PROJEKTSKIZZE

Prävention von Herz-Kreislauf-Erkrankungen

Team Data Dazzlers
Abbas Ammar, Alexander Anane, Damian Johnson, Victoria Zimmermann

15.05.2024

Inhaltsverzeichnis

Beschreibung der Problematik

Hintergrund

Im Rahmen des Wahlpflichtmoduls Data Science and Analytics an der Hochschule Mannheim beschäftigen wir uns als Team Data Dazzlers, bestehend aus fünf Studierenden, intensiv mit dem Thema der Prävention von Herzkrankheiten. Diese Krankheiten zählen weltweit zu den führenden Todesursachen, wobei laut Angaben der Weltgesundheitsorganisation (WHO) bis zu 18 Millionen[1] Menschen jährlich an ihnen sterben. In Deutschland allein sind jährlich etwa 350.000 Todesfälle[2] aufgrund von Herz-Kreislauf-Erkrankungen zu verzeichnen, wie durch das statistische Bundesamt belegt wird.

Die Prävention dieser Erkrankungen spielt eine entscheidende Rolle im Gesundheitsmanagement. Hierbei kommen moderne Technologien und insbesondere Wearables wie Smartwatches ins Spiel. Diese ermöglichen es einem Individuum, seine Gesundheitsdaten kontinuierlich zu messen und auf Grundlage dieser Informationen entsprechend zu handeln.

Das Team Data Dazzlers konzentriert sich darauf, eine Präventionsstrategie für Herzkrankheiten zu entwickeln und zu optimieren. Ein zentraler Ansatzpunkt dabei ist die Untersuchung physikalischer Risikofaktoren, die durch Wearables messbar sind. Durch die Analyse und Auswertung dieser Daten können frühzeitig potenzielle Risiken erkannt und präventive Maßnahmen eingeleitet werden. Hierbei kommt modernste Data-Science-Technologie zum Einsatz, um die Zusammenhänge zwischen den gemessenen physikalischen Parametern und dem Risiko für Herzkrankheiten zu erforschen und zu verstehen. Das Ziel ist es, präzise und individualisierte Empfehlungen für die Gesundheitsvorsorge zu entwickeln, die auf den individuellen Risikoprofilen der Nutzer basieren.

Mögliche Lösungen durch das Problem

1. Prävention und Früherkennung von Herz-Kreislauf-Erkrankungen

Für die Prävention und Früherkennung von Herz-Kreislauf-Erkrankungen nutzen wir ein innovatives Machine-Learning-Konzept, das auf wissenschaftlichen Datensätzen basiert. Diese Datensätze umfassen Informationen über Risikofaktoren, Krankheitsverläufe und Präventionsstrategien. Mit Hilfe von Machine-Learning analysieren wir diese Daten, um Muster und Zusammenhänge zu identifizieren, die auf ein erhöhtes Risiko für Herz-Kreislauf-Erkrankungen hinweisen können.

Der Algorithmus funktioniert wie ein Punktesystem, welches das individuelle Risiko einer Person für Herz-Kreislauf-Erkrankungen bewertet. Dieses System vergibt auf Basis der Trainingsdaten sowie der persönlichen Daten des Nutzers Punkte.

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Todesursachen/_in halt.html

Durch die Bewertung verschiedener Risikofaktoren wird ein Gesamtrisikowert ermittelt, der anzeigt, wie hoch die Wahrscheinlichkeit ist, dass eine Person in der Gegenwart, oder womöglich in der Zukunft an einer Herz-Kreislauf-Erkrankung leiden könnte.

Gleichzeitig können mithilfe von Wearables kontinuierlich Vitalwerte wie Herzfrequenz, Blutdruck und Aktivitätsniveau von Personen erfasst werden. Diese Daten werden in Echtzeit erfasst und könnten mit unserem Machine-Learning-Modell analysiert werden. Auf Grundlage dieser Analyse werden personalisierte Empfehlungen zur Verbesserung des Lebensstils und zur Prävention von Herz-Kreislauf-Erkrankungen generiert.

Diese Empfehlungen können je nach Umsetzbarkeit von den Betroffenen selbst oder mithilfe des eigenen behandelnden Arztes realisiert werden. So können Betroffene aktiv ihren Lebensstil anpassen und gesundheitsschädigende Verhaltensmuster vermeiden. Diese proaktive Herangehensweise ermöglicht es, Herz-Kreislauf-Erkrankungen effektiv zu verhindern oder zumindest ihre Entwicklung zu verlangsamen.

2. Vermeidung/Aufklärung von Volkskrankheiten

Das Gesundheitssystem konzentriert sich immer mehr auf die kurative, sprich die heilende, als auf die präventive, sprich vorbeugende Medizin.[3] Dadurch erkranken Menschen, die durch präventive Maßnahmen eigentlich nicht erkranken hätten müssen.

Durch Projekte, die sich auf die Prävention von Krankheiten konzentrieren, kann also nicht nur dem Einzelnen geholfen werden, ein gesünderes und krankheitsfreies Leben zu führen, sondern auch der Bevölkerung als Gemeinschaft. Denn vor allem Volkskrankheiten, wie Diabetes, Adipositas und Herz-Kreislauf-Erkrankungen belasten nicht nur die Betroffenen, sondern auch unser Gesundheitssystem massiv[4], denn diese Erkrankungen sind meist nur der Anfang, auf den dann Folgeerkrankungen oder Sekundärerkrankungen folgen, durch die die Menschen noch kränker werden.

Indem man sich also nicht auf die Heilung der Krankheiten, sondern stattdessen auf die Prävention und Früherkennung von Krankheiten fokussiert, erspart man nicht nur vielen Menschen das Leiden an einer Krankheit oder ermöglicht ihnen einen besseren Krankheitsverlauf, sondern man entlastet vor allem auch das überforderte deutsche Gesundheitssystem.

3. Sensibilisierung des Bewusstseins für die eigene Gesundheit

Dadurch, dass man in Arztpraxen nur noch schwer Termine bekommt und die Ärzte, wenn man dann mal einen Termin hat, unter massivem Zeitdruck stehen, fühlen sich viele Menschen im Gesundheitssystem übersehen und auch allein gelassen mit ihren Fragen und Sorgen.

[3] https://www.pedocs.de/frontdoor.php?source_opus=10361 DOI: 10.25656/01:10361
[4] https://www.uni-paderborn.de/fileadmin/psychisch-stark-am-arbeitsplatz/pdf/BDP-Bericht-2012.pdf Seite 11

Durch Projekte wie diese, bei denen Betroffene auf einfache und bequeme Art und Weise einen Überblick über die eigene Gesundheit erhalten können, wird ihnen die Angst genommen und auch eine Richtung vorgegeben, an der sie sich orientieren können. Es kann sich sehr befreiend und stärkend anfühlen, die Kontrolle über die eigene Gesundheit zu haben und nicht auf fremde Hilfe angewiesen zu sein. Durch die Einfachheit solcher Anwendungen und die einfache Integration in den Alltag der Menschen beschäftigen sich diese wahrscheinlich mehr und auch lieber mit ihrer Gesundheit, als wenn solche Projekte nicht existieren würden und dadurch wird eine allgemein gesündere Bevölkerung angestrebt und gefördert.

Projektziele

a. Grundlegende Prävention von Herzkreislauferkrankungen

Das Ziel ist es, eine ganzheitliche und effektive Präventionsstrategie für Herz-Kreislauf-Erkrankungen zu entwickeln. Dabei sollen nicht nur Risikofaktoren identifiziert und adressiert werden, sondern auch Maßnahmen zur Förderung eines gesunden Lebensstils und zur Reduzierung von Risikoverhalten ergriffen werden. Die Präventionsstrategie soll sowohl auf individueller Ebene als auch auf der Ebene der Gemeinschaft und der öffentlichen Gesundheit wirksam sein. Es sollen evidenzbasierte Interventionen entwickelt werden, die auf den spezifischen Bedürfnissen und Risikoprofilen der Zielgruppen basieren.

b. Bewertung der Aussagekraft von Wearables in Bezug auf Prävention von Herz-Kreislauf-Erkrankungen

Das Ziel ist es, die Wirksamkeit und Zuverlässigkeit von Wearables zur Prävention von Herz-Kreislauf-Erkrankungen zu untersuchen und zu bewerten. Dabei sollen verschiedene Arten von Wearables, wie Smartwatches und Fitness-Tracker, auf ihre Eignung hin analysiert werden, relevante Gesundheitsdaten zu erfassen und präventive Maßnahmen zu unterstützen.

c. Top 2-3 Risikofaktoren ausgeben

Das Ziel ist es, die wichtigsten Risikofaktoren für den Betroffenen zu identifizieren. Durch umfassende Datenanalysen und Machine-Learning-Modelle sollen die Risikofaktoren ermittelt werden, die den größten Einfluss auf die mögliche Entstehung und Entwicklung einer Herz-Kreislauf-Erkrankung haben. Dadurch kann der Betroffene sich gezielt auf diese Faktoren konzentrieren und ihnen entgegenwirken.

d. Verlaufskurve für das weitere Risiko

Das Ziel ist es, eine prognostische Risikokurve für die zukünftige Entwicklung des Risikos von Herz-Kreislauf-Erkrankungen zu erstellen. Durch die Integration von historischen

Gesundheitsdaten, aktuellen Gesundheitswerten und prädiktiven Modellen sollen individuelle Risikoprofile erstellt werden.

Diese Risikokurven sollen kontinuierlich aktualisiert und verfeinert werden, um zu zeigen, wie sich Verbesserungen oder Verschlechterungen der Risikofaktoren auf den Verlauf des Gesamtrisikos über die Zeit auswirken.

SMART-Ziele

Spezifisch: Ein Algorithmus soll mit hoher Wahrscheinlichkeit das Risiko, an einer Herz-/Kreislauferkrankung zu erkranken, vorhersagen können, wenn ihm die letzten 4 Wochen an Daten eines Nutzers vorliegen.

Messbar: Die Anzahl der gemiedenen Herz-Kreislauf-Erkrankungen soll gemessen werden. Entweder indem bereits Erkrankte ihre Daten bis zum Zeitpunkt der Erkrankung in das Modell einspielen und testen, ob ein hohes Risiko ausgegeben wird, oder indem Risikobewertungen, die das Modell ausgegeben hat, nachverfolgt werden.

Attraktiv: Nutzer haben mehr Kontrolle über die eigene Gesundheit und können diese leichter steuern und überwachen.

Realistisch: Die Ziele des Projekts sind mit den vorhandenen Ressourcen erreichbar, da das gesamte Semester an Zeit zur Verfügung steht und der Umfang des Projekts flexibel angepasst werden kann, indem Must-haves, wie eine Risikobewertung und Nice-to-haves, wie eine Risikokurve, eine GUI, etc. definiert wurden.

Terminiert: Das Projekt endet mit Abschluss des Moduls Data Science and Analytics.

Erste technische Dokumentation

Auswahl der Quellen

Entscheidung für die Datenquellen

Die Entscheidung über die Datenquellen orientierte sich an der „Checkliste für Datenakquise“. Es war wichtig, eine Mischung aus primären und sekundären Datenquellen zu erhalten. Leider gestaltete sich der Zugriff auf medizinische Daten schwierig, wodurch die Auswahl der Datensätze generell sehr eingeschränkt war. Dabei wurde berücksichtigt, möglichst viele Attribute in den Datensätzen zu haben, um viele Risikofaktoren einzubeziehen und unterschiedlich gewichten zu können. Es war auch wichtig, keine leeren Beobachtungen in den Datensätzen zu haben, da sie den Datensatz verunreinigen können.

Auch ein Bias kann einen Datensatz verunreinigen, jedoch war es schwierig, Bias vollständig zu vermeiden. Im zweiten Datensatz gibt es einen Response-Bias, da es sich um eine Befragung handelt und nicht um wissenschaftlich erhobene Daten.

Dieser Bias tritt auf, wenn die Antworten in einer Umfrage oder Studie systematisch von bestimmten Faktoren beeinflusst werden, was zu einer Verzerrung der Ergebnisse führen kann. In diesem Fall könnten persönliche Fragen zur sportlichen Aktivität, zum Gewicht und anderen Gewohnheiten die Befragten dazu bringen, unehrlich zu antworten und somit zu einem Response-Bias führen.

Beschreibung der Datenquellen

Datenquelle 1:

https://www.kaggle.com/code/georgyzubkov/heart-disease-exploratory-data-analysis/notebook

Datum: 14.05.2024

Erstellungsjahr: 2023

Art der Datensammlung: Jährliche Telefonumfrage

Quelle: Kaggle

Mögliche Arten von Bias: Response-Bias

Vorteile:

+ Viele Datensätze

+ Attribute

Nachteile:

- Stark subjektive Attribute, wie „Sportliche Aktivität“ zum Beispiel

Attributbeschreibung:

Aufgrund des Bias muss bei der Vorverarbeitung dieser Datenquelle darauf geachtet werden, den Bias miteinzubeziehen und betroffene Attribute entweder heraus zu filtern oder gegebenenfalls anders zu gewichten.

Spaltenname	Beschreibung	Datenart	Trifft Bias zu?	Relevanz
HeartDisease	Liegt eine Herz-Erkrankung vor?	Kategorisch (0/1)	Nein	Hoch
BMI	Body-Mass-Index	Numerisch	Ja	Mittel
Smoking	Raucht der Befragte?	Kategorisch (0/1)	Ja	Mittel
AlcoholDrinking	Trinkt der Befragte?	Kategorisch (0/1)	Ja	Mittel
Stroke	Liegt ein Schlaganfall in der Vergangeheit?	Kategorisch (0/1)	Nein	Hoch
PhysicalHealth	Wie oft im Monat fühlt der Befragte sich „physisch ungesund“?	Numerisch	Ja	Mittel
MentalHealth	Wie oft im Monat fühlt der Befragte sich „mental ungesund“?	Numerisch	Ja	Niedrig
DiffWalking	Schwierigkeit, Treppen zu steigen.	Kategorisch (0/1)	Ja	Mittel
Sex	Geschlecht	Kategorisch (female/male)	Nein	Hoch
AgeCategory	Alters-Kategorie	Kategorisch	Nein	Hoch
Race	Ethnie	Kategorisch	Nein	Niedrig
Diabetic	Liegt eine Diabetes-Erkrankung vor?	Kategorisch (0/1)	Nein	Mittel
PhysicalActivity	Wurde Sport in den letzten 30 Tagen gemacht?	Kategorisch (0/1)	Ja	Mittel
GenHealth	Wie „gesund“ fühlt sich der Befragte?	Kategorisch (0/1)	Ja	Mittel
SleepTime	Wie viele Stunden Schlaf?	Numerisch	Ja	Hoch
Asthma	Liegt diagnostiziertes Asthma vor?	Kategorisch (0/1)	Nein	Unsicher
KidneyDisease	Liegt eine Nieren-Erkrankung vor?	Kategorisch (0/1)	Nein	Unsicher
SkinCancer	Liegt diagnostizierter Hautkrebs vor?	Kategorisch (0/1)	Nein	Unsicher

Datenquelle 2:

https://www.kaggle.com/code/desalegngeb/heart-disease-predictions

Datum: 14.05.2024

Erstellungsjahr: 1988

Art der Datensammlung: medizinische Erhebung

Quelle: Kaggle

Mögliche Arten von Bias: keiner

Vorteile:

+ Wird von 64+ Papern genutzt = zuverlässig

+ Viele Attribute, die zu der Fragestellung passen

Nachteile:

- Aktualität der Daten

- Wenig Ortsvarianz

Attributbeschreibung:

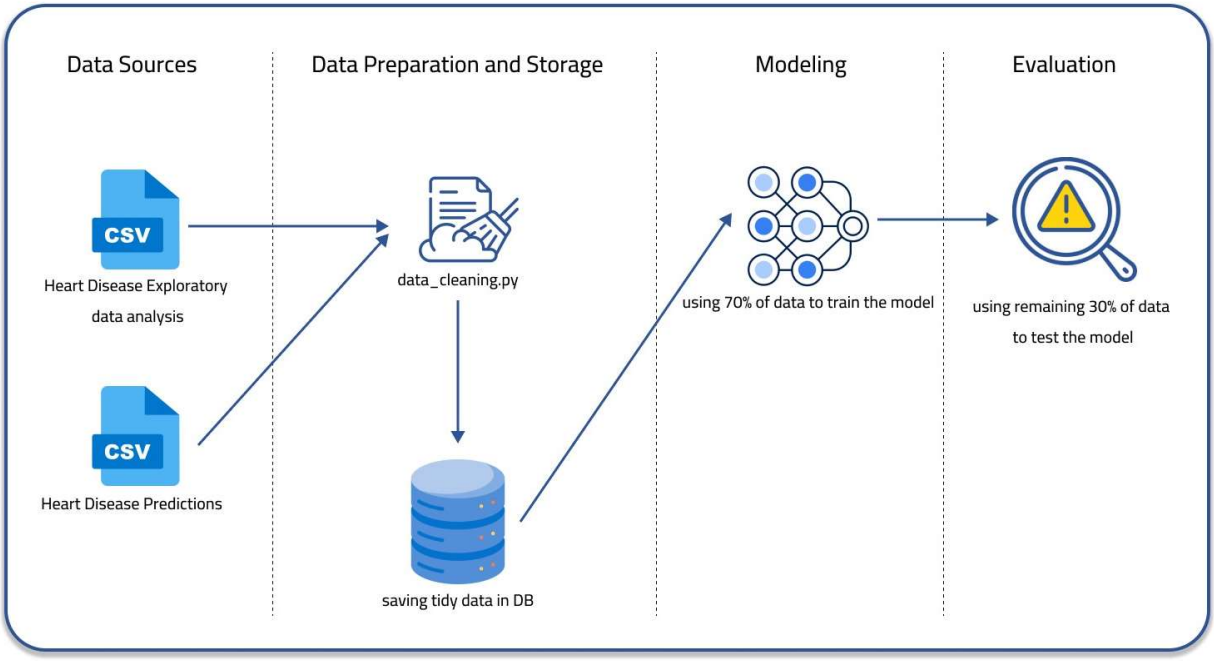
Da in dieser Quelle die Attribute mehr medizinischen Hintergrund fordern, muss hier noch genauer recherchiert werden, inwiefern Attribute einen Einfluss auf das Risiko an einer Herz-Kreislauf-Erkrankung zu erkranken haben und wie sie miteinander zusammenhängen.

Spaltenname	Beschreibung	Datenart	Trifft Bias zu?	Relevanz
age	Alter in Jahren	Numerisch	Nein	Hoch
sex	Geschlecht	Kategorisch (0/1)	Nein	Hoch
cp	Typischer Brust-schmerz	Kategorisch	Nein	Mittel
trestbps	Blutdruck in der Ruhe	Numerisch	Nein	Hoch
chol	Serum Choleste-rin in mg/dl	Numerisch	Nein	Unsicher
fbs	Nüchtern Blutzu-cker > 120 mg/dl	Kategorisch (0/1)	Nein	Mittel
restecg	Ruhe-Kardio-gramm	Kategorisch	Nein	Unsicher
thalach	Maximale Herz-rate	Numerisch	Nein	Hoch
exang	Schmerzen in der Brust durch Trai-ning (angina)	Kategorisch (0/1)	Nein	Unsicher
oldpeak	ST-Senkung unter Belastung im Ver-gleich zur Ruhe	Numerisch	Nein	Unsicher
slope	Steigung der ST-Strecke am Peak der Belastung	Kategorisch	Nein	Unsicher
ca	Anzahl der gro-ßen Gefäße (0-3), gefärbt durch Flu-oroskopie	Kategorisch	Nein	Unsicher
target (the lable)	Liegt eine Herz-Erkrankung vor?	Kategorisch (0/1)	Nein	Hoch

Architektur-Diagramm

In diesem Data-Science-Projekt zur Prävention von Herz-Kreislauf-Erkrankungen durchlaufen wir folgende Schritte:

1. Zunächst werden die Gesundheitsdaten aus den CSV-Dateien gesammelt und in ein geeignetes Format überführt.
2. Anschließend werden die Daten auf fehlende Werte und Ausreißer überprüft und nötige Korrekturen und Transformationen vorgenommen. Die gesäuberten Daten werden zusätzlich abgespeichert. Um die spätere Modellbewertung zu ermöglichen, werden die Daten in Trainings- und Testdatensätze aufgeteilt.
3. Mit den vorbereiteten Trainingsdaten trainieren wir ein Machine-Learning-Modell. Verschiedene Modelle können verglichen und dasjenige mit der besten Leistung ausgewählt werden.
4. Das ausgewählte Modell wird anhand des Testdatensatzes evaluiert, um seine allgemeine Vorhersagegenauigkeit zu bestimmen. Geeignete Metriken werden verwendet, um die Modellleistung zu quantifizieren.
5. Basierend auf den Ergebnissen des Modells interpretieren wir die Zusammenhänge zwischen den Risikofaktoren und der Erkrankungswahrscheinlichkeit. Das Modell kann dann eingesetzt werden, um neue Daten vorherzusagen oder weiterführende Erkenntnisse zu gewinnen.



Code-Repository-Link

https://github.com/dj0910/DSA_Project

Umsetzung

In diesem Projekt wird Python gegenüber R für Data Science und Machine Learning aufgrund seiner Vielseitigkeit, aktiven Community und leicht erlernbaren Syntax bevorzugt.

Zusätzlich ermöglichen uns leistungsfähige Bibliotheken und Frameworks wie NumPy, Pandas, Seaborn, Plotly und Scikit-Learn, unsere Analysen effizient durchzuführen und Modelle zu entwickeln. Diese Tools bieten eine solide Grundlage für Datenmanipulation, Visualisierung und maschinelles Lernen, was unsere Arbeitsabläufe optimiert und die Qualität unserer Ergebnisse verbessert.

Falls sich die Möglichkeit ergibt, wird im Rahmen des Projekts eine simple grafische Oberfläche erstellt, die die Bedienung vereinfachen und visualisieren soll.