MATEMATIKCENTRUM | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2024-06-12

EXAMENSARBETE Diagnosis of Bloodstream Infections Using Machine Learning
STUDENTER Hector Jakobsson, Erik Rydengård
HANDLEDARE Ida Arvidsson, Johanna Engman, Gustav Torisson, Oskar Ljungquist
EXAMINATOR Alexandros Sopasakis

Diagnostisering av infektioner i blodbanan med hjälp av AI

POPULÄRVETENSKAPLIG SAMMANFATTNING **Hector Jakobsson, Erik Rydengård**

Varje år drabbas 1,200,000 människor av infektioner i blodbanan, bara i Europa. Blododlingar används för att diagnostisera dessa men metoden är både resurs- och tidskrävande. En datadriven metod för diagnostisering kan bidra till att fler patienter får rätt vård, samtidigt som belastningen på vården minskar.

Blododlingar är den vanligaste metoden för att diagnostisera infektioner i blodbanan. Det är dock en metod som är kantad av nackdelar. Den är tidskrävande på så sätt att det kan ta flera dagar att få tillbaka provsvaret. Detta är värdefull tid som går till spillo. De potentiellt ödesdigra konsekvenserna som en infektion kan medföra för patienten innebär att läkare ofta tar det säkra före det osäkra och sätter in antibiotika innan svaret från blododlingen har kommit tillbaka, vilket bidrar till ökad antibiotikaresistens.

För att adressera dessa problem har vi i det här examensarbetet använt maskininlärning för att kunna förutspå svaret från en blododling, baserat på tidigare angiven information. Vi kom fram till att man med hjälp av en modell kallad XGBoost kan minska antalet provtagningar med hela 29%. Att förlita sig helt och hållet på denna kan dock vara vanskligt, då den i cirka 1% av fallen missar ett positivt provsvar. Detta innebär att en sådan modell i nuläget framförallt borde ses som ett supplement till den befintliga arbetsmetoden, i form av ett hjälpmedel till den enskilde läkaren.

I vårt arbete utvärderades tre olika modeller, ovan nämnda XGBoost samt två typer av artificiella neurala nätverk (ANN). Kortfattat är XGBoost baserat på en samling av så kallade beslutsträd. Ett träd är uppbyggt av noder där dessa förgrenar sig baserat på beslut utifrån datan, och där löven, de sista noderna i trädet, ger de slutgiltiga utfallen. ANN är å sin sida strukturerade på ett sätt som efterliknar den mänskliga hjärnan, då de är uppbyggda av artificiella neuroner som förmedlar information sinsemellan.

Ett vanligt problem med medicinsk data är att vård är behovsdriven, alltså att prover eller mätningar bara görs om det finns skäl till det. Detta gör att tillgänglig information skiljer sig från patient till patient. För att hantera detta har vi valt att endast inkludera variabler som finns tillgängliga för åtminstone 20% av patienterna. Om en patient efter detta saknar värden måste dessa imputeras, vilket innebär att de estimeras utifrån andra patienters värden. Baserat på tester av olika metoder visade det sig att imputering med median och mest frekvent förekommande värden gav bäst resultat.

Avslutningvis tycker vi att det hade varit intressant att undersöka fler metoder för behandling av datan. Hur hade resultatet påverkats om urvalet av variabler hade begränsats ytterligare? Samtidigt går det att fråga sig om det finns andra modeller, eller rent av en kombination av flera, som är bättre lämpade för uppgiften. Detta och mycket annat lämnar vi till framtida forskning.