Projekt:

„Identifizierung von Risikofaktoren in der Entstehung von Lungenkrebs mithilfe von maschinellem Lernen in R.“

Sabrina Frank

19.01.2024

CAT

„Identifizierung von Risikofaktoren in der Entstehung von Lungenkrebs mithilfe von maschinellem Lernen in R.“

Kann … // Algorithmus x // exact … // (Kunden)-Problem … berechnen / lösen?

Der implementierte Random Forest Algorithmus ist in der Lage, mit Hilfe einer Auswahl von 23 Risikofaktoren eine Risikostratifizierung mit einer Genauigkeit von 100% vorherzusagen. Dies ermöglicht eine Ressourcenersparnis und vergrößerte Kapazität für das Screening von potentiell gefährdeteren Personen und damit ebenso die Früherkennung und damit bessere Behandlung von Lungenkrebs.

Big Data / Data Mining / Data Science Kernaussage:

(P) roblem):
[Welcher Frage hat für die Lösung des Kunden / Auftraggeber die größte Bedeutung?]:

Lungenkrebs ist eine der tödlichsten und die prävalenteste Tumorerkrankungen weltweit[1], wobei Betroffene von einer frühen Erkennung und Behandlung massiv profitieren[2]. Um Personen mit erhöhtem Risiko besser identifizieren und entsprechende Früherkennungsmaßnahmen engmaschiger gestalten und anwenden zu können, könnte ein Algorithmus implementiert werden, der auf der Basis normierter Risikofaktoren eine Risikoprädiktion erstellen kann.
Kann also anhand der Risikofaktoren verlässlich ein Risikolevel für die Entwicklung eines Lungenkrebses mit Hilfe von Machine Learning in R vorhergesagt werden?

(I) ntervention:
(Bibliotheken- und Algorithmen-Auswahl, … z.B. pandas für Finanzdaten …)
[Welche Berechnung erwäge ich vornehmlich?]:

Der Machine Learning Algorithmus wurde vollständig in R implementiert.
Für das Laden des Datensatzes sowie dessen Bereinigung wurde data.table, tidyr, dplyr genutzt. Für die Visualisierung wurden die Pakete ggplot2, GGally, gridExtra, ggcorplot und lattice verwendet. Die Erstellung, das Training und die Testung des Machine Learning Algorithmus wurde mit der caret Bibliothek umgesetzt.
Die Übersicht über die genutzten Packages wurde mit Hilfe von NCmisc realisiert.
Für die Versionskontrolle des Projekts wurde Git mit einem Remote Repository auf GitHub genutzt (https://github.com/SabiFrank/R_Projektarbeit_12_2023.git).

[1] https://de.statista.com/statistik/daten/studie/1201305/umfrage/prognostizierte-anzahl-von-krebstodesfaellen-weltweit/ (Stand 18.01.2024, 13:31)
[2] https://www.krebsdaten.de/Krebs/DE/Content/Krebsarten/Lungenkrebs/lungenkrebs_node.html (Stand 18.01.2024, 13:31)

(K) ontrollintervention
(falls erforderlich: Bibliotheken- und Algorithmen-Auswahl … z.B. scikit-learn für Finanzdaten …)
[Was ist die andere Möglichkeit?]:

Die Kontrollintervention besteht im Testen des Klassifizierungsalgorithmus mit einem Testdatensatz.
Andere Algorithmen (Support Vector Machines, Neuronale Netzwerke etc.) erzielen womöglich bessere Ergebnisse.

(E) rgebnismaß (Zielgröße(n)) – Die Evidence
[Was möchte ich / der Kunde erreichen? Z.B. Prädiktor oder Klassifikator erstellen …]:

Das Ergebnismaß ist die Genauigkeit des Klassifizierungsalgorithmus. Wünschenswert ist eine möglichst hohe Genauigkeit der Klassifizierung und eine möglichst kleine Rate an falsch negativen Klassifizierungen.

Anmerkungen:

Ein Paper von Ahmad et al. aus 2020 nutzt den hier verwendeten Datensatz und schlägt damit die Risikofaktoren und einen Random Forest Prädiktor vor, der in der Lage sein soll, eine solche Risikostratifizierung durchzuführen[3]. An dem Datensatz zu kritisieren ist, dass die Features bis auf das Alter keine Einheit besitzen. Diese sind alle vorskaliert, wobei unklar ist, wie die Skalen zustande gekommen sind. Gleiches gilt für die Werte der Level für das Risiko der Erkrankung als Targets. Außerdem nutzen die Features unterschiedliche Skalen; einige reichen nur bis sieben, andere bis maximal neun. Die fast ausschließlich kategorische Natur der Features beeinflusst außerdem das Handling der Daten, wie hier beschrieben: https://www.playerzero.ai/advanced/r-faqs/how-to-handle-categorical-variables-in-r-a-step-by-step-guide.

Literaturhinweise:

Datensatz: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data, [Database] Lung cancer database. URL: https://data.world/cancerdatahp/lung-cancer-data (January 10, 2018)
Paper: Ahmad AS, Mayya AM. A new tool to predict lung cancer based on risk factors. Heliyon. 2020 Feb 26;6(2):e03402. doi: 10.1016/j.heliyon.2020.e03402. PMID: 32140577; PMCID: PMC7044659.
Guide für Random Forest Algorithmus in R: https://www.r-bloggers.com/2022/02/beginners-guide-to-machine-learning-in-r-with-step-by-step-tutorial/
Korrelationsplot für kategorische Features: https://stackoverflow.com/questions/52554336/plot-the-equivalent-of-correlation-matrix-for-factors-categorical-data-and-mi
Machine Learning in R: https://www.r-bloggers.com/2022/02/beginners-guide-to-machine-learning-in-r-with-step-by-step-tutorial/

[3] Ahmad AS, Mayya AM. A new tool to predict lung cancer based on risk factors. Heliyon. 2020 Feb 26;6(2):e03402. doi: 10.1016/j.heliyon.2020.e03402. PMID: 32140577; PMCID: PMC7044659.

Die Suche nach der besten Evidenz

1. Problem

Lungenkrebs ist eine der tödlichsten und die prävalenteste Tumorerkrankungen weltweit[1], wobei Betroffene von einer frühen Erkennung und Behandlung massiv profitieren[2]. Bisher gibt es noch keine standardisierte Früherkennungsdiagnostik, meist entstehen Lungenkrebsbefunde erst in späten Stadien aufgrund der schwerwiegenden Symptome oder als Zufallsbefunde[2]. Um Personen mit erhöhtem Risiko besser identifizieren und entsprechende Früherkennungsmaßnahmen engmaschiger gestalten und anwenden zu können, soll ein Algorithmus implementiert werden, der auf der Basis normierter Risikofaktoren eine Risikoprädiktion erstellen kann.

2. Definition einer wichtigen suchbaren Frage

Kann anhand der ausgewählten Risikofaktoren verlässlich ein Risikolevel für die Entwicklung eines Lungenkrebses mit Hilfe eines Random Forest Algorithmus in R vorhergesagt werden?

3. Auswahl der wahrscheinlichsten Quelle für diese Evidenz

Der Algorithmus soll auf Grundlage eines Datensatzes zu Lungenkrebs aus dem Paper von Ahmad et al. aus 2020 trainiert werden. Dieser enthält Informationen zu Alter, Geschlecht, Luftverschmutzung, Alkoholkonsum, Stauballergie, beruflichen Risiken, genetischem Risiko, chronischen Lungenerkrankungen, ausgewogener Ernährung, Fettleibigkeit, Raucherstatus, Passivraucherstatus, Brustschmerzen, Bluthusten, Müdigkeit, Gewichtsverlust, Kurzatmigkeit, Keuchen, Schluckbeschwerden, Klumpen der Fingernägel, häufigen Erkältungen, trockenem Husten und Schnarchen des Patienten. Zusätzliche Informationsquellen sind Websites offizieller Einrichtungen und wissenschaftliche Paper zu inhaltlichen Fragen und StackOverflow und offizielle Dokumentationen von entsprechenden Packages bzw. Bibliotheken.

4. Erstellung einer Suchstrategie

Der ausgewählte Datensatz wird anfänglich exploriert, gereinigt und schließlich mit den passenden Plots visualisiert. Auf Feature Reduktion wird verzichtet, um keinen Bias bei der Auswahl der Features zu schaffen[4]. Nach dem Aufsplitten der Daten in Trainings- und Testdatenset werden die Daten entsprechend Präprozessiert und in das Machine Learning Model zum Training gespeist. Schließlich wird das Model mit dem Testdatensatz geprüft und dessen Leistungsstärke noch mit Hilfe des Klassifikationsreports und einer Konfusionmatrix beurteilt.

[4] https://stats.stackexchange.com/questions/586153/feature-selection-with-categorical-variables (Stand 18.01.2024, 13:40)

5. Zusammenstellung der Evidenzausbeute

Der Random Forest Algorithmus erreicht eine Genauigkeit von 100% im Training mit 800 Samples, 23 Prädiktoren und 3 Target Klassen: 'Low', 'Medium', 'High'. Für die Validierung des Trainings wurde eine 5-fache Kreuzvalidierung genutzt.

6. Anwendung der Evidenz

In der Applikation der Testdaten erreicht das Random Forest Model eine Genauigkeit von ebenfalls 100%. Dies scheint unwahrscheinlich, allerdings erreichen die Autoren des Papers eine ebenso hohe Genauigkeit, auch was andere Metriken angehen. Möglicherweise ist das Datenset mit seinen wichtigen Features und hauptsächlich ordinalen Daten besonders gut für einen Random Forest Algorithmus geeignet.