Pima Indians Diabetes ML Dashboard

1. Ziel des Dashboards

Das Dashboard dient der explorativen Datenanalyse (EDA) und der Modellbewertung für das Diabetes-Datenset der Pima Indians.
Es kombiniert deskriptive Statistiken, Vergleichsvisualisierungen, Korrelationsanalysen, Clustering-Methoden und Klassifikationsmetriken, um:

- relevante Einflussfaktoren für Diabetes zu identifizieren
- Zusammenhänge zwischen Merkmalen und der Zielvariable sichtbar zu machen
- die Leistungsfähigkeit von Machine-Learning-Modellen zu bewerten

2. KPI-Übersicht (Statistische Kennzahlen)

Visualisierung:

Kennzahlenkarten (oben im Dashboard)

Angezeigte Kennzahlen:

- Gesamtanzahl Patient:innen: 768
- Nicht-diabetisch: 500 (65 %)
- Diabetisch: 268 (35 %)
- Prädiktive Features: 8

Zweck der Visualisierung:
Die KPIs liefern einen schnellen Überblick über:

- die Größe des Datensatzes,
- die Klassenverteilung (relevant für ML-Modelle),
- die Komplexität des Merkmalsraums.

Beantwortete Fragen:

- Ist der Datensatz ausgeglichen oder unausgeglichen?
- Wie hoch ist der Anteil diabetischer Patientinnen?
- Mit wie vielen Variablen arbeitet das Modell?

3. Kreisdiagramm: Diabetes-Outcome-Verteilung

Visualisierung:
Kreisdiagramm zur Outcome-Verteilung

Darstellung:

- Grün: Nicht-diabetisch (65 %)
- Rot: Diabetisch (35 %)

Zweck der Visualisierung:
Das Diagramm stellt die Zielvariable visuell dar und erleichtert:

- die Einschätzung des Klassenungleichgewichts,
- die Begründung möglicher Modellanpassungen (z. B. SMOTE, Class Weights).

Interaktivität:
Beim Klick auf ein Segment werden zwei PCA-Cluster-Visualisierungen angezeigt:

- diabetische Patient:innen
- nicht-diabetische Patient:innen

Beantwortete Fragen:

- Wie stark ist Diabetes im Datensatz vertreten?
- Ist mit Klassenungleichgewicht zu rechnen?
- Wie unterscheiden sich Subgruppen innerhalb der Klassen?

Erkenntnis:
Das Datenset ist moderat unausgeglichen (65 % vs. 35 %), was ML-Modelle beeinflussen kann.

4. Altersverteilung nach Outcome

Visualisierung:
Gruppiertes Balkendiagramm (Diabetisch vs. Nicht-diabetisch)

Zweck:
Analyse des Zusammenhangs zwischen Alter und Diabetes.

Beantwortete Fragen:

- In welchen Altersgruppen tritt Diabetes häufiger auf?
- Unterscheiden sich die Altersprofile der beiden Gruppen?

Erkenntnis:
Diabetes tritt häufiger in höheren Altersgruppen auf → Alter ist ein relevanter Risikofaktor.

5. Glukosevergleich nach Klassen

Visualisierung:
Balkendiagramm (durchschnittlicher Glukosewert)

Zweck:
Vergleich eines der wichtigsten medizinischen Prädiktoren.

Beantwortete Fragen:

- Unterscheidet sich der durchschnittliche Glukosespiegel signifikant?
- Ist Glukose ein starker Prädiktor?

Erkenntnis:
Diabetische Patientinnen zeigen deutlich höhere Glukosewerte → hohe prädiktive Relevanz.

6. Feature-Vergleich nach Klassen

Visualisierung:
Mehrere kleine Balkendiagramme

Analysierte Features:

- Glucose
- BMI
- Age
- Pregnancies
- BloodPressure
- Insulin
- SkinThickness

Zweck:
Vergleich der durchschnittlichen Feature-Ausprägungen zwischen den Klassen.

Beantwortete Fragen:

- Welche Merkmale unterscheiden Diabetikerinnen am stärksten?
- Welche Features sind erklärungsstark?

Erkenntnisse:

- Glucose und BMI zeigen deutliche Unterschiede
- Age und Pregnancies haben moderaten Einfluss
- BloodPressure ist weniger trennscharf

7. Feature-Korrelation mit Outcome

Visualisierung:
Korrelations-Heatmap

Zweck:
Analyse linearer Zusammenhänge zwischen Features und der Zielvariable.

Beantwortete Fragen:

- Welche Merkmale korrelieren am stärksten mit Diabetes?
- Gibt es Multikollinearität zwischen Features?

Erkenntnisse:

- Glucose: stärkste positive Korrelation mit Outcome
- BMI und Age: mittlere Korrelation
- Teilweise Korrelationen zwischen Features (z. B. BMI & SkinThickness)

8. PCA-Cluster-Verteilung

Visualisierung:
Scatterplot (PC1 vs. PC2)

Zweck:
Dimensionsreduktion zur visuellen Exploration möglicher Cluster.

Beantwortete Fragen:

- Gibt es natürliche Gruppen im Datensatz?
- Lassen sich Klassen im reduzierten Raum trennen?

Erkenntnisse:

- PCA reduziert die Varianz sinnvoll auf zwei Hauptkomponenten
- Teilweise Cluster-Trennung sichtbar, jedoch keine perfekte Separation

9. KMeans-Analyse: Elbow & Silhouette Score

Visualisierung:
Liniengrafiken:

- Elbow Method (Inertia)
- Silhouette Score

Zweck:
Bestimmung der optimalen Anzahl von Clustern.

Beantwortete Fragen:

- Wie viele Cluster sind sinnvoll?
- Wie gut sind die Cluster voneinander getrennt?

Erkenntnis:
Ein Clusterbereich von k = 2 bis 3 ist sinnvoll und konsistent mit der PCA-Analyse.

10. Random-Forest-Modellbewertung

Visualisierungen:

- Konfusionsmatrix
- Klassifikationsbericht

Zweck:
Bewertung der Modellperformance.

Beantwortete Fragen:

- Wie gut erkennt das Modell Diabetes?
- Gibt es viele False Negatives (kritisch im medizinischen Kontext)?

Erkenntnisse:

- Gute Gesamtgenauigkeit (~76 %)
- Ausgewogene Precision-Recall-Balance
- Random Forest ist gut geeignet für dieses Problem

11. Zusammenfassung – Zentrale Erkenntnisse

- Glucose ist der stärkste Prädiktor für Diabetes
- BMI und Alter erhöhen das Risiko signifikant
- Der Datensatz ist leicht unausgeglichen → Modellanpassung empfehlenswert
- PCA und Clustering zeigen Struktur, aber keine perfekte Trennung
- Random Forest liefert robuste Klassifikationsergebnisse

12. Nutzen des Dashboards

Das Dashboard ermöglicht:

- medizinisch interpretierbare Feature-Analysen
- transparente Bewertung von ML-Modellen
- fundierte Entscheidungen zur Feature Selection
- verständliche Kommunikation der Ergebnisse an Nicht-ML-Experten