Besvara nedanstående teoretiska frågor koncist.

1. Lotta delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?

Tex om 3 olika modeller används så används träning för att köra "fit metoden" på alla modeller.
Sen predikteras y värden fram med hjälp av valideringsdatats X. där de predikterade y värdena tillsammans med valideringsdatans y värden används för att se hur bra modellen är.
Modellerna jämförs mellan varandra och en modell väljs som är bäst och antas ge bäst prediktioner.
Den valda modellen kan tränas om på träning och validerings datat.
Sen så körs en prediktering på den valda modellen med test datats X värde.
Det predikterade y värdet tillsammans med test datats y värde används nu för att se hur bra modellen är igen och om den håller lika bra som innan på valideringsdatat och inte är overfitted eller andra problem.
Varför den predikteras med både validering och test är för att säkerställa att inte den slumpmässigt valda datan för validering endast är bra för träningsdatan.
Omträning med träning och validering är för att modellen ska få bättre resultat och kan generalisera bättre på osedd data.

2. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding.

Ordinal encoding
Är när värden har inbördes mening mellan varandra. Tex om värdena låg, mellan och hög ska användas så skulle de kunna anvädas så här

0 = låg
1 = mellan
2 = hög
Det skulle tex kunna betyda ju högre värde desto "högre" upp.

one-hot encoding
i det här fallet skulle 3 kolumner i datat skapas. En för varje värde.
så om de ligger i den här ordningen i datat låg, mellan, hög så skulle varje värde representeras av att det finns en etta i dess kolumner

	låg	mellan	hög
låg =	1	0	0
mellan =	0	1	0
hög =	0	0	1

dummy variable encoding
Här tas en column bort och tex låg skulle kunna representeras av att alla kolumner har värdet noll.
Det skulle då se ut så här

	mellan	hög
låg =	0	0
mellan =	1	0
hög =	0	1

detta används för att undvika något som heter multicollinearity.

3. Göran påstår att datan antingen är ”ordinal” eller ”nominal”. Julia säger att detta måste
tolkas. Hon ger ett exempel med att färger såsom {röd, grön, blå} generellt sett inte har någon inbördes ordning (nominal) men om du har en röd skjorta så är du vackrast på festen (ordinal) – vem har rätt?

Båda har lite rätt. Datat kan vara ordinalt eller nominalt men det är också någonting som måste tolkas.
Det är vi som tar fram modeller som måste avgöra om det finns samband mellan värdena i datat.

För den som inte tycker om rött så är kanske personen med röd skjorta inte vakrast på festen.
Lite "context is king" på det här.

4. Läs följande länk: https://stackoverflow.com/questions/56107259/how-to-save-a-trainedmodel-by-scikit-learn (speciellt svaret från användaren som heter ”sentence”) som beskriver ”joblib” och ”pickle”.
Det är alltså ett sätt att spara modeller och innebär att man kan träna en modell och sedan
återanvända den för att göra prediktioner utan att behöva träna om modellen. Detta kommer ni ha nytta av om ni satsar på VG delen.
Svara på frågan: Vad används joblib och pickle till?

Båda används till att spara och ladda data men de har lite olika användningsområden beroende på den datan som ska sparas/laddas.
pickle är mer generell och kan spara de flesta datastrukturerna men fungerar sämre för större mängder data och kan ha lite säkerhetsbrister pga att den tillåter sån variation av data att kunna laddas in.
joblib är rekommenderad för ML data (scikit-learn) och kan hantera större mängder data bättre med kompression och optimering för numerisk data.

Modellera MNIST delen ligger i filen create_mnist_model_for_streamlit.ipynb
Streamlit appen ligger i filen app.py