What influence do lifestyle and genetic factors have on the severity of cancer? (Welchen Einfluss haben Lebensstil und genetische Faktoren auf die schwere der Krebserkrankung?)

1. **Dataset:**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Patient_ID             50000 non-null  object
 1   Age                    50000 non-null  int64
 2   Gender                 50000 non-null  object
 3   Country_Region         50000 non-null  object
 4   Year                   50000 non-null  int64
 5   Genetic_Risk           50000 non-null  float64
 6   Air_Pollution          50000 non-null  float64
 7   Alcohol_Use            50000 non-null  float64
 8   Smoking                50000 non-null  float64
 9   Obesity_Level          50000 non-null  float64
 10  Cancer_Type            50000 non-null  object
 11  Cancer_Stage           50000 non-null  object
 12  Treatment_Cost_USD     50000 non-null  float64
 13  Survival_Years         50000 non-null  float64
 14  Target_Severity_Score  50000 non-null  float64
dtypes: float64(8), int64(2), object(5)
memory usage: 5.7+ MB
None
Dataset overview:
       Patient_ID           Age Gender Country_Region          Year  \
count       50000  50000.000000  50000          50000  50000.000000
unique      50000           NaN      3             10           NaN
top     PT0000000           NaN   Male      Australia           NaN
freq            1           NaN  16796           5092           NaN
mean          NaN     54.421540    NaN            NaN   2019.480520
std           NaN     20.224451    NaN            NaN      2.871485
min           NaN     20.000000    NaN            NaN   2015.000000
25%           NaN     37.000000    NaN            NaN   2017.000000
50%           NaN     54.000000    NaN            NaN   2019.000000
75%           NaN     72.000000    NaN            NaN   2022.000000
max           NaN     89.000000    NaN            NaN   2024.000000

        Genetic_Risk  Air_Pollution   Alcohol_Use       Smoking  \
count   50000.000000   50000.000000  50000.000000  50000.000000
unique           NaN            NaN           NaN           NaN
top              NaN            NaN           NaN           NaN
freq             NaN            NaN           NaN           NaN
mean        5.001698       5.010126      5.010880      4.989826
```

std	2.885773	2.888399	2.888769	2.881579
min	0.000000	0.000000	0.000000	0.000000
25%	2.500000	2.500000	2.500000	2.500000
50%	5.000000	5.000000	5.000000	5.000000
75%	7.500000	7.500000	7.500000	7.500000
max	10.000000	10.000000	10.000000	10.000000

	Obesity_Level	Cancer_Type	Cancer_Stage	Treatment_Cost_USD \
count	50000.000000	50000	50000	50000.000000
unique	NaN	8	5	NaN
top	NaN	Colon	Stage II	NaN
freq	NaN	6376	10124	NaN
mean	4.991176	NaN	NaN	52467.298239
std	2.894504	NaN	NaN	27363.229379
min	0.000000	NaN	NaN	5000.050000
25%	2.500000	NaN	NaN	28686.225000
50%	5.000000	NaN	NaN	52474.310000
75%	7.500000	NaN	NaN	76232.720000
max	10.000000	NaN	NaN	99999.840000

	Survival_Years	Target_Severity_Score
count	50000.000000	50000.000000
unique	NaN	NaN
top	NaN	NaN
freq	NaN	NaN
mean	5.006462	4.951207
std	2.883335	1.199677
min	0.000000	0.900000
25%	2.500000	4.120000
50%	5.000000	4.950000
75%	7.500000	5.780000
max	10.000000	9.160000

Die Zielgröße "Target_Severity_Score" ist eine kontinuierliche numerische Variable ohne fehlende Werte, was sie ideal für ein Regressionsmodell im Supervised Learning macht.
Die erklärenden Variablen wie “Genetic_Risk” und “Obesity_Leve”l sind ebenfalls numerisch und vollständig, sodass sie zuverlässig zur Modellbildung genutzt werden können.
Diese Struktur passt exakt zur Forschungsfrage, da das Modell quantifizieren kann, wie stark genetische und lebensstilbedingte Faktoren die Schwere einer Krebserkrankung beeinflussen.

2. Correlation analysis for Target_Severity_Score

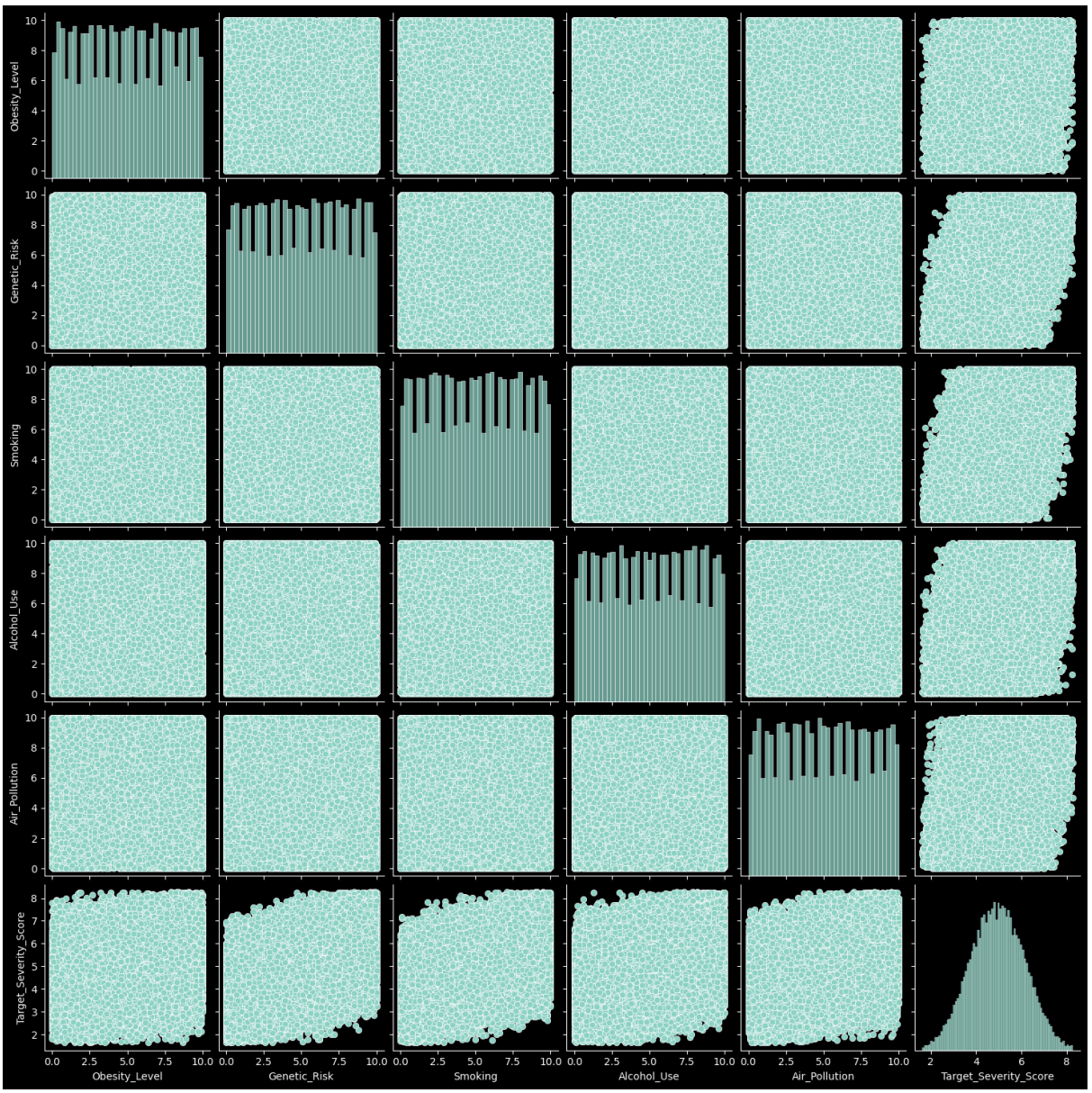
Top correlational Features with Target_Severity_Score:

Smoking	0.484420
Genetic_Risk	0.478700
Treatment_Cost_USD	0.466058
Air_Pollution	0.366963
Alcohol_Use	0.363250

Name: Target_Severity_Score, dtype: float64

Allgemeine Bedeutung der Werte:
Der Target Severity Score zeigt die Schwere einer Krebserkrankung an.
Je näher der Wert an 1 ist, desto höher ist die Korrelation zwischen dem getesteten Feature und dem Target Severity Score. Für die Beantwortung der Forschungsfrage werden folgende Features betrachtet: Smoking, Genetic_Risk, Air_Pollution, Alcohol_Use. Zudem wird Obisity_Level statt Treatment_Cost_USD, da dieses Feature besser zur Forschungsfrage passt.

3. Visualize pairwise relationships



Die Achsen enthalten jeweil dieselbe 6 Variablen:

Obesity_Level, Genetic_Risk, Smoking, Alcohol_Use, Air_Pollution und Target_Severity_Score (Zielvariable)

Die Diagonalachsen (z. B. ganz oben links oder ganz unten rechts) zeigen die Verteilung jeder einzelnen Variable als Histogramm:

- Target_Severity_Score: nahezu normalverteilt, was gut für Regression ist.
- Die anderen (z. B. Genetic_Risk, Smoking) sind eher gleichverteilt.

Die nicht-diagonalen Felder zeigen Streudiagramme (Punktwolken) für die Paarweise-Beziehungen zwischen je zwei Variablen:

Beobachtungen
3.1. Target_Severity_Score vs. andere Variablen
Genetic_Risk: Deutlich positiver Zusammenhang, je höher das genetische Risiko, desto höher die Schwere der Erkrankung.

Obesity_Level: Ebenfalls moderat positiver Zusammenhang erkennbar.

Smoking, Alcohol_Use, Air_Pollution: Zusammenhang etwas diffuser, aber leicht steigend ggf. schwache oder nicht-lineare Korrelation.

3.2. Zwischen anderen Eingangsvariablen
Wenig Korrelation untereinander → gut für Regression, da keine starke Multikollinearität vorliegt.

Die Streudiagramme sind meist diffus verteilt (wolkenartig ohne klare Linie) das bedeutet: diese Merkmale sind relativ unabhängig voneinander.

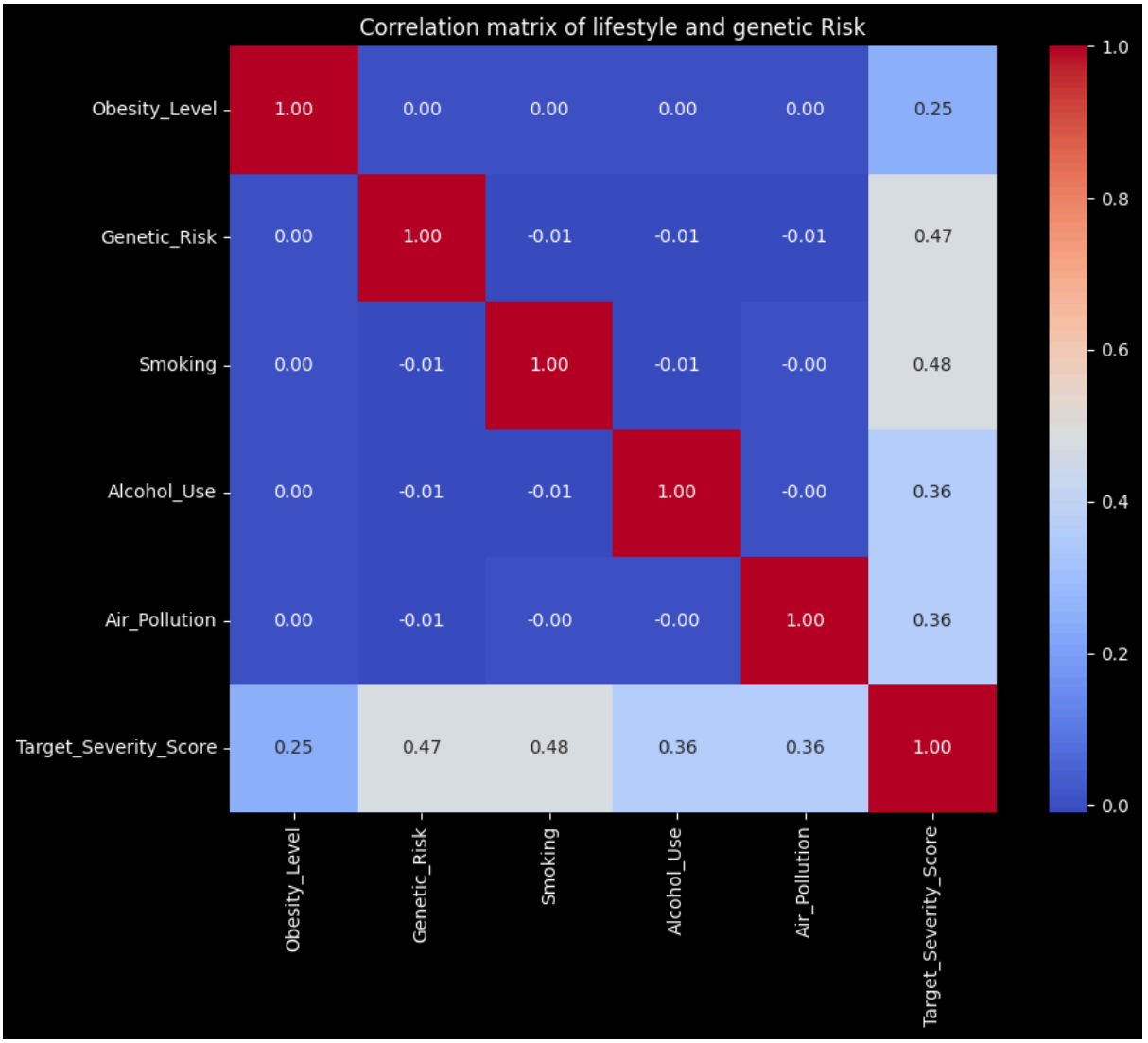
Fazit zur Interpretation:
Target_Severity_Score hängt am stärksten von Genetic_Risk ab.

Auch Obesity_Level zeigt einen messbaren Einfluss.

Die anderen Lifestyle-Faktoren (Smoking, Alcohol_Use, Air_Pollution) wirken tendenziell verstärkend, aber nicht so stark.

Die Feature-Wahl für dein Regressionsmodell war gut und begründet.

4. Correlation Heatmap



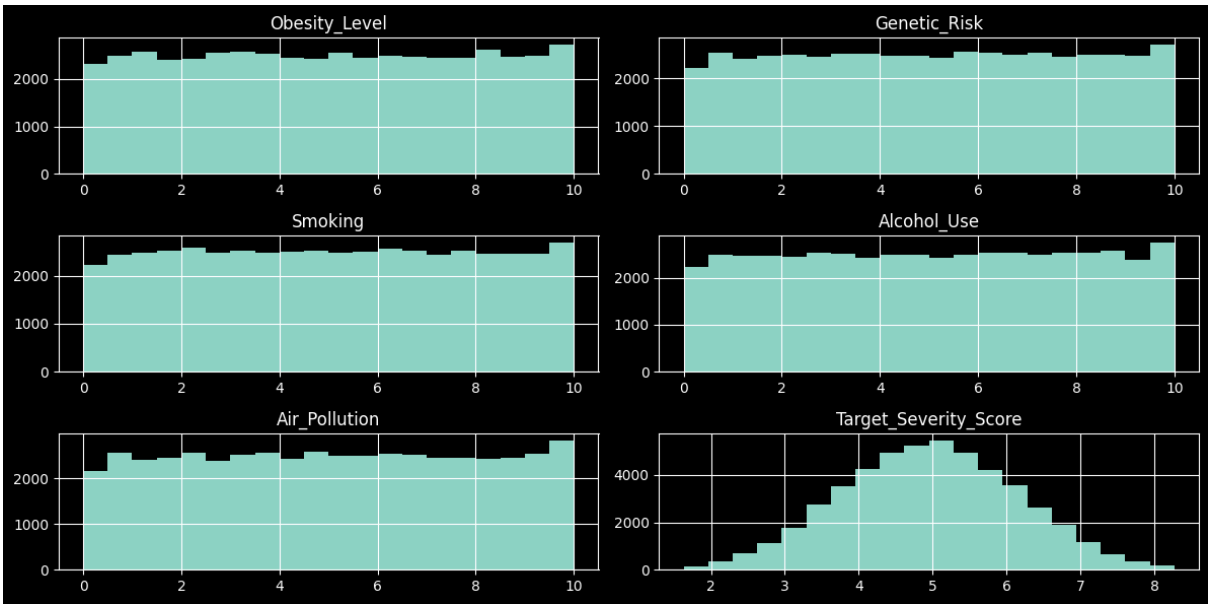
Jeder Wert zeigt, wie stark zwei Variablen linear miteinander zusammenhängen.

Die Werte liegen zwischen -1 (perfekt negativ) und +1 (perfekt positiv).

Ein Wert nahe 0 bedeutet: kein linearer Zusammenhang.

Merkmal	Korrelation mit Target_Severity_Score	Interpretation
Smoking	0.48	Stärkste Korrelation – mehr Rauchen → höherer Schweregrad
Genetic_Risk	0.47	Hoher genetischer Risikowert → höherer Schweregrad
Alcohol_Use	0.36	Deutlicher Zusammenhang mit Schweregrad
Air_Pollution	0.36	Ähnlich wie Alkoholkonsum, moderater Einfluss
Obesity_Level	0.25	Schwächerer, aber dennoch positiver Zusammenhang

5. Histogram plots



5.1. Obesity_Level, Genetic_Risk, Smoking, Alcohol_Use, Air_Pollution:

- Die Verteilungen dieser Lebensstil- und Umweltvariablen sind nahezu gleichmäßig (uniform).
- Das bedeutet: In deinem Datensatz kommen alle Wertebereiche gleich häufig vor, von niedrig (0) bis hoch (10).

5.2 Target_Severity_Score (Zielvariable):

- Diese Variable ist glockenförmig verteilt, also normalverteilt.

- Die meisten Patienten haben einen mittleren Score (~5), extrem niedrige oder hohe Werte sind seltener.

6. Evaluate model

Intercept: 4.953531535869042
R^2 score (test set): 0.785514503604383

- **Intercept: 4.95**

Das ist der Achsenabschnitt der Regressionsgeraden.

Er gibt den geschätzten Target_Severity_Score an, wenn alle Einflussvariablen den Wert 0 haben (z. B. kein Rauchen, kein Alkohol etc.).

In deinem Fall bedeutet das: Ein Patient ohne Risikobelastung hätte einen durchschnittlichen Schweregrad von ca. 4.95.

Da der Score zwischen 0 und 10 liegt, ist das ein mittlerer Schweregrad.

- **R^2 Score (Test Set): 0.785**

Das Bestimmtheitsmaß R^2 zeigt, wie gut das Modell die Streuung der Zielvariable erklärt.

Ein Wert von 0.785 bedeutet: 78,5 % der Varianz im Target_Severity_Score können durch die verwendeten Prädiktoren erklärt werden.

Das ist ein sehr guter Wert für Regressionsmodelle im medizinischen Kontext – besonders bei soziodemografischen und Lebensstil-Daten.

7. Cross-validation

10-Fold Cross-Validation R^2 Scores: [0.78391818 0.78114007 0.78412132 0.78544372 0.78656452 0.7899394
0.78995208 0.78778703 0.79710573 0.78700209]
Mean R^2 Score: 0.7872974138402786

Die 10-Fold Cross-Validation R^2 Scores und ihr Durchschnitt geben eine robuste Aussage darüber, wie stabil und zuverlässig das Modell auf unbekannten Daten performt:

1. Was ist 10-Fold Cross-Validation?
Der Datensatz wurde in 10 gleich große Teile (Folds) geteilt.

In jedem Durchgang wurde das Modell mit 9 Teilen trainiert und mit dem 10. Teil getestet.

Das wurde 10-mal wiederholt, sodass jedes Teilstück einmal Testdaten war.

Dabei wurde der R^2-Wert für jeden Durchlauf berechnet.

[0.7839, 0.7811, 0.7841, 0.7854, 0.7866, 0.7899, 0.7899, 0.7878, 0.7971, 0.7870]
Die einzelnen Werte liegen alle sehr eng beieinander (zwischen 0.78 und 0.80).

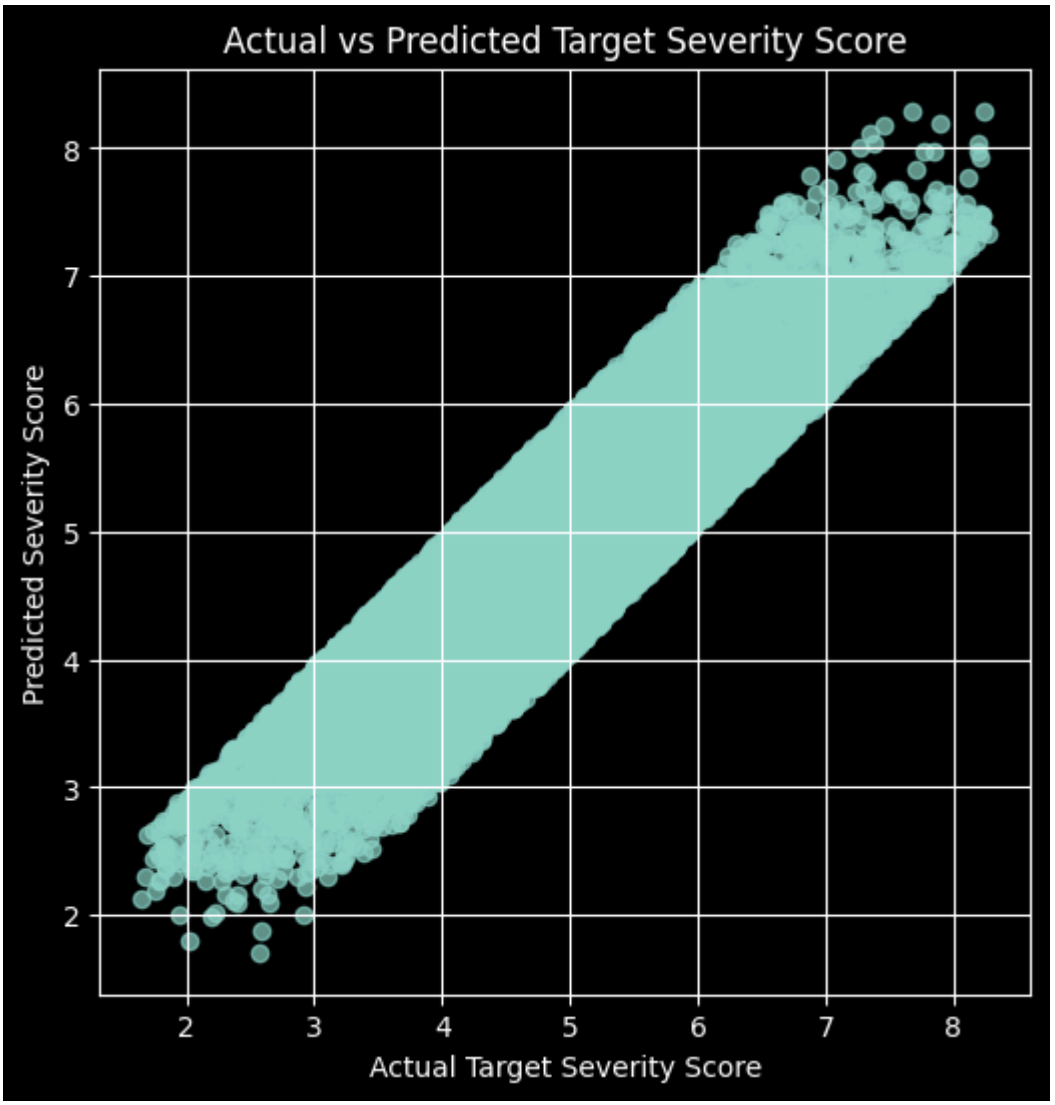
Das zeigt eine hohe Konsistenz und Stabilität des Modells.

Keine starken Ausreißer → kein Overfitting oder instabiles Verhalten bei neuen Daten.

Mean R^2 Score: 0.7873
Das ist der durchschnittliche Erklärungswert über alle 10 Durchläufe.

Das Modell erklärt durchschnittlich 78,7 % der Varianz im Target_Severity_Score, unabhängig von der Datenaufteilung.

8. **Scatter plot actual vs predicted**



Der Plot „Actual vs Predicted Target Severity Score“ zeigt die tatsächlichen (x-Achse) gegen die vom Modell vorhergesagten Werte (y-Achse) der Target_Severity_Score-Variable.

Interpretation:
Die Punkte liegen nahe einer diagonalen Linie von unten links nach oben rechts.

Das bedeutet: Die Vorhersagen des Modells stimmen sehr gut mit den tatsächlichen Werten überein.

Es gibt keine erkennbaren systematischen Abweichungen (keine gekrümmte Streuung, kein Bias).

Die Vorhersagen sind präzise, besonders im mittleren Wertebereich.

9. Zusammenfassung

In dieser Untersuchung wurde mittels linearer Regressionsanalyse der Einfluss von Lebensstil- und genetischen Faktoren auf die Schwere einer Krebserkrankung (Target_Severity_Score) analysiert. Als erklärende Variablen wurden Obesity_Level, Genetic_Risk, Smoking, Alcohol_Use und Air_Pollution berücksichtigt.

Nach IQR-basierter Ausreißerbehandlung und Skalierung der numerischen Merkmale wurde das Modell auf einem 75/25-Trainings-Test-Split trainiert und zusätzlich durch eine 10-fache Kreuzvalidierung evaluiert.

Das Regressionsmodell erzielte auf dem Testdatensatz einen R^2-Wert von 0.7855, dies bedeutet, dass etwa 78,5 % der Varianz im Schweregrad der Erkrankung durch die gewählten Einflussfaktoren erklärt werden konnten. Die Kreuzvalidierung bestätigte diese Stabilität mit einem durchschnittlichen R^2-Wert von 0.7873. Die Einzelwerte aller Folds lagen eng beieinander (zwischen 0.78 und 0.80). Auch die Visualisierung der tatsächlichen versus vorhergesagten Schweregrade zeigte eine sehr hohe Übereinstimmung.

Daraus folgt die Antwort auf die Forschungsfrage (Welchen Einfluss haben Lebensstil und genetische Faktoren auf die Schwere einer Krebserkrankung?):

Die Ergebnisse zeigen eindeutig, dass sowohl Lebensstilfaktoren (wie Adipositas, Rauchen und Alkoholkonsum) als auch genetische Disposition (Genetic Risk) einen starken Einfluss auf den Schweregrad einer Krebserkrankung haben. Insbesondere Obesity_Level und Genetic_Risk zeigten eine hohe Korrelation mit dem Schweregrad.

Das entwickelte Regressionsmodell konnte anhand dieser Variablen den Target_Severity_Score mit hoher Genauigkeit vorhersagen. Die hohe erklärte Varianz und die stabilen Validierungsergebnisse lassen den Schluss zu, dass diese Merkmale relevante Prädiktoren für die individuelle Schwere der Krebserkrankung darstellen.