[Stroke Prediction / Schlaganfallermittlung Dokumentation]

[Derrick Kateu Fonga L., Laura Schönherr, Marty Stemme]

[Technische Hochschule Brandenburg]

Gruppe 4 Abschlussarbeit, Predictive Analytics and Big Data

Dokumentation vom Big Data Projekt

[0 Vorwort 3](#_Toc76128365)

[1 Datensatz & Parameter 3](#_Toc76128366)

[2 Library 4](#_Toc76128367)

[2.1 Implementation und Data Reinigung 4](#_Toc76128368)

[2.2 Trainings- und Testdaten 4](#_Toc76128369)

[3 K- Means 4](#_Toc76128370)

[4 Predictive 5](#_Toc76128371)

[5 Konfusionsmatrix, Precision und Accuracy 5](#_Toc76128372)

[6 Quellen 6](#_Toc76128373)

[Literaturverzeichnis: 6](#_Toc76128374)

# 0 Vorwort

Ziel der vorliegenden Arbeit ist die Vorhersage eines Herzinfarkts auf Basis von medizinischen und allgemeinen vorherrschenden Daten.

Als Kontext hierzu, lässt sich anführen, dass Herzinfarkte die zweithäufigste Todesursache weltweit sind. Sie sind für ungefähr 11% aller Todesfälle verantwortlich, bestätigt die WHO. [[1]](#footnote-2) Allein in Deutschland gab es im Jahr 2019 knapp 331.000 Tote mit ischämischen Herzkrankheiten und Herzinfarkten, dies entspricht in etwa einem Drittel der knapp 940.000 Todesfälle[[2]](#footnote-3). Diese Dokumentation soll dabei helfen vorherzusagen, wie wahrscheinlich es für eine Person ist an einem Herzinfarkt zu erleiden abhängig von verschiedenen Parametern wie ihr Alter, Geschlecht, Vorerkrankungen und ob sie rauchen oder nicht.

# 1 Datensatz & Parameter

In dem Datensatz werden folgende Parameter untersucht:

1. id: eindeutiger Identifikator
2. gender: “Male” männlich, “Female” weiblich oder “Other” divers
3. age: Alter der Person
4. hypertension: 0, wenn die Person nicht an Bluthochdruck leidet, 1 wenn die Person an Bluthochdruck leidet.
5. heart\_disease: 0, wenn die Person nicht an Herzerkrankungen leidet, 1, wenn die Person an Herzerkrankungen leidet.
6. ever\_married: Gibt an, ob die Person verheiratet ist, “No” oder “Yes”
7. work\_type: Gibt die Art von Arbeit an, welche die Person ausübt. “children” die Person hat Kinder und betreut diese zuhause, “Govt\_jov” die Person arbeitet im Government Sektor,”never\_worked” die Person ist arbeitslos und hatte nie Arbeit, “private” die Person arbeitet im privaten Sektor oder “Self-employed” die Person ist selbstständig.
8. Residence\_type: Gibt an, wie die Person wohnt entweder “Rural” in einer ländlichen Gegend oder “Urban” in einer Stadt oder einer Siedlung.
9. avg\_glucose\_level: Gibt den durchschnittlichen Blutzuckerwert an.
10. bmi: Body Mass Index
11. smoking\_status: “formerly smoked” für ehemalige Raucher, “never smoked” für Nichtraucher und “Unknown” wo dieser Status nicht bekannt ist.

Der ausgesuchte Datensatz beschreibt Merkmale und Umstände bestimmter Personengruppen. Diese bilden das grundlegende Gerüst für die spätere Prognosebildung.

Um den Datensatz zu bearbeiten, wurden alle Merkmale die als Text vorliegen in Zahlen umgewandelt. Als Bearbeitungsprogramm wird Jupyter Notebook verwendet, da sich dort andere Bibliotheken verwenden lassen, welche in Spyder eventuell nicht gegeben sind.

# 2 Library

## 2.1 Implementation und Data Reinigung

Für die Implementation werden die Bibliotheken pandas, pyplot, seablrn, sklearn.cluster, einschließlich k-Means und sklearn.preprocessing einschließlich MinMaxScaler benötigt. Pandas dient zum Hochladen des Datensets, einer csv Datei. Die Ausgabe besteht aus einer Wertetabelle mit dem Inhalt der jeweiligen Daten für die Attribute. Für die Daten Reinigung wird Excel, mit bereits vordefinierten Werten benutzt. Hier ist die Funktion Find and Replace (STR+H) angewandt.

## 2.2 Trainings- und Testdaten

Zu weiteren Verarbeitungszwecken werden die vorliegenden Daten in zwei Datensätze unterteilt. Diese bestehen zu 70% aus Trainingsdaten und zu 30% aus Testdaten. Die Testdaten werden zur Vorhersagung eines möglichen Schlaganfalls genutzt.

# 3 K- Means

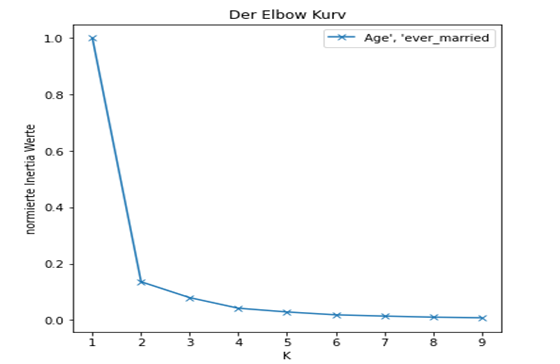
Zunächst sollen dann die besten Attribute für die Clusterbildung gefunden werden. Dazu wird die Itertools Bibliothek genutzt. Die Attribute werden erst in Zweier- und dann in Fünferkombinationen dargestellt. Danach wird die Inertia normiert und berechnet.

Die zunächst besten Vergleiche:

* Gender
* Heart Disease
* Glucose Levels
* Age
* Residence Type
* Bluthochdruck / Hypertension
* Je verheiratet?

Alle Ellenbogenkurven wurden verglichen und die beste ausgewählt. Diese enthält die Attribute ever-married und age.

Die beste Ellebogenkurve mit den Attributen age und ever-married:



Es ist zu sehen, dass der Ellenbogen bei zwei Clustern liegt. Man kann davon ausgehen, dass diese Cluster „Schlaganfall“ und „kein Schlaganfall“ sind. Deswegen ist bei 2 Clustern die WCSS (Inertia), am kleinsten mit der geringsten Anzahl an Clustern und die Homogenität so hoch wie nötig.

Im Anschluss werden alle unnötigen Attribute aus dem Datensatz gelöscht.

Im nächsten Schritt geht es darum Cluster zu finden. Aus der Ellenbogenkurve lies sich ablesen, dass mit zwei Clustern zu rechnen ist.

Die Clusterzentren müssen gefunden werden. Und dann die Daten den Clustern zugeordnet werden. Dies passiert bei K-Means über den euklidischen Abstand. Danach wurden die Cluster geplottet.

# 4 Datenklassifizierung

Anhand der Trainingsdaten wurde die Wahrscheinlichkeit (Train Predictive) eines Schlaganfalls vorhergesagt. Die Ausgabe gibt hierbei nur aus, ob ein Schlaganfall auftreten wird (1) oder ob der Patient Schlaganfallfrei bleibt (Trainpredicted = 0). Dies dient dazu, um das erstellte Modell zu überprüfen.

Dem Output wird nun das Ereignis, ob ein Schlaganfall aufgetreten ist, gegenübergestellt. Es kann festgestellt werden, dass die Übereinstimmung, zwischen dem Vorausgesagtem und dem wahrhaft eingetretenen Ereignis, sehr hoch ist.

Mit den Testdaten wird anschließend die Test Predictive ermittelt und zu einem neuen Datenframe, df, hinzugefügt.

# 5 Konfusionsmatrix, Precision und Accuracy

Die Testdaten werden zu einem neuen Data Frame mit einer Konfusionsmatrix zusammengefasst. Beim Datensatz allgemein ist zu beachten, dass nur die ersten 250 Dateneinträge von den insgesamt rund 5000 Einträgen einen positiven Wert bei der Schlaganfallermittlung haben. Die Testdaten wurden zufällig ausgewählt. Der Testdatensatz enthält 1870 Datenwerte, die kein Schlaganfall sind, und 156 mit Schlaganfall. Diese Ungleichverteilung kann zu den sehr hohen positiven Werten bei der Accuracy und Precision führen. Das aus der Konfusionsmatrix erhaltene Array wird grafisch dargestellt, mit den Testdaten auf der Abszisse und den Trainingsdaten auf der Ordinate. Um die Matrix übersichtlicher darzustellen, wurde sie mit Excel dargestellt.

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

trueYes: kein Schlaganfall

predYes: keine Schlaganfälle, die vorausgesagt wurden, sind auch keine Schlaganfälle

predNo: keine Schlaganfälle, die nicht als „keine Schlaganfälle“ erkannt wurden

trueNo: Schlaganfall

predYes: Schlaganfälle, die als „keine Schlaganfälle“ erkannt wurden, sind Schlaganfälle

predNo: Schlaganfälle, die als Schlaganfälle eingestuft wurden, sind auch wirklich Schlaganfälle

recall Berechnung

TrueYes: 1 à 100%

TrueNo: 0,596 à 59,6%

Das Modell hat in diesem Fall alle „nicht Schlaganfall“ – Ereignisse richtig zugeordnet. Bei den Schlaganfällen jedoch gab es Probleme und in nur 59,6% der Fälle wurde der Schlaganfall richtig erkannt.

Precision ist das Verhältnis der korrekt identifizierten positiven Fälle zu allen vorhergesagten positiven Fällen, d. h. den richtig und den falsch als positiv vorhergesagten Fällen. Präzision ist der Anteil der gefundenen Dokumente, die für die Anfrage relevant sind. Mit der Bibliothek “Numpy” lässt sich gut die Precision berechnen. Die Werte der Diagonale werden mit der Summe aller Werte der Konfusionsmatrix auf der Axe null verrechnet.

Accuracy

Die Accurracy stellt die Fähigkeit des Modells dar, sowohl die positiven als auch die negativen Ergebnisse aller Vorhersagen korrekt vorherzusagen. Mathematisch gesehen stellt sie das Verhältnis der Summe der wahren Positiven und der wahren Negativen Werte aus allen Vorhersagen dar.

Accuracy ist hoch für „keine Schlaganfall“ Erkennung

Aber im Umkehrschluss sind bei der Erkennung von Schlaganfällen nur circa 60% erkannt wurden.

Bei den Testdaten sind 156 Personen mit Schlaganfall vorhanden, daraus lässt sich sehen…. Circa 100 Personen im Trainingsdatensatz mit Schlaganfall à schlechtes Training

# 6 Fazit

Besseres Verhältnis der Schlaganfälle im Datensatz

Nicht so toller Datensatz

# 7 Quellen

## Literaturverzeichnis:

* (1) Stroke Prediction Dataset, fedesoriano, <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset> 20.05.2021 19:24
* (2) Häufigste Todesursache in Deutschland, Rainer Radtke, https://de.statista.com/themen/ 69/todesursachen/
* Stack-Overflow

1. Stroke Prediction Dataset, fedesoriano, <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset> 20.05.2021 19:24 [↑](#footnote-ref-2)
2. Häufigste Todesursache in Deutschland, Rainer Radtke, https://de.statista.com/themen/ 69/todesursachen/ [↑](#footnote-ref-3)