Kursplan – Data Scientist, Utbildnings nr YH01458 - 2021 – 2		
Kurs: Pythonprogrammering och Statistisk Dataanalys samt yrkesrollen	**Poäng:** 45 yhp	**Utgåva:**
Framtagen av UL granskad av RUC: Antonio Progmet	**Språk:** Svenska/Engelska	**Datum:**
Förkunskaper: Inga förkunskaper	**Granskad/Fastställd av:**	

Kursens huvudsakliga innehåll

Syftet med kursen är att den studerande ska få de kunskaper, färdigheter och kompetenser inom Python-programmering som krävs för att kunna arbeta med kvalificerad dataanalys. Den studerande kommer att få bekanta sig med bibliotek så som Numpy, Pandas och Matplotlib. De studerande kommer även att arbeta med grunderna i linjär algebra och statistik innefattande sannolikhetsteori och inferensteori samt implentera detta i Python. De studerande kommer även att få en introduktion i agila arbetssätt, github och datacamp

Kursen omfattar följande moment:

- Användning av Pythons inbyggda typer
- Organisera och strukturera kod enligt kodstandard
- Objektorienterad programmering såsom nyttjandet av klasser
- Versionshantering och GitHub
- Grunderna i sannolikhetsteori och statistisk inferensteori
- Grunderna i linjär algebra
- Paket för dataanalys såsom, numpy, pandas, matplotlib
- Utföra beräkningar och visualisera data
- Agila arbetssätt
- Datacamp

Kursens mål/läranderesultat

Målet med kursen är att den studerande ska behärska pythonprogrammering och de i branschen vanligt förekommande biblioteken. De studerande ska även behärska grunderna i sannolikhetsteori och statistisk inferensteori. De studerande ska ha även ha grundläggande kunskaper om linjär algebra i sin verktygslåda. Kunskaper och färdigheter från denna kurs är nödvändiga för att utföra dataanalys och implementera maskininlärningsmodeller.

Efter genomförd kurs ska den studerande kunna:
Kunskaper:

1. Redogöra för pythons inbyggda datatyper, kontrollstrukturer, funktioner och grundläggande objektorientering samt de i branschen vanligt förekommande biblioteken
2. Förklara grunderna i sannolikhetsteori såsom att förstå och exemplifiera: diskreta och kontinuerliga fördelningar som t.ex. binomialfördelningen och normalfördelningen, väntevärde, varians och kovarians.
3. Redogöra för grunderna i statistisk inferensteori såsom konfidentintervall och hypotesprövning
4. Redogöra för grunderna i linjär algebra såsom matriser och vektorer

Färdigheter:

5. Tillämpa grundläggande pythonprogrammering för att skriva program som utför beräkningar och visualisering data.
6. Använda sannolikhetsteori och statistisk inferensteori som verktyg för att analysera och dra slutsatser i olika beslutssituationer
7. Använda de i branschen vanliga biblioteken för dataanalys och maskininlärning på ett effektivt sätt

Kompetenser:
Inga kompetenser

Former för undervisning

Kursen kommer att genomföras med blended learning med inspelningar och aktiva lektioner. Under kursens gång erbjuds även s.k. Open Office Hours, där studenterna har ytterligare möjlighet att få hjälp av kursledarna genom att ställa frågor.

Former för kunskapskontroll

Examination kommer att ske genom:

- 1 inlämningsuppgift som görs i grupp (IG/G)
- 2 inlämningsuppgifter (IG/G/VG).

Betygsskala

Följande betygsskala tillämpas:
VG = Väl Godkänd, G = Godkänd, IG = Icke Godkänd

Principer för betygssättning

Läranderesultat	Inlämningsuppgift 1 (G)	Inlämningsuppgift 2 (G/VG)	Inlämningsuppgift 3 (G/VG)
1	X		
2		X	
3			X
4	X		
5	X		
6		X	X
7	X		

För betyget Godkänd ska den studerande

- Ha nått samtliga läranderesultat för kursen

För betyget Väl Godkänd ska den studerande:

- Uppnått kraven för betyget Godkänd
- Använda sannolikhetersteori och statistisk inferensteori för att lösa uppgifter med hög säkerhet och väl underbyggda resonemang

Icke Godkänt ges till studerande som har fullföljt kursen men inte nått alla mål för kursen.

Utbildare

Namn: Antonio Prgomet

E-post: Omniway.

Tillgänglighet: Möjligheten att ställa frågor och diskutera med utbildaren sker på lektionstid och vid behov via mejl funktionaliteten på Omniway. Försök att nyttja lektionstiden framför mejl för att kontakta mig.

Kursmaterial

Typ av material	**Kommentar**
Kursens GitHub sida: https://github.com/AntonioPrgomet/python_stat.git	Allt kursmaterial finns på GitHub länken.
Matrix Algebra for Engineers, Chasnov: https://www.math.hkust.edu.hk/~machas/matrix-algebra-for-engineers.pdf	Varje kapitel innehåller länk för videoföreläsningar.
Körner, S. & Wahlgren, L. (2015). Statistisk dataanalys. (5. uppl.). Lund: Studentlitteratur.	ISBN: 9789144108704 Kan köpas t.ex. här: https://www.adlibris.com/se/bok/statistisk-dataanalys-9789144108704 OBS: Införskaffas inför kursstart.
Körner, S. & Wahlgren, L. (2016). Tabeller och formler för statistiska beräkningar. (3. uppl.). Lund: Studentlitteratur.	ISBN: 9789144114545. Kan köpas t.ex. här: https://www.adlibris.com/se/bok/tabeller-och-formler-for-statistiska-berakningar-9789144114545 OBS: Införskaffas inför kursstart.
Referenslitteratur för den som vill läsa extra om Python. "Python Tutorial": https://docs.python.org/3/tutorial/index.html .	
https://www.datacamp.com/groups/shared_links/deb0e5b147765982489c40c9088d0214ccedc3a817d173b340ea82cb99e0a85a Vi kommer I denna kursen inte använda DataCamp, men vill man bli väldigt duktig på Python så kan man t.ex. göra följande kurer vid tillfälle (71 timmar och 18 kurser så endast för den som har tid): https://app.datacamp.com/learn/career-tracks/python-programmer)	**DataCamp online kurser. Registrera dig i DataCamp med din skolmejl (t.ex. malte.maltesson@utb.ecutbildning.se) annars fungerar det inte.**

Kunskapskontroll - Regler

Resultat och betyg registreras senast inom 10 arbetsdagar från deadline.

Om man inte kan utföra examinationen på utsatt deadline (till exempel på grund av allvarlig sjukdom) eller behöver utföra en komplettering så är ”andra” examinationstillfället senast en vecka efter att den studerande fått sin uppgift rättad eller utsatt deadline om man inte gjort examinationen. Du kan lämna in examinationen när du vill under denna period.

Det ”tredje” (och sista) examinationstillfället är senast tre veckor efter att den studerande fått sin uppgift rättad vid senaste tillfället eller serenaste deadline om ingen inlämning har gjorts. Har du missat samtliga examinationer måste du kontakta din utbildningsledare snarast.

Kunskapskontroll – Information

Kursen innehåller två kunskapskontroller där den andra innehåller två delar. Se veckoplaneringen längst ned när kunskapskontrollerna görs.

1. Den första är en inlämningsuppgift där ni kommer arbeta med programmering i biblioteken NumPy, Matplotlib och Pandas. Ni kommer börja arbeta individuellt med den och därefter diskutera hur ni angripit uppgifterna med er studieggrupp.
2. Den andra kunskapskontrollen kommer vara en skriftlig examination och innehålla två delar, en del i sannolikhetsteori och en del i statistisk inferensteori. Görs helt individuellt och endast formelsamlingen samt miniräknare får användas.

Upplägg på Föreläsningar / Lektioner

Lektionerna kommer fokusera på genomgångar och lösningar av uppgifter. Ert behov styr vad vi går igenom så skriv ned frågor som uppstår när ni studerar inför lektionerna så tar vi dem på lektionstid. Viktigt att man *inför* lektionerna arbetat med det material som förväntas enligt planeringen.

Schema:

	Förmiddag: 08.15 – 12.00.	Eftermiddag: 13.15 – 17.00
Måndag	Egenstudier för samtliga orter	Egenstudier för samtliga orter
Tisdag	**Helsingborg / Malmö**	**Stockholm**
Onsdag	**Göteborg**	**Helsingborg / Malmö**
Torsdag	**Stockholm**	**Göteborg**
Fredag	Egenstudier för samtliga orter	Egenstudier för samtliga orter

Veckoplanering – Vad skall jag göra varje arbetsdag?

I detta avsnitt så framgår i detalj vad som skall göras varje dag. Se nästa sida. Viktigt att du följer schemat.

Kursmaterial finns på kursens GitHub sida: https://github.com/AntonioPrgomet/python_stat.git

Veckoplanering

Exakta instruktioner på vad ni skall arbeta med varje arbetsdag framgår här.

	Kursvecka 1 (v.40): Introduktion till Python Programmering
Mån	Kolla på inspelad föreläsning och experimentera med tillhörande kod: https://www.youtube.com/watch?v=M7bnYJyCx0Q&t=4478s från tiden 0:00 till 1:02:20. Innan du kollar på videon om mängder (”Sets” på engelska) vid tidpunkt [39:38] så läs igenom texten och kolla videorna i följande länk: https://www.matteboken.se/lektioner/matte-5/mangdlara . Arbeta med "python_exercises_1_part_1".
Tis	FM: -Kolla på följande video om versionshantering med Git och Github: https://www.youtube.com/watch?v=SWYqp7iY_Tc&t=1290s EM: Lektion 13.15 - 17.00.
Ons	Kolla på inspelad föreläsning: https://www.youtube.com/watch?v=M7bnYJyCx0Q&t=4478s från 1:02:20 till 1:14:59. Arbeta med "python_exercises_1_part_2".
Tor	FM: Lektion 08.15-12.00. EM: Repetition.
Fre	Arbeta med "python_exercises_1_part_1" och "python_exercises_1_part_2".

	Kursvecka 2 (v.41): Linjär Algebra, NumPy, Matplotlib
Mån	- Kolla videos och läs Kapitel 1-5 i boken "Matrix Algebra for Engineers" av Jefferey R. Chasnov. - Efter lektion 2, kolla igenom följande video som bl.a. vid tidpunkten 1.10 visar en bra grafisk minnesregel på när man kan utföra matrismultiplikation: https://www.youtube.com/watch?v=nSNebx6C5Vg Arbeta med uppgifterna: - Kap 1: Uppgift 1 - Kap 2: Uppgift 1-3 - Kap 3: Uppgift 1, 2 (bara första delen, skippa beviset), 3 (bara första delen, skippa beviset). - Quiz på sid.15, uppgift 1,2. - Kap 4: Inga uppgifter. - Kap 5: Uppgift 1.
Tis	FM: Kapitel 6 i boken. Kapitel 10-12 i boken. Arbeta med uppgifterna: - Kap 6: Uppgift 1, 2 (uppgift 2 bygger på ett "trick" som är lätt när man väl gjort det en gång. Annars är det en svår uppgift). - Quiz på sid.32, uppgift 1,3. - Kap 10: Uppgift 1. - Kap 11: Uppgift 1. - Kap 12: Uppgift 1. ~~- Quiz på sid.40, uppgift 1-3.~~ EM: Lektion 13.15 - 17.00.
Ons	- Kolla igenom föreläsningarna kopplat till NumPy: https://www.youtube.com/watch?v=M7bnYJyCx0Q [1:14:59 till 1:17:30]. -Läs länkarna nedan och experimentera med koden: - https://numpy.org/doc/stable/user/quickstart.html - https://numpy.org/doc/stable/user/basics.broadcasting.html - Arbeta med programmeringsuppgifterna i "kunskapskontroll_1_1_NumPy".
Tor	FM: Lektion 08.15-12.00. EM: Repetition.

Fre	Kolla igenom föreläsningarna kopplat till Matplotlib: https://www.youtube.com/watch?v=M7bnYJyCx0Q [1:17:30 till 1:23:31]. Läs länkarna och experimentera med koden: - https://matplotlib.org/stable/tutorials/introductory/usage.html#sphx-glr-tutorials-introductory-usage-py - https://matplotlib.org/stable/tutorials/introductory/pyplot.html#sphx-glr-tutorials-introductory-pyplot-py - Arbeta med programmeringsuppgifterna i "kunskapskontroll_1_2_matplotlib".

	Kursvecka 3 (v.42): Pandas, Kunskapskontroll 1, Sannolikhetsteori
Mån	Kolla på följande video om Pandas: https://www.youtube.com/watch?v=ZoNFPQUUsyk&t=28s och arbeta med tillhörande kod från videon. Läs dokumentation "10 minutes to Pandas" och experimentera med koden: https://pandas.pydata.org/docs/user_guide/10min.html - Arbeta med programmeringsuppgifterna i "kunskapskontroll_1_3_Pandas".
Tis	FM: Repetition. EM: Lektion 13.15 - 17.00 med Márk Mészáros där ni kommer arbeta med Pandas.
Ons	Arbeta och diskutera kunskapskontroll 1 tillsammans med din grupp.
Tor	FM: Lektion 08.15-12.00. EM: **Deadline "Kunskapskontroll 1" kl: 23.59, lämnas in i Omniway.**
Fre	https://www.youtube.com/playlist?list=PLgzaMbMPEHEwkc-XVv3gpprOk7y2IHWLJ - Kolla video 1: Introduktion - Sannolikhetsteori & Statistisk Inferens - Kolla video 2: Introduktion - Sannolikhetsteori Läs kapitel 1 i kursboken. - Avsnitt 1.4 "Odds" är mindre viktigt, läs det översiktligt. Arbeta med uppgifterna: - 101-104 - 106-107 - 109-114 - 115-121

	Kursvecka 4 (v.43): Sannolikhetsteori
Mån	Kolla video 3: Sannolikhetsbegreppet. Läs kapitel 2 i kursboken. Arbeta med uppgifterna: - 201-203 - 204-207 - 208-213 - 214-217 - 218-221 - 222-227
Tis	FM: Repetition. EM: Lektion 13.15 - 17.00.
Ons	Kolla video 4 "Diskret slumpvariabel". Läs kapitel 3 i kursboken. Arbeta med uppgifterna: - 301 - 304, 305 - 306 - 307-313 - 314-316 - 317 - 318-320
Tor	FM: Lektion 08.15-12.00. EM: Repetition.
Fre	Repetition.

	Kursvecka 5 (v.44): Sannolikhetsteori
Mån	Kolla video 5: "Tvådimensionell Slumpvariabel". Läs kapitel 4 i kursboken. Arbeta med uppgifterna: - 401-404, 407-412, 413 (I uppgift 413 härleder ni väntevärdet och variansen för Binomialfödelningen som är en summa av Tvåpunktsfördelade/Bernoulli slumpvariabler), 414-416.
Tis	FM: Repetition. EM: Lektion 13.15 - 17.00.
Ons	Kolla video 10 "Kod Demonstration - Statistik" fram till 04:32. Arbeta med koden "sannolikhetsteori_python"
Tor	FM: Lektion 08.15-12.00. EM:
Fre	Kolla video 6: "Normalfördelningen". Läs kapitel 5 i kursboken. Arbeta med uppgifterna: - 501-508 - 509-512 - 513-517

	Kursvecka 6 (v.45): Sannolikhetsteori, Statistisk Inferensteori
Mån	Kolla video 6: "Normalfördelningen". Läs kapitel 5 i kursboken. Arbeta med uppgifterna: - 501-508 - 509-512 - 513-517
Tis	FM: Repetition. EM: Lektion 13.15 - 17.00.
Ons	Kolla video 7 "Slumpmässigt urval och punktskattning". Läs kapitel 6 i kursboken. Arbeta med uppgifterna: - 601-609
Tor	FM: Lektion 08.15-12.00. EM: Repetition.
Fre	Kolla video 8 "Konfidensintervall" och läs kapitel 7 i kursboken. Arbeta med uppgifterna: - 701, 702 - 703, - 704, 705 - 706 - 710

	Kursvecka 7 (v.46): Statistisk Inferensteori
Mån	Kolla video 8 "Konfidensintervall" och läs kapitel 7 i kursboken. Arbeta med uppgifterna: - 701, 702 - 703, - 704, 705 - 706 - 710
Tis	FM: Repetition. EM: Lektion 13.15 - 17.00.
Ons	Kolla video 9: Hypotesprövning Läs kapitel 8 kursboken. Arbeta med uppgifterna: - 801-804 - 805 - 806 - 807 - 808, 809 - 810, 811 - 812 - 816
Tor	FM: Lektion 08.15-12.00. EM: Repetition.
Fre	Kolla video 9: Hypotesprövning Läs kapitel 8 kursboken. Arbeta med uppgifterna: - 801-804 - 805 - 806 - 807 - 808, 809 - 810, 811 - 812 - 816

	Kursvecka 8 (v.47): Studier inför Kunskapskontroll
Mån	- Kolla video 10 "Kod Demonstration - Statistik". - Arbeta med koden "inferens_python" Repetition.
Tis	FM: Repetition. EM: Lektion 13.15 - 17.00.
Ons	Studera inför kunskapskontrollen i Sannolikhetsteori och Statistisk Inferens.
Tor	FM: Lektion 08.15-12.00. EM: Repetition.
Fre	Studera inför kunskapskontrollen i Sannolikhetsteori och Statistisk Inferens.

	Kursvecka 9 (v.48): Kunskapskontroll
Mån	Studera inför kunskapskontrollen i Sannolikhetsteori och Statistisk Inferens.
Tis	FM: Studera inför kunskapskontrollen i Sannolikhetsteori och Statistisk Inferens. EM: Lektion 13.15 - 17.00.
Ons	Studera inför kunskapskontrollen i Sannolikhetsteori och Statistisk Inferens.
Tor	FM: Lektion 08.15-12.00. EM: Studera inför kunskapskontrollen i Sannolikhetsteori och Statistisk Inferens.
Fre	Kunskapskontroll 2, "Sannolikhetsteori och Statistisk Inferensteori". Deadline Kl: 17.00 och lämnas in i Omniway.

Videos Kopplat till Statistisk Dataanalys

I veckoplaneringen längre ned så är videos jag spelat in inkluderade. Här är fler videos som ni kan kolla på för att få bredare och djupare förståelse. Kapitel hänvisningarna nedan refererar till boken ”Statistisk Dataanalys” av Körner & Wahlgren (2015).

Jag rekommenderar att ni kollar på videorna som ett komplement till de jag gjort så får ni två förklaringar på många av koncepten.

Kapitel 1-2

För kapitel 1-2, kolla följande spellista som en ”introduktion”:
https://www.youtube.com/watch?v=B1v9OeCTlu0&list=PLvxOuBpazmsOGOursPoofaHyz_1NpxbhA&index=1

1. Video 5: *What Does Independence Look Like on a Venn Diagram?* Kan du skippa.
2. En extra video om Kombinatorik (kap 1.6):
   https://www.youtube.com/watch?v=XJnIdRXUi7A&t=79s

Kapitel 3:

För kapitel 3 kolla igenom följande spellista innehållande 13 videos:
https://www.youtube.com/watch?v=oHcrna8Fk18&list=PLvxOuBpazmsNIHP5cz37oOPZx0JKyNszN

3. Video 3 som handlar om Bernoulli fördelningen är det boken kallar för ”Tvåpunktsfördelad variabel”.
4. På s.91 i boken så framgår det att man kan approximera binomialfördelningen med Poissionfördelningen, för detta se video 9 och 10, video 10 bevisar det och är lite ”överkurs” men det är ett vackert bevis så kolla gärna på det.
5. Video 6 handlar om ”Geometric Distribution” som boken kallar för ffg-fördelning.
6. Video 11 kopplat till ”Negative Binomial Distribution” ingår inte i kursen och kan skippas.
7. Video 12 kopplat till ”Multinomial Distribution”, är likt trinomial fördelningen som dyker upp först i kapitel 4.6 i boken. Förstår man den ena så förstår man den andra.

Hur man använder Tabell 1: Binomialfördelningen i formelsamlingen kan ses här:
https://www.youtube.com/watch?v=gfDWDujLtfM&t=71s

Kapitel 4:

Kapitel 4.4

- https://www.youtube.com/watch?v=KDw3hC2YNFc
- https://www.youtube.com/watch?v=85Ilb-89sjk

Kapitel 4.5

- Härledning av räkneregler för varians: https://www.youtube.com/watch?v=zdhkXWyy0K0 , i videon används den generella egenskapen av förväntans värdet att:
  $E[aX + bY + c] = aE[X] + bE[Y] + c$, dvs. förväntans värdet är en linjär operation.

Kapitel 4.6

Multinomial fördelningen är väldigt lik trinomial fördelningen, förstår man den ena så förstår man den andra. I följande video förklaras multinomial fördelningen:

https://www.youtube.com/watch?v=syVW7DgvUaY&t=382s

Kapitel 5:

Kolla först på video 1-7 i följande spellista om kontinuerliga sannolikhetsfördelningar:

https://www.youtube.com/watch?v=OWSOhpS00_s&list=PLvxOuBpazmsPDZGwqhhjE3KkLWnTD34R0

- Vi är hittills vana vid diskreta fördelningar och vi vet t.ex. att $\sum_i P(x_i) = 1$ eftersom summan av alla sannolikheter är 1. I det kontinuerliga fallet så använder man integraler istället för summor, motsvarande formel blir: $\int_{-\infty}^{\infty} f(t)\,dt = 1$ där vi integrerar från minus oändligheten till oändligheten. I det diskreta hade vi en ”Probability mass function” $p(x_i)$ medan vi i det kontinuerliga fallet hade en ”probability density function” $f(t)$.
- På liknande sett vet vi att väntevärdet för diskreta fördelningar kan räknas ut enligt följande formel: E[X] = $\sum_i x_i p(x_i)$ (d.v.s. utfall multiplicerat med sannolikhet). För kontinuerliga fördelningar får vi: E[X] = $\int_{-\infty}^{\infty} tf(t)\,dt$.
  Notera att vi kan använda t, x eller någon annan integrationsvariabel, det spelar ingen roll: $\int_{-\infty}^{\infty} tf(t)\,dt = \int_{-\infty}^{\infty} xf(x)\,dx$.
- Kontinuitetskorrektion (se sid. 137 i boken) gås igenom i video 7 vid tiden: 5:42. Bra förklarat på något som kan upplevas förvirrande i början.

Hur man använder normalfördelningstabellen (tabell 3a i formelsamlingen): https://www.youtube.com/watch?v=p_KApjpyBHE.

För kapitel 5.3, se följande spellista: https://www.youtube.com/watch?v=Zbw-YvELsaM&list=PLvxOuBpazmsP7UN00cNZX64N1o_8635ds

Kapitel 6:

6.3: https://www.youtube.com/watch?v=xJlwSkyeP0k

Kapitel 7

Se hela spellistan som är kopplad till kapitel 7:

https://www.youtube.com/watch?v=27iSnzss2wM&list=PLvxOuBpazmsMdPBRxBTvwLv5Lhuk0tuXh

Kapitel 8:

Se hela spellistan som är kopplat till kapitel 8:

https://www.youtube.com/watch?v=tTeMYuS87oU&list=PLvxOuBpazmsNo893xlpXNfMzVpRBjDH67