Aufgabe für das Projekt: Entwickle eine Fragestellung anhand gegebener oder eigener Daten, und Versuche die Fragestellung möglichst effektiv zu beantworten.

Anhaltspunkte zu einzelnen Schritten

1. Klärung der Aufgabenstellung

- Was ist das Ziel?
- Was sind die Qualitätskriterien an denen die Lösung gemessen wird?
- Ist dies eine Aufgabe für überwachtes Lernen oder für nicht-überwachtes Lernen?
- (Welche Lösungen für ähnliche Probleme existieren schon im Unternehmen?)

2. Daten beschaffen

(Klärung:)

aus welchen Quellen kommen die Daten?
Wie viel Platz benötigen Sie?
Wie umständlich ist der Download?
Wie viel Zeilen und Spalten enthalten die Daten?

Beschaffung

- verschiedene Daten eventuell in eine einheitliche Datei schreiben
- Daten in ein Format umwandeln womit sich arbeiten lässt

3. Daten beschreiben anhand einer Teilmenge

- Größe und Typ
- Welche Features sind unwichtig? Welche sind Zielgröße?
- Sorte: Kategorien oder Kontinuum, Wertebereich, Text?
- Art der Verteilung
- wo fehlen Daten?
- Korrelationen untersuchen
- überlegen welche Transformationen nötig sind
- visualisieren

4. Daten umformen

- Arbeitskopie anlegen
- standardisieren, Ausreißer beseitigen ,fehlende Werte ersetzen, eventuell metrische Werte in Kategorien zusammenfassen ,Attribute umformen(z.B. Datumsangaben)
- Funktionen schreiben könnte, um das Bereinigen der Daten automatisch ablaufen zu lassen,

5. Daten analysieren

- eine Teilmenge der Daten aussondern

- Schnell einige (3-4) Modelle ausprobieren
- Leistung messen

6. Analyse verfeinern

    - überlegen ob die Parameter geeignet sind
    - andere Parameter probieren ,
    - danach über GridSearch oder Randomized Search beste Parameter suchen
    - zusätzlich Ensemble Methoden benutzen

7. Testdatensatz ausprobieren
    - falls akzeptabel benutzen , falls nicht akzeptabel, von vorne anfangen

8. (Programme für Wartungsarbeiten und Überwachung der Daten schreiben)

9. Bericht schreiben
    - Wie bin ich vorgegangen und warum?
    - Was habe ich gelernt
    - Was ging schnell? Was ging mühsam?
    - Haben sich die gewählten Klassen geeignet?
    - Hätte ich andere Daten gebraucht