**Programmentwurf Data Science Prototyp v1.0**

Es ist ein stark vereinfachter und abgeänderter Hausdatensatz gegeben, in dem 17 verschiedene Merkmale von Häusern gegeben sind mit 900 Datenpunkten sowie eine Beschreibung der 17 Merkmale. Andere Merkmale sind ausdrücklich nicht Teil der Aufgabe.

**Teil 1 (8 + 5 P):** Der Datensatz ist mit Hilfe statistischer Standardfunktionen zu analysieren, zu interpretieren und ggfs. zu bereinigen. Nutzen Sie visuelle, in die Programmierumgebung eingebettete Darstellungsarten, um wichtige Zusammenhänge zu visualisieren. Beschreiben Sie die Ergebnisse.

**Teil 2 (10 + 5 P):** Verwenden Sie drei verschiedene Vorhersageverfahren, um den Preis (SalePrice) vorherzusagen. Eines davon soll eine „verständliche“ Regression sein, die Inferenz* ermöglicht. Hierzu ist eine Beispielrechnung an einem Testbeispiel als Textkommentar direkt in IPython durchzuführen.

Gehen Sie vor wie in Data Science Projekten üblich ist. Stellen Sie jeweils die Ergebnisse Ihrer Verfahren vergleichend in Grafiken dar. Diskutieren Sie, für welche Lösung Sie sich entscheiden. Optimieren Sie in einer Iteration die gewählte Lösung, falls eine Verbesserung möglich ist. Kommentieren Sie abschließend in Textform eine weitere Analyseidee.

Schreiben Sie mehrere Methoden: Eine, die es erlaubt, einen Dateinamen zu übergeben (im Originalformat wie in der gegebenen csv) zu übergeben. Diese liest die Datei aus und erstellt einen Dataframe, der die zweite Methode aufruft: Diese führt ein vorher trainiertes Modell, welches übergeben wird, auf einem übergebenen Dataframe auf und berechnet und gibt gut lesbar als Text/Zahl folgende Werte aus (keine Grafik): $R^2$, MSE, RMSE, MAPE, MAX*. Rufen Sie die letztere Methode einmal auf dem Trainingsdatensatz und einmal auf dem Testdatensatz auf sowie einmal mit der gesamten Datei.

**Teil 3 (7 + 5 P):** Nutzen Sie drei Klassifikationsverfahren, um vorherzusagen, ob es eine Klimaanlage gibt (CentralAir). Eine davon soll ein Entscheidungsbaum sein, der auch grafisch ausgegeben werden kann (ggfs. separat in einem pdf). Gehen Sie **analog** zu Teil 2 vor, mit folgenden Unterschieden (siehe *): A, Statt Inferenz – Entscheidungsbaum beschreiben, B, Statt der o.g. Metriken (statt $R^2$ etc.): Korrektklassifikationsrate (Accuracy), False-Positive und False-Negative-Rate.

**Technische Vorgaben:** Python (v 3.7 Anaconda 2 2019.10, Seaborn 0.10.0 (führt zu 4.8.1-py37_0, conda-4.8.1), Lauffähigkeit ist Pflicht (virtuelle Maschinen zum Testen der Lauffähigkeit werden gestellt), csv liegt im gleichen Ordner, diese ist vom Format vorgegeben.

**Bewertungskriterien**

1. **Fachliche Bewertung (25 Punkte):** Lösungsqualität und Eleganz sowie Klarheit und Umfang der Betrachtung, Umsetzung von Data Science wie in der Vorlesung gelehrt in einem Code-Prototyp, korrekte Verwendung von wichtigen Funktionen / Bibliotheken, Güte der Endlösung, Nutzung der erworbenen Kenntnisse aus der Vorlesung, Vollständigkeit der Lösung konkret zu der Beschreibung oben

2. **Dokumentation (15 Punkte):** Dokumentation des Vorgehens der Datenauswertung im Sinne von Data Science, Angabe aller verwendeten Bibliotheken mit Versionsnummern (vollständig), Codekommentare wie in der Informatik üblich, Codequalität, Eleganz und Lesbarkeit

**Abgabe**

Bearbeitung in Gruppen mit jeweils **genau 2 Personen** bis zum **8. April 2019 18:00 Uhr einzureichen über das Moodle Lernsystem.**

**Abzugeben sind:**

1. **Programm:** Quellcode in genau einer Datei, Quelldatei im gleichen Ordner, lauffähig, klare Markierung der Aufgabenteile 1, 2 und 3 (auf Teil 1 kann im Text verwiesen werden wo nötig), Dokumentation (direkt als Text enthalten im IPython Notebook, Beschriftungen direkt an Diagrammen, Codekommentare wo notwendig), Matrikelnummer statt Name nutzen, Angabe ob Zusatz- oder Wahlfach)
2. **pdf-Ausdruck des kompletten Notebooks** mit Grafiken, ggfs. erweitert um eine Zusatzseite mit dem Entscheidungsbaum aus Teil 3 (darf ggfs. als separates pdf abgegeben werden)