

УДК 004.89:641.1

**АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ПИЩЕВЫХ ПРОДУКТОВ
ПО НУТРИЕНТНОМУ ПРОФИЛЮ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ**

А.В. Лемешевский, магистрант

Научный руководитель – **В.В. Шумак**, профессор, д.сх.н., доцент

Полесский государственный университет

Базы данных о составе пищевых продуктов (FCDB) объединяют разнородные источники [1, 2], однако единая категориальная разметка в них зачастую отсутствует, а ручная верификация при объемах свыше десяти тысяч записей является трудоёмкой. Цель работы – классификация пищевых продуктов по 32 категориям на основе нутриентного состава с использованием ансамблевых методов машинного обучения.

Информационную базу составили три источника: USDA FoodData Central [3], французская база CIQUAL и таблицы под редакцией И. М. Скурихина [2]. Совокупный массив включает 15 530 записей, из которых 12 373 использованы для обучения и 3 157 – для тестирования. Исходные 460 подкатегорий, представленные в различных языках и классификационных системах, были консолидированы в 32 укрупнённые группы: мясо, птица, рыба, молочные продукты, овощи, фрукты и другие категории.

Входные признаки организованы в виде трёх иерархических уровней (Таблица 1). Первый уровень включает шесть макронутриентов (энергетическая ценность, белки, жиры, углеводы, вода, пищевые волокна), обеспечивающих первичное разделение основных групп. Второй уровень представлен 14 микронутриентами, выступающими биологическими маркерами для различения смежных категорий. Третий уровень включает семь производных коэффициентов – отношений пар нутриентов с добавлением константы 0,1 в знаменатель

Таблица 1. – Структура признакового пространства

Уровень	Количество	Состав	Роль
I. Макронутриенты	6	Энергетическая ценность, белки, жиры, углеводы, вода, волокна	Первичная сепарация
II. Микронутриенты	14	Сахара, насыщ. жиры, холестерин, Na, K, Ca, Fe, Mg, P, Zn, вит. C, A, D, Se	Биологические маркеры
III. Коэффициенты	7	Белок/жир, белок/углев., жир/углев., сахар/углев., насыщ.жир/жир, Ca/P, Na/K	Масштабная инвариантность
Итого	27		

В рамках эксперимента оценены три ансамблевых алгоритма: Random Forest (200 деревьев, глубина 15), XGBoost (200 деревьев, глубина 8, скорость обучения 0,1) и LightGBM (300 деревьев, глубина 10, скорость обучения 0,05). Выбор градиентного бустинга обусловлен высокой долей пропущенных значений в нутриентных профилях: удаление неполных записей сокращает обучающую выборку с 12 373 до 5 128 (потеря 58 % данных). LightGBM нативно обрабатывает пропуски, учитывая их при построении деревьев решений, что позволило использовать полный объём данных без предварительной импутации.

Апробация проводилась в четыре итерации с последовательным наращиванием признаков и оптимизацией обработки данных (таблица 2). Качество оценивалось по метрикам Ассигасу (доля верно классифицированных объектов) и F1-макро (среднее гармоническое точности и полноты по всем категориям, учитывающее дисбаланс классов).

Первая итерация (6 макронутриентов, 67 категорий) дала точность 55,3 %, что значительно выше уровня случайного угадывания (1,5 %), однако недостаточно для практического применения. Вторая итерация (добавление 14 микронутриентов) обеспечила прирост +24,1 п.п., при этом минералы и витамины выступили биохимическими маркерами, разделяющими классы, неразличимые по макросоставу. Третья итерация (консолидация категорий 460 → 32) повысила точность до 85,7 %. На четвёртой итерации переход от удаления неполных записей к нативной обработке NaN в LightGBM [4] увеличил выборку в 2,4 раза; при снижении Ассигасу на 0,5 п.п. показатель F1-макро вырос с 0,703 до 0,721, что свидетельствует об улучшении качества распознавания малочисленных категорий.

Таблица 2. – Динамика метрик качества по итерациям (лучшая модель)

Итерация	Признаки	Категорий	Обучающая выборка	Пропущенные значения	Точность	F1-макро
1	6	67	8 740	Отбрасываются	55,3 %	0,435
2	20	52	5 086	Отбрасываются	79,4 %	0,698
3	20	32	5 128	Отбрасываются	85,7 %	0,703
4	27	32	12 373	Учитываются	85,2 %	0,721

Сравнительный анализ трёх алгоритмов на итерации с консолидированными категориями (Таблица 3) выявил устойчивое преимущество архитектуры LightGBM [4].

Таблица 3 – Сравнение алгоритмов (итерация 3, 32 категории)

Алгоритм	Точность	F1-macro	F1-weighted
Random Forest	85,1 %	0,697	0,837
XGBoost	84,9 %	0,696	0,841
LightGBM	85,7 %	0,703	0,849

Анализ метрик по категориям (таблица 4) показал, что наивысшее качество ($F1 > 0,90$) достигается для категорий с выраженными биохимическими маркерами: мясо (холестерин и кобаламин, $F1 = 0,975$), овощи (клетчатка и аскорбиновая кислота, $F1 = 0,926$), детское питание (фортификация, $F1 = 0,928$). Наибольшие затруднения возникают при классификации многокомпонентных готовых блюд ($F1 = 0,500$).

Таблица 4. – Категории с наивысшим и наименьшим качеством распознавания

Категория продуктов	F1	Ключевые маркеры
Мясо	0,975	Белок, холестерин, витамин B12
Мясные полуфабр.	0,945	Натрий, насыщенные жиры
Овощи	0,926	Вода, клетчатка, витамин С
Зерновые и крупы	0,920	Углеводы, низкий жир
Фрукты	0,915	Сахара, вит. С, калий
Готовые блюда	0,500	Суперпозиция ингредиентов
Вегетарианские продукты	0,462	Имитация мясного профиля

Методология имеет прикладное значение в контексте Доктрины национальной продовольственной безопасности Республики Беларусь до 2030 года [5]: автоматизированная категоризация обеспечивает оперативную верификацию данных при наполнении национальных баз пищевой ценности и позволяет выявлять аномалии в составе продуктов как инструмент контроля фальсификации сырья (например, присутствие растительных жиров в молочных продуктах по аномальному соотношению насыщенных жиров и кальция/фосфор).

По итогам исследования предложен подход к классификации пищевых продуктов по 32 категориям на основе трёхуровневой архитектуры из 27 нутриентных параметров, обеспечивающий точность 85,7 % ($F1\text{-macro} = 0,721$). Алгоритм LightGBM продемонстрировал преимущество над Random Forest и XGBoost за счёт нативной обработки пропусков, что позволило увеличить обучающую выборку в 2,4 раза и повысить точность на 30,4 п.п. Введение микронутриентов обеспечило прирост точности на 24,1 п.п. по сравнению с моделью на макронутриентах. Сформирована система масштабно-инвариантных коэффициентов, повышающая устойчивость классификации переработанных продуктов. Показана применимость подхода для автоматизации национальных баз пищевой ценности и задач контроля фальсификации продовольственного сырья в Республике Беларусь.

Список использованных источников

1. Бессонов, В. В. Базы данных химического состава пищевых продуктов в эпоху цифровой нутрициологии / В. В. Бессонов [и др.] // Вопросы питания. – 2020. – № 4. – С. 7–17.
2. Скурихин, И. М. Таблицы химического состава и калорийности пищевых продуктов : справочник / И. М. Скурихин, В. А. Тутельян. – М. : ДеЛи принт, 2008. – 276 с.
3. FoodData Central [Электронный ресурс] / U.S. Department of Agriculture. – Режим доступа: <https://fdc.nal.usda.gov>. – Дата доступа: 10.04.2026.
4. Ke, G. LightGBM: A Highly Efficient Gradient Boosting Decision Tree / G. Ke [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 3146–3154.
5. Доктрина национальной продовольственной безопасности Республики Беларусь до 2030 года : постановление Совета Министров Респ. Беларусь от 15.12.2017 № 962.