

Labbrapport i Statistik

Rapport X

Kurskod, Grupp Y

Författare 1 Namn

Författare 2 Namn

Avdelningen för Statistik och maskininlärning

Institutionen för datavetenskap

Linköpings universitet

20XX-XX-XX

Innehåll

1	Inledning	1
1.1	Bakgrund	1
1.2	Syfte	1
1.3	Problemformulering (Frågeställning)	1
2	Data	2
2.1	Databearbetning	2
3	Metod (VIKTIGT!!!)	3
3.1	Teori	3
3.2	Praktisk metod	3
4	Resultat	5
5	Diskussion	6
6	Slutsats	7
7	Referenser	8
8	Bilaga	9
8.1	Kod	9
8.2	Redogörelse för bidrag	9
8.3	Generativ AI	9

1 Inledning

Detta är enklare mall för hur akademiska rapporter ska skrivas. Den kan även användas som stöd när övriga rapporter ska skrivas.

Notera: det finns instruktioner i denna mall löpande till och med Bilaga.

Ha som riktmärke att rapporten ska vara 10-20 sidor (exkludera referenser och bilaga från den siffran). Det är viktigt att det inte ska vara för långt för opponenter och examinator att läsa.

Se till att använda ett korrekt språk i rapporterna, utan stavfel och andra grammatiska misstag. Tänk hela tiden att ni ska kunna visa upp det ni presterar för en framtida arbetsgivare.

Referenser till källor, både elektroniska och fysiska, måste följa en korrekt och enhetlig stil. Det finns flera olika att välja mellan men de vanligaste är APA, Harvard eller Oxford. Det spelar ingen roll vilken av dessa system som ni väljer att använda så länge ni använder en konsekvent och korrekt.

Denna mall använder en bibtex-fil med referenser (report_ref.bib) Exempel på referenser:

- Vi kan läsa om K-means klustering i James et al. (2013).
- En annan referens är (Zaki and Meira Jr 2019)

Läs *här* för mer information om referenser i Rmarkdown.

Notera: Alla figurer, tabeller och liknande ska vara numrerade och ska refereras från brödtexten. Alltså inga orefererade figurer/tabeller! När ni ska kommentera/analysera figurer/tabeller etc, referera till figuren/tabellens nummer: "i figur 1 ses det att...", undvik formuleringar som "i tabellen nedan/ovan..."

Notera: **Ingen R-kod** eller **råa R-utskrifter** ska synas i rapporten (med undantag för om de är i Bilaga under rubriken Kod). Alla tabeller ska vara ordentliga rapporttabeller.

1.1 Bakgrund

I bakgrunden ges en introduktion till ämnet och motivering till problemställningen i uppsatsen.

1.2 Syfte

Här beskrivs syftet med undersökningen samt eventuella avgränsningar. Det är ok att slå ihop Bakgrund och Syfte till en rubrik.

1.3 Problemformulering (Frågeställning)

Detta kapitel är smidigt för att i punktform presentera de frågeställningar som rapporten ska försöka besvara.

2 Data

Detta kapitel beskriver den datamängd som rapporten baserar sig på. Definitioner av variabler och rapportspecifika termer som används, vilken källa (visa referens vid behov!) de kommer ifrån osv.

2.1 Databearbetning

Här ska all databearbetning presenteras, det vill säga ifall ni har behövt transformera, skapa nya variabler eller på annat sätt ändrat i datamaterialet från dess ursprungliga form som ni laddat ner från källan.

Allmänt om data:

- Variabeldefinitioner
- Ange tydligt vad som är reponsvariabel och vad som är era förklarande variabler
- Visa metadata: dimensioner på data, skaltyper på variabler etc
- Standardiseringar när det behövs
- Visa beskrivande statistik, lämpliga plottar och tabeller på data
- Grundläggande variabeltransformationer
- Hantera saknade värden
- Undersök om det finns outliers, oblanterade klasser mm
- Datauppdelning: Träning, validering, test

Undvik onödigt stora plottar/figurer.

Om man har många variabler som man vill plotta → överväg att lägga en del av dessa i Bilagan och skriv något i stil “att fler plottar finns i bilagan...”

Om man har ett antal figurer som är mycket lika varandra → visa en i själva rapporten och lägg resten i Bilagan och skriv att resterade plottar finns i Bilagan. Detta gäller alla delar av rapporten.

3 Metod (VIKTIGT!!!)

3.1 Teori

Detta kapitel beskriver de statistiska metoderna som används i rapporten. Hänvisa till litteraturen då metoderna förklaras och visa formler när det är relevant.

Ha som riktmärke att sektionen Metod ska vara max 6 sidor.

Metoderna ska beskrivas **generellt** i detta kapitel. Det räcker med en **kortfattad** beskrivning. Ni ska inte ta upp alla detaljer på varje metod. Tänk max 1 sida per större metod/modellfamilj. Kan vara kortare.

Notera: Vanlig linjär regression och linjär logistisk regression behöver inte förklaras men ska nämnas om det användas. Om man använder någon utökning av linjär (logistisk) regression så ska det förklaras. I de flesta fall så vill man jämföra med enkel baslinjemetod (baseline), och då brukar oftast linjär (logistisk) regression vara den bästa alternativet för det. Man kan ha en egen rubrik “Baslinjemetod” där man skriver kortfattad om vilken baslinjemetod man använder. Se nedan för mer info om baslinjemetod.

Börja med att kort förklara syftet eller motivationen med metoden/modellen. Tänk: Varför ska man använda denna metod?

Ta upp nedanstående punkter i den mån de är relevanta för er metod:

- Vilka krav eller antaganden som gäller för att man kan använda metoden. (tex: de flesta metoder i kursen antar att vi har oberoende observationer)
- **Kort** beskrivning av metoden: “vad gör metoden?”
 - Definitioner, skattning, optimering mm
- Modeller/modellval, hyperparameterar, modellkomplexitet: Förklara de viktiga hyperparameterarna.
- Överanpassning/generalisering
 - Träning, validering, test
- Utvärderingsmått och liknande

Notera att i de flesta fall så är det lämpligt att jämföra två eller fler modeller/metoder på samma problem. Det kan vara så att man har samma metodfamilj, men med olika hyperparametrar.

- Inom klassificering /regression: Om ni tänker använda någon icke-linjär metod, då är det ofta rimligt att jämföra med linjär regression/linjär logistisk regression (men även andra modeller kan vara intressant att jämföra med). Tänk frågan: “Vad vinner vi med en mer komplicerad modell?”
- Det är ok att ha en metod som sin “huvudmetod” som beskrivs mer ingående i metodkapitlet, och en eller flera jämförande metoder som beskrivs kortare.
- Om ni vill hålla er till endast en metod så bör ni motivera det och den metoden bör utforskas mer ingående med avseende på olika inställningar etc. (Ett exempel kan vara: ni arbetar med bilddata och använder bara olika typer av neurala nätverk)
- Ha som grund: minst 2 olika typer av modeller + en baslinjemetod

3.2 Praktisk metod

Först beskriver ni era metoder/modeller teoretiskt i generella termer. (sektionen ovan)

Sen ska ni beskriva era “praktiska metod”: hur ni applicerar era metoder/modeller på ert specifika problem. Ni ska alltså ha ett strukturerat arbetssätt och ni ska beskriva det övergripande här. Sen under Resultat visar ni resultatet av er applicerade metod.

- När man analyserar data är det naturligt att ha en initial fas där man explorativt utforskar data och olika metoder kopplat till problemet. Denna fas behöver man oftast inte skriva om under Metod. Notera:

denna fas ska vara relativt kort i tid.

- Sen när ni fått ungefärlig bild av vad som ni behöver göras för att analysera er frågeställning så ska ni bestämma ett strukturerat upplägg på hur ni ska använda era metoder/modeller på ert problem.
 - Ska data delas upp? Hur? (träning, validering, korsvalidering, test)
 - Vilka modellval/inställningar/hyperparametrar ska göras? Detta är ofta en svår och viktig fråga.
 - * Ofta behöver man fixera några inställningar/hyperparametrar (pga man behöver avgränsa och tidsbegränsningar)
 - * Andra inställningar/hyperparametrar vill man hitta så optimala värden för som möjligt (ofta med hjälp av korsvalidering/valideringsdata) → ni vill göra en hyperparametersökning för era olika metoder/modeller
 - * Ange tydligt vilka hyperparametrar som ni fixerar och vilka som ni vill försöka optimera för olika metoder/modeller. De som ni optimerar: ange tydligt vilka olika hyperparametervärden som ni väljer mellan för varje specifik hyperparameter. Ex: “För KNN används korsvalidering med tio uppdelningar för att bestämma k . Följande värden på k testas: $k = 5, 7, 9, \dots, 25$. Det k som ger minst genomsnittligt MSE i korsvalideringen används.”
 - Hur metoder/modellerna ska utvärderas och jämföras.
- Det är ok att revidera sitt praktiska upplägg vid behov efter att man har börjat med själva analysen (se till att uppdatera beskrivningen av upplägget)

Förklara vilka mjukvaror/program/paket som ni använder för att tillämpa er metod.

4 Resultat

Detta kapitel presenterar resultatet av beräkningar, diagram och tabeller som sen tolkas och analyseras.

Här anger ni vilka faktiska val ni har gjort med avseende på modellval, hyperparameterar mm.

Varje diagram och tabell som presenteras i rapporten ska vara numrerat och ha en figur/tabelltext som kort beskriver innehållet. Avrunda till lämpligt antal decimaler.

- Ta inte med någon tabell eller diagram som ni inte kommenterar eller tolkar. Alla figurer, tabeller och liknande ska ska refereras från brödtexten.

5 Diskussion

Här ingår en kritisk granskning av resultaten (inte analys eller tolkning utan mer fokuserad på trovärdighet) samt en utvärdering av lämpligheten av de använda metoderna. Här görs även kopplingar tillbaka till tidigare studier som gåtts igenom i bakgrunden med referenser i texten.

En del av diskussion bör också försöka sätta resultatet i sitt sammanhang, vad kan samhället ta lärdom av dessa resultat, hur kan man fortsätta undersöka ämnet (t.ex. andra metoder eller andra faktorer).

Skriv kort om etiska överväganden kopplat till ert datamaterial och frågeställning. (etiska problem med data? etiska problem med er modell?)

Om vill kan man ha “Resultat och diskussion” som en gemensam sektion.

6 Slutsats

Detta kapitel sammanfattar det som presenterats i med huvudfokus på att besvara de frågeställningar som angivits i Inledning.

7 Referenser

- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.
- Zaki, Mohammed J, and Wagner Meira Jr. 2019. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press.

8 Bilaga

Slutligen ska ni även ha en bilaga. Här kan bifoga diagram eller tabeller som tolkas i rapporten men inte anses vara alltför viktiga, men som ni ändå vill ha med. Exempel kan vara:

- Figurer/tabeller som tar mycket plats
- Om ett flertal figurer är mycket lika varandra, visa då en under Resultat och de övriga här i bilagan.
- Beräkningar som tar mycket plats

Notera att alla tolkningar och analyser bör finnas med i brödtexten, ni kan referera till figurer/tabellerna som finns här.

8.1 Kod

Här ska ni ha relevant R-kod

- Visa de viktigaste funktionerna/anropen/inställningarna
- **Ta inte med all kod**

8.2 Redogörelse för bidrag

Här skriver ni en kort redogörelse för hur gruppens medlemmar har bidragit till arbetet och rapporten. Kan vara i form av en punktlista.

8.3 Generativ AI

Se den generella instruktionen för mer info.

Har generativ AI används?

Allmänt:

Sektion per sektion: