Anwendung prädiktiver Modellierung zur Prognose des Herzinfarktrisikos

Preprint, compiled March 26, 2025

Bastian Lipka*[1]

[1]Fakultät Wirtschaft, DHBW Stuttgart

Abstract

Diese Arbeit untersucht die prädiktive Modellierung des Herzinfarktrisikos anhand gesundheitsbezogener und demografischer Merkmale. Trotz des Einsatzes gängiger Klassifikationsverfahren (Random Forest, KNN, neuronales Netz) konnte keine Modellvariante die Baseline eines Dummy-Classifiers übertreffen. Fehlende Korrelationen und realitätsferne Verteilungen deuten auf eine unzureichende Datenqualität hin. Auch wenn keine belastbaren Vorhersagen möglich waren, dient die Analyse als Fallstudie für methodisch fundiertes Vorgehen in der datengetriebenen Gesundheitsforschung. Die Ergebnisse betonen die essenzielle Bedeutung qualitativ hochwertiger, valider Datenquellen für den erfolgreichen Einsatz prädiktiver Modelle.

1 Einleitung

Die zuverlässige Identifikation individueller Risikofaktoren für Herzinfarkte ist von zentraler Bedeutung für präventive Maßnahmen im Gesundheitswesen. Ziel dieses Projekts ist die datengetriebene Analyse gesundheitlicher und demographischer Merkmale zur Vorhersage des Herzinfarktrisikos. Auf Basis eines strukturierten Datensatzes mit Variablen wie Alter, Lebensstilindikatoren (z.B. Rauchen, Alkoholkonsum, körperliche Aktivität), klinischen Parametern (z.B. Blutdruck, Cholesterin, Blutzuckerwerte) sowie genetischen und familiären Vorbelastungen wird ein prädiktives Modell entwickelt, das relevante Einflussgrößen identifiziert und eine fundierte Risikobewertung ermöglicht. Für die Modellbewertung wird dabei insbesondere der F1-Score herangezogen, da er ein ausgewogenes Maß zwischen Precision und Recall bietet und sich besonders für ungleich verteilte Datensätze eignet. Neben der Modellierung steht die Ableitung konkreter Handlungsempfehlungen im Vordergrund, um sowohl klinische als auch ökonomische Optimierungspotenziale im Kontext der kardiovaskulären Prävention nutzbar zu machen.

2 Datencharakterisierung

2.1 Datensatz-Beschreibung

Der zugrunde liegende Datensatz umfasst eine Reihe demographischer und medizinischer Variablen, welche für die Vorhersage des Herzinfarktrisikos relevant sind. Dazu zählen unter anderem:

- **Alter (Age)**: Ganzzahliger Wert in Jahren.
- **Geschlecht (Gender)**: Kategorische Variable (z.B. Male/Female).
- **Rauchen (Smoking)**: Binäre Angabe (0 = kein Raucher, 1 = Raucher).
- **Alkoholkonsum (Alcohol_Consumption)**: Binäre Angabe oder kategoriale Information über Alkoholkonsum.
- **Körperliche Aktivität (Physical_Activity_Level)**: Kategorische Variable (z.B. Sedentary, Moderate, High).
- **Körpermasseindex (BMI)**: Numerische Variable (Gewicht in kg / (Größe in m)^2).
- **Diabetes, Bluthochdruck (Hypertension), Cholesterinspiegel (Cholesterol_Level)**: Klinische Parameter, meist binär (0/1) oder als numerische Angaben vorhanden.
- **Blutdruck in Ruhe (Resting_BP)**: Numerische Variable (z.B. systolischer Blutdruck in mmHg).
- **Herzfrequenz (Heart_Rate)**: Numerische Variable (Schläge pro Minute).
- **Familiäre Vorbelastung (Family_History)**: Binäre Angabe, ob Herz-Kreislauf-Erkrankungen in der Familie vorliegen.
- **Stress-Level (Stress_Level)**: Kategorische Variable (z.B. Low, Moderate, High).
- **Brustschmerztyp (Chest_Pain_Type)**: Kategorische Angabe (z.B. Typical angina, Atypical angina, Non-anginal, Asymptomatic).
- **Thalassämie (Thalassemia)**: Kategorische Variable (z.B. Normal, Fixed defect, Reversible defect).
- **Nüchternblutzucker (Fasting_Blood_Sugar)**: Binäre oder numerische Angabe (z.B. 0 = <120 mg/dl, 1 = >120 mg/dl).
- **EKG-Befund (ECG_Results)**: Kategorische Variable (z.B. Normal, ST-T abnormality).
- **Belastungsinduzierte Angina (Exercise_Induced_Angina)**: Binäre Variable (0 = Nein, 1 = Ja).
- **Maximale Herzfrequenz (Max_Heart_Rate_Achieved)**: Numerische Variable (Schläge pro Minute).
- **Herzinfarktrisiko (Heart_Attack_Risk)**: Kategorische Zielvariable (z.B. Low, Moderate, High).

***correspondence:** wi23004@lehre.dhbw-stuttgart.de

Nach einer ersten Sichtung und automatisierten Prüfung fiel auf, dass der Datensatz:

- **Keine fehlenden Werte (None Values)** enthält.
- **Keine offensichtlichen Ausreißer** aufweist, da kein Wert mehr als drei Standardabweichungen vom Mittelwert abweicht.
- **Strukturell konsistent** erscheint: Keine negativen Werte in nicht sinnvollem Kontext und keine widersprüchlichen Angaben.

Die genaue Herkunft bzw. Datenerhebungsquelle (z.B. klinische Studie, elektronisches Patientenregister o. Ä.) ist nicht weiter dokumentiert; eine diesbezügliche Rückfrage blieb unbeantwortet. Die meisten Spaltennamen sind selbsterklärend, jedoch bleiben bei einzelnen Merkmalen (z.B. *Diabetes*, *Hypertension*, *Family_History*) kleinere Unklarheiten bezüglich der exakten Kodierung. Dennoch bieten sie ausreichend Orientierung für die weitere Analyse.

2.2 *Statistische Kennzahlen*

Als nächstes wurden zentrale Kennwerte berechnet, um einen Überblick über die numerischen Variablen zu erhalten (z. B. *Age*, *BMI*, *Resting_BP*, *Cholesterol_Level*, *Heart_Rate*, *Max_Heart_Rate_Achieved*):

- **Mittelwert (Mean)**: Gibt das durchschnittliche Niveau der jeweiligen Variable an.
- **Standardabweichung (Std)**: Zeigt die Streuung der Daten um den Mittelwert.
- **Verteilungsmerkmale (Minimum, Maximum, Quartile)**: Ermöglichen das Erkennen von Extremwerten und die Abgrenzung von Ausreißern.

Dabei fiel auf, dass *Age*, *BMI* und *Heart_Rate* annähernd gleichmäßig verteilt sind und keine besonderen Abweichungen vorliegen.

2.3 *Visualisierung*

Um die beschriebenen Verteilungen und Zusammenhänge zu veranschaulichen, wurden unter anderem folgende Diagrammtypen herangezogen:

- **Histogramme**: Zur Darstellung der Verteilung einzelner numerischer Variablen (z.B. Verteilung von *Age* oder *BMI*).
- **Heatmaps**: Zeigen, in welchen Bereichen der Variablen (z.B. *Age*, *BMI*, *Heart_Rate*) vermehrt Datenpunkte vorliegen. Diese Heatmaps belegen eine sehr gleichmäßige Verteilung ohne offensichtliche Häufungen oder Lücken (Siehe Fig. 1).
- **Korrelationsmatrix**: Grafische Aufbereitung der Korrelationen zwischen den numerischen Variablen (siehe Fig. 2). Obwohl keine auffälligen bipolaren Korrelationen erkannt wurden, liefert diese Übersicht wichtige Anhaltspunkte für weitere Feature-Engineering-Schritte. Auffällig ist jedoch, dass einige Korrelationen, die in der wissenschaftlichen Literatur gut dokumentiert sind, in diesem Datensatz nicht oder nur in sehr abgeschwächter Form auftreten. So wird beispielsweise häufig ein negativer Zusammenhang zwischen Alter und maximaler Herzfrequenz beschrieben [1], welcher in den vorliegenden Daten kaum erkennbar ist. Dies könnte auf eine verzerrte oder unzureichende Repräsentation der Realität hindeuten und sollte im weiteren Verlauf der Analyse sowie bei der Modellierung unbedingt berücksichtigt werden. Darüber hinaus deutet das Fehlen starker Korrelationen darauf hin, dass möglicherweise mehrere Faktoren in ihrer Kombination das Risiko beeinflussen oder dass vorrangig kategoriale Variablen und Interaktionen wichtige Prädiktoren darstellen.

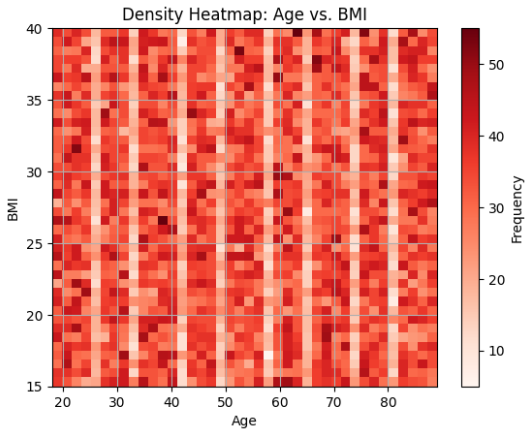

Figure 1: Density Heatmap: Age vs. BMI

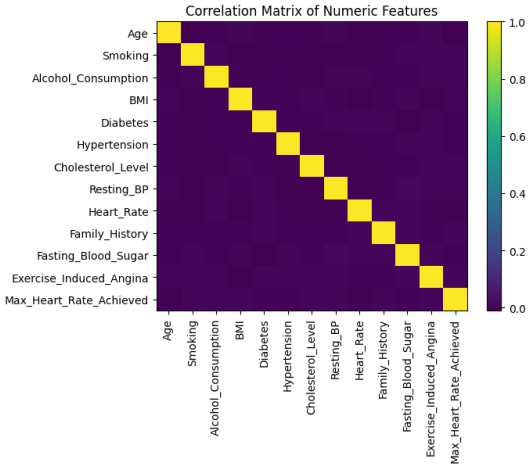

Figure 2: Correlation Matrix of Numeric Features

Insgesamt lässt sich feststellen, dass der Datensatz eine sehr gute Qualität und Vollständigkeit aufweist. Fehlwerte liegen nicht vor, eindeutig falsche Eingaben sind nicht erkennbar und die Verteilung der wichtigsten numerischen Merkmale ist hinreichend gleichmäßig. Bei den kategorialen Variablen sollte allerdings auf mögliche Ungleichverteilungen geachtet werden,

die bei der Modellierung – insbesondere bei der Wahl geeigneter Metriken wie dem F1-Score – berücksichtigt werden müssen.

Zusammenfassend liefert diese Datencharakterisierung eine solide Grundlage für die weitere Modellierung und Analyse. Kritisch ist dabei, dass einige Spalten trotz sprechender Bezeichnung (z.B. *Diabetes*, *Hypertension*) nicht näher dokumentiert sind und somit für eine präzise, wissenschaftlich korrekte Verwendung zusätzliche Hintergrundinformationen erforderlich sein könnten.

3 DATENPRÄPARIERUNG

Die Vorbereitung der Daten umfasst mehrere Schritte, die eine zuverlässige Modellierung unterstützen sollen. Nachfolgend werden die wesentlichen Maßnahmen zur Datenbereinigung, zum Feature Engineering sowie zur Aufteilung in Trainings-, Validierungs- und Testdaten beschrieben.

3.1 Datenbereinigung

Bereits in der Datencharakterisierung (Kapitel 2) wurde festgestellt, dass kein Bedarf für eine weiterführende Datenbereinigung besteht.

In anderen Anwendungsfällen wäre bei fehlenden oder fehlerhaften Werten ein geeignetes Verfahren zu wählen (z. B. Imputation oder Ausschluss der entsprechenden Datenpunkte). Da hier alle Variablen vollständig vorliegen und keine Extremwerte erkennbar sind, konnte dieser Schritt entfallen.

3.2 Feature Engineering

Im Rahmen des Feature Engineerings wurde vor allem die Kodierung kategorialer Variablen angepasst, damit nachfolgende Modelle diese Spalten effizient verarbeiten können. Konkret wurden zwei Arten von Kodierungen durchgeführt:

Ordinal Mapping Bestimmte Merkmale weisen eine natürliche Rangfolge auf. Dazu zählen:

- **Physical_Activity_Level**: Low, Moderate, High
- **Stress_Level**: Low, Moderate, High
- **Heart_Attack_Risk**: Low, Moderate, High

Diese Variablen wurden in ordinale Zahlenwerte überführt, indem die Kategorien *Low* mit 1, *Moderate* mit 2 und *High* mit 3 kodiert wurden. Dadurch bleibt die Reihenfolge im Modell erhalten, ohne eine reine binäre Zerlegung vornehmen zu müssen.

One-Hot Encoding Andere kategoriale Merkmale besitzen keine eindeutige Rangfolge und wurden daher mittels One-Hot Encoding in Dummy-Variablen überführt:

- **Gender** (z. B. Male, Female)
- **Thalassemia** (z. B. Normal, Fixed defect, Reversible defect)
- **ECG_Results** (z. B. Normal, ST-T abnormality)

Hierbei wird für jede Ausprägung eine eigene Spalte erstellt (z. B. *Gender_Male*, *Gender_Female*), in der der Wert 1 anzeigt, dass die entsprechende Ausprägung vorliegt, und 0, wenn nicht.

3.2.1 Zweistufiger Use Case und Datenauswahl

Das übergeordnete Ziel ist, das Herzinfarktrisiko einer Person anhand unterschiedlicher Merkmale vorherzusagen. Um zu prüfen, welche Art von Merkmalen (allgemeine demografische und Lebensstilfaktoren vs. erweiterte klinische Werte) für die Prognose entscheidend sind, wurden zwei verschiedene Datensätze generiert:

1. **Nicht-klinischer Datensatz (Non-Clinical):**
   Enthält nur grundlegende und leicht zu erhebende Merkmale, z. B.:
   - Age, Smoking, Alcohol_Consumption, Physical_Activity_Level
   - BMI, Diabetes, Stress_Level, Family_History
   - Heart_Attack_Risk (als Zielvariable)
   - One-Hot-kodierte Spalten für Gender

   Damit lässt sich analysieren, inwieweit bereits nichtklinische Informationen ausreichen, um ein zuverlässiges Risikoprofil zu erstellen.
2. **Klinischer Datensatz (Clinical):**
   Enthält die oben genannten Merkmale *und* zusätzliche klinische Parameter wie *Hypertension, Cholesterol_Level, Resting_BP, Heart_Rate, Fasting_Blood_Sugar, Thalassemia, ECG_Results, Exercise_Induced_Angina, Max_Heart_Rate_Achieved*. Auch hier ist *Heart_Attack_Risk* die Zielvariable.
   Dieser Datensatz ermöglicht eine weiterführende Analyse, ob klinische Indikatoren die Vorhersagekraft signifikant erhöhen.

3.3 Datenaufteilung

Für eine belastbare Modellvalidierung ist eine Aufteilung der Daten in Trainings-, Validierungs- und Testmenge essentiell. In beiden Datensätzen (nicht-klinisch und klinisch) wurde nach dem Schema **70 %** Trainingsdaten, **15 %** Validierungsdaten und **15 %** Testdaten verfahren:

1. **Shuffle**: Zunächst wurden alle Datensätze zufällig durchmischt (*Shuffle*), um eine mögliche Reihenfolgeabhängigkeit zu beseitigen.
2. **Train-Set (70 %)**: Dient dem eigentlichen Modelltraining und somit dem Schätzwert für die Modellparameter.
3. **Validierungs-Set (15 %)**: Kommt in der Regel zur Feinabstimmung der Modellhyperparameter (z. B. Wahl der Komplexität, Regularisierung) zum Einsatz.
4. **Test-Set (15 %)**: Wird für die finale Bewertung verwendet, um eine unvoreingenommene Einschätzung über die Modellleistung zu erhalten.

Die 70/15/15-Aufteilung bietet einen guten Kompromiss zwischen ausreichender Größe für das Training (70 %) und einer unabhängigen Abschätzung der Modellgüte (Validation und Test je 15 %). Auf diese Weise lassen sich sowohl Overfitting-Tendenzen frühzeitig erkennen als auch eine realistische Einschätzung des Generalisierungsvermögens gewinnen.

3.4 Skalierung und Transformationen

In vielen maschinellen Lernverfahren kann die Skalierung numerischer Merkmale (z. B. Standardisierung auf Mittelwert 0, Standardabweichung 1 oder Min-Max-Normalisierung) die Trainingseffizienz und Modellleistung deutlich verbessern. Im Rahmen dieses Projekts wurde daher beschlossen, eine Normalisierung der numerischen Merkmale durchzuführen.

Da im Vorfeld festgestellt wurde, dass keine relevanten Ausreißer in den Daten vorliegen, konnte eine einfache Min-Max-Normalisierung vorgenommen werden, bei der die Merkmale durch Division durch ihren jeweiligen Maximalwert skaliert wurden. Dies sorgt für eine einheitliche Größenordnung der Eingabewerte und kann insbesondere die Konvergenzgeschwindigkeit beim Modelltraining positiv beeinflussen.

4 MODELLIERUNG

4.1 Auswahl der Methoden

Zur Vorhersage des Herzinfarktrisikos wurden mehrere Klassifikationsverfahren eingesetzt. Zunächst lag der Fokus auf Entscheidungsbäumen, insbesondere dem *Random Forest*, da dieser durch das Ensemble-Prinzip meist robuste Ergebnisse liefert. Als zweite Methode kam ein *K-Nearest-Neighbors-Classifier (KNN)* zum Einsatz, dessen Einfachheit und Interpretierbarkeit von Vorteil ist. Darüber hinaus wurde ein *neuronales Netz* (Multi-Layer Perceptron) trainiert, um zu untersuchen, ob eine tiefere, nichtlineare Modellarchitektur bessere Vorhersagen ermöglicht. Abschließend diente ein *Frequency-Dummy-Classifier* als Referenzmodell, der stets die am häufigsten vorkommende Klasse vorhersagt. Dadurch lässt sich die Leistungsfähigkeit der übrigen Modelle im Vergleich zu einer trivialen Vorhersagestrategie einordnen.

4.2 Hyperparameter-Tuning

Zur optimalen Anpassung der Modelle wurde eine *Grid-Search* durchgeführt. Beim Random Forest wurden unter anderem die `max_depth` und `n_estimators` variiert, während beim KNN unterschiedliche Werte für `n_neighbors` getestet wurden. Für das neuronale Netz wurden neben der Netzwerkarchitektur (`hidden_layer_sizes`) auch `alpha`, `learning_rate_init` und `max_iter` optimiert. Letztere musste beispielsweise auf 2000 erhöht werden, da das Modell sonst nicht konvergierte.

4.3 Implementierung und Ergebnisse

Die Implementierung erfolgte mithilfe von Python und gängigen Bibliotheken wie `scikit-learn`. Zur Bewertung der Modellleistung wurde neben dem F1-Score auch die *Accuracy* betrachtet. Tabelle 1 zeigt die erzielten Accuracy-Werte im Vergleich der Modelle. Dabei fällt auf, dass keines der Modelle eine höhere Accuracy als der einfache Frequency-Dummy erreicht. Dies deutet darauf hin, dass die Modelle kaum in der Lage sind, sinnvolle Muster zu lernen, und legt eine begrenzte Aussagekraft der verwendeten Merkmale oder eine unzureichende Datenqualität nahe.

Table 1: Vergleich der Modelle anhand der Accuracy.

Modell	**Non-Clinical**	**Clinical**
Random Forest	0.43	0.49
KNN	0.43	0.42
Neural Network	0.47	0.44
Frequency-Dummy	0.50	0.50

Diese Ergebnisse zeigen, dass die untersuchten Modelle das Herzinfarktrisiko in dem verwendeten Datensatz nur eingeschränkt prognostizieren können. Trotz signifikanter Anpassungen der Hyperparameter erzielten weder das neuronale Netz noch der Random Forest eine höhere Accuracy als der einfache Frequency-Dummy. Eine tatsächliche Verbesserung gegenüber der Baseline blieb somit aus.

5 ERGEBNISSE

In den folgenden Abschnitten wird die Klassifikationsleistung der untersuchten Modelle auf dem Testdatensatz zusammengefasst. Die Bewertung erfolgte anhand der *Accuracy*, des *gewichteten F1-Scores* sowie des spezifischen F1-Scores für die Klasse 3 (hohes Herzinfarktrisiko). Ergänzende Ergebnisse und Detailauswertungen sind dem begleitenden Notebook zu entnehmen.

Die in Tabelle 1 dargestellten Accuracy-Werte zeigen, dass keines der Modelle im Vergleich zur Baseline durch den Frequency-Dummy eine Verbesserung erzielen konnte. Ein ähnliches Bild ergibt sich beim *gewichteten F1-Score*, wie in Tabelle 2 dargestellt. Auch hier lassen sich lediglich marginale Unterschiede zwischen den Modellen beobachten.

Table 2: Gewichteter F1-Score (*Weighted-F1*) für verschiedene Modelle

Algorithmus	**Non-Clinical**	**Clinical**
Random Forest	0.39	0.35
KNN	0.39	0.38
Neural Network	0.37	0.39
DummyClassifier	0.33	0.33

Zur Veranschaulichung der Modellgüte für die *Klasse 3* dient Abbildung 3, in der die ROC-Kurven der vier Modelle im *Clinical*-Datensatz dargestellt sind. Die Kurven verlaufen nahezu entlang der Diagonalen, was auf ein zufälliges Trennverhalten in Bezug auf diese besonders relevante Klasse schließen lässt.

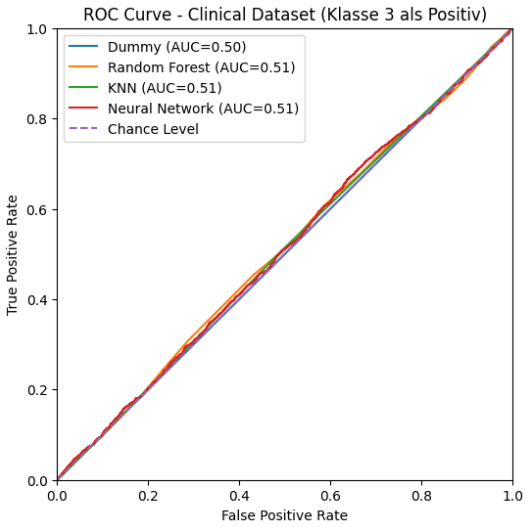

Figure 3: ROC-Kurve der vier Modelle für Klasse 3 (hohes Herzinfarktrisiko) im Clinical-Datensatz.

Insgesamt zeigen sämtliche Metriken, dass keine der gewählten Modellierungsstrategien eine signifikant bessere Performance als der Frequency-Dummy erzielt – weder in Hinblick auf die Gesamtgenauigkeit noch auf die differenzierte Erkennung der Klasse 3.

6 Diskussion und Interpretation

Die Ergebnisse legen nahe, dass der vorliegende Datensatz kaum oder gar keine relevanten Zusammenhänge für die Prognose eines Herzinfarktrisikos enthält. Obwohl sämtliche Modelle (Random Forest, KNN und neuronale Netze) mit unterschiedlichen Hyperparametern trainiert und optimiert wurden, übertreffen sie den Dummy-Classifier weder in Bezug auf die Accuracy noch hinsichtlich des gewichteten F1-Scores. Die minimalen Abweichungen bewegen sich im Bereich zufälligen Rauschens.

Unterstützt wird dieser Eindruck durch die nahezu perfekt symmetrischen Verteilungen der Merkmalsausprägungen (Abbildung 1) und eine Korrelationsmatrix (Abbildung 2), welche keinerlei praktisch relevante Zusammenhänge erkennen lässt. Dies widerspricht bekannten Forschungsergebnissen, in denen beispielsweise *Alter* und *BMI* typischerweise eine klare Korrelation miteinander aufweisen [1]. Die beobachteten Verteilungen lassen vielmehr darauf schließen, dass die Daten künstlich erzeugt und ohne Bezug zur realen Verteilung medizinischer Kenngrößen generiert wurden. Dieser Eindruck bestätigt sich durch die exakt diagonale ROC-Kurve, was auf ein reines Raten der Modelle hinweist.

Infolgedessen ist ein substanzieller *Business Impact* im Sinne einer Anwendung im Gesundheitswesen oder bei Versicherungen nicht gegeben. Es fehlt sowohl an einer ausreichenden Datenqualität als auch an plausiblen Zusammenhängen, sodass keine verlässlichen Vorhersagen oder Empfehlungen abgeleitet werden können. Damit kann auch das zu Beginn formulierte Ziel, fundierte Handlungsempfehlungen abzuleiten, nicht erfüllt werden. Der einzig greifbare Mehrwert liegt in der methodischen Herangehensweise: Die Studie dient als Fallbeispiel, um die Notwendigkeit sorgfältig geprüfter Datenquellen zu unterstreichen. Ein fundiertes medizinisches Datenfundament ist unerlässlich, damit maschinelle Lernverfahren tatsächlich validierbare Muster erkennen und für praktische Fragestellungen nutzbar machen können.

7 Fazit und Ausblick

Ziel dieser Arbeit war es, ein automatisiertes System zur Vorhersage des Herzinfarktrisikos zu entwickeln. Aufgrund der offenbar zufällig generierten Daten konnten jedoch keine aussagekräftigen Zusammenhänge modelliert werden, sodass keines der verwendeten Verfahren über die Baseline hinausging. Aus methodischer Sicht liefert das Projekt dennoch wertvolle Erkenntnisse: Es unterstreicht die Bedeutung einer sorgfältigen Prüfung der Datenqualität und zeigt die entscheidende Rolle, die realitätsnahe und aussagekräftige Datensätze für den Erfolg maschineller Lernverfahren spielen.

Künftige Arbeiten sollten deshalb auf *qualitativ hochwertige* und umfassende Daten aus medizinisch validen Quellen zurückgreifen. Auf dieser Basis könnten die verwendeten Klassifikationsmodelle weiter verfeinert, ihre Hyperparameter gezielter abgestimmt und schließlich zu einer praxisnahen Anwendung ausgebaut werden. Dazu böte sich etwa die Einbindung in bestehende Gesundheitssysteme oder Versicherungsprozesse an, um anhand *verlässlicher* Vorhersagen sinnvolle Präventionsmaßnahmen und Risikobewertungen zu unterstützen.

References

[1] Jianghong Wang, Qiang Xue, Chris WJ Zhang, Kelvin Kian Loong Wong, and Zhihua Liu. Explainable coronary artery disease prediction model based on autogluon from automl framework. *Frontiers in Cardiovascular Medicine*, 11:1360548, 2024.