Kunskapskontroll – AI2 Del 1

I denna kunskapskontrollen kommer du använda Maskininlärning för att skapa en prediktionsmodell för huspriser.

Arbetet är en **gruppuppgift**, där ni kommer vara cirka 3 personer per grupp. Alla i gruppen skall hjälpas åt och alla i gruppen skall kunna allting som gjorts i arbetet. Ni kan lämna önskemål till Antonio om ni är några som vill samarbeta. Antonio kommer dela in grupperna. Precis som i verkligheten så uppmuntras ni till att samarbeta och använda alla hjälpmedel så länge arbetet ni gör är ert egna och ni förstår vad ni gjort.

I kunskapskontrollen så kommer ni att:

1. Besvara teoretiska frågor.
2. Skapa en modell för att prediktera huspriser.
3. Skriva en rapport som gruppen gör tillsammans, d.v.s. gruppen lämnar in en rapport (den inkluderar även en självutvärdering där ni reflekterar över ert arbete).
4. Göra en muntlig presentation.

Se kursplan för betygskriterier, vi kommer gå igenom dessa i detalj på lektionen.

Arbetet (rapport, kod, presentation) lämnas in på LearnPoint. Deadline fredag den 14 juni kl: 23.59. Alla i gruppen lämnar in allt trots att ni genomfört arbetet tillsammans, detta för att jag skall kunna administrera betygssättningen i LearnPoint.

Vid frågor / funderingar prata med Antonio. Skall bli väldigt spännande att följa och läsa era arbeten. Lycka till.

/Antonio

1. Teoretiska frågor

ChatGPT får inte användas när ni besvarar nedanstående teoretiska frågor. Svara koncist.

1. Vad är en CSV fil?

2. Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?
3. Julia delar upp sin data i träning och test. På träningsdatan så tränar hon tre modeller; ”Linjär Regression”, ”Lasso regression” och en ”Random Forest modell”. Hur skall hon välja vilken av de tre modellerna hon skall fortsätta använda när hon inte skapat ett explicit ”validerings-dataset”?
4. Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?
5. Hur kan du tolka RMSE och vad används det till:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2}$$

6. Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?

7. Vad är Streamlit för något och vad kan det användas till?

Läs följande länk som förklarar hur man kan inkorporera text/kategorisk data i modeller såsom kön eller färg: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/ innan du besvarar resterande frågor. Kategorisk data kan också hanteras direkt i Pandas vilket kan underlätta om man arbetar med Pandas DataFrames, se t.ex. https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

8. Vad kännetecknar en nominell variabel?
9. Vad kännetecknar en ordinal variabel?
10. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding.
11. Om vi använder vanlig linjär regression, skall vi använda one-hot-encoding eller dummy-variable encoding?

2. Modellering – Prediktera huspriser

- Skapa en modell för att prediktera huspriser, du skall använda ”housing” data filen (finns uppladdad på LearnPoint). Gör hela ML flödet från datainsamling, EDA, modellering till utvärdering på test datan. Se t.ex. appendix B i kursboken för hur ett ML/AI arbetsflöde kan se ut.
- Se kapitel 2 i kursboken för exempel hur datan modellerats.

- Har gruppen tid och är väldigt ambitiös så kan man skapa en Streamlit applikation. Detta är endast för de som vill, önskar och har tid.

3. Rapport

- Använd mallen ”rapport_mall” när ni skriver er rapport. Läs guiden ”rapport_guide” för hur man skriver en rapport. Se dokumentet ”exempel_rapport” på hur ett examensarbete kan se ut. Ni ska inte skriva så mycket, 3-6 sidor räcker. Detta är bra och rolig träning inför ert slutgiltiga examensarbete.

4. Muntlig presentation

- Använd t.ex. PowerPoint för er presentation. På lektionerna kommer Antonio ge tips för hur man presenterar och skapar en presentation.
- Presentationen skall vara upp till 20 minuter lång. Träna på den innan så ni håller tiden.