# Teknisk Rapport: Klassificering av hjärt- och kärlsjukdom

## Introduktion

### Syfte och Målsättning

Syftet med denna studie är att utforska och analysera olika maskininlärningsmetoder för att klassificera risken för hjärt- och kärlsjukdomar baserat på tillgängliga hälsodata. Fokus ligger på en djupgående dataanalys och en systematisk undersökning av olika prediktionsmodeller. Detta arbete syftar till att ge en fördjupad förståelse för datahantering och modellutveckling.

## Metod och Datahantering

### Dataset och Datakällor

Det dataset som används i denna studie innehåller 70 000 patientposter med ”features” relaterade till hjärt- och kärlsjukdomar. De insamlade variablerna inkluderar demografiska data, fysiska mått samt biokemiska och livsstilsrelaterade markörer. Datakvalitet har analyserats genom utforskande dataanalys (EDA) för att identifiera mönster, avvikelser och eventuella brister i datan.

För att förbättra datasetets användbarhet genomfördes flera förbehandlingssteg:

* Normalisering och standardisering av numeriska variabler.
* Hantering av saknade värden genom imputeringstekniker.
* Kategorisering av kontinuerliga variabler för att underlätta analys.
* Borttagning av outliers baserat på domänkunnande och statistiska metoder.

### Modellutveckling och Utvärdering

Två huvudsakliga maskininlärningsmetoder undersöktes:

1. **Random Forest** - En ensemblemetod som bygger på beslutsträd och ger god robusthet mot överanpassning.
2. **Logistisk Regression** - En linjär modell använd för att analysera samband mellan variabler och förutspå sannolikheten för hjärt- och kärlsjukdom.

Hyperparameteroptimering genomfördes med grid search för att förbättra modellernas prestanda. Modellernas noggrannhet, precision, recall och F1-score analyserades genom en femfaldig korsvalidering.

Vid sidan av de klassiska modellerna utforskades även content-based och collaborative filtering som alternativa metoder för att förstå och modellera patientdata. Dessa tekniker används vanligtvis för rekommendationssystem men undersöktes här för att avgöra om de kunde bidra till förbättrad riskprediktion.

## Resultat och Analys

### Modellprestanda

Random Forest presterade bäst av de undersökta metoderna med en accuracy på 73,2 % och en F1-score på 0,731. Analysen visade att de viktigaste variablerna för prediktion var ålder, blodtryck, BMI och kolesterolnivå.

Vidare experiment med collaborative filtering visade att denna metod kunde användas för att identifiera grupper av patienter med liknande riskprofiler. Detta indikerar att metoden kan vara användbar för att identifiera patienter som sannolikt har liknande sjukdomsrisker.

### Diskussion och Begränsningar

Denna studie visade att maskininlärning kan vara ett kraftfullt verktyg för att identifiera risker för hjärt- och kärlsjukdomar. Samtidigt har vissa begränsningar identifierats:

* Datasetets obalans i könsfördelning kan ha påverkat modellens generaliserbarhet.
* Vissa variabler kan vara proxies för andra underliggande faktorer, vilket kan leda till feltolkning av resultat.
* Metodernas användbarhet i klinisk praxis kräver vidare validering med externa dataset.

### Slutsatser

Denna rapport har undersökt olika maskininlärningsmodeller och deras potential för att klassificera patienter med hjärt- och kärlsjukdomar. Random Forest visade sig vara mest tillförlitlig, men experiment med collaborative filtering visade på intressanta alternativa ansatser. Vidare arbete bör inkludera större dataset, utökad hyperparameteroptimering och validering i verkliga kliniska miljöer.