Factores de riesgo de cáncer

Proyecto de Fundamentos de la Ciencia de los Datos

Jean Cuervo

Table of contents

Chapter 1

Portada

Factores de riesgo de cáncer: análisis exploratorio y explicativo con datos públicos de Kaggle

Proyecto de la asignatura Fundamentos de la Ciencia de los Datos

Autor: Jean Cuervo Universidad Alfonso X el Sabio (UAX) Grado en Inteligencia Artificial y Computación

Curso académico 2025–2026

Chapter 2

Introducción

2.1 Contexto y problema de negocio

El cáncer es una de las principales causas de morbilidad y mortalidad en el mundo. Más allá del diagnóstico clínico individual, resulta clave comprender **qué combinación de factores de estilo de vida, características demográficas y antecedentes** se asocia con un mayor riesgo, para apoyar la **prevención y la educación sanitaria** a nivel poblacional.

Desde la perspectiva de la ciencia de los datos, este proyecto se sitúa inicialmente en las fases de **Comprensión del Negocio** y **Comprensión de los Datos** del ciclo de vida de un proyecto de data science, siguiendo la filosofía de metodologías como **CRISP-DM**. El objetivo no es construir un sistema clínico real, sino utilizar un conjunto de datos sintético y realista para entender mejor:

- Cómo se estructura la información disponible sobre factores de riesgo.
- Qué patrones básicos aparecen en los datos cuando los exploramos de forma sistemática.
- Cómo empezar a contar una **historia con datos** (*data storytelling*) que permita comunicar estos hallazgos a perfiles no técnicos (por ejemplo, responsables de salud pública).

En este contexto, nos planteamos la siguiente **decisión de negocio**:

> ¿Cómo podemos utilizar los datos disponibles sobre estilo de vida, demografía y antecedentes para **identificar grupos de población con mayor riesgo de cáncer** y, en consecuencia, **orientar mejor los recursos de prevención y educación sanitaria**?

2.2 Objetivos del proyecto

2.2.1 Objetivo general

Desarrollar un proyecto de ciencia de datos que permita **explorar, describir y explicar** el riesgo de cáncer a partir de diversas variables de estilo de vida, demográficas y de antecedentes, identificando **patrones y perfiles de riesgo** que sirvan como ejemplo de aplicación de la metodología de la asignatura.

2.2.2 Objetivos específicos

- Definir el **problema de negocio** y las **preguntas de análisis** relacionadas con el riesgo de cáncer.
- **Comprender y organizar el conjunto de datos**, identificando su estructura, tipos de variables y posibles problemas de calidad.
- Realizar un **Análisis Exploratorio de Datos (EDA)**:
  - Análisis **univariado** de variables seleccionadas (por ejemplo, edad, IMC, score de riesgo).
  - Análisis **bi/multivariado** de relaciones relevantes (por ejemplo, tabaquismo y consumo de alcohol frente al riesgo).
  - Detección y comentario de **datos faltantes, valores atípicos y posibles problemas de calidad**.
- Aplicar, en fases posteriores del proyecto, **modelos explicativos y predictivos** sencillos:
  - Modelos de **regresión** con `Overall_Risk_Score` como variable objetivo.
  - Modelos de **regresión logística** con `Risk_Level` como variable objetivo categórica.
  - Métodos de **clusterización** para identificar perfiles de riesgo.
- Elaborar un **informe reproducible en Quarto** que combine código en R, resultados (tablas y gráficos) y explicaciones en **lenguaje de negocio**, conectando los análisis con la toma de decisiones en prevención y educación sanitaria.

2.3 Preguntas de negocio y de análisis de datos

A partir del contexto anterior, planteamos un conjunto inicial de preguntas que guiarán el análisis:

1. **Distribución del riesgo**
   - ¿Cómo se distribuye el `Overall_Risk_Score` en la población del dataset?
   - ¿Qué proporción de individuos se encuentra en los niveles de riesgo `Low`, `Medium` y `High` (`Risk_Level`)?
2. **Factores individuales de riesgo (análisis univariado)**
   - ¿Cuál es la distribución de variables como `Age`, `BMI`, `Smoking`, `Alcohol_Use` o `Physical_Activity`?
   - ¿Existen outliers evidentes o valores extremos que debamos comentar?
3. **Relaciones entre factores y riesgo (análisis bi/multivariado)**
   - ¿Cómo cambia el `Overall_Risk_Score` según el nivel de tabaquismo y consumo de alcohol?
   - ¿Hay diferencias claras en el nivel de riesgo según la presencia de antecedentes familiares (`Family_History`) u otros factores genéticos/infecciosos?
4. **Perfiles de riesgo**
   - ¿Podemos identificar, a partir de los datos, ciertos **perfiles típicos** de riesgo, como:
     - Riesgo elevado principalmente asociado al **estilo de vida**.
     - Riesgo elevado más relacionado con **antecedentes familiares o factores genéticos**.
     - Riesgo bajo o moderado con estilos de vida más saludables.

Estas preguntas se irán refinando a lo largo del EDA y del modelado, siguiendo el enfoque iterativo típico de la ciencia de los datos.

2.4 Metodología y ciclo de vida

El proyecto se inspira en el estándar **CRISP-DM (Cross-Industry Standard Process for Data Mining)**, que estructura un proyecto de datos en fases:

1. **Comprensión del Negocio**
2. **Comprensión de los Datos**
3. **Preparación de los Datos**
4. **Modelado**
5. **Evaluación**

6. **Implantación / Comunicación de resultados**

En este proyecto se recorrerán estas fases de forma progresiva:

- **Comprensión del Negocio:** definición del problema, objetivos y preguntas de análisis.
- **Comprensión de los Datos:** estudio de la fuente de datos, estructura, tipos de variables y calidad de la información.
- **Preparación de los Datos:** tratamiento de valores faltantes, recodificación de variables categóricas, creación de nuevas variables útiles, etc.
- **Modelado y Evaluación:** aplicación de modelos de regresión, regresión logística y clusterización, junto con su evaluación.
- **Comunicación de resultados:** elaboración de gráficos, tablas y explicaciones en lenguaje natural, siguiendo principios de *data storytelling*.

El **Análisis Exploratorio de Datos (EDA)** tendrá un papel central: combinará estadísticas descriptivas, visualizaciones y comentarios para construir una visión clara de "qué nos dicen los datos" antes de pasar al modelado.

2.5 Fuentes de datos y estructura del dataset

2.5.1 Dataset principal: *Cancer Risk Factors* (Kaggle)

El proyecto utiliza el dataset **"Cancer Risk Factors"**, disponible públicamente en **Kaggle**. Se trata de un conjunto de datos sintético diseñado para fines educativos, que simula información de pacientes y sus factores de riesgo asociados al cáncer.

Características principales del dataset:

- Aproximadamente **2.000 registros**, cada uno correspondiente a un paciente.
- En torno a **20 variables**, entre ellas:

Variables demográficas

- `Age` – edad.
- `Gender` – género.
- `Patient_ID` – identificador del paciente.

Variables de estilo de vida

- `Smoking` – nivel de tabaquismo.
- `Alcohol_Use` – consumo de alcohol.
- `Obesity` – indicador relacionado con obesidad.
- `Diet_Red_Meat` – consumo de carne roja.
- `Diet_Salted_Processed` – consumo de alimentos salados o procesados.
- `Fruit_Veg_Intake` – consumo de frutas y verduras.
- `Physical_Activity` – nivel de actividad física.

Factores ambientales, laborales y genéticos

- `Air_Pollution` – exposición a contaminación del aire.
- `Occupational_Hazards` – exposición a riesgos laborales.
- `Family_History` – antecedentes familiares relacionados con cáncer.
- `BRCA_Mutation` – presencia de mutaciones genéticas específicas.
- `H_Pylori_Infection` – información sobre infección por *H. pylori*.
- `Calcium_Intake` y otras variables relacionadas.

Variables derivadas y de salida

- `BMI` – Índice de Masa Corporal.
- `Overall_Risk_Score` – **score continuo** que resume el riesgo global en función de los factores anteriores.
- `Risk_Level` – **nivel de riesgo categórico** (`Low`, `Medium`, `High`).

En la terminología de la asignatura, trabajaremos con una mezcla de **variables numéricas** (continuas/discretas) y **variables categóricas** (nominales/ordinales). Esta clasificación será clave para elegir los gráficos adecuados y las técnicas estadísticas más apropiadas en el EDA.

2.6 Alcance del análisis

El proyecto completo abarcará:

1. **Comprensión del negocio y de los datos**

   - Contexto, problema de negocio, objetivos y preguntas de análisis.
   - Descripción de la fuente de datos y de la estructura del dataset.
2. **Análisis Exploratorio de Datos (EDA)**
   - Análisis univariado y bi/multivariado.
   - Evaluación de la calidad de los datos (valores faltantes, duplicados, outliers).
   - Identificación de patrones relevantes y de posibles perfiles de riesgo.
3. **Modelado y evaluación**
   - Modelos de regresión y regresión logística para explicar y predecir el riesgo.
   - Análisis de clusterización para identificar perfiles de pacientes.
   - Evaluación e interpretación de los resultados desde una perspectiva de negocio.
4. **Comunicación de resultados**
   - Elaboración de visualizaciones y narrativas que permitan responder de forma clara a la pregunta:
     > **¿Qué nos dicen los datos sobre los factores de riesgo de cáncer en este conjunto de pacientes?**

2.7 Tecnología y herramientas

Para alinear el proyecto con las herramientas trabajadas en clase, se utilizarán:

- **Lenguaje:** R.
- **Entorno de desarrollo:** RStudio.
- **Informes reproducibles:** Quarto, generando una web estática en formato HTML.
- **Publicación:** GitHub Pages, utilizando el repositorio: anthony-cuervo23.github.io.

Librerías principales de R

- `readr` para lectura de ficheros CSV.
- `dplyr` y el ecosistema **tidyverse** para manipulación y transformación de datos.
- `ggplot2` para la visualización de datos (gráficos para EDA y *data story-*

telling).

En fases posteriores se podrán incorporar librerías adicionales para modelado y evaluación, siempre siguiendo la metodología vista en la asignatura.

Part I

Datos

Chapter 3

Comprensión de los datos

3.1 1. Lectura y organización de los datos

En esta sección se realiza la **lectura del fichero CSV** con R y se obtienen las primeras características básicas del conjunto de datos:

- Número de registros (filas).
- Número de variables (columnas).
- Vista rápida de la estructura de la tabla.

```
library(readr)
library(dplyr)
library(tidyr)
library(tibble)

# Lectura del dataset desde la carpeta 'data'
cancer <- read_csv("data/cancer-risk-factors.csv")

n_filas <- nrow(cancer)
n_columnas <- ncol(cancer)

cat("Número de registros:", n_filas, "\n")
```

```
Número de registros: 2000
```

```
cat("Número de variables:", n_columnas, "\n")
```

```
Número de variables: 21
```

A continuación se muestran las primeras filas de la tabla, para tener una idea

inicial del contenido:

```
head(cancer, 6)
```

```
# A tibble: 6 x 21
  Patient_ID Cancer_Type   Age Gender Smoking Alcohol_Use Obesity Family_History
  <chr>      <chr>       <dbl>  <dbl>   <dbl>       <dbl>   <dbl>          <dbl>
1 LU0000     Breast         68      0       7           2       8              0
2 LU0001     Prostate       74      1       8           9       8              0
3 LU0002     Skin           55      1       7          10       7              0
4 LU0003     Colon          61      0       6           2       2              0
5 LU0004     Lung           67      1      10           7       4              0
6 LU0005     Lung           77      1      10           8       3              0
# i 13 more variables: Diet_Red_Meat <dbl>, Diet_Salted_Processed <dbl>,
#   Fruit_Veg_Intake <dbl>, Physical_Activity <dbl>, Air_Pollution <dbl>,
#   Occupational_Hazards <dbl>, BRCA_Mutation <dbl>, H_Pylori_Infection <dbl>,
#   Calcium_Intake <dbl>, Overall_Risk_Score <dbl>, BMI <dbl>,
#   Physical_Activity_Level <dbl>, Risk_Level <chr>
```

Y una vista general de la estructura (tipos de cada columna, ejemplos de valores):

```
glimpse(cancer)
```

```
Rows: 2,000
Columns: 21
$ Patient_ID              <chr> "LU0000", "LU0001", "LU0002", "LU0003", "LU000~
$ Cancer_Type             <chr> "Breast", "Prostate", "Skin", "Colon", "Lung",~
$ Age                     <dbl> 68, 74, 55, 61, 67, 77, 59, 74, 71, 55, 63, 82~
$ Gender                  <dbl> 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1~
$ Smoking                 <dbl> 7, 8, 7, 6, 10, 10, 10, 8, 9, 7, 10, 8, 9, 10,~
$ Alcohol_Use             <dbl> 2, 9, 10, 2, 7, 8, 10, 6, 0, 1, 4, 0, 9, 9, 8,~
$ Obesity                 <dbl> 8, 8, 7, 2, 4, 3, 0, 2, 3, 2, 3, 5, 8, 1, 2, 6~
$ Family_History          <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0~
$ Diet_Red_Meat           <dbl> 5, 0, 3, 6, 6, 6, 9, 3, 10, 0, 0, 1, 4, 0, 5, ~
$ Diet_Salted_Processed   <dbl> 3, 3, 3, 2, 3, 0, 4, 3, 4, 4, 10, 1, 6, 2, 0, ~
$ Fruit_Veg_Intake        <dbl> 7, 7, 4, 4, 10, 6, 0, 2, 6, 2, 10, 1, 3, 7, 5,~
$ Physical_Activity       <dbl> 4, 1, 1, 6, 9, 2, 1, 8, 10, 5, 8, 5, 2, 4, 8, ~
$ Air_Pollution           <dbl> 6, 3, 8, 4, 10, 10, 10, 8, 8, 9, 8, 5, 2, 4, 5~
$ Occupational_Hazards    <dbl> 3, 3, 10, 8, 9, 7, 9, 7, 3, 9, 4, 10, 10, 6, 3~
$ BRCA_Mutation           <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ H_Pylori_Infection      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ Calcium_Intake          <dbl> 0, 5, 6, 8, 5, 0, 5, 1, 5, 5, 4, 4, 4, 5, 1, 2~
$ Overall_Risk_Score      <dbl> 0.3986961, 0.4242990, 0.6050822, 0.3184487, 0.~
$ BMI                     <dbl> 28.0, 25.4, 28.6, 32.1, 25.1, 25.1, 32.3, 29.1~
$ Physical_Activity_Level <dbl> 5, 9, 2, 7, 2, 1, 2, 9, 5, 1, 7, 0, 6, 1, 9, 5~
$ Risk_Level              <chr> "Medium", "Medium", "Medium", "Low", "Medium",~
```

Esta información permite comprobar que los datos se han cargado correctamente

y que las variables tienen tipos razonables (numérico, carácter, etc.) para poder trabajar con ellas en las siguientes fases.

3.2 2. Sobre la fuente de datos y el diccionario de variables

El dataset **"Cancer Risk Factors"** procede de la plataforma Kaggle y ha sido diseñado con fines educativos. Cada fila representa un paciente y cada columna recoge información relacionada con:

- **Demografía** (por ejemplo, `Age`, `Gender`, `Patient_ID`).
- **Estilo de vida** (por ejemplo, `Smoking`, `Alcohol_Use`, `Obesity`, `Diet_Red_Meat`, `Fruit_Veg_Intake`, `Physical_Activity`).
- **Factores ambientales y genéticos** (por ejemplo, `Air_Pollution`, `Occupational_Hazards`, `Family_History`, `BRCA_Mutation`, `H_Pylori_Infection`).
- **Variables agregadas de salida** (`BMI`, `Overall_Risk_Score`, `Risk_Level`).

Para facilitar la comprensión, se construye un pequeño **diccionario de variables** con algunas columnas clave, indicando su rol en el proyecto.

```
diccionario <- tribble(
  ~variable,             ~descripcion,                                          ~tipo_datos,      ~ro
  "Patient_ID",          "Identificador único del paciente",                  "categórica",     "ide
  "Age",                 "Edad del paciente",                                 "numérica",       "exp
  "Gender",              "Género del paciente",                               "categórica",     "exp
  "BMI",                 "Índice de Masa Corporal",                           "numérica",       "exp
  "Smoking",             "Nivel de tabaquismo",                               "categórica",     "exp
  "Alcohol_Use",         "Nivel de consumo de alcohol",                       "categórica",     "exp
  "Obesity",             "Indicador relacionado con obesidad",              "categórica",     "expl
  "Physical_Activity",   "Nivel de actividad física",                         "categórica",     "exp
  "Family_History",      "Antecedentes familiares de cáncer",                 "categórica",     "exp
  "Air_Pollution",       "Nivel de exposición a contaminación ambiental",   "numérica",       "exp
  "Occupational_Hazards","Exposición a riesgos laborales",                    "categórica",     "exp
  "Overall_Risk_Score",  "Score numérico de riesgo de cáncer",                "numérica",       "obj
  "Risk_Level",          "Nivel categórico de riesgo (Low/Medium/High)",    "categórica",     "obje
)

diccionario
```

```
# A tibble: 13 x 4
   variable            descripcion                              tipo_datos rol
   <chr>               <chr>                                    <chr>      <chr>
 1 Patient_ID          Identificador único del paciente         categórica iden~
 2 Age                 Edad del paciente                        numérica   expl~
 3 Gender              Género del paciente                      categórica expl~
```

```
 4 BMI                  Índice de Masa Corporal                 numérica   expl~
 5 Smoking              Nivel de tabaquismo                     categórica expl~
 6 Alcohol_Use          Nivel de consumo de alcohol             categórica expl~
 7 Obesity              Indicador relacionado con obesidad      categórica expl~
 8 Physical_Activity    Nivel de actividad física               categórica expl~
 9 Family_History       Antecedentes familiares de cáncer       categórica expl~
10 Air_Pollution        Nivel de exposición a contaminación am~ numérica   expl~
11 Occupational_Hazards Exposición a riesgos laborales          categórica expl~
12 Overall_Risk_Score   Score numérico de riesgo de cáncer      numérica   obje~
13 Risk_Level           Nivel categórico de riesgo (Low/Medium~ categórica obje~
```

En términos de la asignatura, trabajaremos con una mezcla de:

- **Variables numéricas** (continuas o discretas), como `Age`, `BMI`, `Air_Pollution`, `Overall_Risk_Score`.
- **Variables categóricas** (nominales u ordinales), como `Gender`, `Smoking`, `Alcohol_Use`, `Risk_Level`, etc.

Esta distinción será importante a la hora de elegir los gráficos y las técnicas de análisis en el EDA.

3.3 3. Clasificación de variables por tipo

Para tener una visión más sistemática, se obtiene el **tipo de dato** que R asigna a cada columna:

```r
tipos <- sapply(cancer, function(x) class(x)[1])

tipos_df <- tibble(
  variable = names(tipos),
  tipo_r = unname(tipos)
)

tipos_df
```

```
# A tibble: 21 x 2
   variable             tipo_r
   <chr>                <chr>
 1 Patient_ID           character
 2 Cancer_Type          character
 3 Age                  numeric
 4 Gender               numeric
 5 Smoking              numeric
 6 Alcohol_Use          numeric
 7 Obesity              numeric
 8 Family_History       numeric
 9 Diet_Red_Meat        numeric
```

```
10 Diet_Salted_Processed numeric
# i 11 more rows
```

A partir de esta tabla, se construyen dos listas: **variables numéricas** y **variables categóricas**, que se utilizarán luego en el análisis exploratorio.

```
variables_numericas <- tipos_df %>%
  filter(tipo_r %in% c("numeric", "integer", "double")) %>%
  pull(variable)

variables_categoricas <- tipos_df %>%
  filter(!tipo_r %in% c("numeric", "integer", "double")) %>%
  pull(variable)

cat("Variables numéricas:\n")
```

```
Variables numéricas:
```

```
print(variables_numericas)
```

```
 [1] "Age"                    "Gender"
 [3] "Smoking"                "Alcohol_Use"
 [5] "Obesity"                "Family_History"
 [7] "Diet_Red_Meat"          "Diet_Salted_Processed"
 [9] "Fruit_Veg_Intake"       "Physical_Activity"
[11] "Air_Pollution"          "Occupational_Hazards"
[13] "BRCA_Mutation"          "H_Pylori_Infection"
[15] "Calcium_Intake"         "Overall_Risk_Score"
[17] "BMI"                    "Physical_Activity_Level"
```

```
cat("\nVariables categóricas:\n")
```

```
Variables categóricas:
```

```
print(variables_categoricas)
```

```
[1] "Patient_ID"  "Cancer_Type" "Risk_Level"
```

3.4 4. Calidad de los datos: valores faltantes y duplicados

Antes de profundizar en el EDA, es importante revisar algunos aspectos básicos de **calidad de los datos**:

- ¿Hay valores faltantes (`NA`) en alguna columna?
- ¿Existen filas duplicadas?

3.4.1 4.1. Valores faltantes por variable

```
na_resumen <- cancer %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(cols = everything(),
               names_to = "variable",
               values_to = "na_count") %>%
  arrange(desc(na_count))

na_resumen
```

```
# A tibble: 21 x 2
   variable              na_count
   <chr>                    <int>
 1 Patient_ID                   0
 2 Cancer_Type                  0
 3 Age                          0
 4 Gender                       0
 5 Smoking                      0
 6 Alcohol_Use                  0
 7 Obesity                      0
 8 Family_History               0
 9 Diet_Red_Meat                0
10 Diet_Salted_Processed        0
# i 11 more rows
```

Esta tabla indica, para cada variable, **cuántos valores faltantes** hay. Si alguna columna tiene un número significativo de `NA`, se comentará en el EDA y se decidirá cómo tratarla (eliminar filas, imputar valores, etc.).

3.4.2 4.2. Búsqueda de registros duplicados

```
total_registros <- nrow(cancer)
registros_unicos <- nrow(distinct(cancer))

cat("Total de registros: ", total_registros, "\n")
```

```
Total de registros:  2000
```

```
cat("Registros únicos:   ", registros_unicos, "\n")
```

```
Registros únicos:    2000
```

```
cat("Registros duplicados:", total_registros - registros_unicos, "\n")
```

```
Registros duplicados: 0
```

Si el número de registros duplicados es mayor que cero, será necesario analizar si

se trata de:

- Errores de carga de datos.
- Repeticiones legítimas (por ejemplo, pacientes con el mismo patrón de variables).

En función de ese análisis, se decidirá si conviene eliminar o mantener esas filas en el dataset de trabajo.

3.5 5. Resumen descriptivo inicial

Finalmente, se obtiene un **resumen estadístico básico** de las variables numéricas, que servirá como punto de partida para el EDA:

```
cancer %>%
  select(all_of(variables_numericas)) %>%
  summary()
```

```
      Age            Gender          Smoking         Alcohol_Use
 Min.   :25.00   Min.   :0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.:56.00   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.: 2.000
 Median :64.00   Median :0.000   Median : 5.000   Median : 5.000
 Mean   :63.25   Mean   :0.489   Mean   : 5.157   Mean   : 5.035
 3rd Qu.:70.00   3rd Qu.:1.000   3rd Qu.: 8.000   3rd Qu.: 8.000
 Max.   :90.00   Max.   :1.000   Max.   :10.000   Max.   :10.000
    Obesity       Family_History   Diet_Red_Meat    Diet_Salted_Processed
 Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.000   1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.: 2.000
 Median : 6.000   Median :0.0000   Median : 5.000   Median : 4.000
 Mean   : 5.968   Mean   :0.1945   Mean   : 5.189   Mean   : 4.564
 3rd Qu.: 9.000   3rd Qu.:0.0000   3rd Qu.: 8.000   3rd Qu.: 7.000
 Max.   :10.000   Max.   :1.0000   Max.   :10.000   Max.   :10.000
 Fruit_Veg_Intake Physical_Activity Air_Pollution    Occupational_Hazards
 Min.   : 0.000   Min.   : 0.000    Min.   : 0.000   Min.   : 0.000
 1st Qu.: 3.000   1st Qu.: 1.000    1st Qu.: 3.000   1st Qu.: 2.000
 Median : 5.000   Median : 4.000    Median : 5.000   Median : 5.000
 Mean   : 4.928   Mean   : 4.015    Mean   : 5.323   Mean   : 4.979
 3rd Qu.: 8.000   3rd Qu.: 6.000    3rd Qu.: 8.000   3rd Qu.: 8.000
 Max.   :10.000   Max.   :10.000    Max.   :10.000   Max.   :10.000
 BRCA_Mutation    H_Pylori_Infection Calcium_Intake   Overall_Risk_Score
 Min.   :0.0000   Min.   :0.0000     Min.   : 0.000   Min.   :0.02928
 1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.: 1.000   1st Qu.:0.36698
 Median :0.0000   Median :0.0000     Median : 4.000   Median :0.45540
 Mean   :0.0325   Mean   :0.1965     Mean   : 3.941   Mean   :0.45445
 3rd Qu.:0.0000   3rd Qu.:0.0000     3rd Qu.: 6.000   3rd Qu.:0.53978
 Max.   :1.0000   Max.   :1.0000     Max.   :10.000   Max.   :0.85216
      BMI        Physical_Activity_Level
```

```
 Min.   :15.00   Min.   : 0.000
 1st Qu.:23.50   1st Qu.: 2.000
 Median :26.20   Median : 5.000
 Mean   :26.18   Mean   : 4.939
 3rd Qu.:28.70   3rd Qu.: 8.000
 Max.   :41.40   Max.   :10.000
```

Para las variables categóricas más importantes (por ejemplo, `Risk_Level`, `Smoking`, `Alcohol_Use`, `Physical_Activity`), resulta útil ver sus tablas de frecuencias:

```
frecuencias <- cancer %>%
  select(Risk_Level, Smoking, Alcohol_Use, Physical_Activity) %>%
  lapply(table)

frecuencias
```

```
$Risk_Level

  High    Low Medium
   102    324   1574

$Smoking

  0   1   2   3   4   5   6   7   8   9  10
169 204 174 175 166 166 267 103 107 126 343

$Alcohol_Use

  0   1   2   3   4   5   6   7   8   9  10
204 186 210 151 145 126 183 220 195 197 183

$Physical_Activity

  0   1   2   3   4   5   6   7   8   9  10
242 261 225 254 246 173 152 111 117 107 112
```

3.6 6. Conclusión de la fase de comprensión de los datos

En esta sección se ha:

- Leído el fichero CSV con R y comprobado el número de registros y variables.
- Elaborado un pequeño **diccionario de datos** con algunas variables clave y su rol en el proyecto.
- Clasificado las variables en **numéricas** y **categóricas**.

- Revisado la presencia de **valores faltantes** y **posibles duplicados**.
- Obtenido un primer resumen estadístico de las variables numéricas y las frecuencias de algunas variables categóricas.

Con esta base, el siguiente capítulo se centrará en el **Análisis Exploratorio de Datos (EDA)**, utilizando visualizaciones y análisis univariados y bivariados para responder, de forma más visual y narrativa, a la pregunta:

> **¿Qué nos dicen los datos sobre los factores de riesgo de cáncer en este conjunto de pacientes?**

Chapter 4

Análisis exploratorio de los datos (EDA)

4.1 1. Introducción al EDA

En este capítulo se realiza el **Análisis Exploratorio de Datos (EDA)** del conjunto de datos de factores de riesgo de cáncer.

Los objetivos principales son:

- Entender cómo se distribuyen las variables más relevantes (análisis univariado).
- Explorar relaciones entre factores de riesgo y medidas de salida (`Overall_Risk_Score`, `Risk_Level`) (análisis bi/multivariado).
- Detectar posibles problemas de calidad y características del dataset que puedan afectar a los modelos posteriores (desbalance de clases, valores extremos, etc.).

El análisis combina tablas, gráficos y comentarios en lenguaje natural, siguiendo la filosofía de *"¿qué nos dicen los datos?"* trabajada en la asignatura.

4.2 2. Preparación: carga de datos y librerías

```
library(readr)
library(dplyr)
library(tidyr)
```

```
library(tibble)
library(ggplot2)
library(forcats)
library(scales)

# Lectura del dataset
cancer <- read_csv("data/cancer-risk-factors.csv")

# Identificamos tipos de variables
tipos <- sapply(cancer, function(x) class(x)[1])

tipos_df <- tibble(
  variable = names(tipos),
  tipo_r = unname(tipos)
)

variables_numericas <- tipos_df %>%
  filter(tipo_r %in% c("numeric", "integer", "double")) %>%
  pull(variable)

variables_categoricas <- tipos_df %>%
  filter(!tipo_r %in% c("numeric", "integer", "double")) %>%
  pull(variable)

variables_numericas
```

```
 [1] "Age"                     "Gender"
 [3] "Smoking"                 "Alcohol_Use"
 [5] "Obesity"                 "Family_History"
 [7] "Diet_Red_Meat"           "Diet_Salted_Processed"
 [9] "Fruit_Veg_Intake"        "Physical_Activity"
[11] "Air_Pollution"           "Occupational_Hazards"
[13] "BRCA_Mutation"           "H_Pylori_Infection"
[15] "Calcium_Intake"          "Overall_Risk_Score"
[17] "BMI"                     "Physical_Activity_Level"
```

```
variables_categoricas
```

```
[1] "Patient_ID"  "Cancer_Type" "Risk_Level"
```

En el capítulo de **comprensión de los datos** ya se vio que el dataset contiene 2.000 registros y alrededor de 20 variables. Aquí se parte de esos mismos datos para profundizar en sus patrones.

4.3 3. Análisis univariado de variables numéricas

4.3.1 3.1. Resumen estadístico general

```
cancer %>%
  select(all_of(variables_numericas)) %>%
  summary()
```

```
      Age            Gender          Smoking         Alcohol_Use
 Min.   :25.00   Min.   :0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.:56.00   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.: 2.000
 Median :64.00   Median :0.000   Median : 5.000   Median : 5.000
 Mean   :63.25   Mean   :0.489   Mean   : 5.157   Mean   : 5.035
 3rd Qu.:70.00   3rd Qu.:1.000   3rd Qu.: 8.000   3rd Qu.: 8.000
 Max.   :90.00   Max.   :1.000   Max.   :10.000   Max.   :10.000
    Obesity       Family_History   Diet_Red_Meat    Diet_Salted_Processed
 Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.000   1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.: 2.000
 Median : 6.000   Median :0.0000   Median : 5.000   Median : 4.000
 Mean   : 5.968   Mean   :0.1945   Mean   : 5.189   Mean   : 4.564
 3rd Qu.: 9.000   3rd Qu.:0.0000   3rd Qu.: 8.000   3rd Qu.: 7.000
 Max.   :10.000   Max.   :1.0000   Max.   :10.000   Max.   :10.000
 Fruit_Veg_Intake Physical_Activity Air_Pollution    Occupational_Hazards
 Min.   : 0.000   Min.   : 0.000    Min.   : 0.000   Min.   : 0.000
 1st Qu.: 3.000   1st Qu.: 1.000    1st Qu.: 3.000   1st Qu.: 2.000
 Median : 5.000   Median : 4.000    Median : 5.000   Median : 5.000
 Mean   : 4.928   Mean   : 4.015    Mean   : 5.323   Mean   : 4.979
 3rd Qu.: 8.000   3rd Qu.: 6.000    3rd Qu.: 8.000   3rd Qu.: 8.000
 Max.   :10.000   Max.   :10.000    Max.   :10.000   Max.   :10.000
 BRCA_Mutation    H_Pylori_Infection Calcium_Intake   Overall_Risk_Score
 Min.   :0.0000   Min.   :0.0000     Min.   : 0.000   Min.   :0.02928
 1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.: 1.000   1st Qu.:0.36698
 Median :0.0000   Median :0.0000     Median : 4.000   Median :0.45540
 Mean   :0.0325   Mean   :0.1965     Mean   : 3.941   Mean   :0.45445
 3rd Qu.:0.0000   3rd Qu.:0.0000     3rd Qu.: 6.000   3rd Qu.:0.53978
 Max.   :1.0000   Max.   :1.0000     Max.   :10.000   Max.   :0.85216
      BMI        Physical_Activity_Level
 Min.   :15.00   Min.   : 0.000
 1st Qu.:23.50   1st Qu.: 2.000
 Median :26.20   Median : 5.000
 Mean   :26.18   Mean   : 4.939
 3rd Qu.:28.70   3rd Qu.: 8.000
 Max.   :41.40   Max.   :10.000
```

Este resumen confirma rangos razonables para las variables numéricas y permite detectar rápidamente valores mínimos/máximos y posibles outliers.

4.3.2 3.2. Histogramas de variables seleccionadas

Se analizan tres variables numéricas clave: `Age`, `BMI` y `Overall_Risk_Score`.

4.3.2.1 3.2.1. Edad (`Age`)

```
ggplot(cancer, aes(x = Age)) +
  geom_histogram(bins = 30) +
  labs(x = "Edad", y = "Frecuencia") +
  theme_minimal()
```

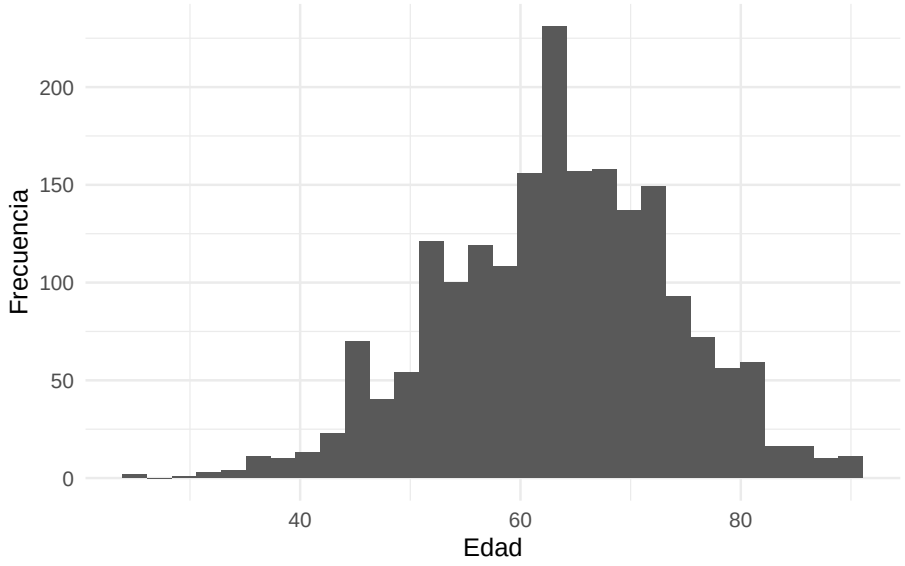

Figure 4.1: Distribución de la edad de los pacientes.

Comentario: La distribución de la edad se concentra principalmente entre los **45 y 80 años**, con un pico alrededor de los **60–65 años**. Hay muy pocos pacientes por debajo de los 40 años. Esto indica que el dataset representa sobre todo a **población adulta de mediana y avanzada edad**, algo coherente con el hecho de que el riesgo de cáncer suele aumentar con la edad.

4.3.2.2 3.2.2. Índice de masa corporal (`BMI`)

```
ggplot(cancer, aes(x = BMI)) +
  geom_histogram(bins = 30) +
```

```
  labs(x = "BMI", y = "Frecuencia") +
  theme_minimal()
```

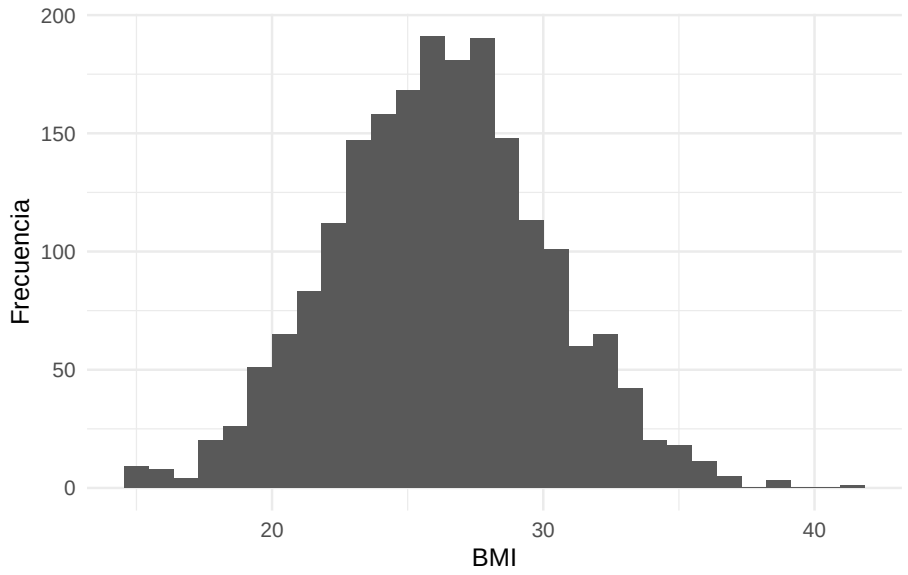

Figure 4.2: Distribución del índice de masa corporal (BMI).

Comentario: El BMI presenta una distribución aproximadamente en forma de campana, centrada en valores de **25–30**. La mayoría de los pacientes se sitúa, por tanto, en rangos de **sobrepeso u obesidad ligera**, con pocos casos de BMI muy bajo o muy alto. Este patrón es coherente con un escenario en el que el exceso de peso es un factor de riesgo frecuente en la población.

4.3.2.3 3.2.3. Score de riesgo global (Overall_Risk_Score)

```
ggplot(cancer, aes(x = Overall_Risk_Score)) +
  geom_histogram(bins = 30) +
  labs(x = "Overall_Risk_Score", y = "Frecuencia") +
  theme_minimal()
```

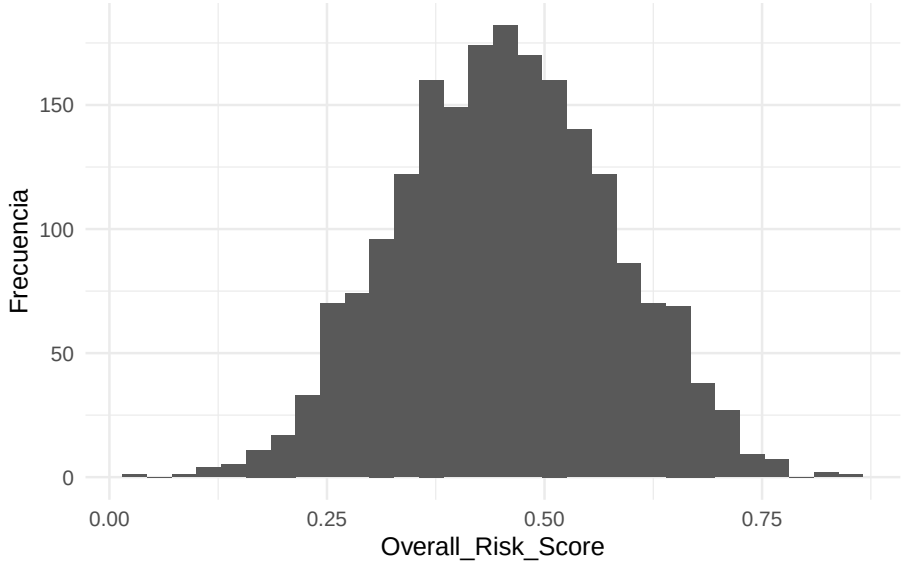

Figure 4.3: Distribución del score de riesgo global (Overall_Risk_Score).

Comentario: La variable `Overall_Risk_Score` tiene una distribución aproximadamente **normal**, centrada cerca de **0,5** y con la mayoría de valores entre ~0,3 y 0,7. Esto sugiere que el dataset está construido de forma que la mayoría de pacientes tienen un **riesgo intermedio**, con pocos casos en los extremos de riesgo muy bajo o muy alto.

4.4 4. Análisis univariado de variables categóricas

En esta sección se exploran las distribuciones de algunas variables categóricas relevantes:

- `Risk_Level`
- `Smoking`
- `Alcohol_Use`
- `Physical_Activity`
- `Family_History`

4.4.1 4.1. Tablas de frecuencias

```r
frecuencias <- cancer %>%
  select(Risk_Level, Smoking, Alcohol_Use, Physical_Activity, Family_History) %>%
  lapply(table)

frecuencias
```

```
$Risk_Level

  High    Low Medium
   102    324   1574

$Smoking

  0   1   2   3   4   5   6   7   8   9  10
169 204 174 175 166 166 267 103 107 126 343

$Alcohol_Use

  0   1   2   3   4   5   6   7   8   9  10
204 186 210 151 145 126 183 220 195 197 183

$Physical_Activity

  0   1   2   3   4   5   6   7   8   9  10
242 261 225 254 246 173 152 111 117 107 112

$Family_History

   0    1
1611  389
```

A partir de estas tablas se observan, por ejemplo:

- `Risk_Level` está claramente **desbalanceada**:
  - La categoría **Medium** concentra la gran mayoría de los casos.
  - **Low** tiene una proporción menor.
  - **High** representa solo una pequeña fracción (unos 100 casos de 2.000, es decir, en torno al 5%).
- `Family_History` tiene muchos más pacientes sin antecedentes (`0`) que con antecedentes (`1`).
- `Smoking`, `Alcohol_Use` y `Physical_Activity` se codifican en niveles **0–10**, cubriendo un amplio rango de intensidad/ frecuencia.

4.4.2 4.2. Gráficos de barras

4.4.2.1 4.2.1. Nivel de riesgo (`Risk_Level`)

```
cancer %>%
  mutate(Risk_Level = as.factor(Risk_Level)) %>%
  ggplot(aes(x = Risk_Level)) +
  geom_bar() +
  labs(x = "Risk_Level", y = "Número de pacientes") +
  theme_minimal()
```

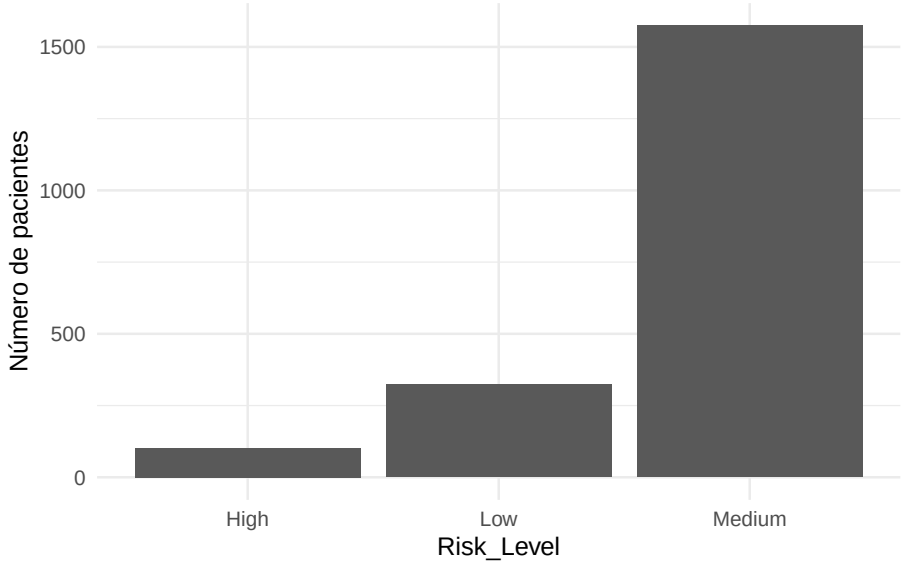

Figure 4.4: Distribución del nivel de riesgo (Risk_Level).

Comentario: La categoría **Medium** domina claramente el gráfico, seguida por **Low** y, en mucha menor medida, **High**. Esto confirma que el nivel de riesgo categórico está **muy desbalanceado**, algo que será importante tener en cuenta cuando se construyan modelos de clasificación, ya que la clase minoritaria (alto riesgo) podría ser más difícil de aprender.

4.4.2.2 4.2.2. Tabaquismo (`Smoking`)

```
cancer %>%
  mutate(Smoking = factor(Smoking, levels = sort(unique(Smoking)))) %>%
  ggplot(aes(x = Smoking)) +
```

```
  geom_bar() +
  labs(x = "Smoking", y = "Número de pacientes") +
  theme_minimal()
```

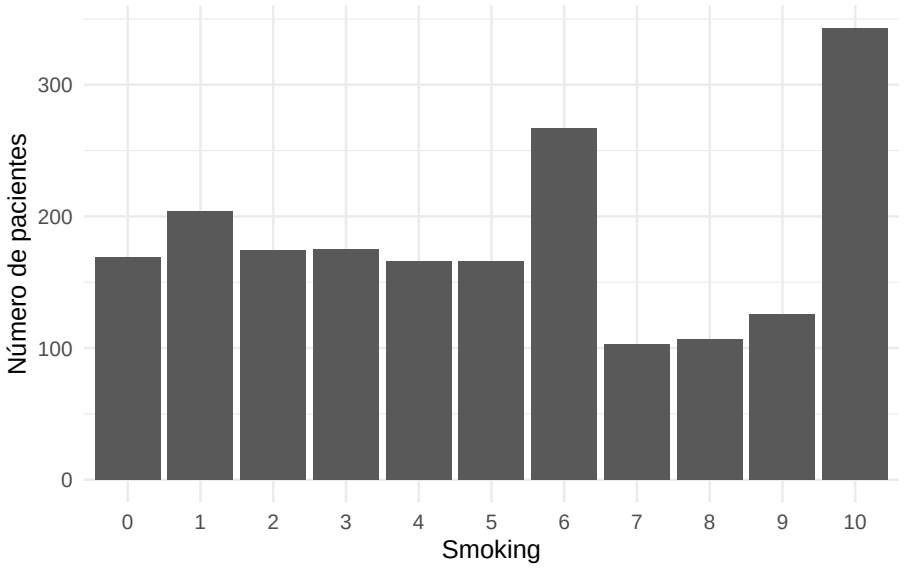

Figure 4.5: Distribución de los niveles de tabaquismo (Smoking).

Comentario: La variable `Smoking` (0–10) presenta pacientes repartidos a lo largo de toda la escala, con una presencia apreciable tanto en niveles bajos como en niveles medios y altos. Esto indica que el dataset contiene **patrones de consumo muy variados**, desde no fumadores hasta fumadores intensivos, lo que resulta interesante para estudiar cómo cambia el riesgo con el tabaquismo.

4.4.2.3 4.2.3. Consumo de alcohol (`Alcohol_Use`)

```
cancer %>%
  mutate(Alcohol_Use = factor(Alcohol_Use, levels = sort(unique(Alcohol_Use)))) %>%
  ggplot(aes(x = Alcohol_Use)) +
  geom_bar() +
  labs(x = "Alcohol_Use", y = "Número de pacientes") +
  theme_minimal()
```

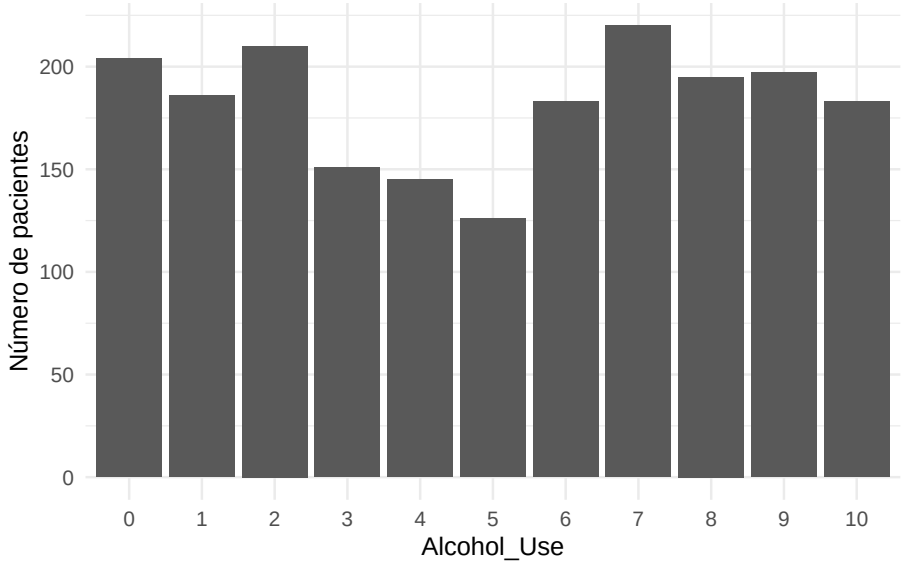

Figure 4.6: Distribución de los niveles de consumo de alcohol (Alcohol_Use).

Comentario: El consumo de alcohol también se distribuye a lo largo de los distintos niveles 0–10, sin una única categoría claramente dominante. Esto sugiere que en el dataset coexisten **distintos patrones de consumo**, lo que permitirá analizar cómo se asocia cada rango con el score de riesgo y con el nivel de riesgo categórico.

```r
cancer %>%
  mutate(Physical_Activity = factor(Physical_Activity, levels = sort(unique(Physical_A
  ggplot(aes(x = Physical_Activity)) +
  geom_bar() +
  labs(x = "Physical_Activity", y = "Número de pacientes") +
  theme_minimal()
```

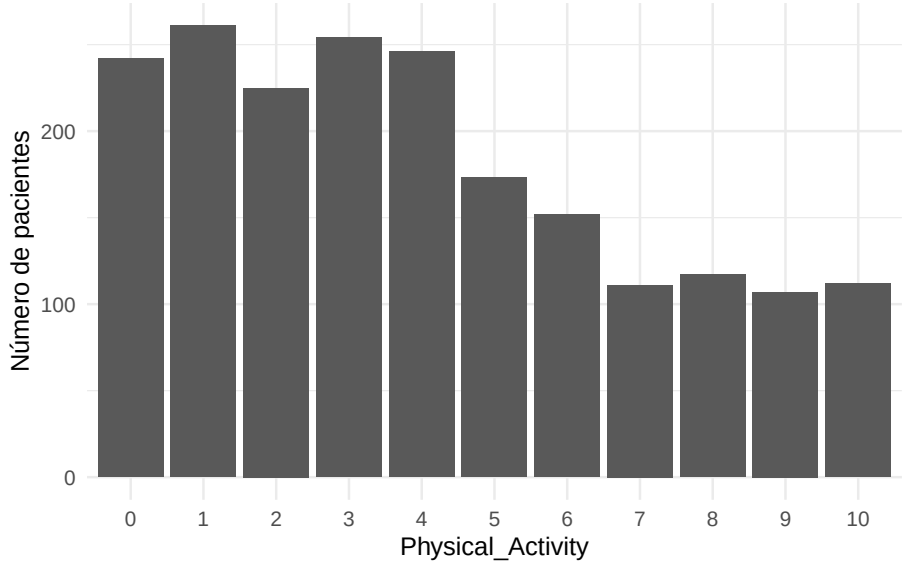

Figure 4.7: Distribución de los niveles de actividad física (Physical_Activity).

Comentario: Los niveles más frecuentes de `Physical_Activity` corresponden a valores **bajos o intermedios**, mientras que los niveles más altos son menos frecuentes. Desde el punto de vista de prevención, esto refleja un predominio de estilos de vida **poco activos**, lo que puede contribuir a un mayor riesgo global.

4.5 5. Análisis bivariado

En esta sección se exploran relaciones entre variables numéricas y entre variables numéricas y categóricas.

4.5.1 5.1. Relaciones numérica vs. numérica

4.5.1.1 5.1.1. `Age` vs `Overall_Risk_Score`

```
ggplot(cancer, aes(x = Age, y = Overall_Risk_Score)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Edad", y = "Overall_Risk_Score") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

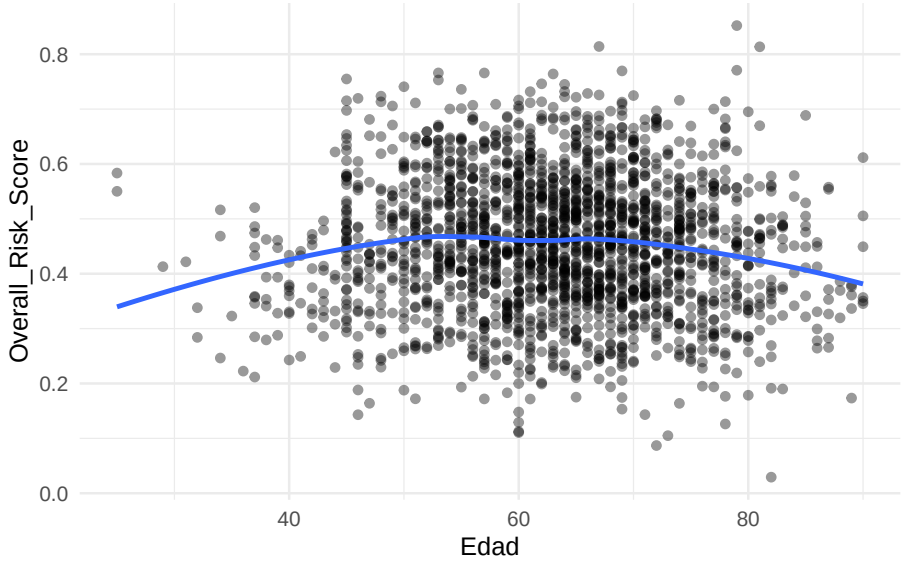

Figure 4.8: Relación entre edad (Age) y score de riesgo (Overall_Risk_Score).

Comentario: La nube de puntos está bastante dispersa y la curva de suavizado (`loess`) apenas muestra una tendencia marcada. Se aprecia, como mucho, una ligera variación del riesgo en edades medias, pero en general **la edad por sí sola no explica gran parte de la variabilidad del score de riesgo** en este dataset sintético.

4.5.1.2 5.1.2 BMI vs Overall_Risk_Score

```
ggplot(cancer, aes(x = BMI, y = Overall_Risk_Score)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "BMI", y = "Overall_Risk_Score") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

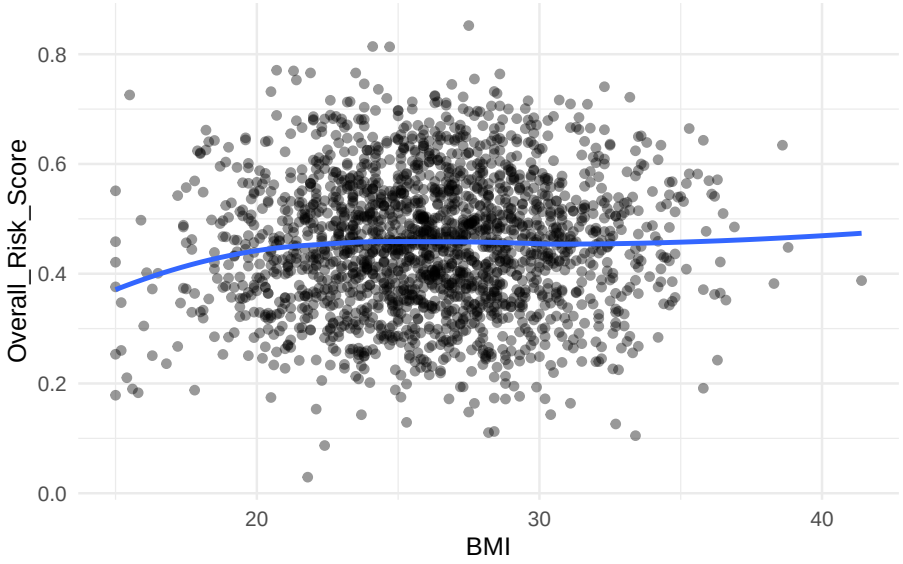

Figure 4.9: Relación entre BMI y score de riesgo (Overall_Risk_Score).

Comentario: En este caso también hay dispersión, pero la curva de suavizado muestra una **tendencia ligeramente creciente**: a medida que aumenta el BMI, el `Overall_Risk_Score` tiende a ser algo mayor. Aunque la relación no es muy fuerte, sí apunta en la dirección esperada desde el punto de vista epidemiológico: **un mayor índice de masa corporal se asocia con un incremento del riesgo global**.

4.5.2 5.2. Relaciones numérica vs. categórica

4.5.2.1 5.2.1. Overall_Risk_Score por nivel de riesgo (Risk_Level)

```
cancer %>%
  mutate(Risk_Level = as.factor(Risk_Level)) %>%
  ggplot(aes(x = Risk_Level, y = Overall_Risk_Score)) +
  geom_boxplot() +
  labs(x = "Risk_Level", y = "Overall_Risk_Score") +
  theme_minimal()
```

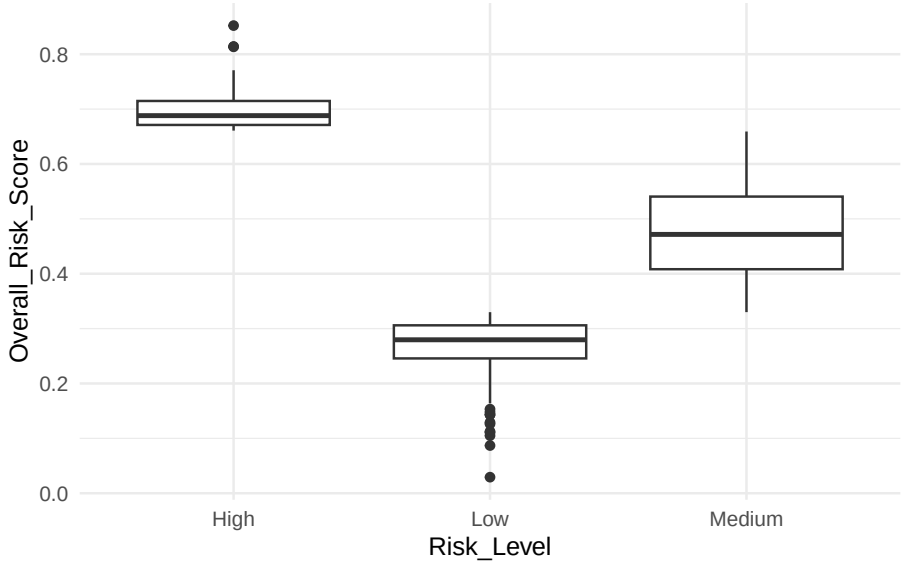

Figure 4.10: Distribución de Overall_Risk_Score según Risk_Level.

Comentario: Los boxplots muestran tres grupos claramente separados:

- Para **Low**, los scores se concentran en torno a valores bajos (~0,25–0,30).
- Para **Medium**, la mediana está en torno a ~0,45–0,50.
- Para **High**, los valores se desplazan hacia la parte alta (~0,70 en adelante).

Hay algo de solapamiento entre cajas, pero en general el `Overall_Risk_Score` **aumenta de forma consistente al pasar de Low a Medium y de Medium a High**, lo que valida la coherencia interna del dataset: la variable continua de riesgo y la categórica de nivel de riesgo están alineadas.

4.5.2.2 5.2.2. `Overall_Risk_Score` por tabaquismo (`Smoking`)

```
cancer %>%
  mutate(Smoking = factor(Smoking, levels = sort(unique(Smoking)))) %>%
  ggplot(aes(x = Smoking, y = Overall_Risk_Score)) +
  geom_boxplot() +
  labs(x = "Smoking", y = "Overall_Risk_Score") +
  theme_minimal()
```

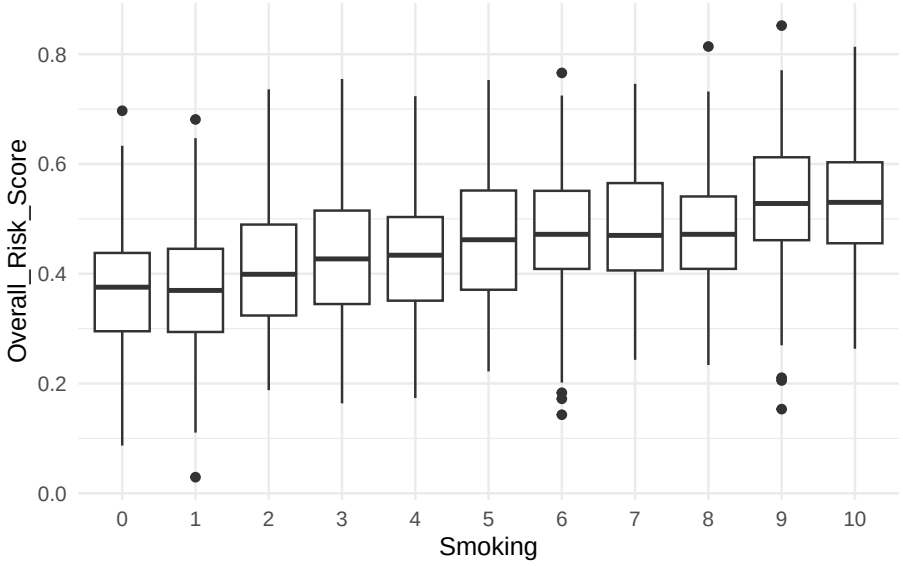

Figure 4.11: Distribución de Overall_Risk_Score según niveles de tabaquismo (Smoking).

Comentario: Al recorrer los niveles de `Smoking` de 0 a 10 se observa que:

- Los niveles bajos tienden a tener **medianas de riesgo algo menores**.
- A medida que aumenta el nivel de tabaquismo, las cajas se desplazan ligeramente hacia valores más altos de `Overall_Risk_Score`, aunque con bastante dispersión dentro de cada grupo.

En conjunto, los boxplots sugieren una **asociación positiva moderada** entre tabaquismo y riesgo global: **más tabaquismo → ligeramente más riesgo**, sin llegar a ser una relación determinista.

4.5.2.3 5.2.3. `Overall_Risk_Score` por consumo de alcohol (`Alcohol_Use`)

```
cancer %>%
  mutate(Alcohol_Use = factor(Alcohol_Use, levels = sort(unique(Alcohol_Use)))) %>%
  ggplot(aes(x = Alcohol_Use, y = Overall_Risk_Score)) +
  geom_boxplot() +
  labs(x = "Alcohol_Use", y = "Overall_Risk_Score") +
  theme_minimal()
```

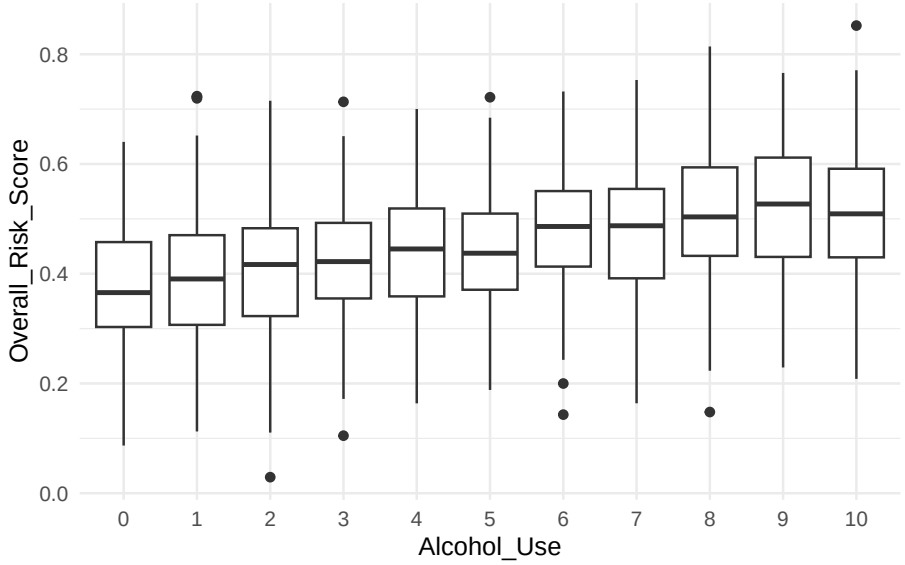

Figure 4.12: Distribución de Overall_Risk_Score según niveles de consumo de alcohol (Alcohol_Use).

Comentario: Los patrones para `Alcohol_Use` son similares a los de `Smoking`: a mayor nivel de consumo de alcohol, la distribución de `Overall_Risk_Score` tiende a situarse en valores algo más altos, aunque con variabilidad interna en cada nivel. Esto sugiere que el dataset refleja la idea de que **un mayor consumo de alcohol se asocia con un riesgo global algo mayor**, pero con la incertidumbre propia de un fenómeno multifactorial.

4.5.3 5.3. Distribución de `Risk_Level` según tabaquismo y consumo de alcohol

Para explorar cómo se reparte el nivel de riesgo según `Smoking` y `Alcohol_Use`, se utilizan gráficos de barras apiladas en proporciones.

4.5.3.1 5.3.1. `Risk_Level` por `Smoking`

```
cancer %>%
  mutate(
    Smoking = factor(Smoking, levels = sort(unique(Smoking))),
    Risk_Level = as.factor(Risk_Level)
  ) %>%
  ggplot(aes(x = Smoking, fill = Risk_Level)) +
```

```
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Smoking", y = "Proporción", fill = "Risk_Level") +
  theme_minimal()
```

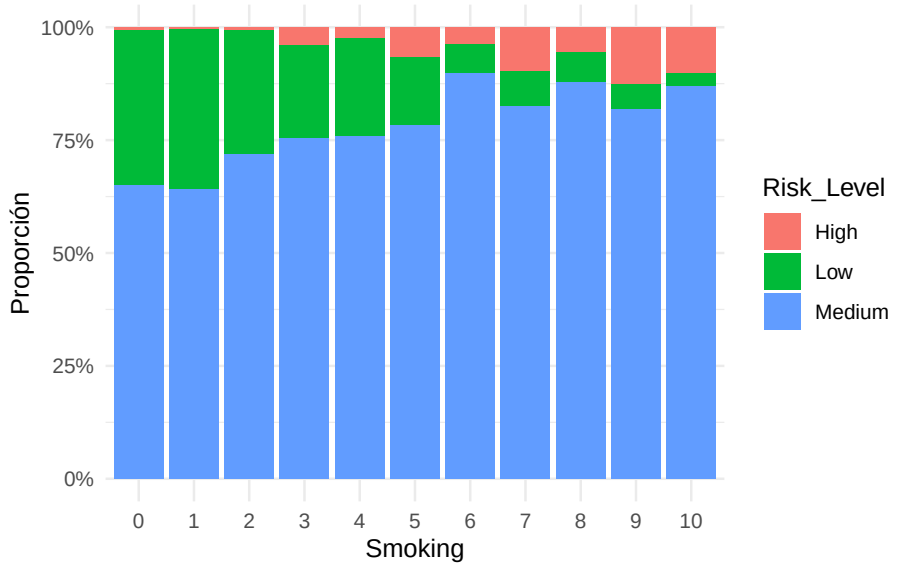

Figure 4.13: Proporción de niveles de riesgo (Risk_Level) según niveles de tabaquismo (Smoking).

Comentario: En este gráfico se aprecia que:

- En los niveles bajos de `Smoking` predomina ampliamente el riesgo **Medium**, con una proporción moderada de **Low** y muy pocos casos de **High**.
- A medida que aumenta el nivel de tabaquismo, la **banda correspondiente a `High` (color rojo)** se hace ligeramente más ancha, mientras que la de `Low` (verde) tiende a estrecharse.

Esto indica que, en este dataset, **los niveles más altos de tabaquismo están asociados con una mayor proporción de pacientes de alto riesgo**, aunque la categoría Medium sigue siendo mayoritaria en todos los niveles.

4.5.3.2 5.3.2. `Risk_Level` por `Alcohol_Use`

```
cancer %>%
  mutate(
```

```
    Alcohol_Use = factor(Alcohol_Use, levels = sort(unique(Alcohol_Use))),
    Risk_Level = as.factor(Risk_Level)
  ) %>%
  ggplot(aes(x = Alcohol_Use, fill = Risk_Level)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Alcohol_Use", y = "Proporción", fill = "Risk_Level") +
  theme_minimal()
```

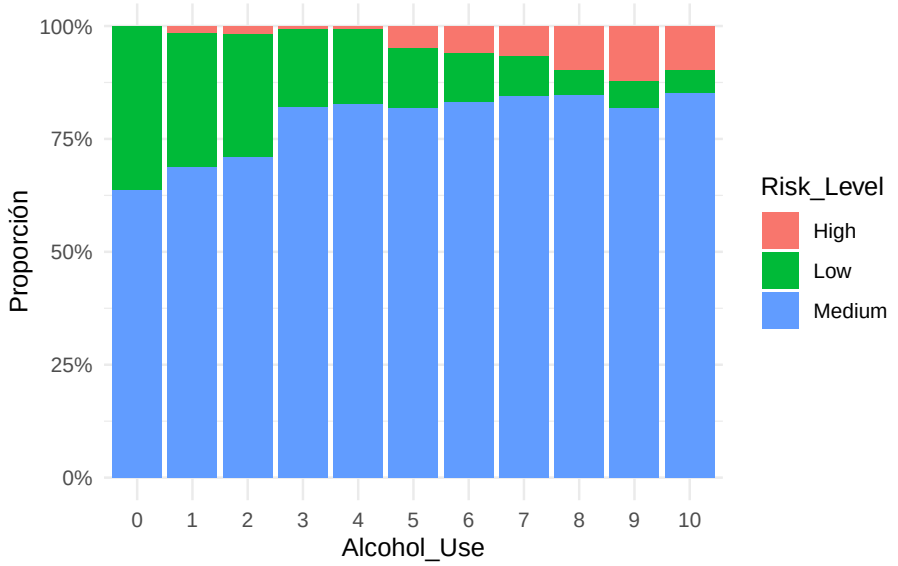

Figure 4.14: Proporción de niveles de riesgo (Risk_Level) según niveles de consumo de alcohol (Alcohol_Use).

Comentario: El patrón es similar al observado para `Smoking`:

- En niveles bajos de consumo de alcohol, la mayor parte de los pacientes se clasifica como riesgo **Medium** y una fracción no despreciable como **Low**.
- En niveles altos de `Alcohol_Use`, la proporción de casos **High** aumenta, mientras que la de **Low** disminuye.

De nuevo, esto refuerza la idea de que el dataset refleja una **relación cualitativa razonable** entre estilos de vida menos saludables (más tabaquismo, más alcohol) y un mayor nivel de riesgo.

4.6 6. Correlación entre variables numéricas

Como resumen de las relaciones entre variables numéricas, se visualiza una **matriz de correlación** para algunas variables seleccionadas.

```
vars_num_seleccionadas <- c("Age", "Air_Pollution", "BMI", "Overall_Risk_Score")

datos_num <- cancer %>%
  select(all_of(vars_num_seleccionadas)) %>%
  drop_na()

matriz_cor <- cor(datos_num)

matriz_cor_df <- as.data.frame(matriz_cor) %>%
  mutate(var1 = rownames(matriz_cor)) %>%
  pivot_longer(
    cols = -var1,
    names_to = "var2",
    values_to = "correlacion"
  )

ggplot(matriz_cor_df, aes(x = var1, y = var2, fill = correlacion)) +
  geom_tile() +
  scale_fill_gradient2(
    limits = c(-1, 1),
    midpoint = 0
  ) +
  labs(x = "", y = "", fill = "Correlación") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

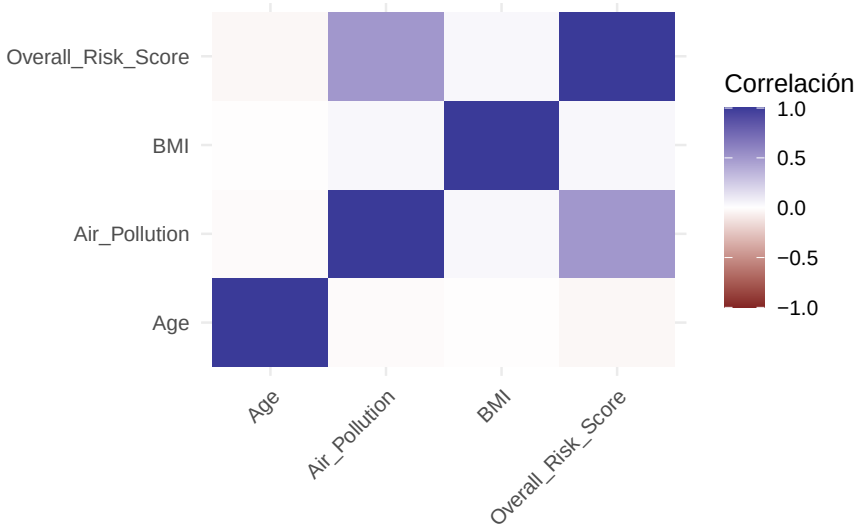

Figure 4.15: Matriz de correlación entre variables numéricas seleccionadas.

Comentario: La matriz muestra, de forma cualitativa, que:

- `Overall_Risk_Score` guarda **correlaciones positivas moderadas** con algunas variables como `BMI` y `Air_Pollution`.
- La correlación con `Age` parece más débil.
- No se observan correlaciones extremadamente altas (cercanas a 1 en valor absoluto) entre las variables seleccionadas.

Esto sugiere que, al construir modelos, es poco probable que haya problemas graves de **colinealidad** entre estas variables numéricas, aunque será conveniente revisarlo con más detalle en la fase de modelado.

4.7 7. Resumen del análisis exploratorio

A partir del EDA realizado se pueden extraer las siguientes ideas principales:

1. **Perfil de la población del dataset**
    - Predominan pacientes adultos de **mediana y avanzada edad** (45–80 años).
    - La mayoría presenta valores de **BMI en rango de sobrepeso u obesidad moderada**.
2. **Distribución del riesgo**

- El `Overall_Risk_Score` se distribuye de forma aproximadamente normal alrededor de 0,5, indicando un predominio de **niveles de riesgo intermedio**.
- La variable categórica `Risk_Level` está **desbalanceada**, con muchos casos en **Medium**, menos en **Low** y pocos en **High**.

3. **Hábitos y estilo de vida**
   - Las variables `Smoking`, `Alcohol_Use` y `Physical_Activity` cubren un abanico amplio de niveles (0–10), con un número apreciable de pacientes en niveles medios y altos de tabaquismo y consumo de alcohol, y niveles bajos/intermedios de actividad física.
4. **Relaciones entre factores y riesgo**
   - `Age` muestra poca relación directa con `Overall_Risk_Score` en este dataset sintético.
   - `BMI`, `Smoking` y `Alcohol_Use` presentan **relaciones positivas moderadas** con el score de riesgo y con la proporción de casos de alto riesgo.
   - Los boxplots de `Overall_Risk_Score` por `Risk_Level` indican una **coherencia interna clara**: el score aumenta al pasar de Low a Medium y de Medium a High.
5. **Correlaciones numéricas**
   - No se observan correlaciones excesivamente altas entre las variables numéricas analizadas, lo que es una buena señal para el posterior uso conjunto en modelos.

En conjunto, el EDA muestra que el dataset de **Cancer Risk Factors** está construido de forma **coherente y razonable** desde el punto de vista estadístico y epidemiológico: los factores de estilo de vida menos saludables se asocian con niveles de riesgo más altos, y las variables continua y categórica de riesgo (`Overall_Risk_Score` y `Risk_Level`) están alineadas.

Estos resultados proporcionan una base sólida para avanzar hacia las **fases de modelado, evaluación y segmentación de perfiles de riesgo** que se abordarán en la siguiente parte del proyecto.