# DA – Machine Learning – huiswerk week 4

Doorloop de volgende stappen in een Jupyter Notebook.

1. Download de data set ‘lending-club-data.csv’
2. Beperk de dataset tot de onderstaande kolommen:  
   grade: grade of the loan  
   sub\_grade: sub-grade of the loan  
   short\_emp: one year or less of employment  
   emp\_length\_num: number of years of employment  
   home\_ownership: home\_ownership status: own, mortgage or rent  
   dti: debt to income ratio  
   purpose: the purpose of the loan  
   term: the term of the loan  
   last\_delinq\_none: has borrower had a delinquency  
   last\_major\_derog\_none: has borrower had 90 day or worse rating  
   revol\_util: percent of available credit being used  
   total\_rec\_late\_fee: total late fees received to day  
   bad\_loan: [0: loan is paid back in full, 1: loan has defaulted]
3. Welk percentage van de dataset betreft ‘goede leningen’ (die zijn terugbetaald)?  
   welk percentage van de dataset betreft ‘slechte leningen’ (die *niet* zijn terugbetaald)?
4. De goede leningen zijn oververtegenwoordigd in de dataset. Gebruik de sample-functie (pandas dataframe) om gelijke aantallen goede en slechte leningen te verkrijgen in de dataset. Deze techniek heet ‘undersampling’, zie ook [wikipedia](https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis).
5. Sommige kolommen zijn nominaal/ordinaal (tekstueel) (grade, sub\_grade, home\_ownership, purpose, term) en daar kan sklearn.tree.DecisionTreeClassifier niet mee omgaan.  
   Gebruik one-hot-encoding om deze kolommen te vervangen door numerieke kolommen.
6. Splits de dataset naar een trainingset (80%) en een testset.
7. Fit een DecisionTreeClassifier m.b.v. Gini index en maximale diepte=6.
8. Fit een DecisionTreeClassifier m.b.v. cross-entropy en maximale diepte=6.
9. Vergelijk de accuracy van de trees uit de vorige twee stappen.
10. Fit een DecisionTreeClassifier m.b.v. cross-entropy en maximale diepte=2 en nog een met maximale diepte=10.
11. Vergelijk de accuracy van de (cross-entropy) trees met dieptes 2, 6 en 10.
12. Maak een confusion matrix voor cross-entropy tree met diepte 6.
13. (optioneel) Visualiseer de cross-entropy trees met diepte 2 en 6.