Wine Quality Predictor

Gruppe 11
Marcus Rongved Molland
Maria Aleksandra Giske Haukeberg
Joakim Ekerhovd Kivijärvi

2.11.2025

1: BESKRIV PROBLEMET

OMFANG / *SCOPE*

Målet med prosjektet er å utarbeide en god og effektiv måte å finne ut av kvaliteten på vin. I løsningen vår trenger en bare å plotte inn informasjonen inn i nettsida og så vil den gi en omtrentlig prediksjon på vinens kvalitet. I dag er måten man bestemmer kvaliteten på vin ganske vagt. Ifølge en artikkel fra wineethusiast.com [1] skal man se på “Balance, Length, Intensity and Expressiveness og Complexity” for å avgjøre om kvaliteten på en vin er god.
Dette er som sagt ganske vagt, og dermed vil det å kunne finne kvaliteten på vin med en enkel nettside være enklere.

Målgruppen til prosjektet vårt er i hovedsak de som ikke har så mye vinkunnskap, men ønsker å lære og/eller bare ønsker å sikre seg en god vin. Den utfordrende delen av prosjektet vil nok være å hente inn informasjonen om vinen, men når det er gjort trenger en bare å plotte det inn på nettsiden og maskinlæringsmodellen vil gjøre magien sin!

METRIKKER

Vi trente en rekke modeller med ulike ensembler og regnet ut accuracy, recall og precision til alle. Basert på dette kom vi frem til at RandomForestRegressor gjorde det best i alle 3. Deretter regulerte vi dataen for å se om det ville bedre seg om vi tok ut noen av kolonnene, slik som var nevnt i datasettet [2] “Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.”. Vi kom dermed frem til at dersom vi tok ut kolonnene “fixed_acidity”, “density”, “citric_acid“ og “total_sulfur_dioxide” forbedret resultatene til modellen seg (og gjorde at brukeren ikke måtte legge inn like mye info).
Resultatet ble:
accuracy og recall med ca. 66.1% (Ser ut til å være like?), precision på 66.9% og en F1-score på 64.6%.

Slik som den er nå, vil appen være en suksess bare ved noen hundretalls tilfredsstilte brukere. Dette er fordi at appen ikke trenger mye for å opprettholdes og ikke krever mye maskinkraft.

2: DATA

Datasettet som brukes er Wine Quality (UCI Machine Learning Repository, ID:186) [Cortez et al., 2009]. Dette datasettet kombinerer to undersett: red wine med 1599 datapunkt og white wine med 4898 datapunkt, som til sammen blir 6497 datapunkt. Hver observasjon representerer en vinprøve fra Portugal, med 11 kjemiske målinger, samt en sensorisk vurdering av kvalitet gitt av vineksperter. Den sensoriske vurderingen er gitt som quality. I tillegg inneholder datasettet en kategorisk variabel color som angir vintype (red eller white).

Kolonner som fixed_acidity, density, citric_acid og total_sulfur_dioxide ble fjernet etter testing grunnet lav viktighetsgrad ifølge RandomForestRegressoren. Dessverre hadde det en liten negativ påvirkning på accuracy (rundt 2%).
Dataen ble delt i 80% trening og 20% testing.
Det er ikke nødvendig å ta hensyn til personvern, da det ikke finnes noen identifiserbar informasjon i datasettet. Heldigvis var det ingen mangel på data i dette datasettet, så vi måtte ikke ta hensyn til det.

3: MODELLERING

I koden prøvde vi mange forskjellige modeller: LinearRegression, AdaBoostRegressor, HistGradientBoostingRegressor, BaggingRegressor, RandomForestRegressor, StackingRegressor og VotingRegressor. I de to siste brukte vi de andre 5 modellene som input. Vi gjør deretter en test for hver av modellene som henter ut accuracy, precision, recall og F1-score. Vi endte opp med å bruke RandomForestRegressor siden den hadde best resultat overall, men StackingRegressor var en veldig nær andreplass.
Etter å ha søkt rundt på forskjellige sider virker det dessverre som om andre har gjort en bedre jobb med en ganske vanlig 80-90% accuracy. Dette kan være av flere grunner, men det er nok mest på grunn av mangel på erfaring. Vi fant “feature importance” ved å ta i bruk feature_importances_ variablen til RandomForestRegressor og lag dem inn i en stolpegraf i koden. Deretter tar vi ut mindre viktige features for å forbedre modellen + gjøre den mer brukervennlig (siden det blir mindre info å finne).

4: DEPLOYMENT

Modellen blir satt i drift ved bruk av Gradio, som lar oss launche en demo online som varer i en uke. Dessverre er det ikke veldig effektivt å lage en ny demo med en ny link hver uke, så den blir nok ikke oppe svært lenge.

5: REFERANSER

[1] Marshall Tilden III, A 4-Step Checklist to Assess the Quality of Wine, https://www.wineenthusiast.com/basics/how-to-taste/a-4-step-checklist-to-assess-the-quality-of-wine/?srsltid=AfmBOopmTubMXJ3MkOiW8FXF-bsIQTplzBXdTCMY0-6JPY2gz7P_ij0P
[2] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.