Technology
Arts Sciences
TH Köln

Technische Hochschule Köln
Fakultät für Informatik und Ingenieurwissenschaften

PRAXISPROJEKT

Vorhersage der Herz-Kreislauf-Erkrankungen

anhand von Lebensstilfaktoren

Vorgelegt an der TH Köln
Campus Gummersbach
im Studiengang
Medieninformatik

ausgearbeitet von:
Abdelbasset Moujtahid
(Matrikelnummer: 11140900)

Prüfer: Prof. Dr. Daniel Gaida

Gummersbach, im Februar 2024

Zusammenfassung

Das Hauptziel dieses Projektes ist die Entwicklung und der Vergleich von Vorhersagemodellen, die das Herz-Kreislauf-Erkrankungsrisiko (HKE-Risiko) einer Person auf der Grundlage ihrer Lebensgewohnheiten genau einschätzen. Dazu müssen die wichtigsten Lebensmittelfaktoren ermittelt werden, die wesentlich zum HKE-Risiko beitragen, und ein zuverlässiges Modell für die Risikobewertung entwickelt werden. Das Verständnis des Zusammenhangs zwischen Lebensstil und chronischen Herzerkrankungen ist entscheidend für eine personalisierte Gesundheitsversorgung. Es soll ein solider Rahmen zur frühzeitigen Risikoerkennung geschaffen und gezielte Maßnahmen zur Verringerung der Belastung durch Herz-Kreislauf-Erkrankungen (HKE) ermöglicht werden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Gender Hinweis

Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich, divers (m/w/d) verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter.

1 Einleitung

Herzkrankheiten und andere Formen von Krankheiten spielen seit vielen Jahren eine verheerende Rolle im Leben der Menschen. Obwohl die Sterblichkeitsrate in Deutschland in den letzten Jahren seit 1970 gesunken war, erreichte sie im Jahr 2022 wieder ihren Höchststand, hauptsächlich aufgrund von Herzerkrankungen [1]. Im Jahr 2022 sind in Deutschland 1,06 Millionen Menschen gestorben. Wie das Statistische Bundesamt mitteilt waren es 4,27% mehr als im Vorjahr. Die häufigste Todesursache war wie in den Vorjahren eine HKE. Mit gut ein Drittel der Verstorbenen (33,6%), gefolgt von Krebs (21,7%). Herz-Kreislauf-Erkrankungen entwickeln sich unauffällig, können aber durch die Identifizierung von "Risikopersonen"verhindert oder frühzeitig erkannt werden. Die Vorhersage von Herzerkrankungen ist ein heikler, riskanter und wichtiger Faktor. Wenn die Vorhersage richtig ausgeführt wird, kann sie von medizinischen Diensten genutzt werden, um Leben zu retten. Die Diagnose und Behandlung von Herzerkrankungen ist vor allem in Entwicklungsländern sehr komplex, da es an Diagnosegeräten, Ärzten und anderen Ressourcen mangelt, die sich auf die angemessene Vorhersage und Behandlung von Herzpatienten auswirken. In jüngster Zeit werden Computertechnologien und statistische Lernverfahren eingesetzt, um Programme zu entwickeln, die Klinikern helfen sollen, Entscheidungen über Herzkrankheiten im Anfangsstadium zu treffen. Die Früherkennung von Krankheiten und die Vorhersage der Wahrscheinlichkeit, dass eine Person eine Herzkrankheit entwickelt, kann die Sterblichkeit senken.

Mehrere in diesem Projekt erwähnte Forschungsarbeiten konzentrierten sich auf die Entwicklung maschineller Lernmodelle zur Vorhersage von Herzerkrankungen auf der Grundlage der individuellen Risikofaktoren der Patienten.

Die Autoren verwendeten verschiedene Algorithmen für maschinelle Lernmodelle wie lineare Regression, logistische Regression, Entscheidungsbäume, Random-Forest-Klassifizierer usw. auf den vom BRFSS gelieferten Datensatz. Das Behavioral Risk Factor Surveillance System (BRFSS) ist das erste System der Nation für gesundheitsbezogene Telefonumfragen, das staatliche Daten über die Einwohner der USA in Bezug auf ihr gesundheitsbezogenes Risikoverhalten sammelt. Dennoch wurde festgestellt, dass sich die Ergebnisse der verschiedenen Studien stark voneinander unterschieden und dass es keinen Vergleich

zwischen den verschiedenen Algorithmen und ihren Berechnungsgenauigkeiten gab.

Um diese Diskrepanzen zwischen den Quellstudien dieses Projekts zu untersuchen, wurden mehrere Lösungsmethoden bei der Vorhersage von Risiken auf der Grundlage individueller Gewohnheiten von Personen adressiert. Dazu gehören die lineare Regression, die logistische Regression, der Decision Tree Classifier, der Random Forest Classifier und schließlich der K-NN Classifier. Diese Algorithmen werden mithilfe des BRFSS Datensatzes trainiert und evaluiert, der Faktoren enthält, die mit der Gesundheit einer Person und somit mit dem Risiko, an einer Herz-Kreislauf-Erkrankung zu erkranken, in Verbindung stehen. Der gesamte Fokus lag auf dem Vergleich der Genauigkeitswerte der Modelle, um das beste Modell für die Entwicklung eines Vorhersageinstruments während der Bachelorarbeit zu verwenden. Um die Genauigkeitswerte zu verbessern, wurden die Hyperparameter angepasst und die wichtigsten Faktoren bei der Entwicklung der Lösungsmethode identifiziert.

Das Ergebnis dieses Projekts ist in erster Linie die Schaffung einer Vorarbeit für das finale Projekt der Bachelorarbeit, indem ein Vorhersageinstrument entwickelt wird der zur Vorhersage des HKE-Risikos auf der Grundlage der persönlichen Gewohnheiten einer Person dient. Zweitens zeigt es die Notwendigkeit des “hyperparemeter tuning” auf, indem es die Genauigkeit des Modells verbessert und die Bedeutung der “optimalen” Wahl der Parameter beim Training eines Algorithmus hervorhebt. Und schließlich die Identifizierung von gemeinsamen Faktoren im Leben einer Person, die das Risiko für Herz-Kreislauf-Erkrankungen erhöhen.

2 Literatur-Übersicht

Herzkrankheiten sind eine der Hauptursachen für Invalidität und vorzeitigen Tod von Menschen auf der ganzen Welt [2]. Sie kosten jedes Jahr schätzungsweise 17,9 Millionen Menschen das Leben. HKE sind eine Gruppe von Erkrankungen des Herzens und der Blutgefäße und umfassen koronare Herzkrankheiten, zerebrovaskuläre Erkrankungen, rheumatische Herzkrankheiten und andere Erkrankungen. Mehr als vier von fünf Todesfällen durch HKE sind auf Herzinfarkte und Schlaganfälle zurückzuführen, und ein Drittel dieser Todesfälle tritt vorzeitig bei Menschen unter 70 Jahren auf. Diese Krankheiten können durch die Bekämpfung der Risikofaktoren verhindert werden. Die wichtigsten verhaltensbedingten Risikofaktoren für Herzerkrankungen und Schlaganfall sind ungesunde Ernährung, Bewegungsmangel, Tabakkonsum und schädlicher Alkoholkonsum. Die Identifizierung derjenigen, die das höchste Risiko für Herz-Kreislauf-Erkrankungen haben, und die Gewährleistung einer angemessenen Behandlung können vorzeitige Todesfälle verhindern.

Eine genaue Vorhersage des HKE-Risikos auf der Grundlage persönlicher Lebensstilfaktoren ist entscheidend. In den letzten Jahren haben sich Algorithmen des maschinellen Lernens, für die Verbesserung der Risikovorhersage, als leistungsfähige Werkzeuge erwiesen. Es gibt jedoch nach wie voe eine Forschungslücke in Bezug auf die Auswahl geeigneter maschineller Lernmodelle (ML-Modelle). Die Literatur-Übersicht soll einen Überblick über bestehende Studien auf diesem Gebiet geben.

Nusovicini et al. haben ML-Algorithmen und logistische Regression zur Vorhersage des Risikos von Herz-Kreislauf-Erkrankungen (HKE), chronische Nierenerkrankung (CNE), chronisch obstruktiver Lungenerkrankung (COL) und arterielle Hypertonie (HTN) verglichen. Die logistische Regression schnitt bei CNE und COL gut ab, während das neuronale Netz und die Support Vector Machine bei HKE und HTN am besten abschnitten. In dieser Studie wurde die Bedeutung von Risikovorhersagemodellen und das Potenzial von ML für eine verbesserte Genauigkeit verdeutlicht [3].

Goldstein et al. zeigten, dass der Einsatz von ML-Methoden bei der Entwicklung von Risikovorhersagemodellen für die herkömlichen Regressionsmodelle zwar nützlich waren, jedoch Einschränkungen erwiesen, da sie nur eine kleine Anzahl von Vorhersagefaktoren mit gleichmäßigen Auswirkungen berücksichtigen konnten. Die Ansätze des maschinellen Lernens gingen auf Herausforderungen ein, die von Regressionsmethoden nicht angemessen bewältigt werden konnten. In der Studie wurden auch allgemeine Überlegungen bei der Anwendung zur Anwendung des maschinellen Lernens erörtert, wie z. B. die Abstimmung von Parametern, Verlustfunktionen, die Bedeutung von Variablen und der Umgang mit fehlenden Daten. Insgesamt diente sie als einführende Ressource für Forscher im Bereich der Risikomodellierung, um das Gebiet des maschinellen Lernens zu erkunden [4].

Reddy et al. ließen Merkmale evaluieren um anhand signifikanter Merkmale die Leistung von ML-Klassifikatoren zur Vorhersage von Herzerkrankungen zu verbessern. Der SMO-Classifier mit der Chi-Quadrat-Attributsbewertungsmethode erreichte eine bemerkenswerte Genauigkeit. Die Studie unterstreicht die Bedeutung einer geeigneten Auswahl von Merkmalen und der Abstimmung von Hyperparametern. Obwohl zufriedenstellende Ergebnisse erzielt wurden, besteht die Möglichkeit, weitere ML-Algorithmen und -Techniken der Merkmalsauswahl zu erforschen, mehrere Datensätze zu kombinieren und weitere Experimente durchzuführen, um die Vorhersageleistung zu verbessern [5].

3 Theorie

3.1 Methoden der Datenwissenschaften

3.1.1 Einleitung

Die Bewältigung von Herausforderungen durch die Gestaltung von Maschinen, die durch Nutzereingaben und -ausgaben lernen können, prägt den grundlegenden Ansatz der Lerntheorie (Maschinelles Lernen). Über die Jahrzehnte hinweg wurden zahlreiche Teilaspekte identifiziert, die die Extraktion und Auswahl von Merkmalen mit dem Ziel der Dimensionsreduktion einschließen. Die Vielfalt der Repräsentationsräume, die Beliebtheit, die Komplexität und sämtliche Variationen des Lernproblems haben zu einer Vielzahl von Lösungsansätzen geführt. Zur Bewältigung solcher Probleme wurden diverse Methoden entwickelt, darunter auch bekannte Algorithmen wie Klassifikations- und Regressionsmodelle.

Im Rahmen dieses Projektes wird das Problem der Vorhersage von HKE mithilfe von statistischen Lerntechniken gelöst.

Dieses Kapitel bietet einen Überblick über die Datenwissenschaft und erläutert die Unterschiede zwischen Klassifizierung und Regression. Der erste Teil ist daher dem aktuellen Forschungsstand im Zusammenhang mit der Anwendung dieser Methoden erörtert. Abschließend werden die Anwendungen des statistischen Lernens in Erinnerung gerufen und beschrieben.

3.1.2 Definitionen

Künstliche Intelligenz (KI): Es sind das Verhalten und die spezifischen Eigenschaften von Computerprogrammen, die Maschinen dazu bringen, menschliche Fähigkeiten und Arbeitsmuster nachzuahmen. Eine der wichtigsten dieser Eigenschaften ist die Fähigkeit, zu lernen, Schlüsse zu ziehen und auf Situationen zu reagieren, die nicht in die Maschine einprogrammiert wurden. Der Begriff KI ist jedoch umstritten, da es keine endgültige Definition von Intelligenz gibt [6].

Statistische Lernmethoden: Statistisches Lernen ist eines der Studienfelder der KI. Es ist auch eine Wissenschaft die sich mit der Entwicklung und Implementierung automatisierter Verfahren befasst, die es einer Maschine ermöglichen, zu lernen und sich weiterzuentwickeln. Der Mensch neigt dazu, seine Situation zu verbessern, indem er über die Jahre hinweg gesammelten Erfahrungen nutzt. Das Verfahren kann auf Maschinen angewendet werden. Dies wird Maschinelles Lernen genannt, bei dem ein Algorithmus anhand von Beispielen trainiert wird. Diese Technik ermöglicht es der Maschine zu “ler-

nen", ohne die Algorithmen, aus denen sie besteht, ändern zu müssen, sodass sie Aufgaben ausführen kann, für die sie nicht programmiert wurde. Statistisches Lernen lässt sich in drei große Abschnitte unterteilen:

1. Im ersten Abschnitt wird, zur Lösung eines bestimmten Problems, ein statistisches Lernalgorithmus ausgewählt.

2. Im zweiten Abschnitt wird der Algorithmus antrainiert um Fehler zu minimieren.

3. Im dritten und letzten Abschnitt werden neue Eingaben erhalten, um entweder einen quantitativen Wert oder einen qualitativen Wert vorherzusagen. Es ist zu beachten, dass vor dem ersten Schritt eine Vorverarbeitung der Daten notwendig, um auf ein gutes Modell hoffen zu können.

3.1.3 Lernmethoden

Die Lernmethoden werden oft danach eingeteilt, wie die Algorithmen lernen, genauer zu werden, um das Ereignis vorherzusagen, und welche Art von Algorithmus Programmierer wählen, hängt von der Art der Daten ab, die sie vorhersagen möchten:

Überwachtes Lernen (Supervised learning) In der Praxis wird größtenteils Überwachtes Lernen als statistische Lernmethode genutzt. Hier werden Eingabevariablen (x) und Ausgabevariablen (y) verwendet, um die Abbildungsfunktion von der Eingabe zur Ausgabe mithilfe eines Algorithmus zu lernen.

$$y = f(x)$$

Das Ziel ist es, die Abbildungsfunktion so gut zu approximieren, dass, sobald neue Eingabedaten x verfügbar sind, die Ausgabevariablen y für diese Daten vorhergesagt werden können. Dies wird als Überwachtes Lernen bezeichnet, weil der Prozess des Lernens eines Algorithmus aus dem Lerndatensatz als ein Lehrer betrachtet werden kann, der den Lernprozess überwacht. Die richtige Antworten sind beim überwachten Lernen bereits bekannt, der Algorithmus führt iterative Vorhersagen über die Lerndaten und wird gewissermaßen vom Lehrer korrigiert. Der Lernprozess wird beendet, wenn der Algorithmus ein akzeptables Leistungsniveau erreicht hat. Zum überwachten Lernen gehören Klassifizierung und Regression [7].

Halbüberwachtes Lernen (Semi-supervised learning) Programme, bei denen eine riesige Menge an Eingabedaten (x) zur Verfügung gestellt wird und nur einige Daten beschriftet werden (labeled data), erfordern den Einsatz von semi-überwachtem Lernen. Ein gutes praktisches Beispiel ist ein Fotoarchiv, in dem nur einige der Bilder mit Labels versehen

sind, z. B. durch Erkennung (Hund, Katze, Mensch), während die meisten Bilder keine Labels haben. Viele Probleme in der realen Welt fallen in diesen Bereich. So kann es teuer oder zeitaufwändig sein, alle Daten zu beschriften, da dies möglicherweise den Zugang zu Experten auf diesem Gebiet erfordert. Während nicht etikettierte Daten billig und leicht zu speichern sind. In diesem Fall kann die Verwendung von überwachtem und nicht überwachtem Lernen für die Identifizierung von Eingabedaten (x) vorteilhaft sein, oder aber optimale Vorhersagen für nicht etikettierte Daten zu machen, diese Daten wieder in eine überwachte Lernmethode einzubringen, um dann unsichtbare Daten vorhersagen zu können [8].

Unüberwachtes Lernen (Unsupervised learning) Unüberwachtes Lernen ist erforderlich, wenn nur die Eingabedaten (x) bekannt sind. Das Ziel des unüberwachtes Lernens ist es, die zugrunde liegende Struktur oder Verteilung in den Daten zu modellieren, um mehr über die Daten zu erfahren. Diese Verfahren werden als unüberwachtes Lernen bezeichnet, da, im Gegensatz zum überwachten Lernen, die Ergebnisse nicht bekannt sind. Die Algorithmen werden sich selbst überlassen. Die Probleme des unüberwachten Lernens können in Cluster- und Assoziationsprobleme unterteilt werden. Ein gängiges Fallbeispiel wäre, Datensätze mit Kundenkaufverhalten bei Amazon. Ohne vorherige Informationen (Labels), könnte unüberwachtes Lernen verwendet werden können, um spezifische Verhaltensmuster zu verstehen, personalisierte Strategien zu entwickeln oder Produktvorschläge zu machen. Spezifische Algorithmen die häufig verwendet werden sind Clustering-Algorithmen [].

Verstärkungslernen (Reinforcement learning) Verstärkungslernen ist eine Form von unüberwachtes Lernen, indem Algorithmen mit einem klaren Ziel und einem definierten Satz von Regeln programmiert werden, um dieses gegebene Ziel zu erreichen. Der Algorithmus wird so programmiert, dass die Maschine Belohnungen (positive rewards) erhält, wenn etwas Nützliches, das mit dem Endziel zusammenhängt, erreicht wird. Falls nicht, erhalten diese eine Strafe [9]. Diese Lernstrategie dient in der Regel dazu:

1. Robotern beibringen Aufgaben zu erledigen
2. Bots das Spielen von Videospiele beibringen
3. Bei der Planung von Ressourceverwaltungsprozessen

3.1.4 Algorithmen

Es gibt keine perfekten Algorithmen oder einen Algorithmus der besser ist als der andere. Jeder kann je nach Problemdomäne besser sein. Hier kommt der menschliche Faktor ins Spiel, um die richtige Wahl zu treffen, indem der Mensch die verwendeten Daten, den

Gegenstand der Fragestellung und vor allem die Leistungsindikatoren berücksichtigt. Im nächsten Abschnitt werden die bekanntesten Methoden vorgestellt.

Lineare Regression

Die lineare Regression ist eine weit verbreitete statistische Methode des überwachten Lernens, die verwendet wird, um die Beziehung zwischen einer unabhängigen und einer abhängigen Variable zu modellieren. Die Genauigkeit und Zuverlässigkeit dieser Methode hängen von bestimmten Grundannahmen ab, die die Interpretation und effektive Anwendung der Regressionsanalyse unterstützen. Die Hauptannahme besteht darin, dass die lineare Regression eine gerade Linie durch eine Reihe von Datenpunkten passt, um die Beziehung zwischen ihnen am besten zu repräsentieren. Die lineare Gleichung $Y = aX + b$ ist das Grundgerüst der linearen Regression, wobei Y die abhängige Variable, X die unabhängige Variable, a die Steigung der Linie und b der y-Achsenabschnitt ist. Die Steigung und der Achsenabschnitt sind wichtige Elemente der Regression und geben an, wie sich die abhängige Variable ändert bzw. welchen Wert sie annimmt, wenn die unabhängige Variable sich ändert oder Null ist. Der Prozess der linearen Regression umfasst mehrere Schritte, darunter die Datenerfassung, Datenbereinigung, Identifizierung der Variablen, explorative Datenanalyse und Bewertung der Modellleistung. Wichtige Bewertungsmetriken sind der mittlere quadratische Fehler (MSE), die Wurzel des mittleren quadratischen Fehlers (RMSE) und der Bestimmtheitsmaß (R-Quadrat), die die Anpassung des Modells an die Daten quantifizieren. Die Grundannahmen der linearen Regression beinhalten Linearität, Unabhängigkeit und Homoskedastizität der Residuen. Trotz ihrer Nützlichkeit kann die lineare Regression komplexe Beziehungen vereinfachen, durch Ausreißer beeinflusst werden und keine Kausalität implizieren. Dennoch bildet sie eine wichtige Grundlage für die Vorhersagemodellierung und ein Verständnis ihrer mathematischen Grundlagen ist ein guter Ausgangspunkt für komplexe Algorithmen des maschinellen Lernens [11].

Logistische Regression

Die logistische Regression ist eine Methode des überwachten Lernens, die speziell für die Modellierung von binären oder kategorischen abhängigen Variablen entwickelt wurde, im Gegensatz zur linearen Regression, die kontinuierliche Werte vorhersagt. Wesentlich für die logistische Regression sind die Grundannahmen, darunter die lineare Trennbarkeit der Daten, die Unabhängigkeit der Werte und die Homoskedastizität der Residuen.
Ein zentraler Aspekt der logistischen Regression ist die Modellierung der Wahrscheinlichkeit eines Ereignisses als Funktion der unabhängigen Variablen. Dies geschieht mithilfe der logistischen Regressionsfunktion, die die Wahrscheinlichkeit eines Ereignisses als Sigmoidkurve darstellt. Diese Kurve ermöglicht die Interpretation der Vorhersagen in Form von Wahrscheinlichkeiten und die Festlegung von Schwellenwerten für Klassifikationsentscheidungen.

Die wichtigsten Werte, die bei der logistischen Regression berücksichtigt werden, sind die Logikfunktion und die Log-Odds-Transformation. Die Logikfunktion bildet die unabhängigen Variablen auf die Sigmoidkurve ab und berücksichtigt dabei die Beziehung zwischen ihnen. Die Log-Odds-Transformation ist eine zentrale Komponente, die die Wahrscheinlichkeit eines Ereignisses in eine lineare Beziehung zu den unabhängigen Variablen transformiert, was die Anwendung von linearen Regressionstechniken ermöglicht.
Trotz ihrer Einfachheit und Effektivität bleibt die logistische Regression ein wichtiges Werkzeug im überwachten Lernen für binäre Klassifizierungsprobleme. Sie ermöglicht die Interpretation von Vorhersagen als Wahrscheinlichkeiten und die Anpassung von Schwellenwerten für Entscheidungsprozesse[12].

Lasso- Ridge-Regression

Lasso- Ridge-Regression sind eine besondere Form der Regression [10]. Die Ridge-Regression zielt darauf ab, die Größe der Koeffizienten zu verringern, um eine Überanpassung zu vermeiden, aber sie setzt keinen der Koeffizienten auf Null. Die Summe der Quadrate der Koeffizienten müssen unter einem festen Wert liegen. Die Ridge-Regression verbessert die Effizienz, aber das Modell ist aufgrund der potenziell hohen Anzahl von Merkmalen weniger interpretierbar. Die Ridge-Regrssion schneidet besser in Fällen von Multikolinearität ab oder wenn es eine hohe Korrelation zwischen bestimmten Merkmalen gibt bzw. geben kann. Dies liegt daran, dass es die Varianz im Austausch für die Verzerrung reduziert. Dafür muss die Anzahl der Merkmale kleiner als die Anzahl der Beobachtungen sein, da die Ridge-Regression keine Merkmale entfernt und in diesem Fall zu schlechten Vorhersagen führen kann. Diese Methode war die beliebteste Methode, bevor die Lasso-Methode aufkam. Die Idee ist ähnlich, nur das Verfahren ist anders.
Die Lasso-Methode verwendet nur eine Teilmenge der ursprünglichen Merkmale. Dadurch das die Summe der absoluten Werte der Koeffizienten kleiner als ein fester Wert ist, wird die Leistung des Modells verbessert. Zu diesem Zweck wird die Größe der Koeffizienten verringert, was dazu führt, das einige Merkmale einen Koeffizienten von Null haben, was sie im wesentlichen aus dem Modell ausschließt. Auf diese Weise werden die Merkmale gefiltert und das Modell vereinfacht und verständlicht.
Es lässt sich zusammenfassend sagen, dass die Lasso-Methode sich besser eignet, wenn mehrere Merkmale vorhanden sind und das Ziel die Erstellung eines einfachen und verständlichen Modells ist. Jedoch ist diese Methode, im Falle einer hohen Korrelation, nicht optimal. Die Ridge-Regression hingegen funktioniert besser bei wenigen Merkmalen oder bei Merkmalen mit hoher Korrelation, sollte aber ansonsten in den meisten Fällen aufgrund der höheren Komplexität und der geringeren Interpretierbarkeit vermieden werden.

Entscheidungsbäume

Wenn die Beziehung zwischen den Merkmalen und dem Ergebnis nicht linear ist, sind

lineare und logistische Regression nicht immer effektiv. Entscheidungsbäume bieten hier eine alternative Lösung. Sie teilen den Datensatz basierend auf Kriterien wie Informationsgewinn auf, was die Menge an neuen Informationen angibt, die durch diese Aufteilung erhalten werden können. Der Baum besteht aus Knoten, die verschiedene Teilmengen der Daten repräsentieren, und die Vorhersage erfolgt, indem das durchschnittliche Ergebnis der Trainingsdaten in jedem Knoten verwendet wird. Entscheidungsbäume sind für Klassifikation und Regression geeignet und bieten klare Interpretierbarkeit und Visualisierung. Sie erfordern keine Standardisierung der Daten und sind robust gegenüber Ausreißern und fehlenden Werten.
Dennoch haben Entscheidungsbäume auch Nachteile. Überanpassung ist ein Hauptproblem, bei dem der Baum zu komplex wird, um verallgemeinerbare Vorhersagen zu treffen. Er kann instabil sein, da das Hinzufügen eines Datenpunktes die gesamte Baumstruktur beeinflussen kann. Ein wenig Rauschen kann die Vorhersageleistung beeinträchtigen.
Entscheidungsbäume nutzen wichtige Konzepte wie Entropie und Informationsgewinn, um Entscheidungen zu treffen. Entropie misst die Unordnung der Daten und steuert die Aufteilung des Baumes. Informationsgewinn quantifiziert, wie viel Information ein Merkmal über die Klasse liefert, und wird verwendet, um die bestmögliche Aufteilung zu finden [13].

Random Forest Klassifizierung

Der Random Forest Classifier (RFC) ist ein leistungsstarker Algorithmus für maschinelles Lernen, der für Klassifizierungsprobleme verwendet wird. Er erstellt mehrere Entscheidungsbäume und kombiniert sie zu einer endgültigen Vorhersage. Der RFC zeichnet sich durch hohe Vorhersagegenauigkeit und Widerstandsfähigkeit gegenüber Überanpassung aus, erfordert jedoch die Abstimmung seiner Hyperparameter für optimale Leistung.
Jeder Entscheidungsbaum im Random Forest wird auf einer zufälligen Teilmenge der Trainingsdaten und Merkmale erstellt, um Vielfalt und weniger Korrelation zwischen den Bäumen sicherzustellen. Diese zufällige Merkmalsauswahl hilft, Überanpassung zu verhindern und die Modellleistung zu verbessern. Die richtige Anzahl von ausgewählten Merkmalen ist entscheidend, um wichtige Informationen angemessen zu erfassen und gleichzeitig die Komplexität des Modells zu kontrollieren.
Die Erstellung des gesamten Ensembles erfolgt durch die Ensemble-Technik des Bagging, bei der Teilmengen des Trainingsdatensatzes durch Zufallsstichproben mit Ersetzung erstellt werden. Jede Teilmenge wird verwendet, um einzelne Basismodelle unabhängig voneinander zu trainieren. Durch das Zusammenfassen der Vorhersagen mehrerer Modelle reduziert Bagging Überanpassung und erhöht die Modellstabilität.
Bagging spielt auch eine entscheidende Rolle bei der Konstruktion von Random Forests. Jeder Baum im Ensemble wird auf einer zufälligen Teilmenge der Trainingsdaten trainiert, wodurch vielfältige und unkorrelierte Bäume entstehen. Dies trägt zur verbesserten Generalisierung und Robustheit des Ensembles bei, was Random Forests besonders effektiv

für Klassifizierungsaufgaben macht. Zusammenfassend ist Bagging eine Schlüsselkomponente bei der Erstellung von Random Forests, die zu ihrer Fähigkeit beiträgt, komplexe Beziehungen in Daten zu verarbeiten und genaue sowie stabile Vorhersagen zu liefern [14].

K-NN Klassifizierung

Der k-Nearest-Neighbor (K-NN) Algorithmus ist ein einfacher und leistungstarker ML-Algorithmus der für Klassifizierungsproblee verwendet wird. Im Gegensatz zu parametrischen Modellen wie der logistischen Regression oder Entscheidungsbäumen basiert K-NN auf dem Konzept der Ähnlichkeit. Die Vorhersage eines neuen Datenpunktes erfolgt durch die Berücksichtigung einer vordefinierten Anzahl von Trainingsdatenpunkten, die dem neuen Punkt am nächsten liegen, daher der Name "k-Nächster-Nachbar´´ [15].
Ein wichtiger Aspekt von K-NN ist die Wahl des k-Werts, der angibt, wie viele Nachbarn berücksichtigt werden sollen. Die Wahl des richtigen k-Werts ist entscheidend, da ein zu niedriger Wert zu übermäßiger Anfälligkeit gegenüber Rauschen führen kann, während ein zu hoher Wert zu einer groben Generalisierung führen kann.
Ein weiteres Schlüsselelement von K-NN ist das Distanzmaß, das angibt, wie die Ähnlichkeit zwischen Datenpunkten berechnet wird. Die häufigsten Distanzmaße sind der euklidische und der Manhattan-Abstand, aber andere Maße wie der Kosinus-Ähnlichkeitskoeffizient können ebenfalls verwendet werden.
K-NN ist einfach zu verstehen und zu implementieren, erfordert jedoch eine effiziente Speicherung der Trainingsdaten. Die Vorhersagezeit hängt direkt von der Größe des Trainingsdatensatzes ab.
Die Leistung von K-NN kann bei hohen Dimensionen des Merkmalsraums abnehmen, da das Konzept der Nähe in höherdimensionalen Räumen weniger sinnvoll wird. Dies wird als Fluch der Dimensionalität bezeichnet.
Insgesamt ist K-NN ein vielseitiger Algorithmus, der besonders gut für Datensätze mit klaren Clusterstrukturen und relativ geringen Dimensionen geeignet ist.

4 Praxis

4.1 Anwendung der Methoden

Dieses Kapitel widmet sich der Untersuchung der Analyse, um die Leistung der Techniken zu bewerten, die als Lösung für die Vorhersage von Herzerkrankungen vorgeschlagen werden. Es wird eine Vergleichsstudie zwischen Regressionstechniken, Entscheidungsbäumen und Random Forest Classifier in Bezug auf Genauigkeit und andere wichtige Merkmale vorgestellt. Ziel ist es, ein Vorhersagesystem auf der Grundlage von BRFSS-Daten zu entwickeln, um die Risikoquote für Herz-Kreislauf-Erkrankungen vorherzusagen.

4.1.1 Problemraum

Heute sind die meisten Länder mit hohen und steigenden Raten von Herz-Kreislauf-Erkrankungen konfrontiert und solche Erkrankungen sind zu einer der Hauptursachen für Schwächung und Tod weltweit bei Männern und Frauen geworden. Sie werden als zweite Epidemie angesehen, die Infektionskrankheiten als Haupttodesursache ablöst. Eine frühzeitige Diagnose von Herzerkrankungen kann dazu beitragen, die Sterblichkeitsrate zu senken.

Die Diagnose einer Herzerkrankung durch Experten ist sehr zeitaufwendig und fällt mit dem Mangel an Experten zusammen, die über das entsprechende Wissen verfügen. Automatisierte Methoden können daher die Grenzen herkömmlicher Diagnosemethoden auflösen und medizinisches Wissen für Diagnosezwecke bereitstellen.

Das Ziel dieses Kapitels ist es, ein Modell zu entwickeln, das genau vorhersagen kann, ob ein Patient kein Herzproblem hat oder ein solches vorliegt. Die Algorithmen, die dabei zum Einsatz kommen, sind lineare, logistische oder Lasso-Bridge-Regression, Entscheidungsbäume, der K-NN Classifier und schließlich der Random Forest Classifier. Das beste Modell durchläuft dann eine interpretierbare Phase, in der verschiedene Metriken untersucht werden.

4.1.2 Datensatz

Die Gesamtheit der verwendeten Daten stammt aus dem Jahr 2021 und setzt sich aus einer Telefonstudie des Behavioral Risk Factor Surveillance System (BRFSS) zusammen, dem landesweit ersten System gesundheitsbezogener telefonischer Erhebungen, das Daten über die Einwohner der USA zu deren gesundheitsbezogenen Risikoverhaltensweisen, chronischen Gesundheitszuständen und der Inanspruchnahme von Präventionsdiensten erfasst. Der Datensatz besteht aus 303 Spalten und 438693 Reihen. Nur 19 der 303 Spalten und 325821 der 438692 Reihen werden nach der Datenbereinigung beibehalten. Die Daten

wurden vom BRFSS numerisch kategorisiert, welches das Verständnis des Datensatzes beeinträchtigt hat. Die Werte 7, 777, 9, 999, 14 tauchen auf wenn der Studienteilnehmer nicht antworten möchte, die Antwort nicht weiß oder vergessen hat etwas anzugeben.

Merkmal	Bezeichnung	Wertangabe
genhlth	Would you say that in general your health is?	5:Poor, 4:Fair, 3:Good, 2:Very Good or 1:Excellent, 7, 9
checkup1	About how long has it been since you last visited a doctor for a routine checkup?	1: Within the past year, 2: Within the past 2 years, 3: Within the past 5 years, 4: 5 or more years ago or 8:Never, 7, 9
exerany2	During the past month, other than your job, did you participate in any physical acitivities or exercices (such as running, fitness, calisthenics, ...)	1:yes, 2:no, 7, 9
michd	Respondents that reported having coronary heart disease (CHD) or myocardial infarction (MI)	1:yes, 2: no
chcscncr	Ever told you had skin cancer?	1:yes, 2:yes, 7, 9
chcocncr	Ever told you had any other types of cancer?	1:yes, 2:no, 7, 9
addepev3	Ever told you had a depressive disorder (including depression, major depression, dysthymia, or minor depression)?	1:yes, 2:no, 7, 9
diabete4	Ever told you had diabetes?	1:yes, 2:yes but during pregnancy, 3:no, 4:no, prediabetes or borderline diabetes, 7, 9
sex	Male or female?	1:male, 2:female
ageg5yr	Age group the respondent is belonging to.	1:[18-25], 2:[25-29], 3:[30-34], .., 13:[80+]
wtkg3	Reported weight in kg (weights are timed 100)	$x * 100$
bmi5	Reported Body Mass Index (BMI) timed 100	$x * 100$
rfsmok3	Respondents that are current smokers	1:yes, 2:no, 9
alcday5	Calculated total number of days where alcoholic beverages were consumed per month (maximal limit of 30 days)	[101-107] = 1-7 Days/Week, [201-230] = 1-30 Days/Month, 888 = 0 Days, 777, 999
fruit2	Not including juices, how often did you eat fruit?	[101-199] = Fruits/Day, [201-299] = Fruits/Week, 300 = Less than 1 serving / Month, [301-399] = Fruits Month/Year, 555 = Never, 777, 999
fvgreen1	How often did you eat vegetables (potatoes excluded)?	Same as fruit2
frenchf1	How often did you eat potatoes (french fries, fried potatoes, etc. included)?	Same as fruit2 and fvgreen1
havarth5	Ever told you had arthritis?	1: yes, 2: no, 7, 9
htm4	Reported height in cm	x

Tabelle 1: Beschreibung der wichtigsten Merkmale des BRFSS Datensatzes

Diese Daten wurden aufgrund von Komplexität und Unverständlichkeit umgeschrieben und übersichtlicher gestaltet, sodass die Merkmale vom passenden Typ sind.

Tabelle 2 beinhaltet die deskriptiven Statistiken der numerischen Merkmale.

	count	mean	std	min	25%	50%	75%	max
Height_(cm)	325953.0	170.607848	10.676263	91.0	163.0	170.0	178.0	241.0
Weight_(kg)	325953.0	83.334628	21.280496	25.0	68.0	82.0	95.0	293.0
BMI	325953.0	28.513945	6.508437	12.0	24.0	27.0	32.0	99.0
Alcohol_Consumption	325953.0	5.061662	8.213537	0.0	0.0	1.0	6.0	30.0
Fruit_Consumption	325953.0	7.293497	10.082232	0.0	1.0	3.0	12.9	128.0
Green_Vegetables$_{C}$*onsumption*	325953.0	8.572757	9.337580	0.0	1.0	6.0	12.9	128.0
FriedPotato_Consumption	325953.0	5.271619	6.115934	0.0	0.9	4.2	8.7	128.0

Tabelle 2: Deskriptive Statistiken

4.1.3 Datenvorverarbeitung

Die tatsächlichen Werte bestehen aus redundanten Daten und viel Rauschen.Die Daten müssen bereinigt und fehlende Werte aufgefüllt oder gelöscht werden, bevor die Daten zur Erstellung eines Modells eingespeist werden. Im Vorverarbeitungsprozess werden diese Probleme berücksichtigt, damit die Vorhersage genau getroffen werden kann. Sobald die Bereinigung abgeschlossen ist, müssen die Daten umgewandelt werden. Viele Lernalgorithmen arbeiten mit eigenen Daten, sie können nominaler, kategorischer, numerischer,... Art sein. So wird die Datentransformation auf den Datensatz angewendet. Die Datenreduktion wird angewandt, um einen komplexen Datensatz in eine vereinfachte Version umzuwandeln, um das Verständnis des Modells zu verbessern und die Genauigkeit des Modells zu erhöhen.

4.1.4 Visualisierung der Daten

Nachdem die Daten bereinigt wurden werden diese visualisiert um verschiedene Muster zu veranschaulichen. Dies ist ein wichtiger Schritt, um die Komplexität der Daten und die verschiedenen Verbindungen, die zwischen verschiedenen Merkmalen bestehen können, zu verstehen. Nach der Bereinigung der Daten werden diese visualisiert. Dies ist ein wichtiger Schritt, der dazu führt, die Komplexität der Daten und die verschiedenen Beziehungen, die zwischen verschiedenen Merkmalen bestehen können, zu verstehen. Die Visualisierung ermöglicht uns, Muster, Trends, Ausreißer und Beziehungen zu identifizieren, die möglicherweise in den Rohdaten verborgen und sonst nicht leicht sichtbar sind. Welche Informationen können aus den Daten genommen werden?

Aus der Abbildung 1 geht hervor, dass mehr als die Hälfte der Stichprobe an Fettleibigkeit leidet. Laut der WHO ist die Ernährung ein Hauptfaktor für die Erhöhung des HKE-Risikos.

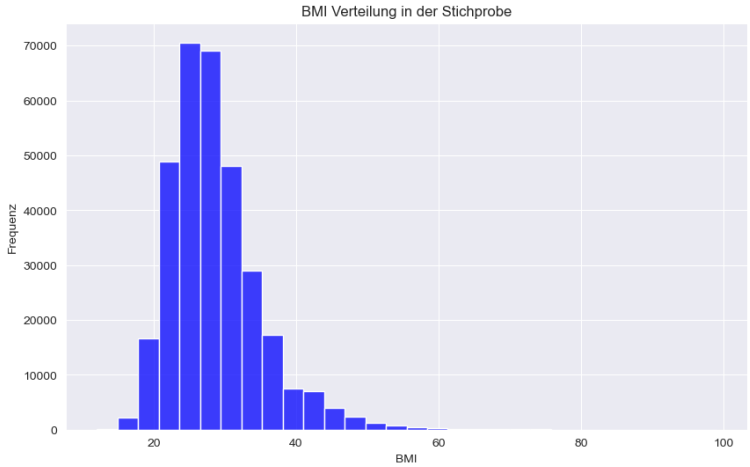

Abbildung 1: Verteilung des BMI in der Stichprobe

Die Boxplot-Grafik der Abbildung 2 ergämzt die BMI Verteilung und veranschaulicht wie sich der Körpergewicht der Menschen in der Stichprobe in Bezug auf ihr Alter verhält. Menschen die zwischen 40-44 Jahre alt sind tendieren übergewichtig zu sein.

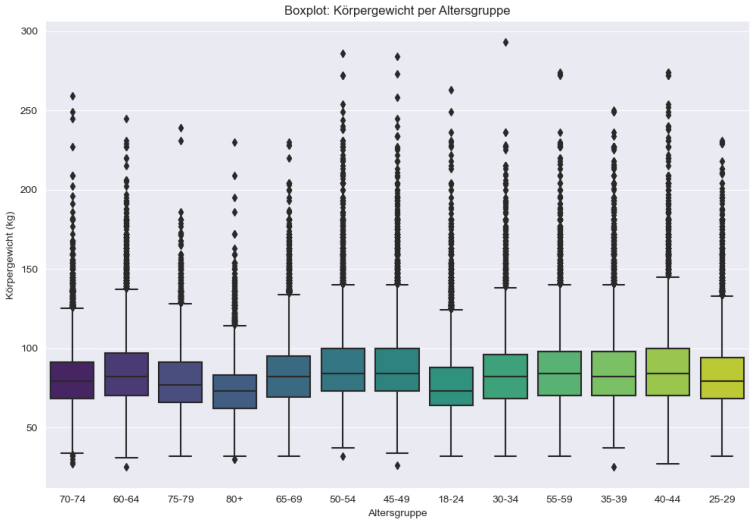

Abbildung 2: Körpergewicht per Altersgruppe

Das Säulendiagramm der Abbildung 3 stellt die Verteilung der einzelnen Gesundheitszustände in der Stichprobe dar. Die meisten Menschen gaben an “sehr gesund” zu sein.

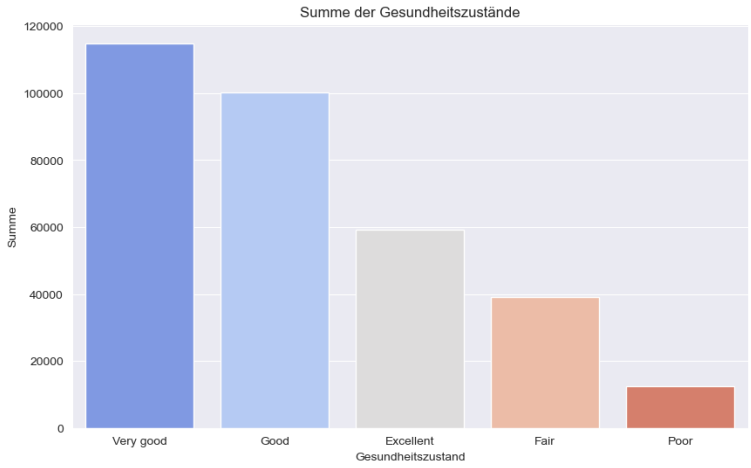

Abbildung 3: Summe der Gesundheitszustände in der Stichprobe

Abbildung 4 illustriert die Distribution des Alkoholkonsums in der Stichprobe, dafür wurden die Methode der Wahrscheinlichkeitsverteilung genutzt. In anderen Worten, wird ermittelt wie sich die Wahrscheinlichkeit des monatlichen Alkoholkonsums (in Tage/Monat) auf die Strichprobe verhält, hierbei spielt die Dichte eine wichtige Rolle, denn diese verdeutlicht diese Wahrscheinlichkeit. Es ist wahrscheinlicher dass ein Teilnehmer der Stichprobe 30 Tage lang kein Alkohol konsumiert, als dass ein Individuum täglich Alkohol trinkt.

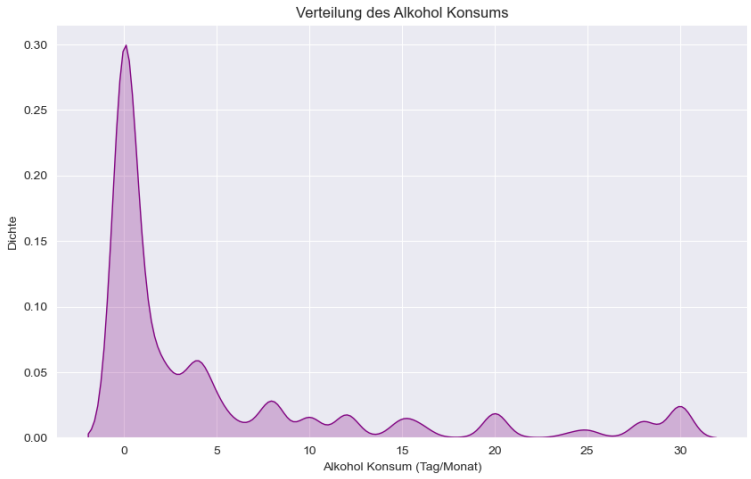

Abbildung 4: KDE-Verteilung des Alkoholkonsums in der Stichprobe

Die Heatmap aus der Abbildung 5 stellt den Zusammenhang zwischen dem Gesundheitszustand der Befragten und der Tatsache, ob sie eine sportliche Aktivität ausüben dar. Die Mehrheit der Individien in der Stichprobe (97697) gaben an einen sehr guten Gesundheitszustand zu haben und in den letzten 30 Tagen sportlich gewesen zu sein. Dabei hätte die Minderheit (4938) trotz der sportlichen Aktivitäten einen sehr schlechten Gesundheitszustand. Der Gesundheitszustand hat keinen Einfluss auf die körperlichen Aktivitäten eines Individuums. Die körperliche jedoch auf den Gesundheitszustand.

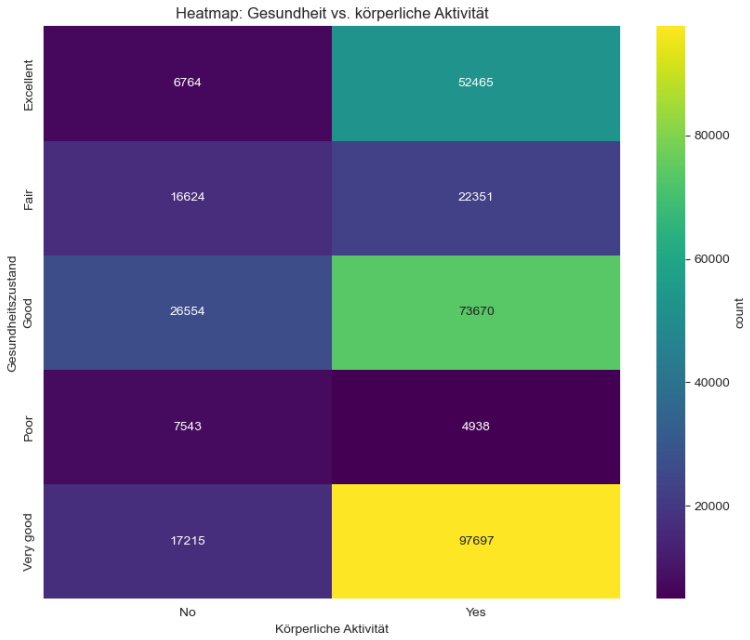

Abbildung 5: Gesundheit vs. Sport

Das Säulendiagramm aus der Abbildung 6 stellt zwei verschiedene Gruppen in der Stichprobe dar, die durch ihr Geschlecht geteilt sind. Auf der einen Seite gibt es Frauen und Männer die angaben noch nie an einer HKE erkrankt zu sein, und auf der anderen Seite Frauen und Männer, die bereits eine hatten oder haben. Es isz zu beobachten, dass in der Gruppe derjenigen, die “Nein” angaben die Mehrheit weiblich ist, während in der anderen Gruppe derjenigen, die “Ja” angaben, die Mehrheit männlich ist. Hierbei handelt es sich um die biologischen Geschlechter bei der Geburt.

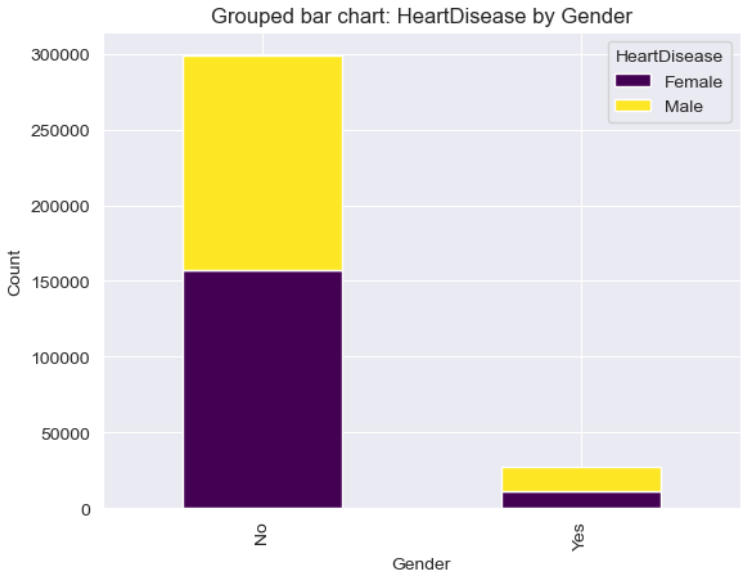

Abbildung 6: HKE-Erkrankungen nach Geschlecht

Aus der Abbildung 7 geht die Verteilung der Gesundheitszustände nach Altersgruppe hervor. Diese Grafik ergänzt die Gesundheitszustand-Verteilung aus der Abbildung 3.

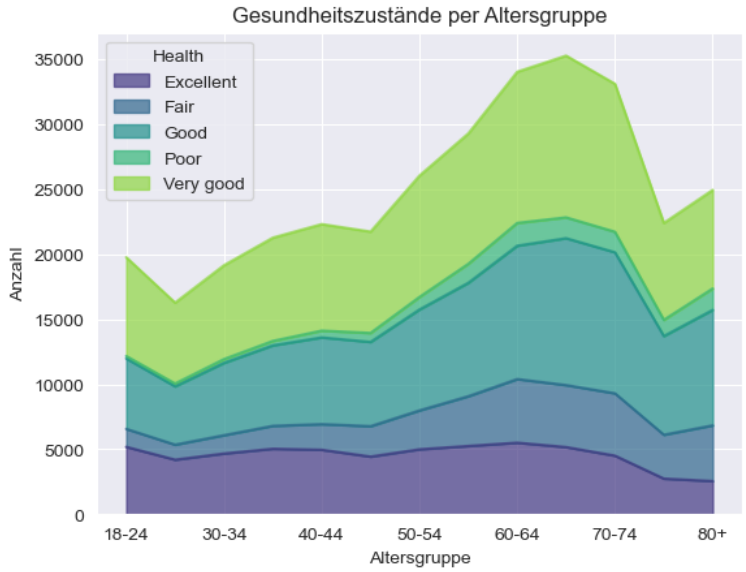

Abbildung 7: Gesundheitszustände nach Altersgruppe

Das Balkendiagramm aus der Abbildung 8 repräsentiert die Verteilung von Diabetes (Typ 1 und Typ 2) und Arthrose in der Stichprobe. Erstaunlicherweise gab dieselbe Anzahl der Menschen an, an beide Krankheiten zu leiden, aber mehr Menschen gaben nicht an Arthritis als Diabetes zu leiden. Menschen, die an Typ-2-Diabetes erkrankt sind, haben ein erhöhtes Arthroserisiko, was wahrscheinlich eher auf Übergewicht zurückzuführen ist - ein Risikofaktor für Typ-2-Diabetes - als auf den Diabetes selbst.

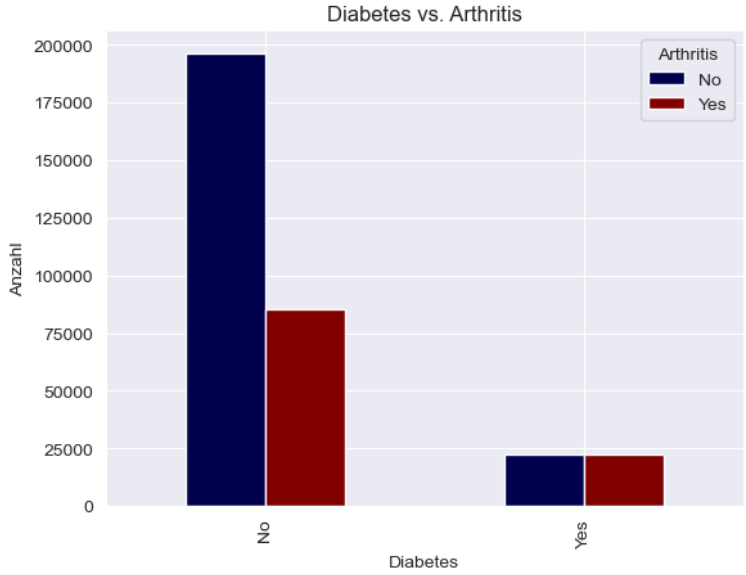

Abbildung 8: Verteilung von Diabetes und Arthrose

4.1.5 Modellentwicklung

Dieser Schritt umfasste die Auswahl des oder der geeigneten ML-Algorithmen für die Daten unter Berücksichtigung der verwendeten spezifischen Variablen und der Forschungsfrage, die behandelt wurde. Diese Entscheidung wurde auf der Grundlage einer sorgfältigen Prüfung der Datenmerkmale und der gewünschten Ergebnisse des Projekts getroffen. Da der Datensatz auch aus kategorischen Daten besteht, wurden diese enkodiert um eine einwandfreie Modellierung zu ermöglichen.
Anhand der Korrelationsmatrix wurden die linearen Beziehungen zwischen den Merkmalen dargestellt. Dabei misst der Korrelationskoeffizient die Stärke und Richtung einer lienaren Beziehung zwischen zwei Variablen. Der Korrelationskoeffizient reicht von -1 (perfekte negative Korrelation) bis 1 (perfekte positive Korrelation). Wenn ein Merkmal zunimmt, nimmt der andere Merkmal proportional zu und man spricht von einer perfekten positiven Korrelation. Wenn ein Merkmal zunimmt, nimmt der andere Merkmal proportional ab und man spricht von einer perfekten negativen Korrelation. Wenn keine lineare Beziehung zwischen den Merkmalen besteht, gibt es keine Korrelation.

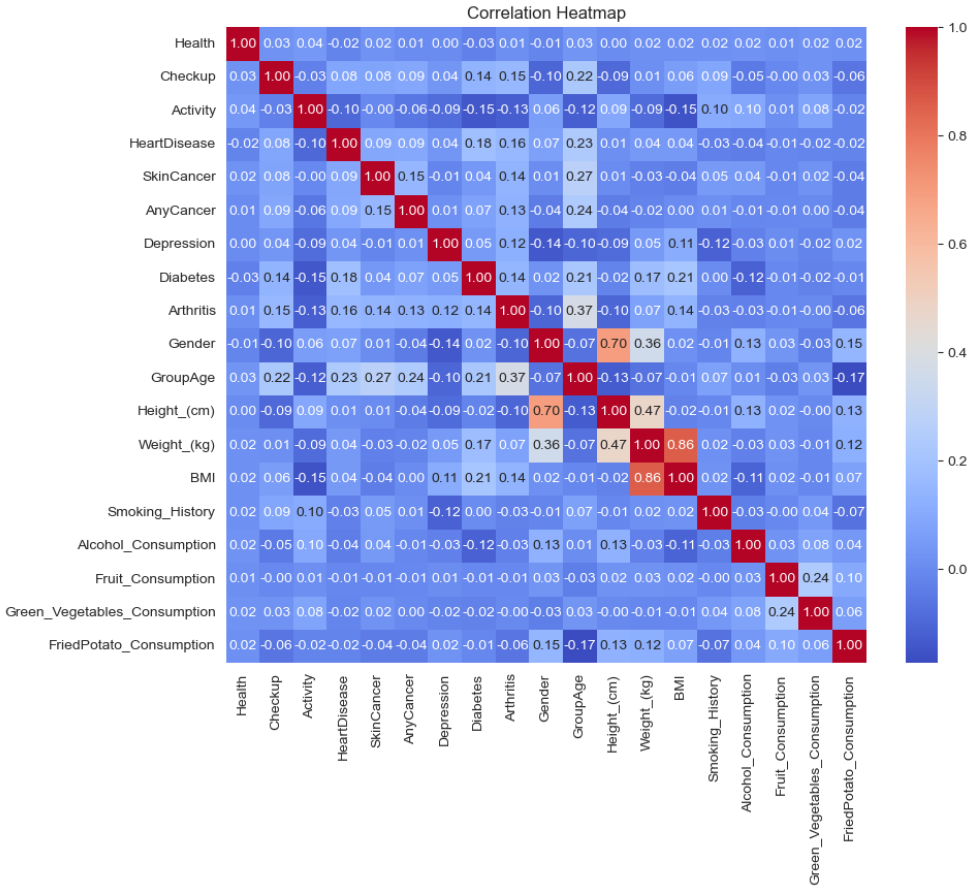

Abbildung 9: Korrelationsmatrix

Die Merkmale des Datensatzes scheinen keine überzeugende lineare Beziehung zu haben. Dies liegt höchstwahrscheinlich daran, dass der Datensatz ein Klassenungleichgewicht hat. Wie bereits erwähnt muss anhand der Lebensstilgwohnheiten ein HKE-Risiko berechnet werden, da aber nur ein kleiner Prozentsatz der Menschen in der Stichprobe tatsächlich an eine HKE gelitten haben oder leiden kommt es zum Klassenungleichgewicht. Es gaben tatsächlich 8% (26989 von 298832) der Menschen in der Stichprobe an, an eine HKE erkrankt zu sein, bzw. zu leiden. Deshalb wird eine Überstichprobe der Minderheitsklasse "Menschen die an eine HKE leiden"gebildet, indem die SMOTE Methode genutzt wird. SMOTE steht für Synthetic Minority Over-Sampling Technique und ist eine Technik mit der das Problem des Klassenungleichgewichts in Datensätzen angengangen wird, insbesondere im Falle dieses Datensatzes welches ein binäres Klassifizierungsproblem ist (Ja/Nein, 1/0, usw.). Die Daten wurden dafür im 80:20 Verhältnis in Trainings- und Testsätze aufgeteilt. Der Trainingssatz bestand aus 80% der Daten, während die verbleibenden 20% zum Testen verwendet wurden.
In Situationen, in denen eine Klasse (die Minderheitsklasse) im Vergleich zu anderen Klasse (der Mehrheitsklasse) unterrepräsentiert ist, können herkömmliche Modelle für maschinelles Lernen Schwierigkeiten haben, Instanzen der Minderheitsklasse genau vorherzusagen. SMOTE zielt darauf ab, dieses Problem durch Oversampling zu mildern und so die Klassenverteilung auzugleichen [16].
Anderseits ist das Oversampling ein Problem. Modelle die auf fiktiven Daten trainiert wurden, können versagen, wenn sie auf reale Probleme angewandt werden. Die grundlegende Schwierigkeit bei Oversampling-Ansätzen besteht darin, dass die synthetisierten Stichproben in einer realen Bevölkerung möglichwerweise nicht wirklich zur Minderheitsklasse gehören.
Um zu vermeiden, dass die Trainingsdaten unverzerrt sind, mussten die Ausreißer entfernt werden. Anhand der Visualisierungen, die während der EDA erstellt wurden, werden die Kategorien identifiziert, die Probleme verursachen könnten. Der Interquartilsbereich (IQR) spielt eine große Rolle bei der Identifizierung von Ausreißern. Der IQR ist der Bereich zwischen dem ersten und dem dritten Quartil, nämlich Q1 und Q3: $IQR = Q3 - Q1$. Die Datenpunkte, die unter $Q1 - 1,5 * IQR$ oder über $Q3 + 1,5 * IQR$ fallen, sind Ausreißer.
Nach der Auswahl der ML-Algorithmen wurden diese anhand des Trainingssatz trainiert und anschließend anhand des Testsatzes auf ihre Leistung hin bewertet. Es wurden verschiedene Algorithmen des maschinellen Lernens eingesetzt, darunter die lineare Regression, die logistische Regression, der Entscheidungsbaum-Classifier, der Random Forest-Classifier und der K-NN-Classifier. Um zu ermitteln, welches Modell am besten abschneidet, wurde – neben einer Kreuzvalidierung und Berechnung der durchschnittlichen Leistung – ein AUC-Vergleich durchgeführt. Ziel des AUC-Vergleichs ist es die Zuverlässigkeit der Modelle zu vergleichen.

Die Koeffizienten des Modells mit dem besten AUC Wert wurden untersucht, um die persönliche Merkmale zu bestimmen, die die Risikovorhersage von HKE signifikant beeinflussten. Die Bedeutung der Merkmale wurde auf deren Größe und Wichtigkeitskoeffizienten abgeleitet. Bei Variablen mit höheren Koeffizienten wurde davon ausgegangen, dass sie einen größeren Einfluss auf die Vorhersage des Modells haben.

4.1.6 Ergebnisse

Das Vorliegende Praxisprojekt untersucht die Wirksamkeit verschiedener ML-Modelle in der Vorhersage von HKE. Die Algorithmen wurden mithilfe von jupyter Notebook und Python mit Lernpaketen wie pandas, numpy, scikit-learn, matplotlib und seaborn implementiert. Die Modelle wurden auf einer Trainingsmenge trainiert und auf einer separaten Testmenge bewertet, um ihre Leistung zu beurteilen.
Um eine ausgewogene Bewertung der Modelle zu gewährleisten, wurde die f1-Bewertung als Metrik gewählt, die sowohl die Genauigkeit (*precision*) als auch die Wiedererkennung (*recall*) berücksichtigt.

	Modell	f1-score
1	Logistic Regression	0.77
2	Decision Tree Classifier	0.89
3	K-NN Classifier	0.87
4	Random Forest Classifier	0.93

Tabelle 3: Durchschnittliche Leistung der ML-Modelle

Auf der Tabelle 3 sind die Leistungen der verschiedenen Modelle unter Verwendung der Kreuzvalidierungsmethode zu sehen die entwickelt wurden. Die accuracy_score Methode wird nicht für Regressionsalgorithmen verwendet da die Ausgabe eine kontunierliche Variable ist. Deshalb wurde die Leistung der lineare Regression ausgelassen. Die Ergebnisse zeigen, dass das Random-Forest-Classifier Modell die höchste durchschnittliche f1-Bewertung von 0.93 erreicht, gefolgt vom Entscheidungsbaum-Classifier mit einem Score von 0.89, während die logistische Regression und der K-NN-Classifier eine niedrigere Bewertung erzielten. Zur weiteren Analyse der Leistung wurden mit Hilfe der scikit-learn Bibliothek Klassifizierungsberichte für den Trainingssatz erstellt.

	precision	recall	f1-score	support
0	0.74	0.75	0.74	59747
1	0.75	0.74	0.74	59786
accuracy			0.74	59786
macro avg	0.74	0.74	0.74	119533
weighted avg	0.74	0.74	0.74	119533

Tabelle 4: Klassifizierungsbericht: Logistic Regression

	precision	recall	f1-score	support
0	0.84	0.85	0.85	59747
1	0.85	0.84	0.85	59786
accuracy			0.85	59786
macro avg	0.85	0.85	0.85	119533
weighted avg	0.85	0.85	0.85	119533

Tabelle 5: Klassifizierungsbericht: Decision Tree Classifier

	precision	recall	f1-score	support
0	0.89	0.94	0.91	59747
1	0.94	0.88	0.91	59786
accuracy			0.91	59786
macro avg	0.91	0.91	0.91	119533
weighted avg	0.91	0.91	0.91	119533

Tabelle 6: Klassifizierungsbericht: Random Forest Classifier

	precision	recall	f1-score	support
0	0.86	0.72	0.78	59747
1	0.76	0.89	0.82	59786
accuracy			0.80	59786
macro avg	0.80	0.80	0.80	119533
weighted avg	0.80	0.80	0.80	119533

Tabelle 7: Klassifizierungsbericht: K-NN Classifier

Die Leistung des Random Forest Classifiers wurde auch anhand einer ROC-Kurve (siehe Abbildung 10) weiter untersucht. Der AUC-Wert, der den Grad der Trennbarkeit zwischen den Klassen misst, wurde mit 0.91 ermittelt. Dies deutet darauf hin, dass das Modell gesunde Personen als gesund und Personen mit einem Risiko für Herz-Kreislauf-Erkrankungen als gefährdet einstuft.

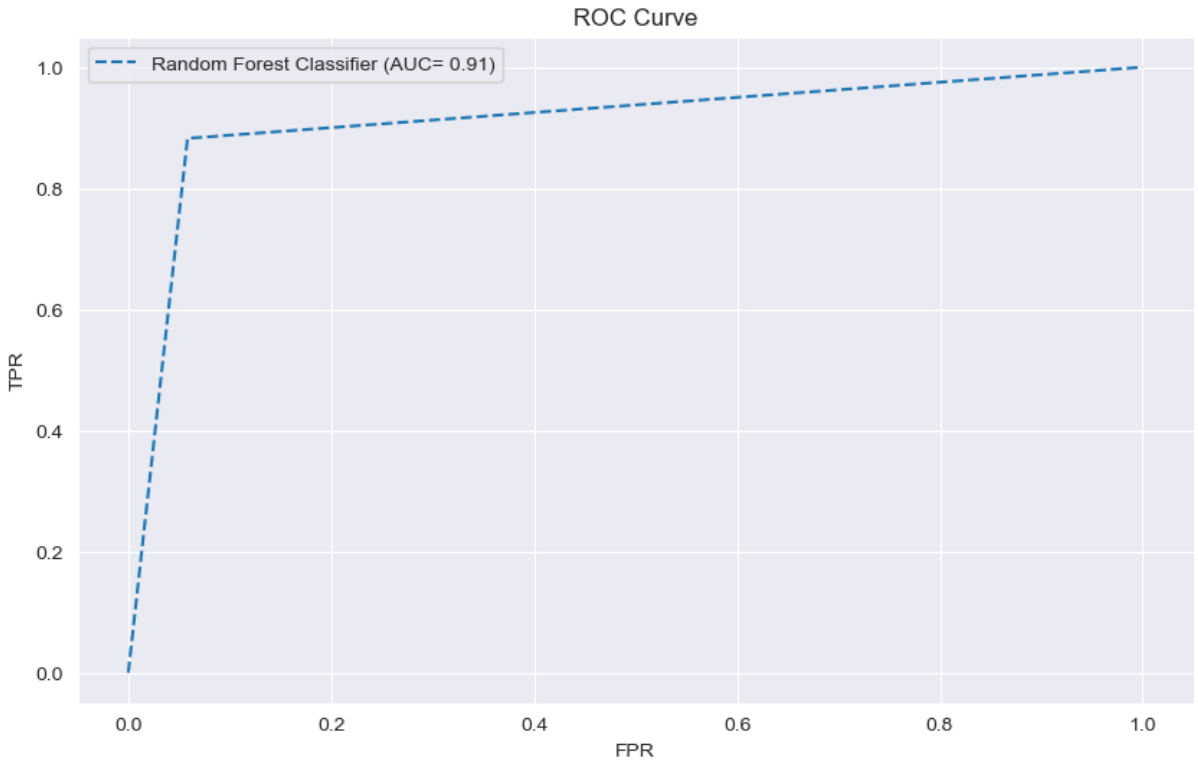

Abbildung 10: ROC-Kurve: Random Forest Classifier

Die Ergebnisse des Random Forest Classifiers wurden weiter analysiert, um die Merkmalsbedeutung zu bestimmen, die Aufschluss über die Variablen gibt, die wesentlich zu den Vorhersagen des Modells beitragen (siehe Abbildung 11). Die Merkmale Altersgruppe ”GroupAge” dicht gefolgt vom Gesundheitszustand ”Health” haben die größten Wichtigkeits-Koeffiziente. Der Random Forest Classifier tendiert mehr Gewicht dem Alter und dem Gesundheitszustand eines Individuum in der Stichprobe zu geben. Menschen die einer älteren Altersgruppen angehören oder einen schlechten Gesundheitszustand wurde in den Vorhersagen des Modells eine größere Bedeutung beigemessen, das Modell neigt dazu, diese Personen auf der Grundlage der Trainingsdaten als Personen mit einem höheren HKE-Risiko einzustufen.

Anschließend wurde der Schwerpunkt der Forschung auf die Abstimmung der Hyperparameter gelegt, um die Leistung zu verbessern, was jedoch nicht wirklich notwendig war, da das Modell bereits hohe Leistungswerte aufweist. Zu diesem Zweck wurde GridSearch mit Cross-Validation verwendet. Die Werte blieben unverändert (siehe Tabelle 8), die durchschnittliche Leistung des Modells lag wieder bei 0.93.

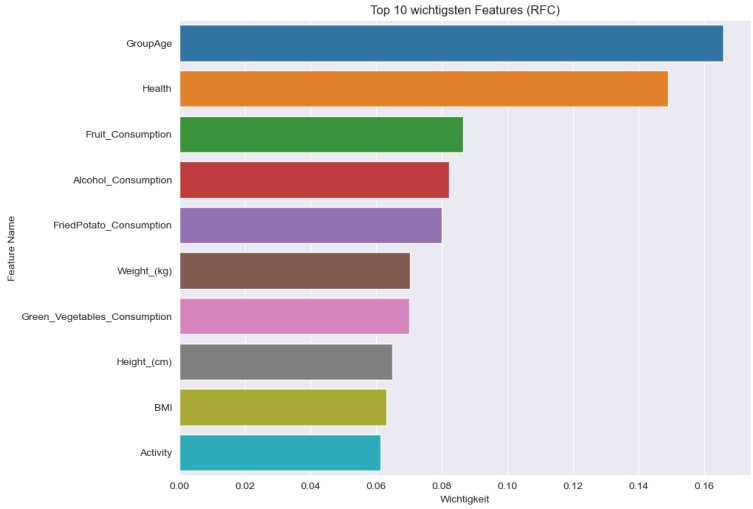

Abbildung 11: Die 10 wichtigsten Merkmale (RFC)

	precision	recall	f1-Bewertung	support
0	0.89	0.94	0.91	59747
1	0.94	0.88	0.91	59786
accuracy			0.91	59786
macro avg	0.91	0.91	0.91	119533
weighted avg	0.91	0.91	0.91	119533

Tabelle 8: Klassifizierungsbericht n°2: Random Forest Classifier

Zur weiteren Bewertung der Leistung des Random Forest Classifier Modell wurde eine Konfusionsmatrix erstellt, wie in Abbildung 12 dargestellt. Es wurden die Werte für richtig positiv (TP), richtig negativ (TN), falsch positiv (Fehler vom Typ 1) und falsch negativ (Fehler vom Typ 2) ermittelt. Aus diesen Werten wurden die Sensitivität (Recall) und Spezifität des Modells berechnet. Von den 119.533 Personen in der Teststichprobe waren 52.615 gesund und 56.250 wurden mit HKE diagnostiziert. Das Modell klassifizierte 52.753 Personen korrekt mit HKE, was einer Sensitivität von 88% entspricht. Zusätzlich klassifizierte das Modell von 52.615 gesunden Personen 45.444 Personen korrekt als Personen mit geringem HKE-Risiko, was einer Spezifität von 94% entspricht. Dies deutet darauf hin, dass das Modell ein hohes Maß an Genauigkeit bei der Identifizierung von Personen mit einem Risiko für HKE und bei der Unterscheidung von gesunden Personen aufweist.

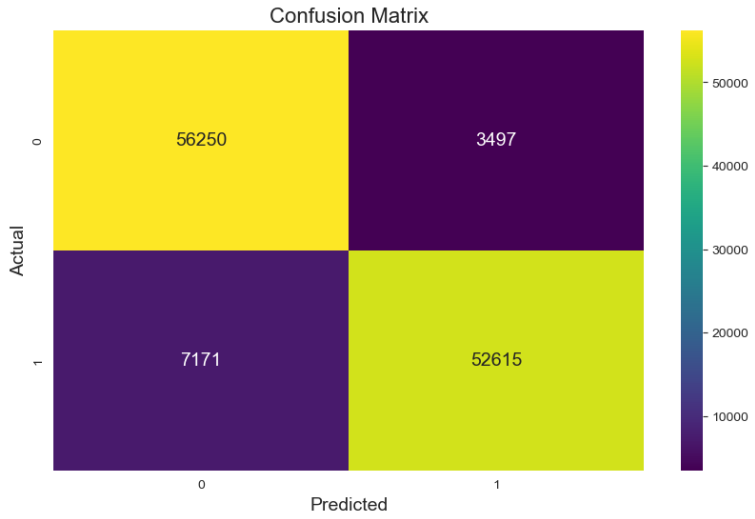

Abbildung 12: Konfusionsmatrix

Insgesamt zeigen die Ergebnisse, dass der Random Forest Classifier mit oder ohne abgestimmte Hyperparameter ein gutes Leistungsniveau bei den Testdaten aufweist. Er sagt das HKE-Risiko erfolgreich voraus, wie die hohen Werte für Sensitivität, Spezifität und AUC zeigen. Die Analyse der Merkmalsbedeutung liefert wertvolle Einblicke in die Variablen, die wesentlich zu den Vorhersagen des Modells beitragen, und hebt die Bedeutung der Altersgruppe und des allgemeinen Gesundheitszustands als Ernährungsaspekt bei der Bewertung des HKE-Risikos hervor. Diese Ergebnisse tragen zu einem besseren Verständnis der mit dem HKE-Risiko verbundenen Faktoren bei und können bei der Entwicklung gezielter Interventionen und Präventionsmaßnahmen helfen.

5 Fazit

Während des Projekts hat sich der Random Forest Classifier auf der Grundlage der durchgeführten Analyse als der leistungsfähigste Klassifikator für HKE-Vorhersagen erwiesen. Eine umfassende Bewertung mehrerer Modelle unter Verwendung von Kreuzvalidierung und f1-Bewertung als Metrik ergab, dass der Random Forest Classifier die höchste f1-Bewertung von 0.93 erreichte und damit andere Modelle wie K-NN, Entscheidungsbaum oder das logistische Regressionsmodell, die niedrigere f1-Bewertungen erreichten, übertraf.

Die weitere Prüfung der Modelle umfasste eine eingehende Untersuchung ihrer Leistung. Während die Modelle gute Leistungen zeigten, war die Diskrepanz zwischen ihren Durchschnittswerten nach der Kreuzvalidierung und davor nicht signifikant; dennoch konzentrierte sich die Forschung auf die Abstimmung der Hyperparameter, um eine Überanpassung oder ein ähnliches Phänomen zu vermeiden. Folglich wurde eine Kreuzvalidierung mit Gittersuche durchgeführt. Durch den Einsatz der GridSearchCV-Technik wurde das beste Modell mit den folgenden Parametern ermittelt: 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200; dies ergab einen mittleren F1-Score von 0.93.

Um die Leistung des Modells zu validieren, wurde das abgestimmte Random Forest Classifier Modell anhand der Testmenge bewertet. Der Klassifizierungsbericht zeigte einen f1-Score von 0.91, was eine Übereinstimmung mit dem beim Training erzielten f1-Score darstellt. Diese Übereinstimmung deutet auf eine gute Generalisierungsfähigkeit des Modells hin, was die Zuverlässigkeit des Modells unterstreicht.

Um Einblicke in die Faktoren zu erhalten, die die Vorhersagen des Random Forest Classifier beeinflussen, wurde eine Analyse der Bedeutung der Merkmale durchgeführt. Vor allem die Variablen “GroupAge” und “Health” erwiesen sich als die einflussreichsten, für die Vorhersage der HKE. Dieses Ergebnis deutet darauf hin, dass Personen, die über einen schlechten allgemeinen Gesundheitszustand den Trainingsdaten zufolge anfälliger für HKE sind.

Zur weiteren Beurteilung der Leistung des RFC-Modells wurde eine Konfusionsmatrix verwendet, die wahr-positive (TP), wahr-negative (TN), falsch-positive (Typ-1-Fehler) und falsch-negative (Typ-2-Fehler) Werte ergab. Anschliessend wurde Sensitivität (Recall) und Spezifizität berechnet. Das Modell zeigte eine beeindruckende Genauigkeit, indem es zu 88% Personen mit HKE und 94% der gesunden Personen richtig klassifizierte, was auf seine Fähigkeit hinweist, Personen mit HKE-Risiken effektiv zu identifizieren und sie von den gesunden Personen zu unterscheiden.

Die Leistung des Modells wurde anhand der Fläche unter der Kurve (AUC) der Receiver-Operating-Characteristics (ROC)-Kurve bewertet und ergab einen AUC-Wert von 0.91. Dieses Ergebnis deutet auf ein hohes Maß an Trennbarkeit zwischen gesunden Personen und Personen mit einem HKE-Risiko hin und bestätigt die Wirksamkeit des Modells bei der genauen Vorhersage dieser Erkrankung.

Seine Robustheit wird durch hohe Werte für Sensitivität, Spezifität und AUC veranschaulicht, die zusammengenommen seine Eignung für die Vorhersage des HKE-Risikos belegen. Die Analyse der Merkmalsbedeutung trägt außerdem zu unserem Verständnis der Schlüsselvariablen bei, nämlich der Altersgruppe und des allgemeinen Gesundheitszustands, die die Vorhersagen des Modells erheblich beeinflussen.Diese Erkenntnisse können eine entscheidende Rolle bei der Entwicklung gezielter Interventionen und Präventivmaßnahmen spielen und so zur Eindämmung und Behandlung von HKE beitragen.

6 Quellenverzeichnis

[BRFSS] 2021 BRFSS Survey Data and Documentation https://www.cdc.gov/brfss/annual_data/annual_2021.html, https://www.kaggle.com/datasets/dariushbahrami/cdc-brfss-survey-2021

[1] Sterberate und Todesursache der Deutschen, Statistisches Bundesamt Deutschland https://www-genesis.destatis.de/genesis/online?operation=abruftabelleBearbeiten&levelindex=2&levelid=1707920074784&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel=werteabruf&code=23211-0001&auswahltext=&werteabruf=Werteabruf#abreadcrumb (2023).

[2] World Health Organization: Cardiovascular diseases. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (2021).

[3] Nusinovici S, Tham YC, Chak Yan MY, Wei Ting DS, Li J, Sabanayagam C, Wong TY, Cheng CY. Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol. 2020 Jun;122:56-69. doi: 10.1016/j.jclinepi.2020.03.002. Epub 2020 Mar 10. PMID: 32169597. https://pubmed.ncbi.nlm.nih.gov/32169597/

[4] Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017 Jun 14;38(23):1805-1814. doi: 10.1093/eurheartj/ehw302. PMID: 27436868; PMCID: PMC5837244. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5837244/

[5] Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN, Pranavanand S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Applied Sciences. 2021; 11(18):8352. https://doi.org/10.3390/app11188352 https://www.mdpi.com/2076-3417/11/18/8352

[6] Gabriela Bonin: Künstliche Intelligenz gibt es eigentlich nicht https://hub.hslu.ch/informatik/kunstliche-intelligenz-gibt-es-nicht-wichtig-ist-digitale-ethik/

[7] Jorge Leonel: Supervised Learning https://medium.com/@jorgesleonel/supervised-learning-c16823b00c13 (2018)

[8] web 3.0 wonders: What is Semi-Supervised Learning? https://medium.com/@mhdmusthak582/what-is-semi-supervised-learning-explained-9ed9d7dd6968 (2023)

[9] Curtis Savage: What is Reinforcement Learning? https://medium.com/ai-for-product-people/what-is-reinforcement-learning-d2d3318c4423 (2023)

[10] Prof. Dr. Dirk Neumann: Ridge Regression and Lasso, Albert-Ludwigs-Universität Freiburg https://www.is.uni-freiburg.de/resources/seminar-papers/Ridge_Regression_LASSO.pdf

[11] Traycerenee: What is linear regression? https://medium.com/mlearning-ai/interview-question-what-is-linear-regression-c53c0d538c35 (2023)

[12] The Data Beast: What is logistic regression? https://medium.com/@thedatabeast/interview-questions-for-logistic-regression-359dd9488cce (2023)

[13] Decision Trees, scikit-learn https://scikit-learn.org/stable/modules/tree.html (2024)

[14] Random Forest Classifier, scikit-learn https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (2024)

[15] The Educative Team: What is k-NN? https://learningdaily.dev/what-is-k-nn-cb9ff55adb21 (2023)

[16] Nonso N, Ioannis K. Efficient treatment of outliers and class imbalance for diabetes prediction. Volumw 104. Artificial Intelligence in Medicine. 2020. https://https://doi.org/10.1016/j.artmed.2020.101815 https://www.sciencedirect.com/science/article/pii/S093336571830681X

Erklärung über die selbständige Abfassung der Arbeit

Ich versichere, die von mir vorgelegte Arbeit selbständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht.
Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

(Ort, Datum, Unterschrift)