1 Qualifikationsaufgabe für das Praktikum Data Science 2021

Im ILIAS finden Sie eine Präsentation (PDF + Video), welche den Ablauf eines typischen Data-Science-Prozesses erläutert. Ziel dieser Aufgabe ist es, Ihr Wissen aus der Vorlesung und den vergangenen Übungen auf ein konkretes Data-Science-Problem anzuwenden. Dies ist gleichzeitig die Qualifikationsaufgabe für das Praktikum "Analyse großer Datenbestände" im Sommersemester 2021. Sie können die Aufgabe natürlich auch lösen, wenn Sie nicht am Praktikum teilnehmen wollen. Es wird keine Musterlösung veröffentlicht.

1.1 Szenario

Anhand physikalischer und ökonomischer Attribute sollen Sie bestimmen, ob ein Stromnetz – zugegebenermaßen ein sehr kleines, also mit nur wenigen Beteiligten – stabil ist oder nicht. Es liegt ein binäres Klassifikationsproblem vor, welches wir mit Matthews Correlation Coefficient[1] auswerten. Im ILIAS finden Sie die Daten, aufgeteilt nach Training/Test sowie Attribute/Klassenlabels. Das UCI-Repository[2] bietet eine Beschreibung der Daten und des Szenarios. Ihre Aufgabe ist es nun, die Daten zu analysieren und eine Vorhersage für die Testdaten zu erstellen. (Logischerweise befinden sich die Klassenlabels für die Testdaten nicht im ILIAS.)

1.2 Bewertung

Ihre Lösung wird anhand folgender Kategorien bewertet:

1. Methodische Qualität, also Einsatz geeigneter Data-Science-Techniken und Interpretation der Ergebnisse.
2. Qualität des Codes.
3. Qualität der Vorhersage, wobei eine gute Lösung ausreichend ist. (Sie müssen Ihre Modelle nicht bis zur letzten Nachkommastelle tunen; die Pipeline sollte maximal innerhalb weniger Minuten auf einem Standard-Notebook durchlaufen.)
4. Reproduzierbarkeit der Ergebnisse und Einhalten des Abgabeformats.

1.3 Abgabeformat

Abgabefrist ist der 14.03.2021. Das Abgabeobjekt befindet sich im ILIAS. Die Aufgabe muss selbstständig gelöst werden. Geben Sie drei Dateien ab: Ein Notebook mit dem Code, eine HTML-Version des Notebooks inklusive aller Ausgaben und eine Vorhersagedatei. Verwenden Sie die Dateinamen `VORNAME_NACHNAME_solution.Rmd` (bzw. `.ipynb`), `VORNAME_NACHNAME_output.html` und `VORNAME_NACHNAME_prediction.csv`.

Sie können R oder Python verwenden, jeweils in der aktuellsten Version. Nutzen Sie ein `Rmarkdown`- oder ein `Jupyter`-Notebook, welches sämtlichen Code und Ihre Interpretation der Ergebnisse enthält. Versehen Sie komplizierte Code-Passagen mit Kommentaren. Entfernen Sie Code, welcher nicht funktioniert oder dessen Ergebnisse Sie nicht interpretieren. Das Notebook soll am Ende natürlich die Vorhersagedatei erzeugen, aber zuvor auch andere Schritte eines Data-Science-Prozesses durchlaufen. Zum Bestehen der Aufgabe sollten folgende Elemente enthalten sein:

1. Exploration mit Statistiken und/oder Plots
2. Erstellen neuer Features
3. Vergleich von einer Baseline, einem einfachen Vorhersagemodell und einem komplizierten Vorhersagemodell

[1] https://en.wikipedia.org/wiki/Matthews_correlation_coefficient
[2] https://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulated+Data+

Das Notebook soll ohne manuelle Intervention vollständig und fehlerfrei ausführbar sein, also mittels "Run all" o.ä. Sofern notwendige Pakete über `CRAN` (R), `pip` (Python) oder `conda` (Python) verfügbar sind, müssen Sie keine expliziten Installationsanweisungen in Ihr Notebook einfügen. Ihr Code soll die Eingabedateien aus einem Ordner `data/` lesen, welcher sich im selben Ordner wie Ihr Notebook befindet. Die Vorhersagedatei soll ebenfalls im Ordner `data/` abgespeichert werden. Die Vorhersagedatei soll dasselbe Format wie die Datei mit den Klassenlabels der Trainingsdaten haben, also eine CSV-Datei mit einer Spalte, dem Spaltennamen in der ersten Zeile und keinen Anführungszeichen um die Werte.