Datenakquisition und –qualität Übung 04

Name: _____

Aufwand in h: _____ Punkte: _____/24

Allgemeine Anmerkungen: Beantworten Sie die Fragestellungen ausführlich und geben Sie, falls möglich, ausführliche Beispiele an. Bei Programmieraufgabenstellungen ist der Source Code zu dokumentieren und eine saubere Lösungsidee zu erstellen. Testfälle sind entsprechend mitzuliefern. Eingabe- und Ausgabedateien sind auch mitzuliefern. In dieser Lehrveranstaltung müssen die Aufgabenstellungen mit Python (Version 3.X) gelöst werden.

Aufgabe 1 (24 Punkte): Arbeiten mit Pandas

In der letzten Übungseinheit wurde das Thema „Data pre-processing mit Pandas" näher erläutert. Auf Basis der dort kennengelernten Werkzeuge soll nun, passend zu dem in der Vorlesung vorgestellten Thema „How to deal with missing values", ein Datensatz verarbeitet werden. Neben der allgemeinen Vorverarbeitung sollen zu den online zur Verfügung gestellten Daten (patient4_fullDays) verschiedenste Strategien angewendet werden, um mit fehlenden Werten umzugehen. Die gegebenen Daten sind Realdaten aus einem Forschungsprojekt. Daten in solcher Form kommen sehr häufig vor.

Folgende Informationen sind gegeben:

1. Die Daten liegen in verschiedenen Dateien vor. Jede Datei beinhaltet Patientendaten für einen Patienten und für einen Tag

2. Die Dateien sind eigentlich csv-Dateien, leider wurden die Endung irrtümlicherweise auf „.xls" geändert.

3. Jede Datei besteht aus mehreren Spalten: Glucosewert, Insulinwert, CH, Date, Minutes und ID

4. Es gibt einige fehlende Werte, da die Messgeräte teilweise keine Daten geliefert haben.

Bearbeiten Sie konkret folgende Aufgabenstellungen. Achten Sie darauf, dass sie alle Schritte in Python durchführen, vorzugsweise mit den aus der Übung bekannten Modulen:

1. Lesen Sie die Daten mit Hilfe von Python ein. Achten Sie hierbei auf die Formatierung der Dateien wie Trennzeichen und Dezimalseparator.
   Fügen Sie anschließend alle Dateien zu einem Datensatz zusammen. Dieser soll die Daten anschließend chronologisch für alle verfügbaren Tage beinhalten.

2. Der erste und wichtigste Schritt in der Datenverarbeitung ist, sich einen Überblick über die zur Verfügung gestellten Daten zu verschaffen. Erstellen Sie dafür eine Statistik, welche den gegebenen Datensatz charakterisiert, also Informationen über fehlende Werte, Mittelwerte, minimale bzw. maximale Werte pro Features, Verteilung der einzelnen Features etc. liefert. Bereiten Sie die ermittelten Zahlen entsprechend auf, z.B. mit Plots und / oder Tabellen. Begründen und beschreiben Sie ihr Vorgehen ausführlich.

3. Die Datenformate sind in der Rohform in der Form „JJJJ/MM/TT" in einer Spalte sowie die Uhrzeit in Minuten seit Tagesbeginn in einer weiteren Spalte aufgeführt. Überführen Sie diese beiden Spalten in eine neue, sinnvoll formatierte Spalte „Timestamp".

4. Der Timestamp sollte durchgängig alle fünf Minuten eine Messung beinhalten. Durch Messfehler kann es sein, dass ein kompletter Datensatz für eine Messung inklusive des Timestamps nicht vorhanden ist – es also größere Sprünge als fünf Minuten gibt. Falls dem so ist, finden Sie diese Fehler und fügen Sie

eine neue Zeile für die fehlenden Zeitschritte mit passendem Timestamp ein. Alle anderen Werte der Zeile sind dann natürlich nicht bekannt.

5. Generieren Sie die „Missingness Matrix“ und plotten Sie diese.

6. Implementieren Sie ein Skript / Programm in dem Sie folgende „Missing Value Strategien“ für den Glukosewert sowie den Insulinwert ausimplementieren:
    a. Zeilenweises Entfernen von Missing Values
    b. Ersetzung durch Mittelwerte
    c. Ersetzung durch einen zufälligen Wert
    d. Interpolation durch Regression

7. Speichern Sie die Dateien für jede der vier Ersetzungsstrategien separat als CSV ab.

Geben Sie alle vorverarbeiteten Datensätze inklusive Skripte mit einer ordentlichen Dokumentation gezippt ab.