# Fallstudie zur Vorlesung Data Mining

## Präventionsoptimierung durch Vorhersage von Herzinfarkten

### Kontext

Erleiden Menschen einen Herzinfarkt, ist dies auf persönlicher Ebene immer ein schwerer Schlag für die Betroffenen, Angehörigen und Freunde. Für die gesetzlichen Krankenkassen sind Herzinfarkte zudem ein großer wirtschaftlicher Kostenpunkt. Ca. 10.000 EUR haben die Versicherungen für die akute Behandlung sowie Rehamaßnahmen pro Fall aufzubringen, was deutschlandweit in Summe 1,841 Milliarden Euro pro Jahr bedeutet. Gezielte Vorbeugung, z.B. durch die Beratung bei einer Präventionsassistenz, Ernährungskurse oder begleitete Sportangebote sind dabei starke Mittel, Leid und Kosten zu verhindern, indem das Herzinfarktrisiko um durchschnittlich 60% gesenkt wird.


Die FVM, ein gesetzlicher Krankenversicherer mit einer Millionen Versicherten aus dem Münsterland, möchte diese Möglichkeiten zukünftig stärker forcieren, um langfristig viel Leid, aber auch hohe finanzielle Belastungen zu vermeiden. Die Herausforderung besteht für die FVM dabei in dem gezielten Anbieten der Präventationsangebote für genau die Versicherten, die im Laufe ihres Lebens wahrscheinlich einen Herzinfarkt erleiden werden, da flächendeckende Präventationsmaßnahmen für alle Versicherten nicht finanzierbar sind. Zu diesem Zweck sollen Versicherte zukünftig freiwillig einen Bogen mit 21 Fragen beantworten können, der Information über den Lebensstil und Vorerkrankungen gibt, um sie damit hinsichtlich ihres Herzinfarktrisikos einzuordnen. Bei ähnlichen Befragungen wurden dank kleiner Prämien Beteiligungsquoten von 80% erzielt. Die Einordnung, ob ein Versicherter im Laufe des Lebens einen Herzinfarkt erleiden wird und damit für gezielte Präventation in Frage kommt, soll von einem Machine Learning Modell getätigt werden. Mit den Training dieses Modells beauftragt die FVM Sie und Ihr Team.

Fokus Ihres Teams soll dabei ein kostenzentriertes Modell sein, welches Einsparungen maximiert. Berücksichtigen Sie für Ihr Klassifikationsmodell Kosten & Erträge aus der nachfolgenden Tabelle.

| | | **Wahrheit** | |
|---|---|---|---|
| | | kein Infarkt | Infarkt |
| **Vorhersage** | kein Infarkt | 0 EUR | -5000 EUR |
| | Infarkt | -1000 EUR | 5000 EUR |

Die Tabelle zeigt, dass Versicherte, die richtigerweise als Infarktgefährdet identifiziert werden, zu Einsparungen in Höhe von 5000 EUR führen, da Kosten durch einen Infarkt mit einer Wahrscheinlichkeit von 60% aufgrund der gezielten Präventation, die mit 1000 EUR einzuberechnen sind, vermieden werden können. Für Versicherte, die fälschlicherweise als risikobehaftet klassifiziert werden, entstehen Opportunitätskosten in Höhe von 1000 EUR durch die Übernahme von Präventationsangeboten, die nicht notwendig sind. Versicherte, die fälschlicherweise als nicht-gefährdet klassifiziert werden, sind nicht nur besonders kritisch, sondern verursachen auch Kosten in Höhe von 5000 EUR, da ihre korrekte Einordnung das Risiko eines Infarkts mit allen verbundenen Kosten auf 60% gesenkt hätte. Bei Versicherten die korrekt als nicht-gefährdet klassifiziert werden, entstehen weder Kosten noch Erträge.

### Daten

Die Datenbasis für Ihr Modell finden Sie unter dem Namen `heart_disease.RData`. Diese Datei enthält 22 Merkmale von 200.000 Teilnehmenden einer wissenschaftlichen Studie zur Erforschung von Risikofaktoren für einen Herzinfarkt. Die dabei erhobenen beschreibenden Merkmale decken sich mit den Informationen, welche die FVM von ihren Versicherten abfragen will. Am Ende dieses Dokuments erhalten Sie zusätzliche Informationen zum Datensatz.

## Aufgabenstellung

Ziel dieser Fallstudie ist, mit Hilfe des Datensatzes ein Modell zu trainieren, das dazu geeignet ist, Versicherte zu erkennen, die im Laufe ihres Lebens einen Herzinfarkt erleiden werden. Die Vorhersage Ihres Modells wird am Ende mit Hilfe eines weiteren Datensatzes mit 50.000 Versicherten geprüft, für den Sie die Zielvariable nicht kennen. Mit diesem Datensatz (`heart_disease_result.RData`) wird mittels der Gesamtkosteneinsparung bewertet, wie gut die Vorhersage Ihres Modells funktioniert. Stellen Sie also sicher, dass beim Training Ihres Modells kein Overfitting vorliegt, da die Vorhersage auf neuen Datensätzen ansonsten zu schlechten Ergebnissen führen wird.

Die konkreten Schritte, welche Sie zur Lösung der Aufgabenstellung in **R** lösen sollen, lauten wie folgt:

1. Verschaffen Sie sich einen Überblick und ein Verständnis der vorliegenden Daten durch deskriptive Analysen und grafische Darstellungen.
2. Säubern Sie die Daten falls notwendig und leiten Sie wenn möglich neue "schlaue" Variablen her, die für die Vorhersagen genutzt werden können.
3. Nutzen Sie die Ihnen bekannten geeigneten Klassifkationsalgorithmen und erstellen Sie Vorhersagen. Tunen Sie gegebenenfalls die Hyperparameter des Modells.
4. Messen Sie die Güte des Modells bzw. vergleichen Sie die Güte der Modelle und wählen ein finales Modell. Nutzen Sie dazu statistische Kennzahlen, aber vor allem die Gesamtkosteneinsparung.
5. Ermitteln Sie, welche Merkmale sich gut zur Vorhersage eignen.
6. Veranschaulichen Sie Ihre Ergebnisse durch Tabellen und Abbildungen und interpretieren Sie diese im betriebswirtschaftlichen Kontext inklusive einer kritischen Reflexion.

## Format

Ihre Abgabe besteht aus drei Teilen:

### R Code

Geben Sie den von Ihnen erstellten R Code zur Datenanalyse und Modellentwicklung sowie Vorhersage ab. Der Code sollte nur die relevanten Teile enthalten, welche die Ergebnisse in Ihrem Bericht erzeugen. Stellen Sie sicher, dass der Code sauber, nachvollziehbar, gut dokumentiert und vor allem auch lauffähig ist. Die aus dem Code resultierenden Ergebnisse sollten denen in Ihrem Bericht entsprechen.

### Vorhersage

Erstellen Sie eine Vorhersage für die in `heart_disease_result.RData` enthaltenen Versicherten. Erstellen Sie dazu eine `*.RData` Datei (mit dem Befehl `save`), die einen Data Frame mit dem Namen `predicted` enthält. Dieser Data Frame enthält die in `heart_disease_result.RData` enthaltenen Spalten, ergänzt um die Spalte `prediction`. Die Spalte `prediction` enthält Ihre Vorhersage. Die Werte in dieser Spalte dürfen ausschließlich die Werte `kein Infarkt` oder `Infarkt` enthalten. Achten Sie unbedingt auf das korrekte Format!

### Fallstudienbericht

Erstellen Sie gemeinsam in Ihrer Gruppe einen Bericht mit maximal 5 Seiten. Wie eine typische wissenschaftliche Arbeit, sollte dieser Bericht folgende Elemente enthalten:

- **Titel:** Name und Beschreibung des Berichts, Namen der Autorinnen und Autoren.
- **Zusammenfassung/Executive Summary:** Eine vollständige, aber sehr kurze Zusammenfassung des gesamten Berichts in 2-5 Sätzen (einschließlich der wesentlichen Ergebnisse/Erkenntnisse).
- **Einleitung:** Beschreibung der konkreten Fragestellungen und Ziele der vorliegenden Arbeit.
- **Methodik:** Beschreibung der verwendeten Methoden und Analyseschritte; sehr knapp, keine theoretischen Details.
- **Ergebnisse:** Präsentation der Ergebnisse der Analyse; diese sollten mit Hilfe von Tabellen und Abbildungen möglichst übersichtlich dargestellt werden (jede Abbildung oder Tabelle benötigt eine Beschriftung; diese wird bei Tabellen oben, bei Abbildungen unten angeführt; Tabellen und Abbildungen werden je fortlaufend nummeriert; im Text wird auf jede Tabelle und jede Abbildung mit der entsprechenden Nummer verwiesen).

- **Diskussion/Ausblick:** Beantwortung der Ausgangsfrage und Ausblick auf weitere mögliche Untersuchungen. Gehen Sie auch gerne auf Verbesserungspotentiale für das Modell ein und diskutieren Sie ethische Aspekte solch eines Vorhabens.
- **Literaturverzeichnis:** Falls Sie zusätzliche Quellen verwenden.
- **Anhang:** Nutzen Sie den Anhang für zusätzliche Abbildungen und Tabellen, welche in dem Bericht keinen Platz finden. Der Anhang darf über den vorgegebenen Umfang des Berichts hinausgehen.

## Abgabe

Abgabe des Codes in Form eines `*.R` Skripts, der `*.RData` Datei und des Berichts als `*.pdf` Datei über ILIAS bis zum **19.01.2023**.

## Bewerungskriterien

Bewertet werden alle Teile der Abgabe nach folgenden Kriterien:

- Korrektheit der fachlichen Lösung und Umfang der durchgeführten Analysen
- Adäquater Einsatz und kritische Reflexion von Methoden und Werkzeugen aus den Vorlesungen
- Übersichtlichkeit und Nachvollziehbarkeit des eingereichten Codes
- Strukturierung, Gestaltung und fachliche Korrektheit des Berichts
- Einordnung und Relexion der Analyseergebnisse im wirtschaftlichen Kontext

## Merkmale des Datensatzes

Jede Zeile des Datensatzes beschreibt je eine befragte Person mit ihren soziodemographischen Merkmalen, Aussagen zur Gesundheit und dem Lebensstil sowie dem Zielmerkmal, ob die Person einen Herzinfarkt hatte oder nicht.

| ID | Spaltenbezeichnung | Erläuterung |
|---|---|---|
| 1 | Herzinfarkt | Indikator ob ein Herzinfarkt bisher auftrat oder nicht (Zielmerkmal) |
| 2 | Hoher_Blutdruck | Indikator ob ein hoher Blutdruck diagnostiziert wurde (1 ja, 0 nein). |
| 3 | Hoher_Cholspiegel | Indikator ob ein hoher Cholisterinspiegel diagnostiziert wurde (1 ja, 0 nein). |
| 4 | Cholspiegel_Check | Indikator ob der Cholisterinspiegel jemals überprüft wurde (1 ja, 0 nein). |
| 5 | BMI | Body Mass Index der befragten Person. |
| 6 | Rauchen | Indikator ob die befragte Person raucht (1 ja, 0 nein). |
| 7 | Schlaganfall | Indikator ob Schlaganfall bisher auftrat (1 ja, 0 nein). |
| 8 | Diabetes | Indikator ob Diabetes vorliegt (2 Diabetes, 1 Prädiabetes, 0 nein). |
| 9 | Phys_Aktivität | Indikator ob befragte Person innerhalb der letzten 30 Tage physisch aktiv war (1 ja, 0 nein). |
| 10 | Essgewohnheit_Obst | Indikator ob täglich Obst gegessen wird (1 ja, 0 nein). |
| 11 | Essgewohnheit_Gemuese | Indikator ob täglich Gemüse gegessen wird (1 ja, 0 nein). |
| 12 | Starker_Alkkonsum | Indikator ob Alkoholabhängigkeit besteht (1 ja, 0 nein). |
| 13 | Gesundheitsvorsorge | Indikator ob befragte Person versichert ist (1 ja, 0 nein). |
| 14 | Fin_Schwierigkeit | Indikator ob befragter Person durch finanzielle Probleme Behandlungen nicht möglich sind (1 ja, 0 nein). |
| 15 | Allg_Gesundheit | Einordnung der allgemeinen Gesundheit. |
| 16 | Ment_Gesundheit | Anzahl der Tage innerhalb des letzten Monats, an denen mentale Probleme auftraten. |
| 17 | Phys_Gesundheit | Anzahl der Tage innerhalb des letzten Monats, an denen physische Probleme auftraten. |
| 18 | Gehstoerung | Indikator ob Gehstörungen auftreten (1 ja, 0 nein). |
| 19 | Geschlecht | Geschlecht der befragten Person. |
| 20 | Alter | Alter der befragten Person. |
| 21 | Bildung | Höchster Bildungsabschluss der befragten Person. |
| 22 | Einkommen | Einkommensklasse der befragten Person. |