-Российский университет
Могилев, Беларусь

База данных (БД) My Food Data представляет собой постоянно обновляемую интегрированную систему Министерства сельского хозяйства США, которая предоставляет расширенную информацию о питательных веществах и других компонентах пищевых продуктов и является надежным источником точных и качественных данных.

БД My Food Data используется в учебном научно-исследовательском DATA SCIENCE проекте Food for pleasure/Plate для ежедневного подсчета калорий (разрабатывается студентами Белорусско-Российского университета) в качестве встраиваемой в мобильное приложение фактографической БД.

Источник внедряемых данных – интернет-ресурс MyFoodData.com. БД представлена на нем в виде электронной таблицы с показателями для порции размером 100 г.

ID	Name	Food Group	Calories	Fat (g)	Protein (g)	Carbohydrate (g)	Sugars (g)
167512	Pillsbury Golden Layer Buttermilk Biscuits Artificial Flavor Refrigerated Dough	Baked Foods	307	13,2	5,88	41,18	5,88
167513	Pillsbury Cinnamon Rolls With Icing Refrigerated Dough	Baked Foods	330	11,3	4,34	53,42	21,34
167514	Kraft Foods Shake N Bake Original Recipe Coating For Pork Dry	Baked Foods	377	3,7	6,1	79,8	NULL
167515	George Weston Bakeries Thomas English Muffins	Baked Foods	232	1,8	8	46	NULL
167516	Waffles Buttermilk Frozen Ready-To-Heat	Baked Foods	273	9,22	6,58	41,05	4,3
167517	Waffle Buttermilk Frozen Ready-To-Heat Toasted	Baked Foods	309	9,49	7,42	48,39	4,41
167518	Waffle Buttermilk Frozen Ready-To-Heat Microwaved	Baked Foods	289	9,4	6,92	44,16	4,5
167519	Waffle Plain Frozen Ready-To-Heat Microwave	Baked Foods	298	9,91	6,71	45,41	5,04
167520	Pie Crust Cookie-Type Graham Cracker Ready Crust	Baked Foods	501	24,8	5,1	64,3	18,13
167521	Pie Crust Cookie-Type Chocolate Ready Crust	Baked Foods	484	22,4	6,08	64,48	26,31

Рис. 1. Поля и записи исходной таблицы

Поле *Name* содержит наименование конкретного национального продукта, который может быть нам несвойственен (по традициям питания, например) или не поставляться в наш регион. После этапа чистки их останется примерно 30 % от исходного количества, в БД – около 12 тыс. уникальных записей.

На этапе чистки данных, кроме удаления большого количества имен собственных наименований продуктов, еще предстоит выполнить перевод имени нарицательного на русский язык (автоматически программой-переводчиком с последующей проверкой корректности перевода, поскольку могут возникнуть вопросы, связанные с синонимией и т. д.).

В поле *Name* возможна ситуация, что из всех разновидностей йогурта (творога) ни одно из наименований не знакомо потребителям нашего региона (региона предполагаемого распространения программного продукта, в который встраивается обработанная БД продуктов питания). В этом случае разработчики предусмотрели градацию продукта по жирности, например, творог с 1-процентным содержанием жира, творог с 3-процентным содержанием жира и т. д. (без указания марки «Савушкин продукт», «Бабушкина крынка»).

Для йогурта дополнительно стоит учитывать наличие/отсутствие сахара.

Хлеб предполагается указать в виде *Пшеничный*, *Ржаной* – с указанием вида муки. Также дополнительно можно вводить параметр *Зерновой* (все без имен собственных).

Поле *Food Group* используется в разрабатываемом приложении для организации поиска по категориям продуктов. Группы продуктов из встраиваемой БД не изменялись: *Молоко и молочные продукты*; *Хлеб и хлебобулочные изделия*; *Жиры, масло и маргарин*; *Крупы*; *Овощи*; *Фрукты*; *Сухофрукты*; *Бобовые*; *Грибы*; *Мясо, птица и субпродукты*; *Колбаса*; *Мясные консервы и копчености*; *Яйца*; *Рыба свежая и морепродукты*; *Орехи*; *Сладости.*

Для поля *Calories* выбран тип целочисленного числа, т. к. с исходными значениями будут проводиться математические операции с возвращением на выходе целого числа в качестве значения поля.

Данные поля *Fat* (*g*) представлены в виде чисел с плавающей запятой, как и у полей *Protein* (*g*)*, Sugars* (*g*) и *Carbohydrate* (*g*).

Для каждого из числовых полей, содержащих количественные параметры БД (кроме счетчика ID), принято решение исключить значение Null и заменить его на нуль (операция деления алгоритмом не выполняется).

На этапе предварительного анализа данных, выявляя аномальные по логике вычислений данные, необходимо учитывать размер порции и содержание калорий в единице продукта: на 1 г белка приходится 4 ккал; на 1 г жира – 9 ккал; на 1 г углеводов – 4 ккал.

С учетом вышеприведенной алгоритмической зависимости общая калорийность продукта не должна превышать 9 ккал, умноженных на массу порции в граммах:

$$Общая\_калорийность\_продукта \le 9ккал * Масса\_порции .$$

Калорийность *Calories* эталонной порции (100 г) будет не более 900 ккал:

$$Калорийность\_продукта \le 9\ ккал * 100 .$$

Еще одним тестирующим критерием является сумма значений в каждом из полей *Fat* (*g*), *Protein* (*g*)*, Sugars* (*g*) и *Carbohydrate* (*g*), отличающаяся от веса пользовательской порции (в граммах):

$$Fat(g) + Protein\ (g) + Sugars\ (g) + Carbohydrate\ (g) + \varepsilon = Вес\_порции .$$

Таким образом, данные конкретно взятой записи будут проверяться на несоответствие друг другу и на нарушение логики вычислений.