Semesterarbeit Predictive Analytics

Autoren	Apicella Nevio Balke Nicolas Fassbind Andrin Pletscher Steven von Arx Matthias
Klasse	IT21b
Dozent	Lorenzo Tanadini
Erstellungsdatum	5. Oktober 2023

Apicella Nevio / Pletscher Steven
Balke Nicolas / von Arx Matthias
Fassbind Andrin

Zürcher Hochschule
für Angewandte Wissenschaften

Inhaltsverzeichnis

Apicella Nevio / Pletscher Steven
Balke Nicolas / von Arx Matthias
Fassbind Andrin

1 Ausgangslage

Diabetes hat schwere Langzeitfolgen. Diese können eine Schädigung der Augennetzhaut, eine allgemeine Nervenschädigung, Herz-Kreislauf-Erkrankungen wie ein erhöhtes Schlaganfallrisiko, allgemeine Gefässschäden oder den "diabetischen Fuss" umfassen. Beim diabetischen Fuss führen Gefässschäden und eine reduzierte Reizleitung der Nerven oft zu grossen Fusswunden.

Diese Probleme können jedoch durch zwei Möglichkeiten verhindert werden. Diese werden in den folgenden Unterkapitel genauer behandelt.

1.1 Ziele dieser Arbeit

In den folgenden zwei Abschnitten werden die Hauptziele dieser Semesterarbeit beschrieben.

1.1.1 Vorbeugung der Krankheit

Diabetes Typ 1 kann mit den heutigen medizinischen Möglichkeiten noch nicht verhindert werden. Doch das Risiko für das Auftreten von Diabetes Mellitus Typ 2 kann durch eine gesunde Lebensweise deutlich reduziert werden.

1.1.2 Früherkennung und Behandlung

Diabetes kann gut durch regelmässige Blutzuckermessung und Behandlung mit oralen Antidiabetika und Insulin behandelt werden. Für die Vorbeugung von Diabetes ist es wichtig, zu wissen, welche Lebensbedingungen das Entstehen dieser Krankheit begünstigen. Hierfür können Daten analysiert werden, um festzustellen, welche Einflüsse bestimmte Faktoren wie z.B. der BMI auf die Erkrankung haben. Auch bei der Früherkennung spielt die Analyse von Prädiktoren, wie zum Beispiel des HbA1c-Werts, eine wichtige Rolle.

1.2 Hypothese

Es wird angenommen, dass der HbA1c-Wert den größten Einfluss auf das Diabetesrisiko hat. Es wird erwartet, dass höhere HbA1c-Werte mit einem proportional höheren Risiko für die Entwicklung von Diabetes korrelieren.

2 Daten

Der Diabetes-Prädiktionsdatensatz ist eine Sammlung medizinischer und demographischer Daten von Patienten sowie deren Diabetesstatus (positiv oder negativ). Er umfasst 100'000 Beobachtungen. Die Daten enthalten Merkmale wie Alter, Geschlecht, Body Mass Index (BMI), Bluthochdruck, Herzerkrankungen, Rauchverhalten, HbA1c und Blutzucker. Dieser Datensatz kann verwendet werden, um maschinelle Lernmodelle zur Vorhersage von Diabetes bei Patienten auf der Grundlage ihrer Krankengeschichte und demographischer Informationen zu erstellen. Dies kann für Angehörige der Gesundheitsberufe nützlich sein, um Patienten mit einem Diabetesrisiko zu identifizieren und personalisierte Behandlungspläne zu entwickeln. Darüber hinaus kann der Datensatz von Forschern genutzt werden, um den Zusammenhang zwischen verschiedenen medizinischen und demografischen Faktoren und der Wahrscheinlichkeit, an Diabetes zu erkranken, zu untersuchen[1].

2.1 Beschreibung der Attribute

#	Attribut	Beschreibung
1	Geschlecht	Das Geschlecht bezieht sich auf das biologische Geschlecht einer Person, das sich auf die Anfälligkeit für Diabetes auswirken kann.
2	Alter	Das Alter ist ein wichtiger Faktor, da Diabetes häufiger bei älteren Erwachsenen diagnostiziert wird, wobei die Altersspanne in unserem Datensatz von 0-80 Jahren reicht.
3	Bluthochdruck	Bluthochdruck ist ein medizinischer Zustand, bei dem der Blutdruck in den Arterien dauerhaft erhöht ist. Er hat die Werte 0 oder 1, wobei 0 bedeutet, dass kein Bluthochdruck vorliegt und 1, dass Bluthochdruck vorliegt.
4	Herzerkrankung	Herzkrankheiten sind eine weitere Krankheit, die mit einem erhöhten Risiko für die Entwicklung von Diabetes verbunden ist. Sie hat die Werte 0 oder 1, wobei 0 bedeutet, dass keine Herzerkrankung vorliegt und 1, dass eine Herzerkrankung vorliegt.
5	Rauchverhalten	Rauchen gilt ebenfalls als Risikofaktor für Diabetes und kann die mit Diabetes verbundenen Komplikationen verschlimmern. In unserem Datensatz haben wir 5 Kategorien, nämlich nicht aktuell, früher, keine Info, aktuell, nie und jemals.
6	Body Mass Index (BMI)	Der BMI (Body-Mass-Index) ist ein Mass für das Körperfett auf der Grundlage von Gewicht und Grösse. Höhere BMI-Werte werden mit einem höheren Diabetesrisiko in Verbindung gebracht. Die Bandbreite des BMI im Datensatz reicht von 10,16 bis 71,55. Ein BMI von weniger als 18,5 gilt als untergewichtig, 18,5-24,9 als normal, 25-29,9 als übergewichtig und 30 oder mehr als fettleibig.
7	HbA1c	Glykosyliertes Hämoglobin (Langzeitzucker) widerspiegelt den durchschnittlichen Blutzuckerwert über die letzten zwei bis drei Monate. Höhere Werte weisen auf ein grösseres Risiko hin, an Diabetes zu erkranken. Meist deutet ein HbA1c-Wert von mehr als 6,5 % auf Diabetes hin.
8	Blutzucker	Der Blutzuckerspiegel ist die Menge an Glukose, die sich zu einem bestimmten Zeitpunkt in der Blutbahn befindet. Ein hoher Blutzuckerspiegel ist ein wichtiger Indikator für Diabetes.
9	Diabetes	Diabetes ist die Zielvariable. 1 = ja, 0 = Nein

Tabelle 1 - Beschreibung der Attribute (https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data)

[1] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

3 Literatursuche

Im folgenden Abschnitt befinden sich drei Studien, welche für das Thema Diabetes relevant sind.

3.1 Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients

Hb1Ac ist die vorhergesagte Halbwertszeit von roten Blutzellen. Die umfassenden Informationen, die ein einzelner HbA1c-Test liefert, machen ihn zu einem zuverlässigen Biomarker für die Diagnose und Prognose von Diabetes.

Wir vermuten, dass der HbA1c-Wert ein guter Prädikator für Diabetes ist, und wollen dieser Vermutung auf den Grund gehen. [2]

3.2 Obesity and diabetes

Adipositas, insbesondere Stammfettsucht, steht in engem Zusammenhang mit der Prävalenz von Diabetes und Herz-Kreislauf-Erkrankungen. Die Plasmakonzentrationen von Leptin, Tumornekrosefaktor-α und nicht veresterten Fettsäuren sind bei Adipositas erhöht und spielen eine Rolle bei der Entstehung von Insulinresistenz. Die glykämische Kontrolle des Diabetes und die Insulinresistenz verbessern sich mit der Verringerung der Adipositas, aber die Behandlung der Adipositas ist schwierig, und eine nachhaltige Gewichtsreduzierung ist mit einer Diät allein kaum zu erreichen.

Diese Studie weist auf eine klare Korrelation zwischen Adipositas und Diabetes hin. Wir erwarten daher eine hohe Signifikanz dieses Prädikators für die Vorhersage von Diabetes.[3]

3.3 Cigarette smoking and Diabetes

Diese Studie ist ausfolgenden Gründen für unsere Arbeit relevant:

Die Studie liefert wissenschaftliche Belege dafür, dass Rauchen das Diabetesrisiko erhöht. Sie unterstreicht, dass Raucher insulinresistent sind und ein erhöhtes Risiko für Typ-2-Diabetes haben. Diese Informationen sind wichtig, um die Bedeutung der Vermeidung von Risikofaktoren wie Rauchen bei der Prävention von Diabetes zu unterstreichen.

Auf Grund der bekannten Korrelationen zwischen Gesundheitsschäden und Rauchen erwarten wir hier ebenfalls einen hohen Zusammenhang zu Diabetes.[4]

[2] https://journals.sagepub.com/doi/full/10.4137/BMI.S38440
[3] https://www.sciencedirect.com/science/article/abs/pii/S1521690X99900179
[4] https://www.sciencedirect.com/science/article/abs/pii/S0033062003000112?via%3Dihub

4 Arbeitsplan

4.1 Schritt 1: Datenvorbereitung und -exploration

- Datensatzbeschaffung: Wir werden das ausgewählte Diabetes-Datenset herunterladen und verstehen.
- Datenbereinigung: String-Daten werden sinnvoll durch Numerische Werte ersetzt. (HotOne-Encoding)
- Explorative Datenanalyse (EDA): Wir werden statistische Analysen und Visualisierungen durchführen, um ein besseres Verständnis für die Daten zu entwickeln.
- F-Tests für individuelle Daten.

4.2 Schritt 2: Modellentwicklung

- Feature Selection: Wir werden relevante Merkmale auswählen, die für die Vorhersage von Diabetes relevant sind. (Falls F-Tests nicht bestanden)
- Modellauswahl: Wir werden verschiedene Klassifikationsalgorithmen wie logistische Regression, Random Forest und neuronale Netze evaluieren, um das am besten geeignetes Modell für unsere Daten zu finden.
- Modelltraining und -validierung: Wir werden das ausgewählte Modell auf unseren Daten trainieren und die Leistung anhand von Metriken wie Genauigkeit, Präzision und F1-Score bewerten.

4.3 Schritt 3: Entwicklung des Data Products

- Design der Benutzeroberfläche: Wir werden eine interaktive Benutzeroberfläche (Shiny App) entwickeln, die es Benutzern ermöglicht, ihre persönlichen Daten einzugeben und eine Vorhersage über ihr Diabetesrisiko zu erhalten.
- Implementierung: Wir werden die entwickelten Modelle in die Benutzeroberfläche integrieren, um Echtzeitvorhersagen zu ermöglichen.
- Testing und Fehlerbehebung: Wir werden das Data Product umfassend testen, um sicherzustellen, dass es zuverlässige Vorhersagen liefert. Etwaige Fehler werden behoben.

4.4 Schritt 4: Dokumentation

- Dokumentation: Wir werden alle Schritte, Entscheidungen und Ergebnisse in einem ausführlichen Bericht festhalten, der den gesamten Prozess der Semesterarbeit dokumentiert.

Apicella Nevio / Pletscher Steven
Balke Nicolas / von Arx Matthias
Fassbind Andrin

5 Datenbereinigung

Zu Beginn beinhalteten die Daten bei den Prädiktoren «sex» und «smoking_history» strings, also «female/male» respektive «never/former/current». Sowohl “sex” sowie “smoking_history” wurde one-hot-encoded. Dies sehen wir als sinnvoll an da “sex” binär ist. Auch der Prädiktor „smoking_history“ wurde one-hot-encoded, hier wurde aufsteigend von „never“ bis „current“ nummeriert, um die intensität zu widerspiegeln. One-hot-encoding durch arbiträre Zahlen kann zu Problemen führen, jedoch muss für eine logistische Regression jeder Prädiktor durch Zahlen repräsentiert werden.

```
> str(diabetes_prediction_dataset)
'data.frame':   100000 obs. of  9 variables:
 $ gender             : chr  "Female" "Female" "Male" "Female" ...
 $ age                : num  80 54 28 36 76 20 44 79 42 32 ...
 $ hypertension       : int  0 0 0 0 1 0 0 0 0 0 ...
 $ heart_disease      : int  1 0 0 0 1 0 0 0 0 0 ...
 $ smoking_history    : chr  "never" "No Info" "never" "current" ...
 $ bmi                : num  25.2 27.3 27.3 23.4 20.1 ...
 $ HbA1c_level        : num  6.6 6.6 5.7 5 4.8 6.6 6.5 5.7 4.8 5 ...
 $ blood_glucose_level: int  140 80 158 155 155 85 200 85 145 100 ...
 $ diabetes           : int  0 0 0 0 0 0 1 0 0 0 ...
```

Abbildung 1 - Die Daten vor der Bereinigung

```
'data.frame':   100000 obs. of  9 variables:
 $ gender             : num  0 0 1 0 1 0 0 0 1 0 ...
 $ age                : num  80 54 28 36 76 20 44 79 42 32 ...
 $ hypertesnion       : int  0 0 0 0 1 0 0 0 0 0 ...
 $ heart_diseas       : int  1 0 0 0 1 0 0 0 0 0 ...
 $ smoking_history    : int  0 1 0 3 3 0 0 1 0 0 ...
 $ bmi                : num  25.2 27.3 27.3 23.4 20.1 ...
 $ HbA1c              : num  6.6 6.6 5.7 5 4.8 6.6 6.5 5.7 4.8 5 ...
 $ blood_glucose_level: int  140 80 158 155 155 85 200 85 145 100 ...
 $ diabetes           : int  0 0 0 0 0 0 1 0 0 0 ...
```

Abbildung 2 - Die Daten nach der Bereinigung

Im Vergleich der Abbildungen 1,2 sieht man, dass bei «smoking_history» der Prädiktor «No Info» zu 1 encodiert wird. Dies wurde beschlossen da «No Info» nach unserer Korrelationsmatrix, siehe Kapitel 6.4, keine Korrelation besitzt und somit höchst wahrscheinlich aus einer Mischung der anderen Auswertungen besteht. Eine Eins ist ein durchschnitts Wert, welcher dies repräsentieren kann.

6 Datenanalyse

Im nachfolgenden Kapitel werden verschiedene Diagramme und deren Interpretation aufgezeigt.

6.1 Balkendiagramm

Das Balkendiagramm zur Geschlechtsverteilung zeigt eine leichte Unausgewogenheit, wobei eine höhere Anzahl von Frauen im Vergleich zu Männern im Datensatz vorhanden ist. In Bezug auf den Diabetesstatus sind ungefähr 90% der Personen nicht-diabetisch, was diese Gruppe zur überwiegenden Kategorie macht. Die Analyse der Geschlechtsverteilung ist wichtig, da bestimmte Gesundheitszustände Geschlechter unterschiedlich beeinflussen können, was zu potenziellen Variationen in der Diabetesprävalenz zwischen Männern und Frauen führen könnte.

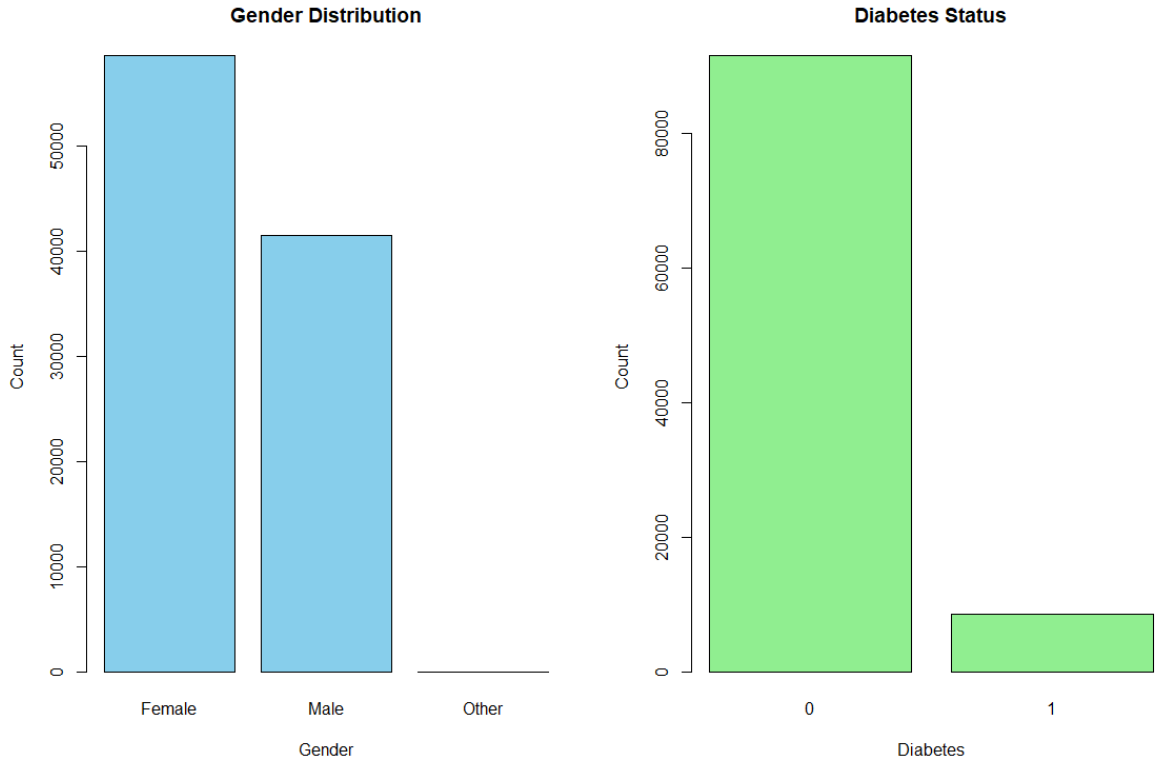

Abbildung 3 - Balkendiagramm Geschlechtsverteilung und Diabetes Status

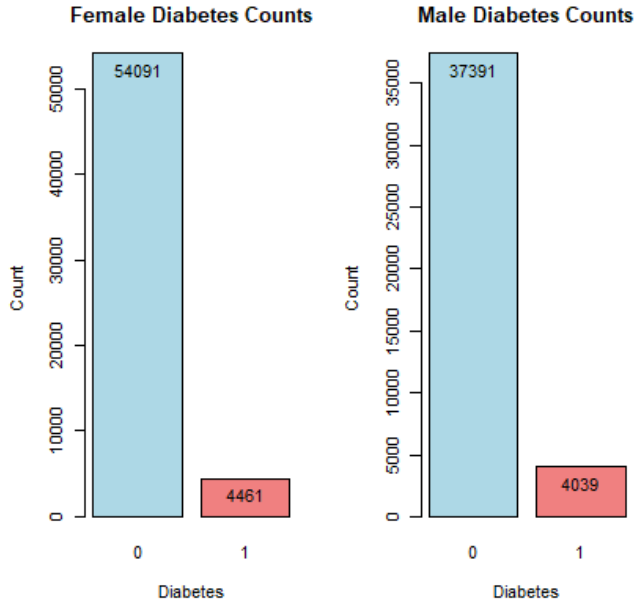

Abbildung 4 - Diabetesstatus nach Geschlecht

Entgegen den Erwartungen, dass im Datensatz deutlich mehr Frauen als Männer bzw. Andere vertreten sind und die Zielvariabel ebenfalls sehr ungleich verteilt ist, ist der Anzahl Personen mit Diabetes gruppiert nach den Geschlechtern in etwa im gleichen Verhältnis. Eine falsche Korrelation zwischen dem Geschlecht und Diabetes sollte somit nicht geschehen.

Das Verhältnis von Diabetes Patienten gruppiert nach dem Prädikator Herzprobleme ist nicht gleichmässig verteilt. Es könnte eine leichte positive Korrelation zwischen Diabetes und Herzproblemen identifiziert werden. Dasselbe gilt für Bluthochdruck (Hypertension).

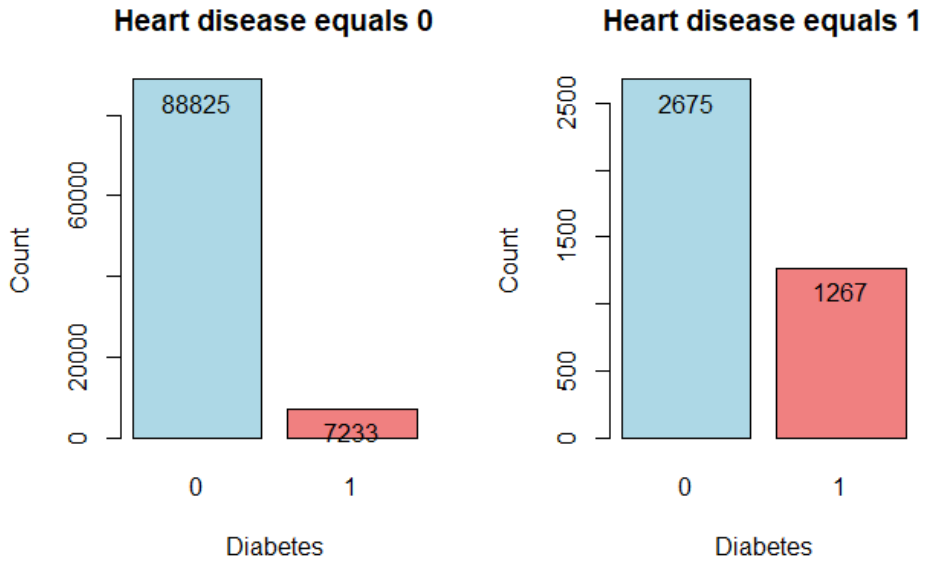

Abbildung 5 - Diabetesstatus nach Herzproblemen

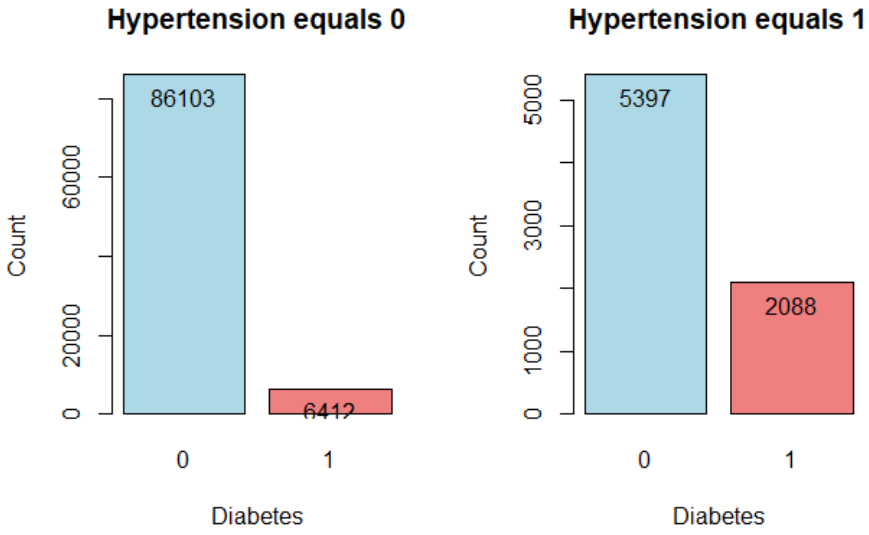

Abbildung 6 - Diabetesstatus nach Bluthochdruck

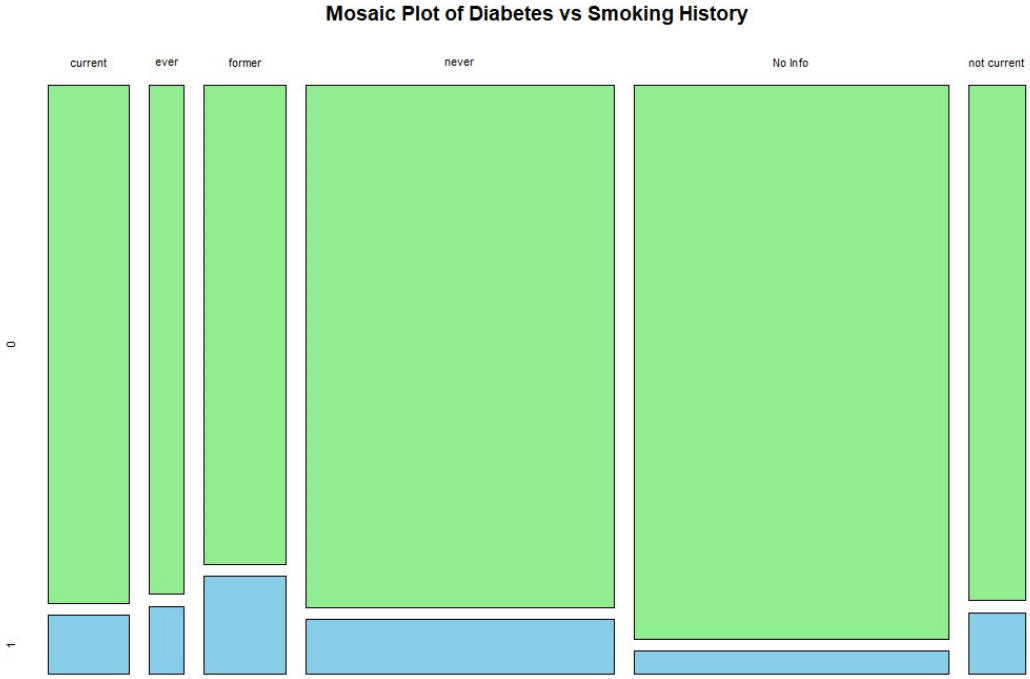

Abbildung 7 - Diabetesstatus nach Raucherhistorie

Die Daten zeigen wie erwartet keine Relation zwischen den unterschiedlichen Raucherkategorien und dem Diabetesstatus.

6.2 Boxplot

Der Boxplot zur Gegenüberstellung des Alters nach Diabetesstatuskategorien offenbart einen klaren Trend: Ältere Personen neigen dazu, eine höhere Diabetesinzidenz aufzuweisen, während mehrere Ausreisser auf junge Personen mit Diabetes hindeuten. Diese Erkenntnis unterstreicht die massgebliche Rolle des Alters als signifikanten Faktor in der Diabetesvorhersage. Die Identifikation von Ausreissern in den jüngeren Altersgruppen ist besonders wichtig, da sie oft auf Diabetes Typ 1 hinweisen, eine Form von Diabetes, die in jungen Jahren häufiger auftritt. Diese Erkenntnis ist von entscheidender Bedeutung für Früherkennung und Präventionsstrategien, insbesondere im Zusammenhang mit früh einsetzendem Diabetes Typ 1.

Im Boxplot, der die Blutzuckerspiegel nach Diabetesstatus darstellt, zeigen sich höhere Werte bei Diabetespatienten, was den Erwartungen entspricht. Diese Beobachtung betont die klinische Relevanz von Blutzuckerspiegeln bei der Diabetesdiagnose. Eine kontinuierliche Überwachung des Blutzuckerspiegels ist für Diabetespatienten entscheidend, um ihren Zustand effektiv zu verwalten. Diese Erkenntnisse aus dem Boxplot sind von grossem Wert für Gesundheitsfachleute und Forscher, die sich mit Diabetesstudien befassen.

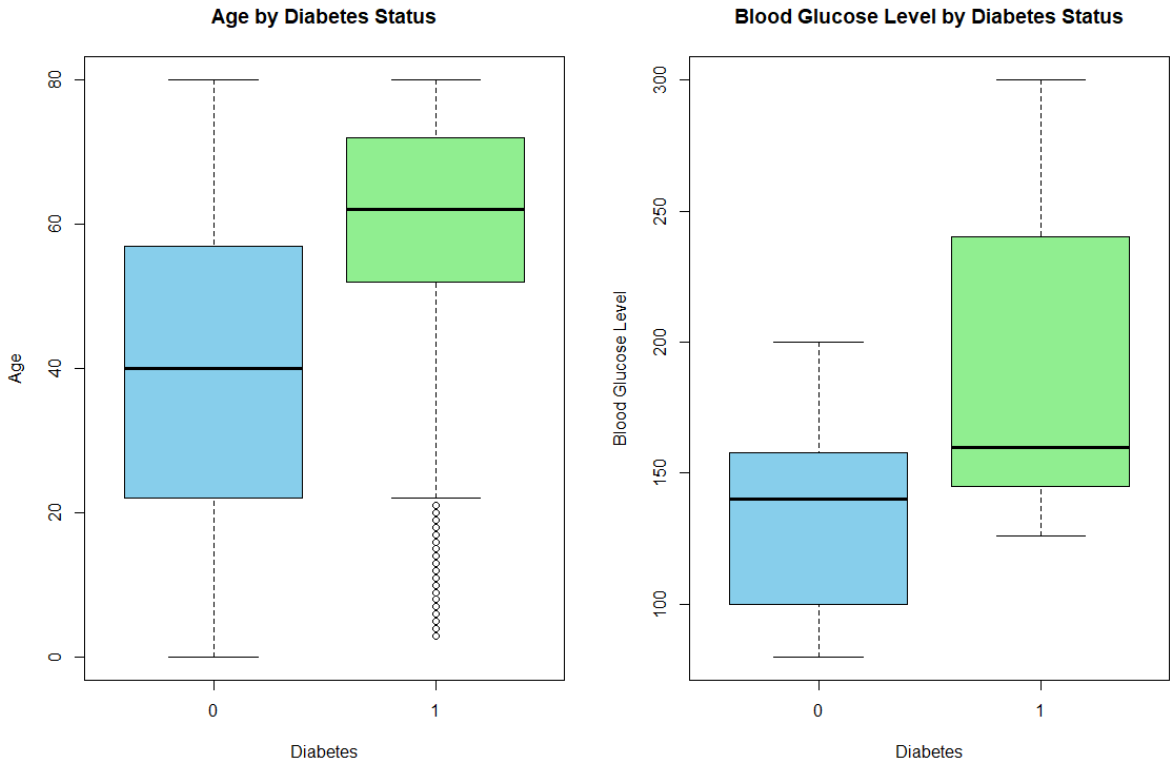

Abbildung 8 - Boxplot Gegenüberstellung des Alters und Blutzuckerspiegel nach Diabetesstatus

6.3 Histogramm

Das Histogramm zur Altersverteilung zeigt eine relativ standardmässige Normalverteilung, was darauf hindeutet, dass die meisten Personen im Datensatz im durchschnittlichen Altersbereich liegen. Es gibt jedoch Ausreisser an beiden Enden des Altersspektrums, was auf das Vorhandensein sehr junger und sehr alter Personen hinweist. Diese Erkenntnis ist wichtig, da das Alter das Diabetesrisiko erheblich beeinflussen kann; ältere Personen haben im Allgemeinen ein höheres Risiko.

Das Histogramm zur BMI-Verteilung zeigt, dass die meisten Personen im Bereich von 25 bis 35 liegen, was auf ein häufiges Vorkommen von Übergewicht bis mässige Fettleibigkeit bedeutet.

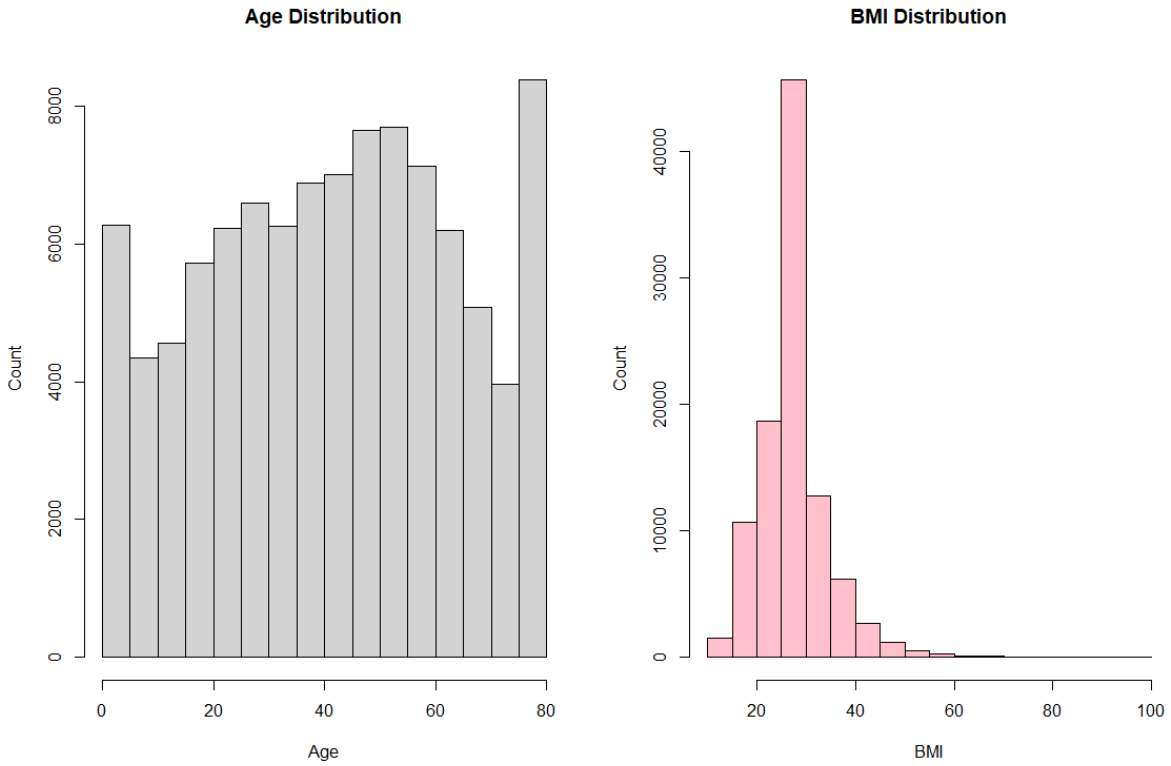

Abbildung 9 - Histogramm zur Alters- und BMI Verteilung

Das Histogramm zu Herzproblemen und Raucher Anteil zeigt ebenfalls eine ungleiche Verteilung der Klassen in den Daten. Im Datensatz sind signifikant mehr Personen mit Herzproblemen.

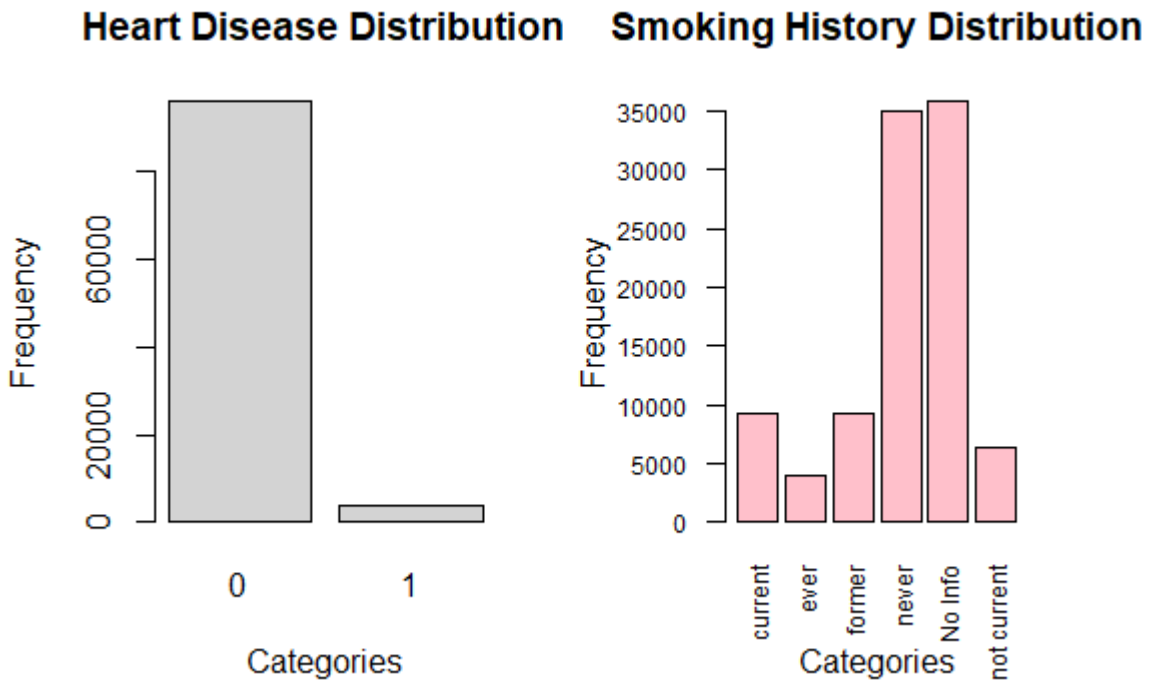

Abbildung 10 - Histogramm Heart Diesease & Smoking History

Bei der HbA1c zeigt sich eine mittelinks Verteilung der Daten mit hoher Varianz in der Zählung.

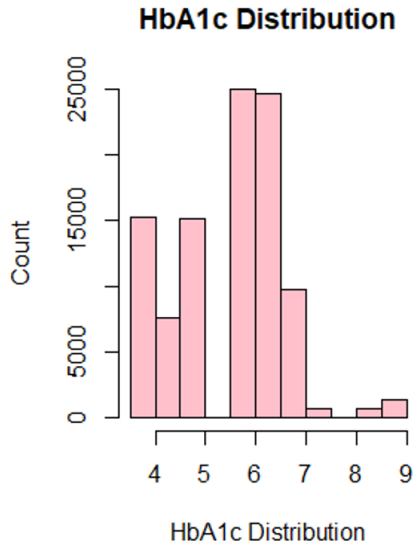

Abbildung 11 - Histogramm HbA1c

6.4 Korrelations-Matrix

Die Prädiktoren Blutzucker (blood_glucose), HbA1c und Alter weisen die höchsten positiven Korrelationen auf. BMI und Bluthochdruck folgen. Es gilt diese daher genauer zu betrachten.

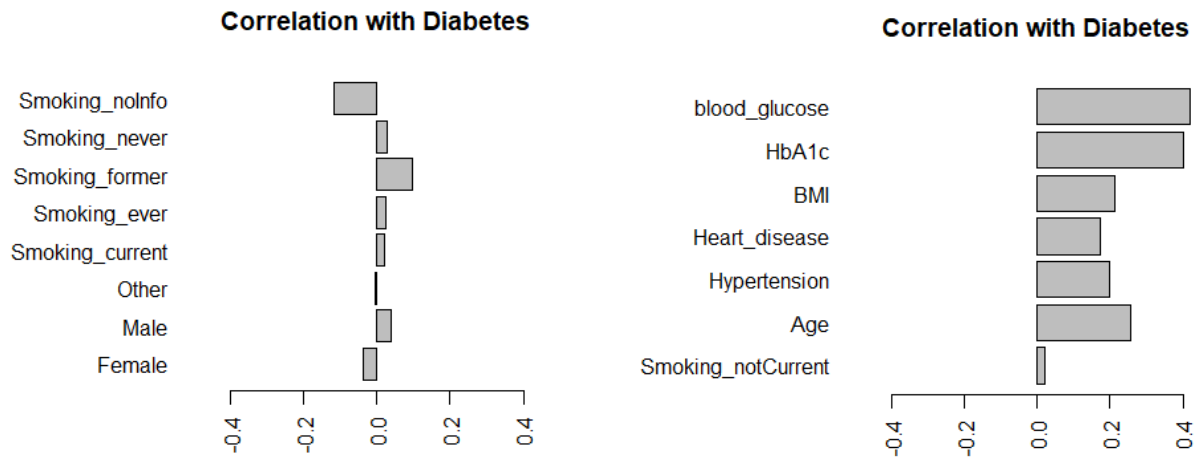

Abbildung 12 - der einzelnen smoking Faktoren

Die Korrelationsmatrix analysiert die Beziehungen zwischen Alter, BMI, HbA1c-Level und Blutzuckerspiegel. Interessanterweise zeigen sich über alle Variablen hinweg nahezu keine Korrelationen, sondern Werte, die nahe null liegen. Diese Ergebnisse deuten auf äusserst schwache oder gar keine linearen Beziehungen zwischen den untersuchten Variablen hin. Insbesondere fällt eine moderate positive Korrelation von 0,33 zwischen Alter und BMI für nicht-diabetische Personen auf. Dies könnte darauf hindeuten, dass ältere Personen in der nicht-diabetischen Gruppe tendenziell leicht höhere BMIs aufweisen, obwohl die Stärke dieser Beziehung als schwach betrachtet werden kann. Das Verständnis dieser Korrelationen, die nahe null liegen, ist von Bedeutung, da sie darauf hinweisen, dass keine signifikanten linearen Zusammenhänge zwischen den analysierten Variablen bestehen.

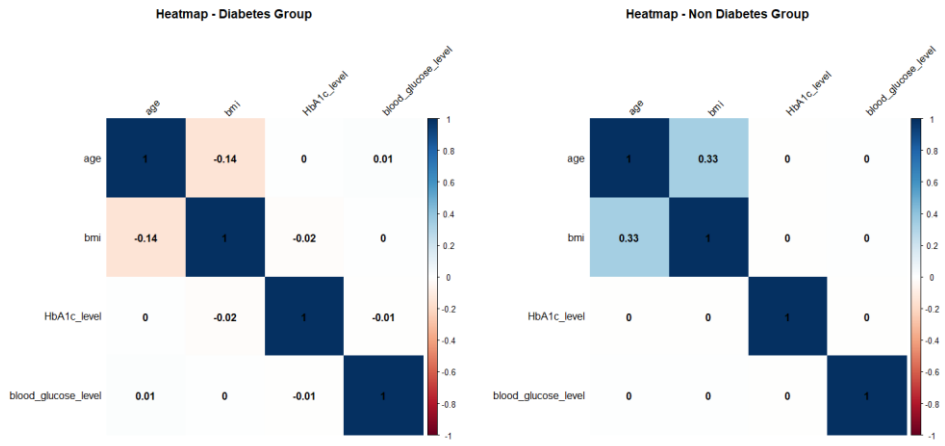

Abbildung 13 - Korrelationsmatrix Diabetes- und Nicht Diabetes Gruppe

7 Modellwahl - Vorhersagemethoden

Für künftige Voraussagen haben wir uns für drei Modelle entschieden. Den Entscheidungsbaum, die logistische Regression und K-nearest-neighbours. In den folgenden Abschnitten wird detailliert auf die einzelnen Modelle eingegangen. Ausserdem wird die Problemstellung, auf welche sich unsere Modelle beziehen, erläutert.

7.1 Problemstellung

Das Ziel ist es, verschiedene Modelle zu bilden, welche anhand unserer Prädiktoren (gender, age hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level) möglichst akkurate Voraussagen machen können ob eine Person an Diabetes erkrankt ist oder nicht. Um die Modelle zu testen, wurden die Daten jeweils in Trainings- und Testdaten unterteilt. Die Modelle wurden anhand der Trainingsdaten erstellt und später mit den Testdaten evaluiert.

7.2 Entscheidungsbaum

Entscheidungsbäume eignen sich für Klassifikationsprobleme. Ein Vorteil von Klassifikationsbäumen ist, dass sie gut interpretierbar sind. Für einen Klassifikationsbaum ist ausserdem keine Dummy-Kodierung nötig. Ein weiterer Vorteil von einem Klassifikationsbaum ist, dass sie robust gegenüber Variablentransformationen sind und man sich daher diesen Schritt sparen kann. Da die Vorhersage in unserem Fall auf eine Klassifikation von Diabetes oder «Ja» oder «Nein» fällt eignet sich ein Entscheidungsbaum sehr gut.

7.2.1 Initialbaum

Da ein Entscheidungsbaum nicht mehr gut interpretierbar ist, wenn er eine hohe Anzahl Aufteilungen aufweist, wurden bei der Erstellung die Parameter minsplit = 300, minbucket = 100, maxdepth = 8 mitgegeben. Diese bilden zusammen die Stopregel. Minsplit gibt dabei an wie viele Beobachtungen mindestens für einen Knoten gemacht werden müssen, dass bei diesem Knoten gesplittet werden kann. Minbucket gibt einen Mindestwert an Beobachtungen pro Knoten vor und Maxdepth die grösste Tiefe des Baumes an. Mit diesen Werten wurde der Baum T0 (Abbildung 14) ist erstellt.

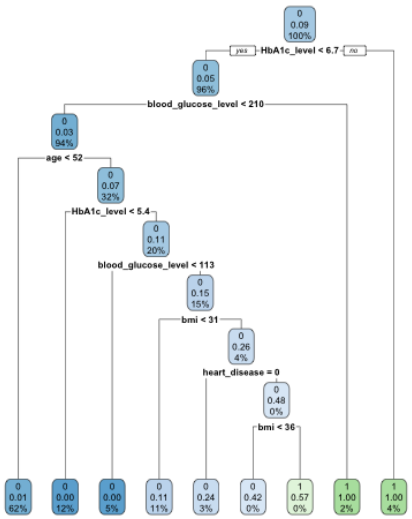

Abbildung 14 - Baum T0 (ohne Pruning)

7.2.2 Pruning

Der Initialbaum ist mit seinen 9 Endknoten noch deutlich zu gross und somit nur schwer interpretierbar. Aufgrund dessen wurde Pruning eingesetzt. Dafür wurden zuerst die CP-Werte der möglichen Bäume berechnet. Im CP-Plot (Abbildung 16) kann deutlich gesehen werden, dass der relative Fehler bei Bäumen bis zu drei Endknoten stetig abnimmt. Bei Bäumen mit mehr als drei Endknoten kann kaum noch eine Verbesserung festgestellt werden. Der Baum T0 wird daher mit dem CP-Wert 0.0059 geprunt. Dabei kommt der Baum T heraus. Dieser Baum hat nur noch 2 Splits und somit 3 Endknoten. Der Finale Baum ist somit sehr einfach interpretierbar.

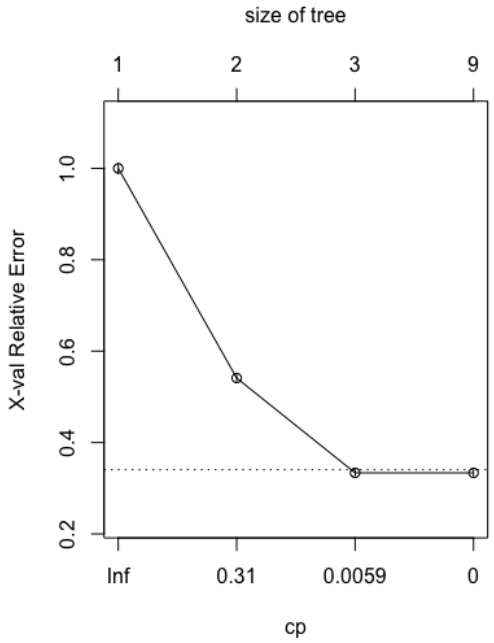

Abbildung 16 - CP-Plot

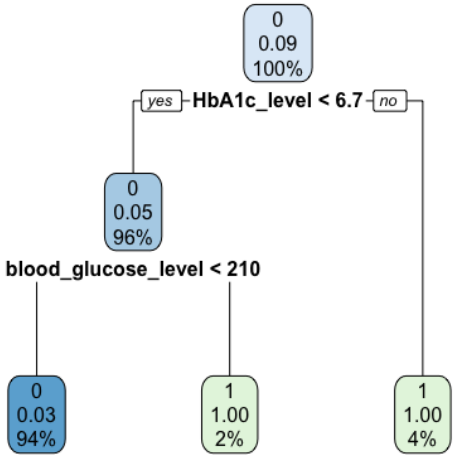

Abbildung 15 - Baum T (geprunt)

7.2.3 Evaluation

Um die Perfomance dieses Baumes zu evaluieren, wurde mit dem Baum eine Voraussage für die Testdaten gemacht und mit den Beobachtungen der Testdaten verglichen.

Die Werte können in einer Konfusionsmatrix zusammengefasst werden. Der Wert Null bedeutet hier, dass die Person gesund bzw. nicht an Diabetes erkrankt ist. Die Fragestellung ist hierbei: «ist die untersuchte Person nicht an Diabetes erkrankt?». Da diese Verneinung eventuell zu Verwirrungen führen kann, wurde die Konfusionsmatrix mit den Interpretationen ergänzt.

Konfusionsmatrix		Beobachtungen	
		1	0
Voraussage	1	TP: 18377 => 0.92% Person hat kein Diabetes und wurde als solches erkannt	FP: 519 => 0.025 % Person ist krank, wurde jedoch fälschlicherweise als gesund eingestuft.
	0	FN: 0 => 0% Person ist gesund, wurde jedoch fälschlicherweise als krank eingestuft.	TN: 1104 => 0.05% Person ist an Diabetes erkrankt und wurde als solche erkannt.

Tabelle 2 - Konfusionsmatrix Entscheidungsbaum

Wie man in Tabelle 2 gut erkennen kann, werden die meisten Personen richtig klassifiziert. Die Genauigkeit (Accuracy) vom Klassifikationsbaum über den Testdaten ist 0.974. Jedoch liegt die No-Information-rate bei 0.9188 (dies wäre die Genauigkeit, wenn einfach jeweils der häufigere Wert vorausgesagt wird. Dies ist bei unseren Daten jedoch nicht überraschend da viel mehr gesunde als erkrankte Personen in den Daten enthalten sind. Die Spezifität für das Erkennen von gesunden Personen ist bei 1.0 sehr hoch. Dies kommt daher, dass keine einzige Person fälschlicherweise als krank eingestuft wurde. Die Sensitivität (Recall) liegt bei 0.68.

7.3 Logistische Regression

Das Logistische Model eignet sich ebenfalls als binäre Klassifikation. Hierbei wurden folgende Modelle generiert.

Das volle Model beinhaltet sämtliche Prädiktoren. Mit der Vorwärts und der Rückwärtselimination wurde dasselbe (volle) Model generiert. Es wurden daher zwei weitere Modelle manuell generiert:

Manuelles Model 1: diabetes ~ age + bmi + HbA1c_level + blood_glucose_level

Dieses enthält die 4 Prädiktoren mit den höchsten Korrelationkoeffizienten.

Manuelles Modell 2: diabetes ~ HbA1c_level + blood_glucose_level

Enthält die 2 Prädiktoren mit den höchsten Korrelationskoeffizienten

Folgende Performance wurde auf einem Testset, welches 20% des vollständigen Datensets entspricht erzielt:

7.3.1 Volles bzw. Vorwerts und Rückwerts selektiertes Modell

Predicted \ Actual	0	1
0	18103	623
1	176	1098

Abbildung 17 - Konfusionsmatrix, volles Modell

AIC	18049
Rel. Klassifikationsfehler	3.995%
Accuracy	96.005%
Präzision	86.185%
Recall	63.8%
F1-Score	73.322%

Tabelle 3 - Kennwerte Volles Modell

7.3.2 Manuelles Modell 1

Predicted \ Actual	0	1
0	18127	647
1	152	1074

Abbildung 18 - Konfusionsmatrix, manuelles Modell 1

AIC	18654
Rel. Klassifikationsfehler	3.995%
Accuracy	96.005%
Präzision	87.602%
Recall	62.406%
F1-Score	72.888%

Tabelle 4 - Kennwerte Modell 1

7.3.3 Manuelles Modell 2

Predicted \ Actual	0	1
0	18148	786
1	131	935

Abbildung 19 - Konfusionsmatrix, manuelles Modell 2

AIC	22672
Rel. Klassifikationsfehler	4.585%
Accuracy	95.415%
Präzision	87.711%
Recall	54.329%
F1-Score	67.097%

Tabelle 5 Kennwerte manuelles Modell 2

7.4 K-nearest-neighbours

Als drittes Klassifikationsmodel wurde ein K-nearest-neighbours (KNN) gewählt. Auch KNN eignet sich bei binären Zielvariablen. Besonders spannend für dieses Project war das Betrachten des Verhaltens von KNN bei einem Datenset welches eine signifikant unausgeglichene Zielvariabel Verteilung aufweist.

KNN wurde immer mit allen Prädiktoren durchgeführt jedoch mit verschiedenen k. Dies bedeutet das bei den Durchläufen verschieden viele Nachbarproben in Betracht gezogen wurden.

7.4.1 Evaluation

Um die Performance zu beurteilen, wurden Accuracy, Preäzision, Recall, F1-Score und Rel. Klassifikationsfehler betrachtet. Accuracy sagt aus wie viele Proben korrekt klassifiziert wurden. Präzision, Recall und F1-Score geben eine genauere aussage darüber welches Label korrekt klassifiziert wurde.

KNN wurde mit folgenden k durchgeführt.

7.4.2 K = 3

	Actual	
Predicted	0	1
0	18150	585
1	191	1074

Abbildung 20 - Konfusionsmatrix, k=3

Rel. Klassifikationsfehler	3.88%
Accuracy	96.12%
Präzision	84.9%
Recall	64.73%
F1-Score	73.46%

Tabelle 6 - Kennwerte für KNN k=3

7.4.3 K = 5

	Actual	
Predicted	0	1
0	18210	608
1	131	1051

Abbildung 21 - Konfusionsmatrix, k=5

Rel. Klassifikationsfehler	3.695%
Accuracy	96.305%
Präzision	88.917%
Recall	63.35%
F1-Score	73.98%

Tabelle 7 - Kennwerte für KNN k=5

7.4.4 K = 7

	Actual	
Predicted	0	1
0	18247	634
1	94	1025

Abbildung 22 - Konfusionsmatrix, k=7

Rel. Klassifikationsfehler	3.64%
Accuracy	96.36%
Präzision	91.599%
Recall	61.784%
F1-Score	73.79%

Tabelle 8 - Kennwerte für KNN k=7

7.4.5 K = 15

	Actual	
Predicted	0	1
0	18282	670
1	59	989

Abbildung 23 - Konfusionsmatrix, k=15

Rel. Klassifikationsfehler	3.645%
Accuracy	96.355%
Präzision	94.37%
Recall	59.614%
F1-Score	73.068%

Tabelle 9 - Kennwerte für KNN k=15

Deutlich erkennbar wird in den Abbildungen 20-23 und den dazu gehörigen Tabellen, dass je höher der Hyperparameter k gesetzt ist, desto höher ist die Präzision. Dies lässt sich darauf zurückführen, dass in den Daten eine signifikante über Repräsentation von Gesunden Proben gegeben ist. Deshalb werden bei höheren k immer mehr Gesunde Proben gesehen und immer mehr daten werden als gesund gelabelt.

7.5 Methodenvergleich

Die «Tabelle 10 - Vergleich der Methoden» gibt einen Überblick über die Performance der drei Methoden: Entscheidungsbaum, Logistische Regression und K-nearest-neighbours. Der Entscheidungsbaum weist die höchste Genauigkeit auf, während die Logistische Regression und K-nearest-neighbours ähnliche Ergebnisse liefern. Die Wahl der Methode kann von anderen Faktoren wie der Interpretierbarkeit des Modells und den Anforderungen der Anwendung abhängen.

Parameter	Entscheidungsbaum	Logistische Regression	K-nearest-neighbours
Rel. Klassifikationsfehler	2.6%	4.0%	3.6%
Accuracy	97.4%	96.0%	96.4%
Precision	92.8%	86.2%	91.6%
Recall	68.0%	63.8%	61.8%
F1 Score	N/A	73.3%	73.8%

Tabelle 10 - Vergleich der Methoden

```
> summary(pred)
    0     1
18902  1098
> model_evaluation(pred, data.test)
         Actual
Predicted     0     1
        0 18341   561
        1     0  1098
[1] "Relativer Klassifikationsfehler:  2.805 %"
[1] "Accuracy:  97.195 %"
[1] "Precision:  100 %"
[1] "Recall:  66.1844484629295 %"
[1] "F1-Score:  79.651795429815 %"
>
```

Abbildung 24 - KNN mit den Prädiktoren HbA1c und blood_glucose

In Abbildung 25 ist zu sehen, dass auch KNN mit nur den zwei Prädiktoren, welche der Entscheidungsbaum herausgefiltert hat, 100% Accuracy erreicht.

7.5.1 Methoden Fazit

Zusammenfassend erbringen alle drei Modelle befriedigende Resultate. Der Entscheidungsbaum liegt deutlich vorne in Bezug auf die Accuracy und Präzision. Sollte ein oder beide Prädiktoren fehlen für den Entscheidungsbaum speichert RStudio weitere Splits auf welche das Modell zurückgreifen könnte. Bei diesen «Backup» Splits würde die Accuracy jedoch höchstwahrscheinlich deutlich sinken, wonach die anderen Modelle bessere Resultate erzielen könnten.

8 Data Product

In diesem Kapitel wird die entwickelte Shiny-App vorgestellt. Dabei wird auf den Nutzen und die Anwendung der Shiny-App einen Fokus gelegt.

8.1 Nutzen der Shiny-App

Die Shiny-App wurde entwickelt, um anhand verschiedener Gesundheitsparameter eine Vorhersage des Diabetesrisikos durchzuführen. Der Benutzer interagiert mit der Anwendung, indem er über Schieberegler auf der Benutzeroberfläche Informationen wie Geschlecht, Alter, Bluthochdruck, Herzkrankheiten, Rauchgewohnheiten, BMI, HbA1c-Level und Blutzuckerspiegel eingibt. Diese Eingaben werden in Echtzeit verarbeitet und in ein Dataframe umgewandelt.

Die App verwendet drei verschiedene Modelle zur Vorhersage von Diabetes. Die logistische Regression den Entscheidungsbaum und der k-Nearest-Neighbor-Algorithmus. Anhand dieser Modelle wird die Wahrscheinlichkeit einer Diabeteserkrankung geschätzt.

Die Vorhersagen der Modelle werden dem Benutzer in Echtzeit auf der Benutzeroberfläche als Text angezeigt. Die Ergebnisse zeigen, ob die Modelle Diabetes als wahrscheinlich oder unwahrscheinlich einschätzen. Der Benutzer hat die Möglichkeit, die Vorhersagen verschiedener Modelle zu vergleichen, um zu verstehen, wie unterschiedliche statistische Ansätze zu unterschiedlichen Ergebnissen führen können.

Die KNN-Vorhersage verwendet einen Datensatz, der zuvor in Trainings- und Testdaten aufgeteilt wurde, wobei die Daten für den Algorithmus skaliert werden. Die Anwendung verwendet zuvor trainierte logistische Regressions- und Entscheidungsbaummodelle, die davor als RDS-Dateien (R Data Serialization) gespeichert wurden.

Insgesamt bietet die Shiny-App eine interaktive Plattform für Benutzer, um ihr individuelles Diabetesrisiko auf der Grundlage verschiedener Gesundheitsparameter zu erkunden.

8.2 Anwendung der Shiny-App

Auf der linken Seite der Benutzeroberfläche finden Sie Schieberegler für verschiedene Gesundheitsparameter. Wenn Sie die Gesundheitsparameter genau einstellen, erhalten Sie detaillierte Vorhersagen und können die Auswirkungen verschiedener Muster auf Ihr Diabetesrisiko nachvollziehen.

8.3 Layout-Entscheidung

Bei der Gestaltung des Layouts der Shiny-App, wurde Wert daraufgelegt, eine benutzerfreundliche Erfahrung bei der Vorhersage des Diabetesrisikos zu gewährleisten. Durch die Verwendung von Schiebereglern kann der Benutzer die Gesundheitsparameter einfach anpassen, was eine intuitive und benutzerfreundliche Interaktion fördert.

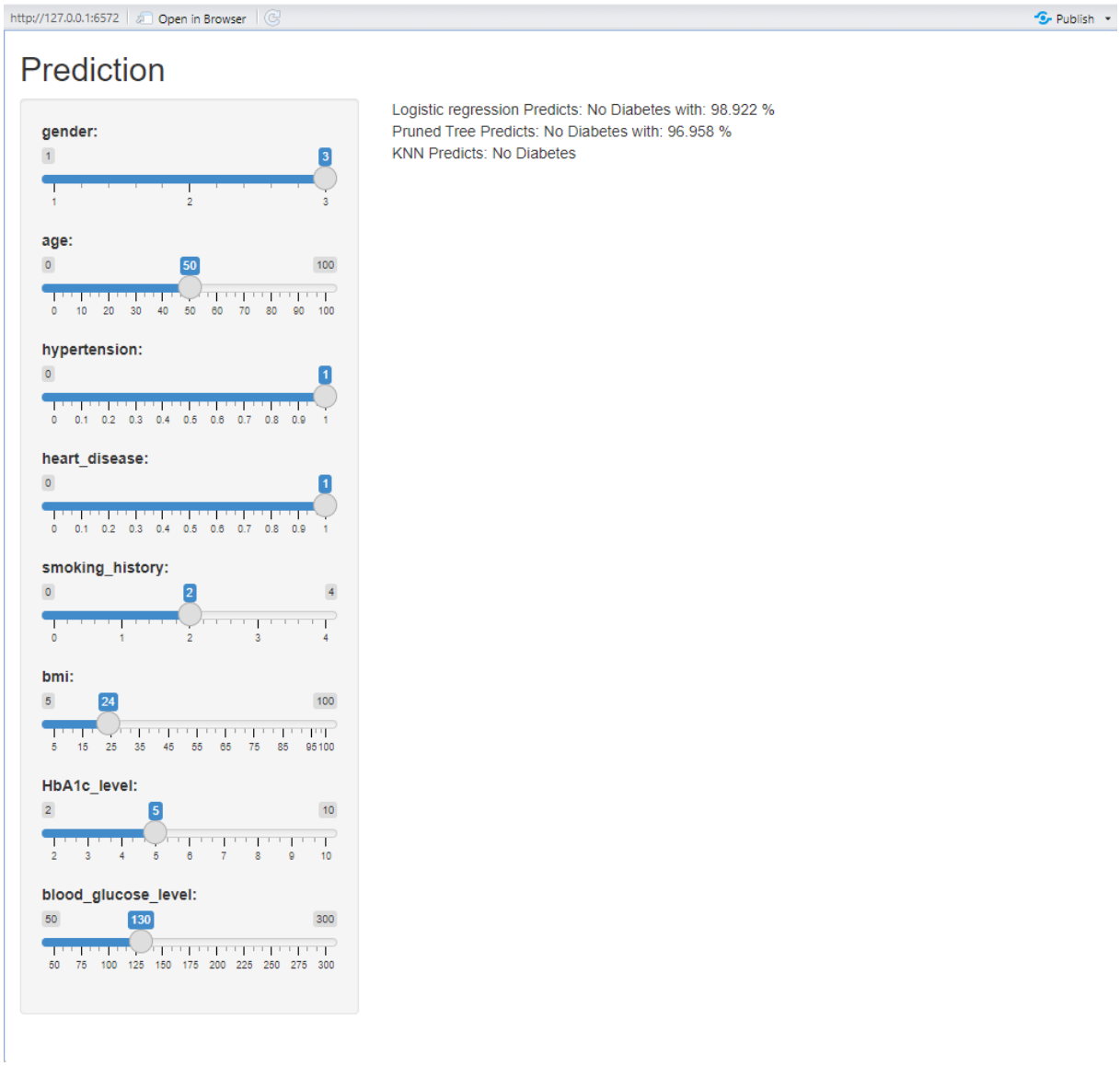

Abbildung 25 - Shiny App Diabetes Prediction

8.4 Potenzielle Kunden

Die Shiny-App zur Diabetesvorhersage ist für verschiedene Gruppen von Interesse, darunter Einzelpersonen, Gesundheitsdienstleister, Forschungsinstitute, Bildungseinrichtungen, Krankenversicherungen und Unternehmen aus dem Gesundheits- und Wellnessbereich. Einzelpersonen können die App nutzen, um ihr individuelles Diabetesrisiko zu verstehen und vorbeugende Massnahmen zu ergreifen. Gesundheitsdienstleister können die App für präventive Gesundheitsberatung und personalisierte Risikoeinschätzung nutzen. Forschungseinrichtungen und Wissenschaftler können die Shiny-App als Werkzeug zur Analyse von Gesundheitsdaten und zur Validierung neuer Ansätze zur Diabetesvorhersage nutzen. Bildungseinrichtungen könnten die App als Lehrmittel für Studierende in den Bereichen Gesundheitswissenschaften und Datenanalyse nutzen. Krankenversicherungen und Unternehmen im Bereich der betrieblichen Gesundheitsförderung könnten die Anwendung nutzen, um präventive Massnahmen zu fördern und personalisierte Gesundheitsinformationen bereitzustellen. Die Anwendung könnte somit ein breites Spektrum von Akteuren aus den Bereichen Gesundheit, Forschung und Bildung ansprechen, die an einer interaktiven Plattform zur Risikobewertung und prädiktiven Analyse von Diabetes interessiert sind.

8.5 Entscheidungsbaum-App

Zusätzlich wurde eine zweite Shiny-App entwickelt. Die App erlaubt das interaktive Erstellen von Entscheidungsbäumen und plottet dabei relevante Graphiken wie die CP-Tabelle. Sie eignet sich daher vor allem für Lehrzwecke. Die Parameter minbucket, minsplit, maxdepth sind die Parameter für das Erstellen des Initialbaums. Dieser Baum kann dann gepruned werden anhand der CP-Tabelle. Dieser Table hat Einträge, auf welche mit einem Index zugegriffen werden kann. Der Initialbuam wird dann anhand dieses Indexes «Prune with CP_index» zum finalen Baum geprunt.

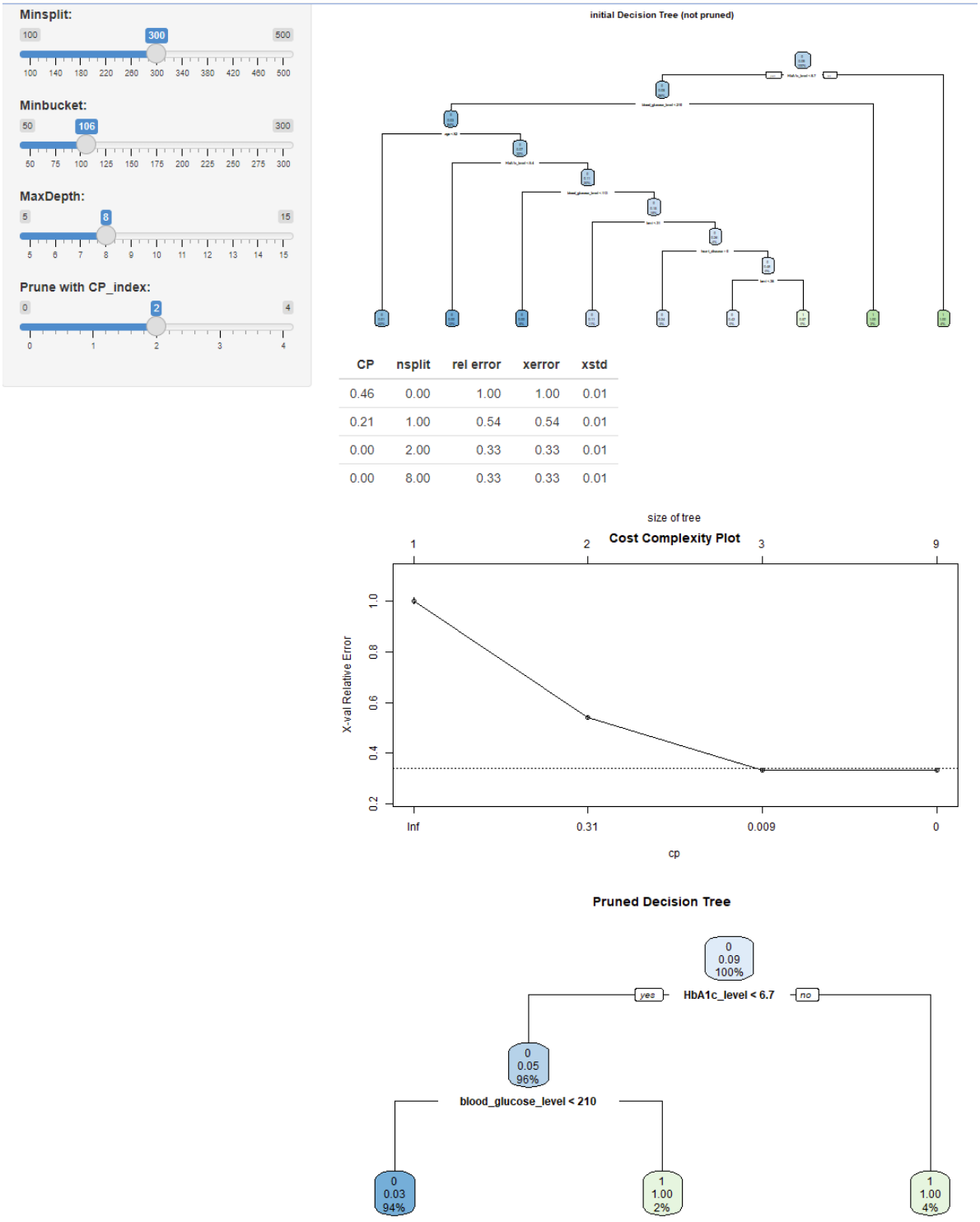

Abbildung 26 - Entscheidungsbaum Shiny App

9 Fazit

Im Rahmen unserer Untersuchung zur Bewertung des Diabetesrisikos und der potenziellen Einflussfaktoren haben wir die Hypothese aufgestellt, dass der HbA1c-Wert den größten Einfluss auf das Diabetesrisiko hat und höhere HbA1c-Werte proportional mit einem erhöhten Diabetesrisiko korrelieren. Unsere Analyse liefert jedoch interessante Erkenntnisse, die eine differenziertere Betrachtung erfordern.

Im Gegensatz zu unserer ursprünglichen Annahme konnten wir feststellen, dass Blutzuckerwerte eine wichtige Rolle im Diabetesrisiko spielen und sogar einen leicht höheren Stellenwert als der HbA1c-Wert haben. Dieses Ergebnis zeigt, dass die Konzentration ausschließlich auf HbA1c nicht genügt, um das komplexe Diabetesrisiko vollständig zu erfassen.

Die Hypothese, welche den HbA1c-Wert als primären Indikator für das Diabetesrisiko hervorhebt, ist falsch. Der HbA1c-Wert spielt zwar eine wichtige Rolle, Blutzucker ist aber mindestens genauso wichtig. Unsere Analyse zeigt, dass Diabetes das Ergebnis eines komplexen Zusammenspiels verschiedener Faktoren ist, mit den Faktoren HbA1c-Wert und Blutzuckerwert als wichtigste Prädiktoren.

Insgesamt zeigt unsere Untersuchung, dass ein umfassendes Verständnis und eine genauere Bewertung von Diabetes eine multifaktorielle Herangehensweise erfordern. Es ist von zentraler Bedeutung, das Zusammenspiel von HbA1c, Blutzuckerspiegel und möglichen weiteren Einflussfaktoren zu berücksichtigen.

9.1 Limitationen

Mit nur 8 Prädiktoren sind alle Modelle relativ begrenzt in Ihrer Vorhersage. Ein weiterer Faktor, welcher die erstellten Modelle limitiert ist, dass die Trainingsdaten zu 90% gesunde Patienten enthalten. So ist anzunehmen, dass alle Modelle auf diese Daten overfitted sind. Weiter sind in den Daten Diabetes Typ 1 und 2 als eine Zielvariable zusammengenommen wurden. Diese zwei deutlich verschieden Erkrankungen hängen von verschiedenen Faktoren ab und sollten voneinander getrennt werden oder von verschiedenen Modellen vorhergesagt werden.

10 Verzeichnisse

10.1 Tabellenverzeichnis

10.2 Abbildungsverzeichnis

Apicella Nevio / Pletscher Steven
Balke Nicolas / von Arx Matthias
Fassbind Andrin

11 Disclaimer

Chat-GPT wurde benutzt, um Vorlagen und Ergänzungen zu schreiben.