

摘要

特殊医学用途配方食品作为一类专为满足特定健康状态下人群营养需求而设计的食品 [1]，在医学营养干预与治疗中占据举足轻重的地位。鉴于其重要性，国内对特医食品的审批与注册流程实施了严格管理 [2]，截至 2024 年 4 月，仅有 182 款特医食品成功通过审批 [3]。本研究致力于深入挖掘这 182 款特医食品的产品信息，通过细致分析产品标签与说明书数据，旨在全面揭示特医食品的生产概况，并运用可视化手段进行直观展示。在此基础上，将进一步探索构建特医食品推荐系统的可行性，为临床决策与患者营养管理提供更为精准、个性化的科学依据，具有重要的实际意义。

针对任务一：由于每款特医食品说明书是.pdf 的后缀，无法对其中的营养成分表直接进行数据提取与处理吗，所以首先调用 python 的 pdfplumber 包设计算法直接对 PDF 中的内容进行读取，其中重点使用 `page.extract_text` 方法提取 PDF 中每一页的内容。然后分别根据各个分任务的要求，首先将提取的营养成分数据写入文件 `result1.xlsx`；再提取 182 款特医食品说明书中【产品类别】、【组织状态】、【适用人群】和【注册证号】的数据，根据各个分任务的要求写入到 `result2.xlsx`，并在报告正文中列出部分数据加以说明。

针对任务二：首先对两种产品来源分别根据生产年份归类计数，得到 2017 年至 2024 年每一年两种来源产品的获批量，利用 ECharts 绘制双折线图，通过分析得到结论：在 2023 年，国产产品的年获批量达到高峰；而进口产品最近 5 年的获批量均不超过 2 个。任务 2.2 先以适用人群类别分类，在两种适用人群类别分类中各自统计产品来源，然后利用 ECharts 绘图工具绘制出旭日图。其中，1 岁以上特医食品中，绝大多数都是国产产品，而特医婴配食品中，国产产品和进口产品占比均为 50% 左右。在任务 2.3 中先对任务 1 中得到的数据根据产品类别进行分类计数，然后使用 `matplotlib.pyplot` 绘制柱状图，其中全营养配方食品和非全营养配方产品获批量最多。对于任务 2.4，使用任务 1.1 得到的各特医食品营养含量分布图，对脂肪和蛋白质两列分别使用 `matplotlib.pyplot` 的 `hist` 方法绘制直方图。最后为了更好的概括适用人群特征，我们去除了“适用人群”信息的重复项，得到 182 款特医食品适用人群的完整特征汇总具体的词云图。其中通用特征为脱水状态：快速补充水和电解质；营养不足：进食受限或有代谢障碍的情况下需补充全面营养。

针对任务三：使用任务 1 中整理得到的有关数据，利用 `pandas` 包的合并操作，合并并在推荐系统中需要考虑的 182 款特医食品的相关信息制成训练数据集，基于 kimi 大语言模型，利用适合的 `prompt` 使大语言模型学习训练数据集，基于数据集进行特意药品的推荐。完成训练后，使用任务 3 指定的客户需求进行训练模型的测试，对推荐模型的输出进行人工检查，得出推荐模型的输出是合理的，构造的智能推荐系统出色地完成了工作。

目录

| | |
|-----------------------------------|----|
| 一、问题重述 | 3 |
| 1.1 问题背景 | 3 |
| 1.2 问题重述 | 3 |
| 二、思路分析 | 3 |
| 三、任务一的求解 | 4 |
| 3.1 任务 1.1 的求解 | 4 |
| 3.2 任务 1.2 的求解 | 6 |
| 3.3 任务 1.3 的求解 | 8 |
| 3.4 任务 1.4 的求解 | 9 |
| 四、任务二的求解 | 9 |
| 4.1 登记年份和产品来源与获批量的可视化分析 | 9 |
| 4.2 人群类别与产品来源的可视化分析 | 10 |
| 4.3 产品类别与获批量的可视化分析 | 11 |
| 4.4 脂肪和蛋白质含量的可视化分析 | 12 |
| 4.5 特医食品适用人群特征的可视化分析 | 13 |
| 五、任务三的求解 | 14 |
| 参考文献 | 17 |

一、问题重述

1.1 问题背景

在特殊营养需求领域，特殊医学用途配方食品扮演着关键角色，专为满足进食受限、消化障碍、代谢紊乱或特定疾病状态人群的营养需求而精心配制，涵盖从 0 月龄至 12 月龄婴儿及 1 岁以上人群的特殊配方。其在医学营养管理与治疗中发挥不可或缺的作用。

为确保产品安全与有效，特殊医学用途配方食品需历经严格的审批与注册流程，我国对此类食品的审核标准极为严苛，因此，对每一款特医食品的信息精准掌握显得尤为关键。

1.2 问题重述

data.xlsx 记录了 182 款特医食品的基本信息; 特医食品说明书文件夹中包含 182 款特医食品说明书;result 文件夹中包含了需要提交的模板文件。

本文基于这些上述数据，通过统计分析，从各种角度完成以下任务：

1. 提取 182 款特殊医学用途配方食品产品标签、说明书中的相关数据，并对其进行数据预处理。
2. 对计 182 款特医食品生产概况进行可视化分析。
3. 使用推荐算法构建特医食品推荐系统。

二、思路分析

根据题目中所给的三个任务，本文概括出的解题思路与答大体步骤如图1所示：

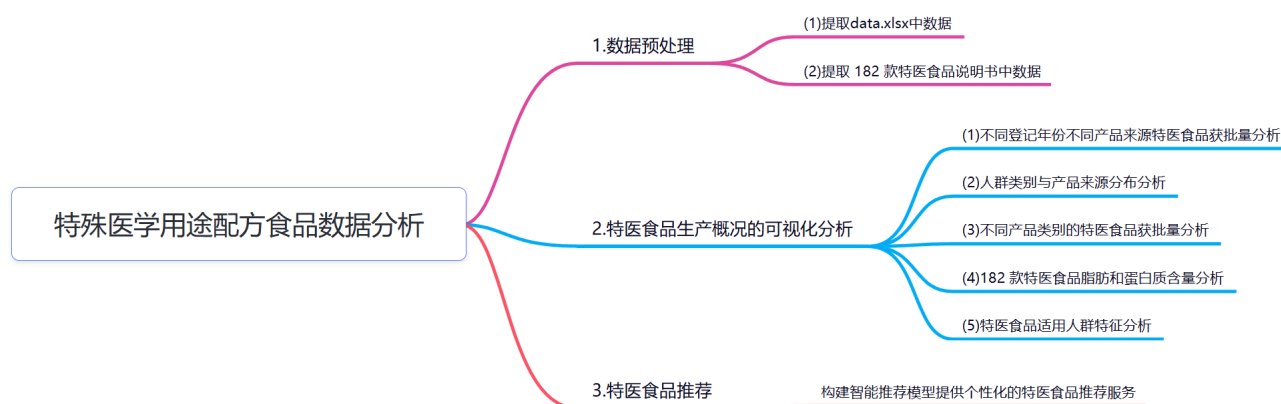


图 1 解题流程的思维导图

基于上述思维导图，可知本文将题目所给任务总体以下三步，首先对 data.xlsx 和 182 款特医食品说明书中数据进行提取和清洗，即数据预处理环节。其次，基于整理与统计的数据，对不同登记年份不同产品来源的特医食品获批量、人群类别与产品来源分布、不同产品类别的特医食品获批量、特医食品脂肪和蛋白质含量和特医食品适用人群特征进行可视化分析，即任务二。最后基于任务一和任务二得到的数据，构建智能推荐模型提供个性化的特医食品推荐服务，即任务三。

三、 任务一的求解

3.1 任务 1.1 的求解

首先通过对特医食品说明书进行查看可以发现：

- 共有 182 款特医食品说明书，数据量庞大。
- 每款特医食品说明书是.pdf 的后缀，无法对其中的营养成分表直接进行数据提取与处理。
- 由于以上两点，若所有特医食品说明书的 PDF 转成 excel 表格再进行数据清洗，工作量庞大，可实现性不高。

基于上述附件中的特征，所以考虑使用编程方法设计可以直接对 PDF 中进行读取和处理的算法。我们调用 python 的 pdfplumber 包，用 pdfplumber 的 open 方法读取指定目录下的特医食品说明书。对于我们需要提取的“注册号”数据项，可以通过获取文件名并且去掉后缀名的方法获得。对于营养成分表的提取，我们首先通过 page.extract_text 方法提取 PDF 中每一页的内容，然后调用 python 字符串查找的方法查找“营养成分表”关键字。当查找到此关键字时，说明此页就是第一个包含营养成分表的页。从这一页开始，调用 extract_tables 方法来提取表格内容，按照上述步骤，依次提取每一个特医食品说明书中的文件夹名和表格内容，

对于提取结果的存储，新建一个 DataFrame 来存储。对于每次提取到的结果，调用 pd.concat 方法来将提取到的内容追加到 DataFrame 中。当完成所有数据的提取后，调用 to_excel 方法将所有的内容将数据写入 result1.xlsx 中。之后调用 nlargest 方法按照蛋白质含量对表格内容最高进行排序，并取蛋白质含量最高的三个特医食品。每 100kJ 中蛋白质含量最高的三种特医食品如表1所示，python 部分关键代码具体如下。

```
with pdfplumber.open(pdf_file) as pdf:
    start_extracting = False # 标记是否开始提取表格
    nutrition_tables = [] # 用于保存营养成分表的所有数据
    table_header = None # 存储表头，仅在第一次提取表格时使用
    # 获取pdf_file文件名不含后缀不含上级目录
    pdf_file_name = pdf_file.split('/')[-1].split('.')[0]
    for page_num, page in enumerate(pdf.pages):
```

```

text = page.extract_text()
# 查找“【营养成分表】”关键字
if "【营养成分表】" in text and not start_extracting:
    start_extracting = True
    # print(f"在第 {page_num + 1} 页找到标签【营养成分表】")
    # 提取表格内容
if start_extracting:
    tables = page.extract_tables()
    for table in tables:
        # 如果是第一页，设置表头
        if table_header is None:
            # if "营养成分" in table[0][0]:
            # 如果 table[0][0] 完全等于 '营养成分'
            if table[0][0] == '营养成分':
                table_header = table[0] # 第一行作为表头
                nutrition_tables.extend(table[1:]) # 添加数据部分
        else:
            # 处理跨页表格，跳过表头行（如果存在）
            if len(table[0]) == len(table_header):
                if table[0] == table_header:
                    nutrition_tables.extend(table[1:]) # 表头重复，跳过第一行
            else:
                nutrition_tables.extend(table) # 没有表头，直接添加所有行

# 定义营养成分列表
nutrients = ["能量(kJ)", "脂肪(g)", "碳水化合物(g)", "蛋白质(g)", "钠(mg)", "氯(mg)",
             "钾(mg)", "磷(mg)"]
# 创建一个空的DataFrame来存储结果
result_df = pd.DataFrame(columns=["注册号"] + nutrients)

# 创建一个新的DataFrame行
row_data = pd.DataFrame([[register_no] + [nutrition_data[nutrient] for nutrient in
    nutrients]], columns=result_df.columns)
# 使用pd.concat追加数据
result_df = pd.concat([result_df, row_data], ignore_index=True)

# 把result_dfb保存到Excel文件,命名为result1.xlsx
for col in result_df.columns[1:]: # 假设第一列是非数值列，可以选择性地排除
    result_df[col] = pd.to_numeric(result_df[col], errors='coerce')
result_df.to_excel('result1.xlsx', index=False)

# 读入result1.xlsx
df_good_nutrition = pd.read_excel('result1.xlsx')
top_3_protein = df_good_nutrition.nlargest(3, '蛋白质(g)')

```

| 注册证号 | 能量 kJ | 脂肪 g | 碳水化合物 g | 蛋白质 g | 钠 mg | 氯 mg | 钾 mg | 磷 mg |
|-----------------|-------|------|---------|-------|------|------|------|------|
| 国食注字 TY20230045 | 100 | 0.04 | 0.007 | 5.9 | 45.0 | 0.0 | 0.0 | 0.0 |
| 国食注字 TY20220002 | 100 | 0.00 | 0.100 | 5.8 | 80.4 | 0.0 | 0.0 | 0.0 |
| 国食注字 TY20230062 | 100 | 0.00 | 0.080 | 5.8 | 24.0 | 0.0 | 0.0 | 0.0 |

表 1 每 100kJ 中蛋白质含量最高的三种特医食品表

3.2 任务 1.2 的求解

(1) 特殊情况处理

由于不同说明书 pdf 中可能有格式、表述的差异，会增加后续数据汇总整理的难度。我们调用了正则化函数 `re.sub()`，用于去除”组织状态“、“适用人群“中不必要或错误的符号，并且对含义相同但表述不同的部分进行了统一。处理前后二者对比如图2所示，左图是处理之前的适用人群，右图为处理之后的适用人群。

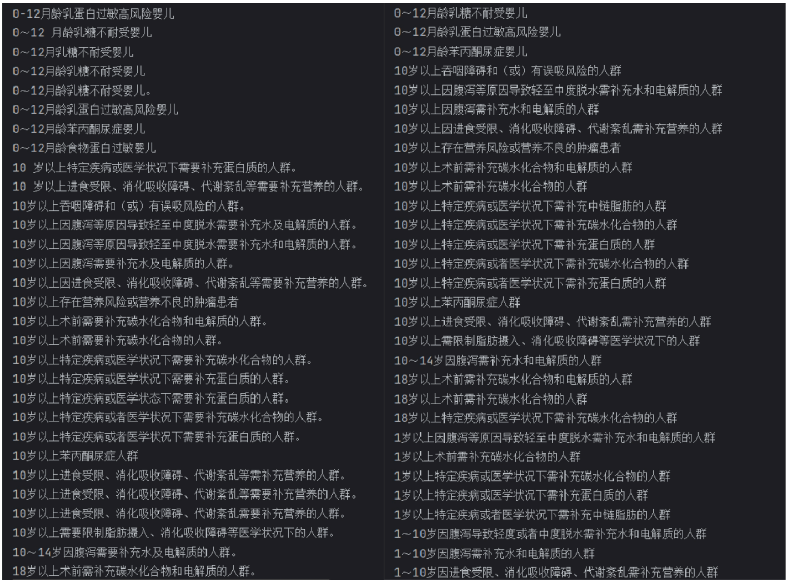


图 2 特殊处理对比图

(2) 分析与求解

在任务 1.1 求解的基础上，我们同样可以调用 python 的 `pdfplumber` 包，然后遍历指定文件夹中的所有 PDF 文件，通过使用正则表达式分别查找“【产品类别】”等内容来提取各个 PDF 中的数据，然后将统计到的数据写入 `result2.xlsx` 中。前 5 款特医食品的具体数据表2所示，python 部分关键代码具体如下。

```
def extract_product_category(pdf_path):
    try:
        # 使用pdfplumber打开PDF文件
        with pdfplumber.open(pdf_path) as pdf:
            full_text = ""
            # 从每一页提取文本
```

```

for page in pdf.pages:
    full_text += page.extract_text() or ""

# 使用正则表达式查找“【产品类别】”和“【”之间的内容
# match = re.search(r"【产品类别】(.*)【", full_text, re.S)
# match = re.search(r"【组织状态】(.*)【", full_text, re.S)
match = re.search(r"【适用人群】(.*)【", full_text, re.S)
if match:
    product_category = match.group(1).strip()
    # 去除空格和句号
    product_category = re.sub(r"[。]", "", product_category)
    product_category = re.sub(r"适用于", "", product_category)
    product_category = re.sub(r"[-~]", "~", product_category)
    product_category = re.sub(r"状态", "状况", product_category)
    product_category = re.sub(r"等需要|等需|需要进行|需要", "需", product_category)
    product_category = re.sub(r"及", "和", product_category)
    product_category = re.sub(r"0~12月龄食物蛋白过敏婴儿", "食物蛋白过敏婴儿",
        product_category)
    product_category = re.sub(r"营养补充", "补充营养", product_category)
    product_category = re.sub(r"岁进", "岁因进", product_category)
    product_category = re.sub(r"\n+", "", product_category)
    return product_category
return None
except Exception as e:
    print(f"读取 {pdf_path} 时出错: {e}") # 处理错误
    return None

# 遍历指定文件夹中的所有PDF文件
for file_name in os.listdir(folder_path):
    if file_name.endswith('.pdf'):
        pdf_path = os.path.join(folder_path, file_name)
        # 从PDF提取产品类别
        product_category = extract_product_category(pdf_path)

        if product_category:
            # 使用文件名(去掉扩展名)作为注册证号
            registration_number = os.path.splitext(file_name)[0]

            # 更新Excel中对应的行
            for row in sheet.iter_rows(min_row=2): # 跳过表头行
                if row[reg_col_idx].value == registration_number:
                    row[product_col_idx].value = product_category
                    break

# 保存更改到Excel文件
workbook.save(xlsx_path)

# 执行更新函数
update_excel_with_category(excel_file_path, pdf_folder_path)

```

| 序号 | 企业名称 | 产品名称 | 注册证号 | 有效期至 | 产品类别 | 组织状态 | 适用人群 |
|----|--------------------------|------------------------|-----------------|------------|-----------|------|-------------------|
| 1 | SHS INTERNATIONAL LTD | 纽康特特殊医学用途婴儿氨基酸配方食品 | 国食注字 TY20175001 | 2027/10/13 | 氨基酸配方 | 粉状 | 食物蛋白过敏婴儿 |
| 2 | ABBOTT LABORATORIES S.A | 雅培亲护特殊医学用途婴儿乳蛋白部分水解配方粉 | 国食注字 TY20175002 | 2027/10/13 | 乳蛋白部分水解配方 | 粉状 | 0~12 月龄乳蛋白过敏高风险婴儿 |
| 3 | ABBOTT LABORATORIES S.A. | 菁挚呵护特殊医学用途婴儿乳蛋白部分水解配方粉 | 国食注字 TY20175003 | 2022/11/19 | 乳蛋白部分水解配方 | 粉状 | 0-12 月龄乳蛋白过敏高风险婴儿 |
| 4 | 杭州贝因美母婴营养品有限公司 | 贝因美特殊医学用途婴儿无乳糖配方食品 | 国食注字 TY20180001 | 2028/1/17 | 无乳糖配方 | 粉状 | 0~12 月龄乳糖不耐受婴儿 |
| 5 | SHS INTERNATIONAL LTD | 纽贝瑞特殊医学用途婴儿苯丙酮尿症配方粉 | 国食注字 TY20185001 | 2028/1/17 | 氨基酸代谢障碍配方 | 粉状 | 0~12 月龄苯丙酮尿症婴儿 |

表 2 前 5 款特医食品的具体数据表

3.3 任务 1.3 的求解

通过任务 1.2 的数据预处理过后，我们可以阅读 result2.xlsx 发现关于适用人群的描述很多，例如：食物蛋白过敏婴儿、早产/低出生体重婴儿、消化吸收不足造成营养不良的 1~10 岁人群、10 岁以上因腹泻等原因导致轻至中度脱水需要补充水及电解质的人群等等。通过观察可以发现适用人群为 0-12 月龄人群的都含有“婴儿”二字；适用人群为 1 岁让以上的都不含有“婴儿”二字。因此我们对适用人群类别中特医婴配食品和 1 岁以上特医食品量中两种类别的区分，只需要检索适用人群中是否含有“婴儿”二字即可。

根据上述的分析，我们只需要读入 result2.xlsx 文件，并遍历检索对应数据项，即适用人群的数据中是否含“婴儿”二字，然后在 result2.xlsx 中新建一个新的列，列名为‘适用人群类别’；再调用 apply 方法将计算‘适用人群类别’列的值的方法 categorize_population 应用到此列上。结果的部分数据如表3所示,python 部分关键代码具体如下。

```
def categorize_population(row):
    if '婴儿' in row['适用人群']:
        return '特医婴配食品'
    else:
        return '1岁以上特医食品'

df_info['适用人群类别'] = df_info.apply(categorize_population, axis=1)
df_info.head()
```

| 序号 | 企业名称 | 产品名称 | 注册证号 | 有效期至 | 产品类别 | 组织状态 | 适用人群 | 适用人群类别 |
|----|--------------------------|------------------------|-----------------|------------|-----------|------|-------------------|--------|
| 1 | SHS INTERNATIONAL LTD | 纽康特特殊医学用途婴儿氨基酸配方食品 | 国食注字 TY20175001 | 2027/10/13 | 氨基酸配方 | 粉状 | 食物蛋白过敏婴儿 | 特医婴配食品 |
| 2 | ABBOTT LABORATORIES S.A | 雅培亲护特殊医学用途婴儿乳蛋白部分水解配方粉 | 国食注字 TY20175002 | 2027/10/13 | 乳蛋白部分水解配方 | 粉状 | 0~12 月龄乳蛋白过敏高风险婴儿 | 特医婴配食品 |
| 3 | ABBOTT LABORATORIES S.A. | 菁挚呵护特殊医学用途婴儿乳蛋白部分水解配方粉 | 国食注字 TY20175003 | 2022/11/19 | 乳蛋白部分水解配方 | 粉状 | 0-12 月龄乳蛋白过敏高风险婴儿 | 特医婴配食品 |
| 4 | 杭州贝因美母婴营养品有限公司 | 贝因美特殊医学用途婴儿无乳糖配方食品 | 国食注字 TY20180001 | 2028/1/17 | 无乳糖配方 | 粉状 | 0~12 月龄乳糖不耐受婴儿 | 特医婴配食品 |
| 5 | SHS INTERNATIONAL LTD | 纽贝瑞特殊医学用途婴儿苯丙酮尿症配方粉 | 国食注字 TY20185001 | 2028/1/17 | 氨基酸代谢障碍配方 | 粉状 | 0~12 月龄苯丙酮尿症婴儿 | 特医婴配食品 |

表 3 结果的部分数据表

3.4 任务 1.4 的求解

由于题目告知我们，特殊医学用途配方食品注册号的格式为：国食注字 TY+4 位年号 +4 位顺序号，顺序号第 1 位数字为“5”表示该食品为进口产品，顺序号第 1 位数字为“0”表示该食品为国产产品；4 位年号为该食品的登记年份，所以我们只需要在任务 1.3 预处理之后的 result2.xlsx 的基础上，查找注册证号中的第 7 位到第 11 位即可。

根据上述的分析，我们只需要读入 result2.xlsx 文件，并遍历检索对应数据项，即判断注册证号中的第 11 位为 5 还是位 0 和获取注册证号中的 7-10 位。然后与任务 1.3 的编程思路一样，新增“产品来源”和“登记年份”，将检索到的数据填入其中。前 5 款特医食品任务 1.3 和任务 1.4 的结果表如表4所示,python 部分关键代码具体如下。

```
def extract_info_from_register_no(register_no):
    year = register_no[6:10]
    source = "进口产品" if register_no[10] == '5' else "国产产品"
    return year, source

df_info['登记年份'], df_info['产品来源'] =
    zip(*df_info['注册证号'].apply(extract_info_from_register_no))
df_info.head()
```

| 序号 | 企业名称 | 产品名称 | 注册证号 | 有效期至 | 产品类别 | 组织状态 | 适用人群 | 适用人群类别 | 登记年份 | 产品来源 |
|----|--------------------------|------------------------|-----------------|------------|-----------|------|-------------------|--------|------|------|
| 1 | SHS INTERNATIONAL LTD | 纽康特特殊医学用途婴儿氨基酸配方食品 | 国食注字 TY20175001 | 2027/10/13 | 氨基酸配方 | 粉状 | 食物蛋白过敏婴儿 | 特医婴配食品 | 2017 | 进口产品 |
| 2 | ABBOTT LABORATORIES S.A | 雅培亲护特殊医学用途婴儿乳蛋白部分水解配方粉 | 国食注字 TY20175002 | 2027/10/13 | 乳蛋白部分水解配方 | 粉状 | 0~12 月龄乳蛋白过敏高风险婴儿 | 特医婴配食品 | 2017 | 进口产品 |
| 3 | ABBOTT LABORATORIES S.A. | 菁挚呵护特殊医学用途婴儿乳蛋白部分水解配方粉 | 国食注字 TY20175003 | 2022/11/19 | 乳蛋白部分水解配方 | 粉状 | 0-12 月龄乳蛋白过敏高风险婴儿 | 特医婴配食品 | 2017 | 进口产品 |
| 4 | 杭州贝因美母婴营养品有限公司 | 贝因美特殊医学用途婴儿无乳糖配方食品 | 国食注字 TY20180001 | 2028/1/17 | 无乳糖配方 | 粉状 | 0~12 月龄乳糖不耐受婴儿 | 特医婴配食品 | 2018 | 国产产品 |
| 5 | SHS INTERNATIONAL LTD | 纽贝瑞特殊医学用途婴儿苯丙酮尿症配方粉 | 国食注字 TY20185001 | 2028/1/17 | 氨基酸代谢障碍配方 | 粉状 | 0~12 月龄苯丙酮尿症婴儿 | 特医婴配食品 | 2018 | 进口产品 |

表 4 前 5 款特医食品任务 1.3 和任务 1.4 的结果表

四、任务二的求解

4.1 登记年份和产品来源与获批量的可视化分析

不同登记年份不同产品来源的特医食品获批量如图3所示。

该图表通过双折线图的形式展示了从 2017 年至 2024 年国产和进口的特医食品获批量的变化情况。

从图中可以看出，国产的特医食品获批量在 2017 年数量为 0，这表明在该年度没有国产特医食品获批；从 2017 年开始，国产特医食品的数量逐年增加；2022 年到 2023 年，国产的特医食品获批量飞速上升并达到峰值 68，但在 2024 年下降到 21。进口的特医食品获批量在 2017 年-2018 年有小幅度的上升，并且到达峰值 12；数字持平了 1 年以后，从 2019 年开始，数量呈现出逐年下降的趋势;2022-2024 年总体上有小幅度的上升，但趋势并不明显。

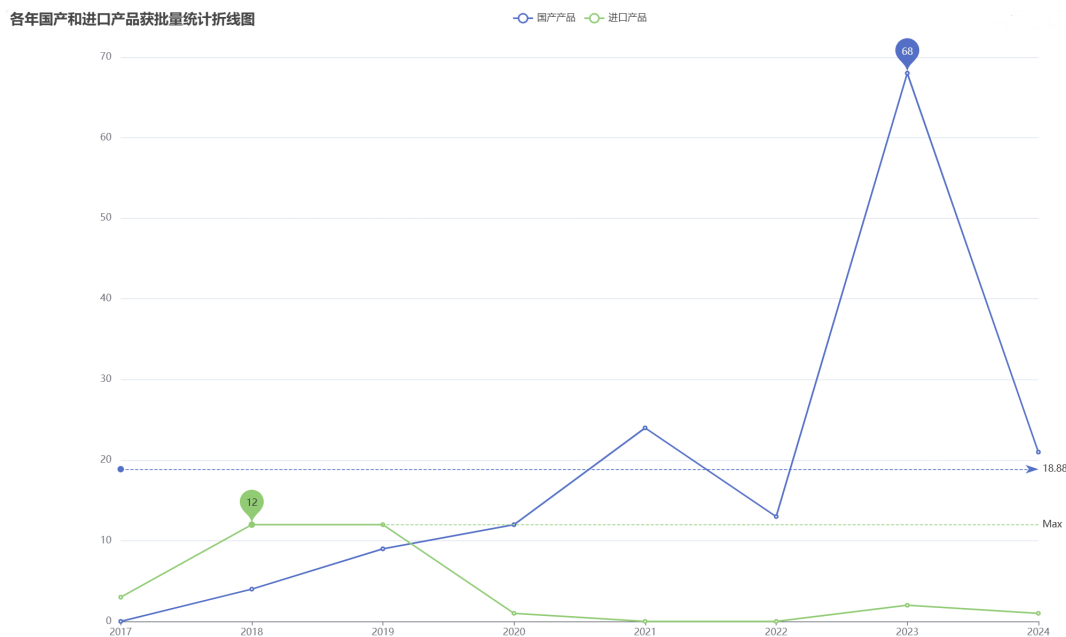


图3 不同登记年份不同产品来源的特医食品获批量图

国产特医食品和进口特医食品的数量变化呈现出相反的趋势，总体上国产特医食品在逐年上升，而进口特医食品在逐年下降；国产特医食品的最大获批量为 68，而进口特医食品最大获批量只为 12；在 2017 年-2024 年中，国产特医食品的年均获批量能够得到 18.88，高于进口特医食品的最大获批量。

4.2 人群类别与产品来源的可视化分析

特医食品产品来源与适用人群类别的旭日图如4所示。

内层为饼图的旭日图清晰地展示了特医食品产品的不同来源以及它们所适用的主要人群类别。从图中我们可以得出以下几点关键分析：

- 从特医食品适用人群类别的数量上来看，1 岁以上特医食品要明显多于特医婴配食品，约能占全部特医食品数量的四分之三。
- 特医食品产品来源的数量上来看，国产产品的数量要明显多于进口产品的数量，并且数量能够达把八成左右。
- 将特医婴配食品的进口与国产产品比例与 1 岁以上特医食品的进口与国产产品作比较可以发现，特医婴配食品的进口产品比例略少于国产产品，相差不大；而 1 岁以上特医食品的进口产品比例远低于国产产品的比例。
- 基于第三点，说明特殊医学用途配方食品审批机构在特医婴配食品方面对国产产品和进口产品的认可度相差不大；但在与 1 岁以上特医食品方面，特殊医学用途配方食品审批机构对国产产品的认可度要远大于进口产品。

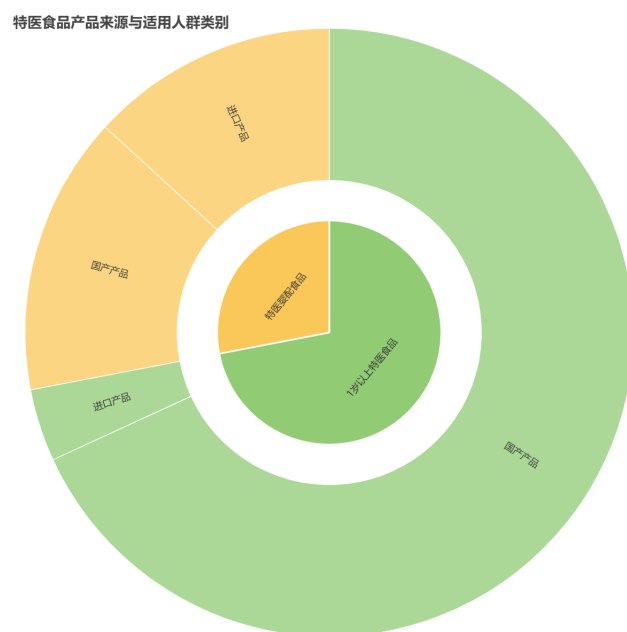


图 4 特医食品产品来源与适用人群类别的旭日图

4.3 产品类别与获批量的可视化分析

任务 2.3 需要统计不同产品类别的特医食品获批量。基于任务 1.2 统计的包含”产品类别”的统计表，对产品类别字段分类计数，得到一个产品类别以及对应获批量的统计表。基于获批量进行降序排列，使用 pyplot 工具绘制柱状图。两条生产线各故障类别的占比双层环形图如图5所示。

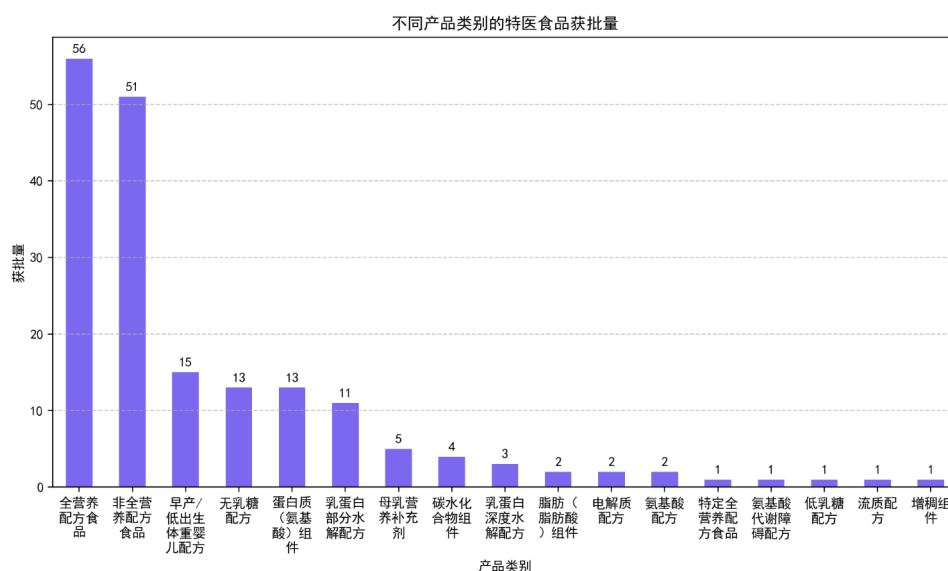


图 5 两条生产线各故障类别的占比双层环形图

柱状图显示，全营养配方食品和非全营养配方产品获批量最多，超均超过 50 项。占比超过一半。其次是无乳糖配方、蛋白质组件等配方产品，面向对特定营养配方有需求的人群。氨基酸配方、电解质配方、增稠组件等类型产品获批量仅为 1-2 项。从该图标

中，可以看出我国某些类型特医食品获批量小，全国仅有 1 件获批食品，鉴于我国人口基数大的情况，可能还需要有更多的产品进入市场，以满足特定人群多样化的饮食需求。

4.4 脂肪和蛋白质含量的可视化分析

基于任务 1.1 中获得的每百千焦营养成分数据，分别提取了脂肪和蛋白质的含量。利用 matplotlib.pyplot 的 hist 方法生成频数分布直方图，并采用不同颜色区分两种成分，使它们呈现在同一图表上，便于对比观察。两条生产线全年故障发生持续时长的叠加直方图分别如图6所示。

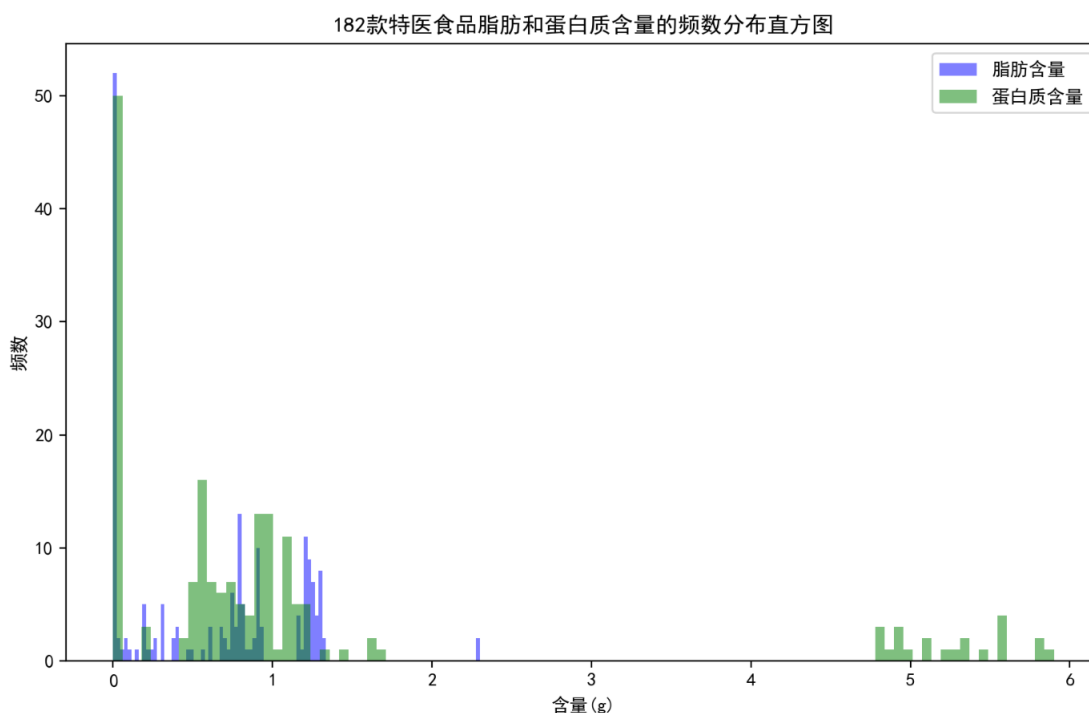


图 6 特医食品脂肪和蛋白质含量图

从图中我们可以看出，无论是脂肪还是蛋白质，含量为零的频次均出现较高的柱状，这表明多数特医食品在设计上是低脂、低蛋白的。低含量的集中可能源于特医食品对特定病患的需求，比如一些需严格控制脂肪摄入的患者群体。此外，我们也观察到脂肪和蛋白质的含量普遍分布在每百千焦 1-2 克之间，这是较常见的剂量区间，可能适合普通人群的营养需求，满足日常摄入的适量补充。

进一步分析显示，在蛋白质含量方面，有不少产品的蛋白质含量超过每百千焦 5 克，这类产品通常被设计为高蛋白补剂，适用于需要额外蛋白质补充的人群，例如术后康复患者、营养不良或高体力需求的病患。然而，脂肪含量并未出现类似的专门补剂产品，这种差异可能源于蛋白质和脂肪在营养摄入中的作用不同。蛋白质是人体必需的营养成分，且个别人群可能因疾病或体质特性对其需求较高，而脂肪的需求在特医食品中则更为谨慎，尤其考虑到脂肪摄入过量可能带来的健康风险。

此外，蛋白质在特医食品中的高含量频次还可能与蛋白质过敏的现象有关。在设计特医食品时，厂家需对成分进行合理调整，以适应不同患者的个体化需求。这种区别也反映出特医食品在成分设计上的个性化与精细化，表明其对不同营养需求和健康考量的权衡。

4.5 特医食品适用人群特征的可视化分析

为了更好的概括适用人群特征，我们去除了“适用人群”信息的重复项，并调用正则化函数 `re.sub()` 滤除含义相同但形式不同的项，最后得到 182 款特医食品适用人群的完整特征汇总具体的词云图如图7所示。



图 7 特医食品的“适用人群”词云图

特医食品适用人群可概括为以下特征：

- 1. 婴幼儿人群（0~12 月龄）
 - 乳糖不耐受或乳蛋白过敏：需要特殊配方奶粉，避免过敏反应；
 - 苯丙酮尿症（PKU）：需低苯丙氨酸饮食；
 - 早产或低出生体重婴儿：需要热量和营养密集的配方。
- 2. 儿童人群（1~10 岁）
 - 消化吸收问题或代谢紊乱：需要易吸收的营养配方；
 - 腹泻导致脱水：需要水和电解质补充；

- PKU：需要持续低苯丙氨酸饮食。

3. 青少年与成人（10 岁以上）

- 营养不足：进食受限或有代谢障碍的情况下需补充全面营养；
- 术前/后营养支持：补充碳水化合物和电解质；
- 特定疾病：某些情况下需补充中链脂肪、蛋白质或碳水化合物 [4]。

4. 老年人（50 岁以上）

- 消化吸收障碍：需补充全面营养防止营养不良；
- 营养不足：进食受限或有代谢障碍的情况下需补充全面营养。

5. 通用特征

- 脱水状态：快速补充水和电解质；
- 代谢性疾病：需要定制化的营养干预以避免加重病情。

特医食品设计针对不同年龄段和病理情况下的特定营养需求，给予个性化的营养支持。

五、任务三的求解

任务三需要基于已有的数据，构建一个智能推荐模型，实现为特定人群的情况推荐合适的特医产品。特定人群可能面临某种营养物质过敏，对某种营养物质需求大，或是对产品组织状态等有要求，每一种特医产品也有针对的适用人群，因此，推荐系统需要对特医产品的适用人群、适用人群类别，各种营养成分的含量，产品类别，组织状态等进行综合的考量，以针对客户的需求做出准确推荐。

实现上，采用 kimi 大模型基于数据集进行训练，构建推荐系统。基于已有的数据，构建出合适的训练数据集，与适当的 prompt 一同输入 kimi 大模型，可以使大模型学习数据集，实现推荐系统的功能。具体步骤如下：

STEP1 数据集的构建：由于推荐系统需要综合考虑产品的类别、组织状态、使用人群等基本信息，对于不同人的营养需求，还需要考虑产品的营养成分含量。因此，将任务 1.1 整理得到的表，与任务 1.4 整理得到的表进行合并，选取注册证号、产品名称、适用人群、适用人群类别、组织状态、产品类别、产品来源，以及碳水、蛋白质等七个营养物质的每 100kJ 含量，输出成为新的表，作为训练数据集。数据集内容具体如图8所示。

STEP2 采用适当的 prompt：使 kimi 大模型完成读入训练数据集进行学习，成为一个推荐模型。使用的 prompt 为：“这是一份包括 182 种特医产品的基本信息，适用人群，组织状态，产品类别，以及提供每 100kJ 能量时，几种主要营养物质的含量。我需要你学习这份数据，扮演一个基于这个数据的推荐模型，当我想你描述客户的状况和需求

| 产品名称 | 适用人群 | 用人群类 | 组织状态 | 产品类别 | 产品来源 | 能量(kJ) | 脂肪(g) | 碳水化合物(g) | 蛋白质(g) | 钠(mg) |
|---------------------------|-------------------------|------|------|------|------|--------|-------|----------|--------|-------|
| 纽康特特殊医学用途婴儿氨基酸配方食品 | 食物蛋白过敏婴儿 | 特医婴配 | 粉状 | 氨基酸配 | 进口产品 | 100 | 1.2 | 2.5 | 0 | 10 |
| 雅培亲护特殊医学用途婴儿乳蛋白部分水解配方粉 | 0~12月龄乳蛋白过敏高风险婴儿 | 特医婴配 | 粉状 | 乳蛋白部 | 进口产品 | 100 | 1.28 | 2.51 | 0.55 | 9.4 |
| 菁挚呵护特殊医学用途婴儿乳蛋白部分水解配方粉 | 0~12月龄乳蛋白过敏高风险婴儿 | 特医婴配 | 粉状 | 乳蛋白部 | 进口产品 | 100 | 1.28 | 2.549 | 0.546 | 9.4 |
| 贝因美特殊医学用途婴儿无乳糖配方食品 | 0~12月龄乳糖不耐受婴儿 | 特医婴配 | 粉状 | 无乳糖配 | 国产产品 | 100 | 1.3 | 2.6 | 0.56 | 8 |
| 纽贝瑞特殊医学用途婴儿苯丙酮尿症配方粉 | 0~12月龄苯丙酮尿症婴儿 | 特医婴配 | 粉状 | 氨基酸代 | 进口产品 | 100 | 1.21 | 0 | 0 | 10 |
| 亲舒特殊医学用途婴儿乳蛋白部分水解配方粉 | 0~12月龄乳蛋白过敏高风险婴儿 | 特医婴配 | 粉状 | 乳蛋白部 | 进口产品 | 100 | 1.3 | 2.5 | 0.54 | 9.4 |
| 喜康宝贝初特殊医学用途早产/低出生体重婴儿配方奶 | 早产/低出生体重婴儿 | 特医婴配 | 液态 | 早产/低 | 进口产品 | 100 | 1.24 | 2.46 | 0.72 | 11.5 |
| 喜康宝贝有特殊医学用途早产/低出生体重婴儿配方奶 | 早产/低出生体重婴儿 | 特医婴配 | 液态 | 早产/低 | 进口产品 | 100 | 1.24 | 2.46 | 0.72 | 11.5 |
| 安儿宝特殊医学用途婴儿无乳糖配方粉 | 0~12月龄乳糖不耐受婴儿 | 特医婴配 | 粉状 | 无乳糖配 | 进口产品 | 100 | 1.3 | 2.5 | 0.5 | 10 |
| 早瑞能恩特殊医学用途早产/低出生体重婴儿配方食品 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 早产/低 | 进口产品 | 100 | 1.24 | 2.49 | 0.69 | 12.5 |
| 乐赋特殊医学用途电解质配方食品 | 10岁以上因腹泻等原因导致轻至中度脱水 | 特医婴配 | 粉状 | 非全营养 | 国产产品 | 100 | 0 | 5.8 | 0 | 2 |
| 乐索特殊医学用途电解质配方食品 | 10岁以上术前需要补充碳水化合物 | 特医婴配 | 粉状 | 非全营养 | 国产产品 | 100 | 0 | 5.9 | 0 | 2 |
| 小安素®特殊医学用途全营养配方食品 | 消化吸收不足造成营养不良的1~11岁以上婴幼儿 | 特医婴配 | 粉状 | 全营养配 | 进口产品 | 100 | 0.89 | 3.1 | 0.78 | 11.5 |
| 优博敏佳特殊医学用途婴儿乳蛋白部分水解配方食品 | 0~12月龄乳蛋白过敏高风险婴儿 | 特医婴配 | 粉状 | 乳蛋白部 | 国产产品 | 100 | 1.26 | 2.5 | 0.55 | 9.4 |
| 雅培喜康宝特殊医学用途早产/低出生体重婴儿配方粉 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 早产/低 | 进口产品 | 100 | 1.315 | 2.336 | 0.627 | 6.2 |
| 纽登星特殊医学用途早产/低出生体重婴儿配方食品 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 早产/低 | 进口产品 | 100 | 1.3 | 2.37 | 0.63 | 10.5 |
| 纽贝瑞特殊医学用途氨基酸代谢障碍配方食品 | 10岁以上苯丙酮尿症人群 | 特医婴配 | 粉状 | 非全营养 | 进口产品 | 100 | 0.74 | 2 | 0 | 31 |
| 蔼儿舒特殊医学用途婴儿乳蛋白深度水解配方食品 | 0~12月龄食物蛋白过敏婴儿 | 特医婴配 | 粉状 | 乳蛋白深 | 进口产品 | 100 | 1.2 | 2.6 | 0.64 | 9.4 |
| 喜康宝贝添特殊医学用途婴儿营养补充剂 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 母乳营养 | 进口产品 | 100 | 0.61 | 3.1 | 1.7 | 11.5 |
| 早启能恩特殊医学用途早产/低出生体重婴儿配方食品 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 早产/低 | 进口产品 | 100 | 1.24 | 2.51 | 0.67 | 11.5 |
| 超启能恩特殊医学用途婴儿乳蛋白部分水解配方食品 | 0~12月龄乳蛋白过敏高风险婴儿 | 特医婴配 | 粉状 | 乳蛋白部 | 进口产品 | 100 | 1.21 | 2.79 | 0.45 | 9.4 |
| 安儿宁能恩特殊医学用途婴儿无乳糖配方食品 | 0~12月龄乳糖不耐受婴儿 | 特医婴配 | 粉状 | 无乳糖配 | 进口产品 | 100 | 1.2 | 2.79 | 0.49 | 7.5 |
| 佳膳佳立畅特殊医学用途全营养配方食品 | 适用于10岁以上进食受限、消化吸收障碍 | 特医婴配 | 粉状 | 全营养配 | 国产产品 | 100 | 0.7 | 3.4 | 0.8 | 8 |
| 优博启能特殊医学用途婴儿营养补充剂 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 母乳营养 | 国产产品 | 100 | 0.31 | 3.6 | 1.6 | 11.5 |
| 小佰太能特殊医学用途全营养配方食品 | 1~10岁因进食受限、消化吸收障碍 | 特医婴配 | 粉状 | 全营养配 | 进口产品 | 100 | 0.9 | 3.2 | 0.7 | 7.5 |
| 伊利®欣活®特殊医学用途全营养配方粉 | 适用于10岁以上进食受限、消化吸收障碍 | 特医婴配 | 粉状 | 全营养配 | 国产产品 | 100 | 0.8 | 3 | 1 | 11.5 |
| 贝新尔特殊医学用途早产/低出生体重婴儿配方食品 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 早产/低 | 进口产品 | 100 | 1.29 | 2.4 | 0.68 | 11.5 |
| 全安素®特殊医学用途全营养配方食品 | 10岁以上进食受限、消化吸收障碍 | 特医婴配 | 粉状 | 全营养配 | 进口产品 | 100 | 0.78 | 3.19 | 0.88 | 11.5 |
| 小佳膳特殊医学用途全营养配方食品 | 1~10岁进食受限、消化吸收障碍 | 特医婴配 | 粉状 | 全营养配 | 进口产品 | 100 | 0.9 | 3.1 | 0.7 | 7.5 |
| 佳膳悠选特殊医学用途全营养配方食品 | 10岁以上进食受限、消化吸收障碍 | 特医婴配 | 粉状 | 全营养配 | 进口产品 | 100 | 0.9 | 2.8 | 1 | 11.5 |
| 爱思诺赋儿嘉特殊医学用途婴儿无乳糖配方食品 | 0~12月龄乳糖不耐受婴儿 | 特医婴配 | 粉状 | 无乳糖配 | 进口产品 | 100 | 1.16 | 2.8 | 0.53 | 6.2 |
| 爱思诺晨而慧®特殊医学用途早产/低出生体重婴儿配方 | 早产/低出生体重婴儿 | 特医婴配 | 粉状 | 早产/低 | 进口产品 | 100 | 1.23 | 2.5 | 0.71 | 11.5 |

图 8 训练数据集图

时，你能够基于推荐适合的产品。你应综合考虑人群的需求，对如果有某种偏好要加以考虑，对与对营养成分有要求的客户，也应考虑产品的营养含量来做出推荐。”

STEP3 向 kimi 大模型上传训练数据集: 输入 prompt 后，kimi 大模型将基于训练集开始训练, 具体的训练过程如图9所示。

这是一份包含182种特医产品的基本信息，适用人群，组织状态，产品类别，以及提供每100kJ能量时，几种主要营养物质的含量。我需要你学习这份数据，扮演一个基于这个数据的推荐模型。当我想描述客户的情况和需求时，你能够基于推荐适合的产品。你应该综合考虑人群的需求，对如果有某种偏好要加以考虑，对与对营养成分有要求的客户，也应考虑产品的营养含量来做出推荐。

好的，我已经明白了您的需求。请上传包含182种特医产品基本信息的文件，我会仔细学习这些数据，并根据您描述的客户状况和需求来推荐合适的产品。您可以上传Excel电子表格、PDF文档、Word文档等格式的文件，只要确保文件的总字数不超过20万字。上传文件后，我将开始分析并准备为您提供推荐。

请给予这份数据进行训练

任务3训练数据 XLSX, 25.18 KB

好的，我已经接收并开始处理您提供的文件。我将分析这些特医产品的详细信息，包括它们的注册号、产品名称、适用人群、适用人群类别、组织状态、产品类别、产品来源、能量含量以及每100kJ能量时几种主要营养物质的含量。通过这些信息，我能够根据客户的具体状况和需求，为他们推荐合适的特医产品。

任务3训练数据.xlsx

| A | B |
|----------------|--------------------------|
| 注册号 | 产品名称 |
| 国食注字TY20175001 | 纽康特特殊医学用途婴儿氨基酸配方食品 |
| 国食注字TY20175002 | 雅培亲护特殊医学用途婴儿乳蛋白部分水解配方粉 |
| 国食注字TY20175003 | 菁挚呵护特殊医学用途婴儿乳蛋白部分水解配方粉 |
| 国食注字TY20180001 | 贝因美特殊医学用途婴儿无乳糖配方食品 |
| 国食注字TY20185001 | 纽贝瑞特殊医学用途婴儿苯丙酮尿症配方粉 |
| 国食注字TY20185002 | 亲舒特殊医学用途婴儿乳蛋白部分水解配方粉 |
| 国食注字TY20185003 | 喜康宝贝初特殊医学用途早产/低出生体重婴儿配方奶 |
| 国食注字TY20185004 | 喜康宝贝有特殊医学用途早产/低出生体重婴儿配方奶 |
| 国食注字TY20185005 | 安儿宝特殊医学用途婴儿无乳糖配方粉 |
| 国食注字TY20185006 | 早瑞能恩特殊医学用途早产/低出生体重婴儿配方食品 |
| 国食注字TY20180002 | 乐赋特殊医学用途电解质配方食品 |
| 国食注字TY20180003 | 乐索特殊医学用途电解质配方食品 |
| 国食注字TY20185007 | 小安素®特殊医学用途全营养配方食品 |
| 国食注字TY20180004 | 优博敏佳特殊医学用途婴儿乳蛋白部分水解配方食品 |
| 国食注字TY20185009 | 雅培喜康宝特殊医学用途早产/低出生体重婴儿配方粉 |
| 国食注字TY20185010 | 纽登星特殊医学用途早产/低出生体重婴儿配方食品 |
| 国食注字TY20185011 | 纽贝瑞特殊医学用途氨基酸代谢障碍配方食品 |
| 国食注字TY20185012 | 蔼儿舒特殊医学用途婴儿乳蛋白深度水解配方食品 |
| 国食注字TY20185013 | 喜康宝贝添特殊医学用途婴儿营养补充剂 |
| 国食注字TY20195001 | 早启能恩特殊医学用途早产/低出生体重婴儿配方食品 |
| 国食注字TY20195002 | 超启能恩特殊医学用途婴儿乳蛋白部分水解配方食品 |
| 国食注字TY20195003 | 安儿宁能恩特殊医学用途婴儿无乳糖配方食品 |
| 国食注字TY20190001 | 佳膳佳立畅特殊医学用途全营养配方食品 |
| 国食注字TY20190002 | 优博启能特殊医学用途婴儿营养补充剂 |
| 国食注字TY20195004 | 小佰太能特殊医学用途全营养配方食品 |
| 国食注字TY20190003 | 伊利®欣活®特殊医学用途全营养配方粉 |
| 国食注字TY20190004 | 贝新尔特殊医学用途早产/低出生体重婴儿配方食品 |
| 国食注字TY20195005 | 全安素®特殊医学用途全营养配方食品 |
| 国食注字TY20195006 | 小佳膳特殊医学用途全营养配方食品 |
| 国食注字TY20195007 | 佳膳悠选特殊医学用途全营养配方食品 |

图 9 智能推荐模型训练过程图

学习训练数据之后，kimi 大模型成为一个推荐系统，当我们给出某位客户的状况描述时，将会在数据集中选取若干种适合的产品，推荐给客户使用。

STEP4 推荐模型的测试: 输入客户状况，观察并验证模型的输出是否准确有效。

输入：客户 1：婴儿、蛋白质过敏。模型输出为的结果推荐了纽康特特殊医学用途婴儿氨基酸配方食品、恩敏舒特殊医学用途婴儿氨基酸配方食品。对着原数据进行检查，这两款产品用于蛋白质过敏的婴儿都是非常适合的。具体的测试过程如图10所示。



图 10 智能推荐模型测试过程图 1

输入：客户 2：10 岁儿童、需要补充蛋白质、乳糖不耐受。模型输出推荐了伊能佳特殊医学用途蛋白质组件配方食品等产品。对比原数据检查，模型对这些产品的描述准确无误，做出的推荐也十分合理。具体的测试过程如图11所示。



图 11 智能推荐模型测试过程图 2

此外，还假设了其他状况人群，如乳糖不耐受的成人，对脂肪过敏的人群等。推荐模型的输出结果都能够准确无误，适合相应人群食用。因此，基于大模型的特医食品推荐系统能够出色地完成任务。

参考文献

- [1] 厉梁秋,王梦婷.200朵“小蓝花”绽放,特殊医学用途配方食品行业稳步发展[J].食品安全导刊,2024,(30):5.DOI:10.16043/j.cnki.cfs.2024.30.049.
- [2] 施京京.保障特医食品营养安全市场监管总局优化三类特医食品注册管理要求[J].中国质量监管,2024,(07):35.
- [3] 胡锡丰.强监管促特殊食品更安全[N].中国质量报,2024-01-04(002).DOI:10.28164/n.cnki.nczlb.2024.000022.
- [4] 蔡军,李曙光,王红,等.10岁以上特医全营养配方食品配方及营养分析[J].食品安全质量检测学报,2024,15(14):242-248.DOI:10.19812/j.cnki.jfsq11-5956/ts.20240509003.
- [5] 施京京.保障特医食品营养安全市场监管总局优化三类特医食品注册管理要求[J].中国质量监管,2024,(07):35.