

A close-up of a logo

Description automatically generated

Mục lục

[I.Thông tin của nhóm và tiến độ đồ án 3](#_heading=h.gjdgxs)

[II. Giai đoạn 1 Data abstraction 3](#_heading=h.30j0zll)

[A. Data forms 3](#_heading=h.1fob9te)

[B.Data Profiling 5](#_heading=h.3dy6vkm)

[1.ID 7](#_heading=h.1t3h5sf)

[2. Gender: 7](#_heading=h.4d34og8)

[3.Age: 8](#_heading=h.2s8eyo1)

[4.Marital Status: 8](#_heading=h.17dp8vu)

[5.Children: 9](#_heading=h.3rdcrjn)

[6.Smoker: 9](#_heading=h.26in1rg)

[7.Employed: 10](#_heading=h.lnxbz9)

[8.Years Worked: 10](#_heading=h.35nkun2)

[9.Income Level: 11](#_heading=h.1ksv4uv)

[10.Social Media: 11](#_heading=h.44sinio)

[11.Online Gaming: 12](#_heading=h.2jxsxqh)

[12.Cancer: 13](#_heading=h.z337ya)

[III. Giai đoạn 2 Task abstraction 14](#_heading=h.3j2qqm3)

[1.Khái niệm: 14](#_heading=h.1y810tw)

[2. Các Task Abstraction: 15](#_heading=h.4i7ojhp)

[- So sánh số ca mắc bệnh ung thư giữa người không hút thuốc và người hút thuốc: 15](#_heading=h.2xcytpi)

[- Tính tỷ lệ số người mắc bệnh ung thư trong số tất cả những người tham gia khảo sát 15](#_heading=h.1ci93xb)

[- So sánh số người mắc bệnh ung thư theo các mức thu nhập: 15](#_heading=h.3whwml4)

[- Thống kê số người mắc bệnh ung thư theo độ tuổi: 15](#_heading=h.2bn6wsx)

[- Thống kê người mắc bệnh ung thư theo năm làm việc (đối với người đã đi làm ): 15](#_heading=h.qsh70q)

[- So sánh số người mắc bệnh ung thư giữa người có con và chưa có con: 15](#_heading=h.3as4poj)

[- Thống kê người mắc ung thư có tỉ lệ social media và online gaming như thế nào: 15](#_heading=h.1pxezwc)

[- Thống kê tỉ lệ mắc ung thư dựa trên tình trạng hôn nhân: 15](#_heading=h.49x2ik5)

[IV. Giai đoạn 3: Idiom and Implementation 15](#_heading=h.2p2csry)

[A. Idiom 15](#_heading=h.147n2zr)

[So sánh số ca mắc bệnh ung thư giữa người không hút thuốc và người hút thuốc 15](#_heading=h.3o7alnk)

[Tính tỷ lệ số người mắc bệnh ung thư trong số tất cả những người tham gia khảo sát 16](#_heading=h.23ckvvd)

[So sánh số người mắc bệnh ung thư theo các mức thu nhập 16](#_heading=h.ihv636)

[Thống kê số người mắc bệnh ung thư theo độ tuổi 16](#_heading=h.32hioqz)

[Thống kê người mắc bệnh ung thư theo năm làm việc (đối với người đã đi làm) 17](#_heading=h.1hmsyys)

[So sánh tỉ lệ mắc bệnh ung thư giữa người có con và chưa có con: 17](#_heading=h.41mghml)

[Thống kê người mắc ung thư có tỉ lệ social media và online gaming như thế nào 18](#_heading=h.2grqrue)

[Thống kê tỉ lệ mắc ung thư dựa trên tình trạng hôn nhân 19](#_heading=h.vx1227)

[B. Implementation 19](#_heading=h.3fwokq0)

[So sánh số ca mắc bệnh ung thư giữa người không hút thuốc và người hút thuốc: 19](#_heading=h.1v1yuxt)

[Tính tỷ lệ số người mắc bệnh ung thư trong số tất cả những người tham gia khảo sát: 20](#_heading=h.4f1mdlm)

[So sánh số người mắc bệnh ung thư theo các mức thu nhập 21](#_heading=h.2u6wntf)

[Thống kê số người mắc bệnh ung thư theo độ tuổi 23](#_heading=h.19c6y18)

[Thống kê người mắc bệnh ung thư theo năm làm việc (đối với người đã đi làm): 23](#_heading=h.3tbugp1)

[So sánh tỉ lệ mắc bệnh ung thư giữa người có con và chưa có con: 24](#_heading=h.28h4qwu)

[Thống kê người mắc ung thư có tỉ lệ social media và online gaming như thế nào: 26](#_heading=h.nmf14n)

[Thống kê tỉ lệ mắc ung thư dựa trên tình trạng hôn nhân: 28](#_heading=h.37m2jsg)

[V. Reference 29](#_heading=h.1mrcu09)

# I.Thông tin của nhóm và tiến độ đồ án

| Bảng phân công | | |
| --- | --- | --- |
| **Công việc thực hiện** | **Người thực hiện** | **Tiến độ hoàn thành** |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS | 21127705 – Từ Phước Toàn | 100% |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS | 21127620 – Trần Hoàng Khải | 100% |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS | 20127361 – Nguyễn Ngọc Diệu Trang | 100% |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS | 20127632 – Trần Đình Thiện | 100% |

# II. Giai đoạn 1 Data abstraction

Dữ liệu được lấy từ File: [Cancer\_prediction\_dataset.xlxs](https://www.kaggle.com/datasets/fdcellat/cancer-prediction-dataset?fbclid=IwAR2tIxWFViGZitdSf4oUFWLOgpUr3IEfbMbTzmLoOF7ae6JDnqRtNuA5B_g)

*Conceptual Model  
 Khái quát về bộ dữ liệu:*Bộ dữ liệu được tiến hành khảo sát bởi một trường đại học, nhằm nghiên cứu về các yếu tố nguy cơ tiềm ẩn cho bệnh ung thư phổi. Cuộc khảo sát bao gồm một loạt các câu hỏi về dân số học, lối sống và sức khỏe.

## Data forms

1. Dataset Type & Data types:

* Dataset Type: Table.
* Loại Dataset: Static. Vì có thể truy xuất một lần  
  được hết tất cả các dữ liệu.
* Data types:
* Mỗi dòng là 1 người được khảo sát có mắc bệnh ung thư hay không.
* Mỗi cột là một attribute bao gồm:  
  { ID, Gender, Age, Marital Status, Children, Smoker, Employed, Years Worked, Income Level, Social Media, Online Gaming, Cancer}.
* Tổng số Thuộc tính: 12.
* Tổn số items: 1000.

2. Attributes

* ID: Một định danh duy nhất gồm 5 chữ số được gán cho mỗi người tham gia khảo sát. Những số này được tạo ngẫu nhiên và không có ý nghĩa thực tế.
* Gender: Giới tính của người tham gia khảo sát, được chỉ định là 'Nữ' hoặc 'Nam'.
* Age: Tuổi của người tham gia khảo sát, dao động từ 18 đến 90 tuổi.
* Marital Status: Tình trạng hôn nhân của người tham gia khảo sát, được phân loại là 'Đã kết hôn', 'Độc thân', 'Góa phụ' hoặc 'Ly thân'.
* Children: Số lượng con cái của người tham gia khảo sát, từ 0 đến 5.
* Smoker: Cho biết liệu người tham gia khảo sát có hút thuốc hay không, với các lựa chọn 'Có' hoặc 'Không'.
* Employed: Tình trạng việc làm của người tham gia khảo sát, được chỉ định là 'Có' đối với người có việc làm và 'Không' đối với người thất nghiệp.
* Years Worked: Tổng số năm người tham gia khảo sát đã làm việc, từ 0 đến 40 năm. Giá trị này được đặt là 0 đối với những người không có việc làm.
* Income Level: Mức thu nhập tự đánh giá của người tham gia khảo sát, được phân loại là 'Cao', 'Trung bình' hoặc 'Thấp'. Người thất nghiệp tự động thuộc danh mục 'Thấp'.
* Social Media: Cho biết liệu người tham gia khảo sát có sử dụng các nền tảng truyền thông xã hội hay không, với các lựa chọn 'Có' hoặc 'Không'.
* Online Gaming: Chỉ ra liệu người tham gia khảo sát có tham gia vào trò chơi trực tuyến hay không, với các lựa chọn 'Có' hoặc 'Không'.
* Cancer: Cho biết liệu người tham gia khảo sát đã được chẩn đoán mắc bệnh ung thư phổi hay không, với các lựa chọn 'Có' hoặc 'Không'. Trường này được điều chỉnh nhân tạo dựa trên sự kết hợp của các yếu tố như tuổi, tình trạng hút thuốc, thời gian làm việc, và lối sống để mô phỏng các yếu tố nguy cơ tiềm ẩn cho bệnh ung thư phổi

3. Attributes Type

| Tên thuộc tính | Attribute types | Ordered attribute type | | Hierarchical | Direction | | | Attribute charecteristics | | Bin |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Interval | Ratio | Squential | Diverging | Cyclic | continuous | Discrete |
|  |
| ID | c |  |  |  |  |  |  |  | X | 996 |  |
| Gender | c |  |  |  |  |  |  |  | X | 2 |  |
| Age | q |  | X |  | X |  |  |  | X | 72 |  |
| Marital Status | c |  |  |  |  |  |  |  |  | 4 |  |
| Children | q |  | X |  | X |  |  |  | X | 6 |  |
| Smoker | c |  |  |  |  |  |  |  |  | 2 |  |
| Employed | c |  |  |  |  |  |  |  | X | 2 |  |
| Year works | q |  | X |  | X |  |  |  | X | 40 |  |
| Income Level | c |  |  |  |  |  |  |  | X | 3 |  |
| Social Media | c |  |  |  |  |  |  |  | X | 2 |  |
| Online Gaming | c |  |  |  |  |  |  |  | X | 2 |  |
| Cancer | c |  |  |  |  |  |  |  | X | 2 |  |

## B.Data Profiling

| Field Name | NULL | Missing | Actual | Completenes | Cardinality | Uniqueness | Distinctness |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 0 | 0 | 1000 | 100% | 996 | 99,60% | 99,60% |
| Gender | 0 | 0 | 1000 | 100% | 2 | 0.2% | 0.2% |
| Age | 0 | 0 | 1000 | 100% | 72 | 7.2% | 7.2% |
| Marital Status | 0 | 0 | 1000 | 100% | 4 | 0.4% | 0.4% |
| Children | 0 | 0 | 1000 | 100% | 6 | 0.6% | 0.6% |
| Smoker | 0 | 0 | 1000 | 100% | 2 | 0.2% | 0.2% |
| Employed | 0 | 0 | 1000 | 100% | 2 | 0.2% | 0.2% |
| years work | 0 | 0 | 1000 | 100% | 40 | 4% | 4% |
| Income level | 0 | 0 | 1000 | 100% | 3 | 0.3% | 0.3% |
| social Media | 0 | 0 | 1000 | 100% | 2 | 0.2% | 0.2% |
| Online Gaming | 0 | 0 | 1000 | 100% | 2 | 0.2% | 0.2% |
| Cancer | 0 | 0 | 1000 | 100% | 2 | 0.2% | 0.2% |

### 1.ID

| **Input Metadata** | |  | **ID (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | ID |  | integer | 1000 | 100% |
| Field Data Type | integer |  | **ID (Top 4 Values)** | **Count** | **Percentage** |
| Field Length | 5 |  | 13290 | 2 | 0.2% |
| **Data Profiling Summary Statistics** | |  | 13955 | 2 | 0.2% |
| NULL | 0 |  | 57526 | 2 | 0.2% |
| Missing | 0 |  | 84442 | 2 | 0.2% |
| Actual | 1000 |  | **ID (Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnnnn | 1000 | 100% |
| Cardinality | 996 |  |  |  |  |
| Uniqueness | 99,60% |  |  |  |  |
| Distinctness | 99,60% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 5 |  |  |  |  |
| Field Length (MAX) | 5 |  |  |  |  |
| Field Value (MIN) | 10149 |  |  |  |  |
| Field Value (MAX) | 99976 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 2. Gender:

| **Input Metadata** | |  | **Gender (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Gender |  | string | 1000 | 100% |
| Field Data Type | string |  | **Gender (Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 6 |  | female | 515 | 51.5% |
| **Data Profiling Summary Statistics** | |  | male | 485 | 48.5% |
| NULL | 0 |  | **Gender (Field Formats)** | **Count** | **Percentage** |
| Missing | 0 |  | nnnn | 485 | 48.5% |
| Actual | 1000 |  | nnnnnn | 515 | 51.5% |
| Completeness | 100% |  |  |  |  |
| Cardinality | 2 |  |  |  |  |
| Uniqueness | 0,20% |  |  |  |  |
| Distinctness | 0,20% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 4 |  |  |  |  |
| Field Length (MAX) | 6 |  |  |  |  |
| Field Value (MIN) | male |  |  |  |  |
| Field Value (MAX) | female |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 3.Age:

| **Input Metadata** | |  | **Age(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Age |  | integer | 1000 | 100% |
| Field Data Type | integer |  | **Age(Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 2 |  | 76 | 21 | 2.1% |
| **Data Profiling Summary Statistics** | |  | 66 | 20 | 2.0% |
| NULL | 0 |  | 56 | 20 | 2.0% |
| Missing | 0 |  | 82 | 20 | 2.0% |
| Actual | 1000 |  | 45 | 19 | 1.9% |
| Completeness | 100% |  | **Age (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 72 |  | nn | 1000 | 100% |
| Uniqueness | 7.2% |  |  |  |  |
| Distinctness | 7.2% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 2 |  |  |  |  |
| Field Length (MAX) | 2 |  |  |  |  |
| Field Value (MIN) | 18 |  |  |  |  |
| Field Value (MAX) | 89 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 4.Marital Status:

| **Input Metadata** | |  | **Marital Status (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Marital Status |  | string | 1000 | 100% |
| Field Data Type | string |  | **Marital Status (Top 4 Values)** | **Count** | **Percentage** |
| Field Length | 9 |  | Single | 267 | 26.7% |
| **Data Profiling Summary Statistics** | |  | Separated | 242 | 24.2% |
| NULL | 0 |  | Widowed | 232 | 23.2% |
| Missing | 0 |  | Married | 259 | 25.9% |
| Actual | 1000 |  | **Marital Status (Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnnnnn | 267 | 26.7% |
| Cardinality | 4 |  | nnnnnnn | 491 | 49.1% |
| Uniqueness | 0.4% |  | nnnnnnnnn | 242 | 24.2% |
| Distinctness | 0.4% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 6 |  |  |  |  |
| Field Length (MAX) | 9 |  |  |  |  |
| Field Value (MIN) | ####### |  |  |  |  |
| Field Value (MAX) | ####### |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 5.Children:

| **Input Metadata** | |  | **Children (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Children |  | integer | 1000 | 100% |
| Field Data Type | integer |  | **Children (Top 4 Values)** | **Count** | **Percentage** |
| Field Length |  |  | 3 | 192 | 19.2% |
| **Data Profiling Summary Statistics** | |  | 1 | 188 | 18.8% |
| NULL | 0 |  | 2 | 167 | 16.7% |
| Missing | 0 |  | 5 | 160 | 16% |
| Actual | 1000 |  | **Children (Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | n | 1000 | 100% |
| Cardinality | 6 |  |  |  |  |
| Uniqueness | 0.6% |  |  |  |  |
| Distinctness | 0.6% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 1 |  |  |  |  |
| Field Length (MAX) | 1 |  |  |  |  |
| Field Value (MIN) | 0 |  |  |  |  |
| Field Value (MAX) | 5 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 6.Smoker:

| **Input Metadata** | |  | **Smoker (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Smoker |  | string | 1000 | 100% |
| Field Data Type | string |  | **Smoker (Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 3 |  | Yes | 512 | 51.2% |
| **Data Profiling Summary Statistics** | |  | No | 488 | 48.8% |
| NULL | 0 |  |  |  |  |
| Missing | 0 |  |  |  |  |
| Actual | 1000 |  | **Smoker (Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnn | 512 | 51.2% |
| Cardinality | 2 |  | nn | 488 | 48.8% |
| Uniqueness | 0.2% |  |  |  |  |
| Distinctness | 0.2% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 2 |  |  |  |  |
| Field Length (MAX) | 3 |  |  |  |  |
| Field Value (MIN) | ####### |  |  |  |  |
| Field Value (MAX) | ####### |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 7.Employed:

| **Input Metadata** | |  | **Employed(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Employed |  | string | 1000 | 100% |
| Field Data Type | string |  | **Employed(Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 3 |  | Yes | 514 | 51.4% |
| **Data Profiling Summary Statistics** | |  | No | 486 | 48.6% |
| NULL | 0 |  |  |  |  |
| Missing | 0 |  |  |  |  |
| Actual | 1000 |  | **Employed(Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnn | 514 | 51.4% |
| Cardinality | 2 |  | nn | 486 | 48.6% |
| Uniqueness | 0.2% |  |  |  |  |
| Distinctness | 0.2% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 2 |  |  |  |  |
| Field Length (MAX) | 3 |  |  |  |  |
| Field Value (MIN) | ####### |  |  |  |  |
| Field Value (MAX) | ####### |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 8.Years Worked:

| **Input Metadata** | |  | **Age(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Year Worked |  | integer | 1000 | 100% |
| Field Data Type | integer |  | **Age(Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 2 |  | 0 | 499 | 49.9% |
| **Data Profiling Summary Statistics** | |  | 35 | 21 | 2.1% |
| NULL | 0 |  | 36 | 20 | 2.0% |
| Missing | 0 |  | 27 | 19 | 1.9% |
| Actual | 1000 |  | 2 | 18 | 1.8% |
| Completeness | 100% |  | **Age (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 40 |  | nn | 430 | 43.0% |
| Uniqueness | 4% |  | n | 570 | 57.0% |
| Distinctness | 4% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 1 |  |  |  |  |
| Field Length (MAX) | 2 |  |  |  |  |
| Field Value (MIN) | 0 |  |  |  |  |
| Field Value (MAX) | 39 |  |  |  |  |
| Field Formats | 2 |  |  |  |  |

### 9.Income Level:

| **Input Metadata** | |  | **Income level (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Income level |  | string | 1000 | 100% |
| Field Data Type | string |  | **Income level (Top 3 Values)** | **Count** | **Percentage** |
| Field Length | 6 |  | low | 669 | 66.9% |
| **Data Profiling Summary Statistics** | |  | high | 166 | 16.6% |
| NULL | 0 |  | medium | 165 | 16.5% |
| Missing | 0 |  |  |  |  |
| Actual | 1000 |  | **Incone level(Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnn | 669 | 66.9% |
| Cardinality | 3 |  | nnnn | 166 | 16.6% |
| Uniqueness | 0.3% |  | nnnnnn | 165 | 16.5% |
| Distinctness | 0.3% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 3 |  |  |  |  |
| Field Length (MAX) | 6 |  |  |  |  |
| Field Value (MIN) | low |  |  |  |  |
| Field Value (MAX) | medium |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 10.Social Media:

| **Input Metadata** | |  | **Social Media (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Social Media |  | string | 1000 | 100% |
| Field Data Type | string |  | **Social Media (Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 3 |  | Yes | 524 | 52.4% |
| **Data Profiling Summary Statistics** | |  | No | 476 | 47.6% |
| NULL | 0 |  |  |  |  |
| Missing | 0 |  |  |  |  |
| Actual | 1000 |  | **Social Media (Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnn | 524 | 52.4% |
| Cardinality | 2 |  | nn | 476 | 47.6% |
| Uniqueness | 0.2% |  |  |  |  |
| Distinctness | 0.2% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 2 |  |  |  |  |
| Field Length (MAX) | 3 |  |  |  |  |
| Field Value (MIN) | no |  |  |  |  |
| Field Value (MAX) | yes |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 11.Online Gaming:

| **Input Metadata** | |  | **Online Gaming(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Online Gaming |  | string | 1000 | 100% |
| Field Data Type | string |  | **Online Gaming(Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 3 |  | Yes | 492 | 49.2% |
| **Data Profiling Summary Statistics** | |  | No | 508 | 50.8% |
| NULL | 0 |  |  |  |  |
| Missing | 0 |  |  |  |  |
| Actual | 1000 |  | **Online Gaming(Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnn | 492 | 49.2% |
| Cardinality | 2 |  | nn | 508 | 50.8% |
| Uniqueness | 0.2% |  |  |  |  |
| Distinctness | 0.2% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 2 |  |  |  |  |
| Field Length (MAX) | 3 |  |  |  |  |
| Field Value (MIN) | no |  |  |  |  |
| Field Value (MAX) | yes |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 12.Cancer:

| **Input Metadata** | |  | **Smoker (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Cancer |  | string | 1000 | 100% |
| Field Data Type | string |  | **Cancer(Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 3 |  | Yes | 776 | 77.6% |
| **Data Profiling Summary Statistics** | |  | No | 224 | 22.4% |
| NULL | 0 |  |  |  |  |
| Missing | 0 |  |  |  |  |
| Actual | 1000 |  | **Cancer(Field Formats)** | **Count** | **Percentage** |
| Completeness | 100% |  | nnn | 776 | 77.6% |
| Cardinality | 2 |  | nn | 224 | 22.4% |
| Uniqueness | 0.2% |  |  |  |  |
| Distinctness | 0.2% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 2 |  |  |  |  |
| Field Length (MAX) | 3 |  |  |  |  |
| Field Value (MIN) | No |  |  |  |  |
| Field Value (MAX) | Yes |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

*a.Đánh Giá Chung:* Dựa vào kết quả của Profiling, ta nhận thấy giá trị null và missing không tồn tại ở các thuộc tính.

#### b.Dựa vào kết quả ta thấy được:

- ID ở trong bảng này không thể dùng làm khóa chính vì có 3 ID lập lại 2 lần nên nó, nó phải là 1 khóa ứng cử viên

- Các ID lập nhau nhưng nội dung không giống nhau, có thể là lỗi hệ thống hoặc có người dùng làm đánh giá nhiều hơn 1 lần

- Cuộc khảo sát được thực hiện với độ tuổi là thanh niên, người đi làm cho tới người cao tuổi với số tuổi cao nhất ghi nhận là 89 tuổi

- Tỉ lệ người tham gia khảo sát với tình trạng kết hôn khá phong phú và đồng đều và người đang độc thân (single) là nhiều nhất

- Khảo sát có để cập số con người được khảo sát để chỉ ra rằng có mối quan hệ giữa việc sinh con và nguy cơ mắc một số loại ung thư như ung thư vú hoặc tử cung

- Khảo sát được thực hiện với những người có hút thuốc lá vì độ liên quan tới ung thư phổi

- Nghề nghiêp là 1 yếu tố khảo sát vì tính ảnh hưởng như môi trường làm việc về tình thần và thể chất cũng ảnh hưởng gây ra ung thư

- Để bổ sung cho giả thiết môi trường làm việc thì việc cần thu thập thâm niên làm việc, mức thu nhập là 2 điều không thể thiếu cho phán đoán này

- người khảo sát ung thư chiếm đến 77.6% tổng số người tham gia mặc dù các lựa chọn đều bình bình ngang nhau

# III. Giai đoạn 2 Task abstraction

## 1.Khái niệm:

Phân tích thành dạng trừu tượng của một hoặc nhiều công việc để tìm ra những điểm giống nhau hoặc khác nhau của những công việc đó. Nếu chỉ dựa vào ngôn ngữ bình thường, mặc dù các công việc đó có thể giống nhau nhưng lại được thể hiện bởi những thuật ngữ khác nhau thì chúng ta sẽ khó tìm ra được điểm giống và khác nhau.

→ Task abstraction là sử dụng một ngôn ngữ chung để thể hiện, từ đó dễ dàng tìm ra những điểm giống nhau và khác nhau của nhiều công việc với nhau.

A screenshot of a computer

Description automatically generated

*Hình 2.1 Hình ảnh về Task Abstraction*

## 2. Các Task Abstraction:

### - So sánh số ca mắc bệnh ung thư giữa người không hút thuốc và người hút thuốc:

Produce -> locate -> compare.

### - Tính tỷ lệ số người mắc bệnh ung thư trong số tất cả những người tham gia khảo sát

Product -> Locate -> Summarize.

### - So sánh số người mắc bệnh ung thư theo các mức thu nhập:

Produce -> explore -> Compare.

### - Thống kê số người mắc bệnh ung thư theo độ tuổi:

Produce -> explore -> Sum.

### - Thống kê người mắc bệnh ung thư theo năm làm việc (đối với người đã đi làm ):

Produce -> explore -> Sum.

### - So sánh số người mắc bệnh ung thư giữa người có con và chưa có con:

Produce -> explore -> Compare.

### - Thống kê người mắc ung thư có tỉ lệ social media và online gaming như thế nào:

Product->lockup -> Summarize.

### - Thống kê tỉ lệ mắc ung thư dựa trên tình trạng hôn nhân:

Product -> locate -> Summarize.

# IV. Giai đoạn 3: Idiom and Implementation

## Idiom

### So sánh số ca mắc bệnh ung thư giữa người không hút thuốc và người hút thuốc

**Derived attribute:** Cancer case ( Tổng số người mắc bệnh ung thư)

| **Idiom** | **Bar Charts** |
| --- | --- |
| Data Type | 1 categorical:  Smoker  1 quantitative: Cancer case |
| Encoding | Mark: line  Channel:   * Position: vertical and horizontal * Express: quantitative (length) * Spatial region: categorical * Separate: horizontal position * Align: vertical position * Order: length * Color: Luminance * Color map: binary |
| Task | Compare |
| Scale | Key attribute: 2 levels |

### Tính tỷ lệ số người mắc bệnh ung thư trong số tất cả những người tham gia khảo sát

**Derived attribute:** Case (Tổng số người tham gia khảo sát)

| **Idiom** | **Bar Charts** |
| --- | --- |
| Data Type | 1 categorical:  Cancer  1 quantitative: Case |
| Encoding | Mark: line  Channel:   * Position: vertical and horizontal * Express: quantitative (length) * Spatial region: categorical * Separate: horizontal position * Align: vertical position * Order: length * Color: Luminance * Color map: binary |
| Task | Lookup Value |
| Scale | Key attribute: 2 levels |

### So sánh số người mắc bệnh ung thư theo các mức thu nhập

**Derived attribute:** cancerTotal(tổng số người mắc bệnh ung thư)

| **Idiom** | **Bar Charts** |
| --- | --- |
| Data Type | 1 categorical:  Income level  1 quantitative: cancerTotal |
| Encoding | Mark: line  Channel:   * Position: vertical and horizontal * Express: quantitative (length) * Spatial region: categorical * Separate: horizontal position * Align: vertical position * Order: length * Color: luminance |
| Task | Compare |
| Scale | Key attribute: 3 levels |

### Thống kê số người mắc bệnh ung thư theo độ tuổi

**Derived attribute:** cancerAgeTotal (số người mắc bệnh ung thư theo độ tuổi)

| **Idiom** | **Line Charts** |
| --- | --- |
| Data Type | 1 ordinal: age  1 quantitative: cancerAgeTotal |
| Encoding | Mark: point & line connection marks between them  Channel:   * Position: horizontal and vertical * Align: lengths ( quantitative) * Separate, order: horizontal region (ordered) |
| Task | Find trend |
| Scale | Key attribute: 72  Value attribute: 1000 |

### Thống kê người mắc bệnh ung thư theo năm làm việc (đối với người đã đi làm)

**Derived attribute:** cancerYearWorkedTotal (số người mắc bệnh ung thư theo độ tuổi)

| **Idiom** | **Line Charts** |
| --- | --- |
| Data Type | 1 categorical:  YearWorked  1 quantitative: cancerYearWorkedTotal |
| Encoding | Mark: point & line connection marks between them  Channel:  ·        Position: horizontal and vertical  ·        Align: lengths ( quantitative)   * Separate, order: horizontal region (ordered) |
| Task | Find Trend |
| Scale | Key attribute: 35  Value attribute: 1000 |

### So sánh tỉ lệ mắc bệnh ung thư giữa người có con và chưa có con:

**Derived Attribute**: TotalCancer (Tổng số lượng người ung thư có con)

| **Idiom** | **Stacked Bar Charts** |
| --- | --- |
| Data Type | 1 categorical:  TotalCancer  2 quantitative: Have Children, Children |
| Encoding | Mark:  ● Line.  ● Glyph: multiple sub-bars are stacked vertically (Children)   Channel:  ● Length (TotalCancer).  ● Bar glyph with length-coded subcomponents of value attribute for each category of secondary key attribute (Sub TotalCancer).  ● Separate bars by category of primary key attribute(Have Children). |
| Task | Compare |
| Scale | Stacked key attribute:5  Main key attribute:2 |

### Thống kê người mắc ung thư có tỉ lệ social media và online gaming như thế nào

**Derived attribute**: Cancer case (Tổng số người mắc bệnh ung thư)

| **Idiom** | **Stacked Bar Charts** |
| --- | --- |
| Data Type | 2 categorical:  Social Media, Online Gaming  1 quantitative: Cancer case |
| Encoding | Mark: bar mark (vertical)   * Glyph: multiple sub-bars are stacked vertically   Channel:   * Position: vertical and horizontal * Length, color HUE * Spatial region: glyph * Align: full glyph, lowest bar component * Separate: horizontal position |
| Task | Lookup Value |
| Scale | Stacked key attribute: 2 levels.  Main key attribute: 2 levels |

### Thống kê tỉ lệ mắc ung thư dựa trên tình trạng hôn nhân

**Derived attribute**: Count of cancer (Tổng số người mắc bệnh ung thư)

| **Idiom** | **Bar Charts** |
| --- | --- |
| Data Type | 1 categorical:  Martial Status  1 quantitative: Count of cancer |
| Encoding | Mark: line  Channel:   * Position: vertical and horizontal * Express: quantitative (length) * Spatial region: categorical * Separate: horizontal position * Align: vertical position * Order: Label (alphabetical) * Color: HUE |
| Task | Compare |
| Scale | Key attribute: 4 levels |

## Implementation

### So sánh số ca mắc bệnh ung thư giữa người không hút thuốc và người hút thuốc:

Tableau:  <https://public.tableau.com/app/profile/t.to.n/viz/Sosnhtlmcbnhungthgiangihtthucvkhnghtthuc/Dashboard2?publish=yes>

A graph with a black rectangle

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout | x |  |
| Perceptual grouping | x |  |

### Tính tỷ lệ số người mắc bệnh ung thư trong số tất cả những người tham gia khảo sát:

Tableau:  <https://public.tableau.com/app/profile/t.to.n/viz/Thngksngibungthkhithamgiakhost/Dashboard1?publish=yes>

A graph with a black rectangle

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout | x |  |
| Perceptual grouping | x |  |

### So sánh số người mắc bệnh ung thư theo các mức thu nhập

Tableau:<https://public.tableau.com/app/profile/trang.nguyen.ngoc.dieu1006/viz/official_17037567117420/Sheet2?publish=yes>

A graph of income level

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout | x |  |
| Perceptual grouping | x |  |

### Thống kê số người mắc bệnh ung thư theo độ tuổi

Tableau: <https://public.tableau.com/app/profile/trang.nguyen.ngoc.dieu1006/viz/official_17037567117420/Sheet3?publish=yes>

A graph with blue lines

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping |  | x |

### Thống kê người mắc bệnh ung thư theo năm làm việc (đối với người đã đi làm):

Tableau: <https://public.tableau.com/views/Thngkngimcbnhungththeonmlmvicivingiilm/Sheet2?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link>

A graph with a line

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout | x |  |
| Perceptual grouping | x |  |

### So sánh tỉ lệ mắc bệnh ung thư giữa người có con và chưa có con:

Tableau <https://public.tableau.com/views/Sosnhsngimcbnhungthgiangicconvchaccon/Sheet1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link>

A graph of a bar chart

Description automatically generated with medium confidence

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout | x |  |
| Perceptual grouping | x |  |

### Thống kê người mắc ung thư có tỉ lệ social media và online gaming như thế nào:

Tableau:  <https://public.tableau.com/app/profile/kh.i.tr.n5082/viz/Thngkngimcungthctlsocialmediavonlinegamingnhthno/Dashboard1>

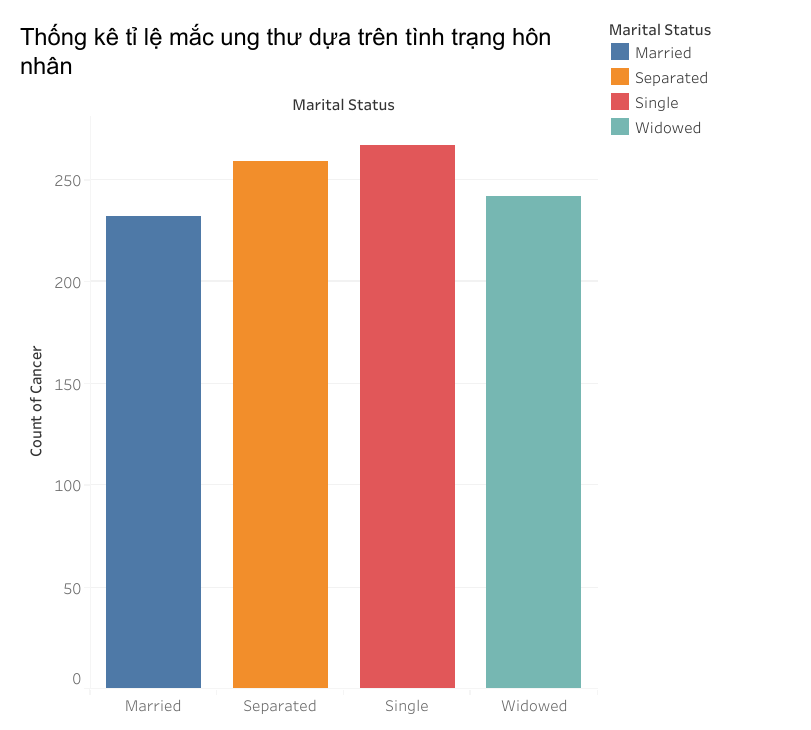
A graph with blue and orange squares

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping | x |  |

### Thống kê tỉ lệ mắc ung thư dựa trên tình trạng hôn nhân:

Tableau:  <https://public.tableau.com/app/profile/kh.i.tr.n5082/viz/Thngktlmcungthvkhngmcungthdatrntnhtrnghnnhn/Dashboard2>



| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout | x |  |
| Perceptual grouping | x |  |

# V. Reference

<https://www.kaggle.com/datasets/fdcellat/cancer-prediction-dataset?fbclid=IwAR2tIxWFViGZitdSf4oUFWLOgpUr3IEfbMbTzmLoOF7ae6JDnqRtNuA5B_g>

<https://drive.google.com/drive/folders/14t8jrlHysQawx5R5opvRfMtF9FRPRXoH>

<https://drive.google.com/file/d/1TBRaHCOakafAyb9We9NvbkT-Dg-MPC9h/view>