Maskininlärning för riskbedömning av hjärt- och kärlsjukdomar



Jacob Andersson

EC Utbildning

Data Science Projekt

2024 - 11

# Abstract

Cardiovascular diseases (CVDs) are a leading global cause of mortality, emphasizing the need for early risk assessment.

This study leverages machine learning to predict CVD risk using health and lifestyle data, employing models such as Random Forest, XGBoost, Logistic Regression, and Support Vector Classifier (SVC). Ensuring balanced data significantly enhanced model performance, particularly in reducing false negatives to better identify individuals at risk. Random Forest achieved the highest AUC score of 0.87, demonstrating effective classification between healthy and at-risk individuals.

Key risk factors included BMI, dietary habits, alcohol consumption, smoking, health check-up frequency, and arthritis prevalence. Complementary clustering analysis, performed using HDBSCAN, identified 52 unique health profile clusters, providing insight into diverse risk profiles and supporting tailored preventive strategies.

While promising, this model serves as an educational tool and would require further validation for clinical application.

Innehållsförteckning

[Abstract 2](#_Toc181386109)

[1 Inledning 1](#_Toc181386110)

[2 Teori 2](#_Toc181386111)

[2.1 Prediktionsmodeller 2](#_Toc181386112)

[2.1.1 Logistisk regression 2](#_Toc181386113)

[2.1.2 Random Forest Classifier (RFC) 2](#_Toc181386114)

[2.1.3 Support Vector Machine (SVM) 2](#_Toc181386115)

[2.1.4 XGBoost 2](#_Toc181386116)

[2.2 Klustermodeller 3](#_Toc181386117)

[2.2.1 K-Means 3](#_Toc181386118)

[2.2.2 K-Prototypes 3](#_Toc181386119)

[2.2.3 HDBSCAN 3](#_Toc181386120)

[3 Metod 4](#_Toc181386121)

[3.1 Datainsamling 4](#_Toc181386122)

[3.2 Exploratory Data Analysis (EDA) 4](#_Toc181386123)

[3.2.1 Grundläggande statistiska analyser 4](#_Toc181386124)

[3.2.2 Visualisering av numeriska variabler 5](#_Toc181386125)

[3.2.3 Analys av kategoriska variabler 6](#_Toc181386126)

[3.3 Datatransformering och förbearbetning 7](#_Toc181386127)

[3.3.1 Omvandling av kategoriska variabler 7](#_Toc181386128)

[3.3.2 Hantering av obalans i datan 7](#_Toc181386129)

[3.4 Träning och utvärdering av prediktionsmodeller 7](#_Toc181386130)

[3.5 Utvärdering av kluster 7](#_Toc181386131)

[3.5.1 K-Means 8](#_Toc181386132)

[3.5.2 K-Prototypes 8](#_Toc181386133)

[3.5.3 HDBSCAN 8](#_Toc181386134)

[3.6 Streamlitapplikation 8](#_Toc181386135)

[3.6.1 Prediktion av hjärt- och kärlsjukdomar 8](#_Toc181386136)

[3.6.2 Klustring 8](#_Toc181386137)

[3.6.3 Exploratory Data Analysis (EDA): 8](#_Toc181386138)

[4 Resultat och Diskussion 9](#_Toc181386139)

[4.1 Resultat för Prediktiva Modeller 9](#_Toc181386140)

[4.1.1 Random Forest Classifier (RFC) 10](#_Toc181386141)

[4.1.2 XGBoost (XGB) 10](#_Toc181386142)

[4.1.3 Support Vector Classifier (SVC) 10](#_Toc181386143)

[4.1.4 Logistic Regression 10](#_Toc181386144)

[4.2 Resultat för Klusteranalys 11](#_Toc181386145)

[4.2.1 KMeans Klustring 11](#_Toc181386146)

[4.2.2 KPrototypes Klustring 11](#_Toc181386147)

[4.2.3 HDBSCAN Klustring 11](#_Toc181386148)

[4.2.4 Val av klustermodell 12](#_Toc181386149)

[4.3 Diskussion 13](#_Toc181386150)

[4.3.1 Diskussion prediktiva modeller 13](#_Toc181386151)

[4.3.2 Diskussion klusteranalys 14](#_Toc181386152)

[5 Slutsatser 16](#_Toc181386153)

[6 Självutvärdering 18](#_Toc181386154)

[Källförteckning 19](#_Toc181386155)

# Inledning

Hjärt- och kärlsjukdomar är en av de ledande orsakerna till dödsfall och allvarliga hälsoproblem världen över (Illustrerad Vetenskap, 2020). Trots medicinska framsteg fortsätter dessa sjukdomar att vara en stor utmaning för både individer och sjukvården. Att tidigt identifiera riskfaktorer och kunna förutse risken för hjärtproblem kan bidra till snabbare insatser och mer effektiv behandling. I takt med att datainsamling och analysmetoder har utvecklats, har möjligheterna att använda prediktionsmodeller inom sjukvården också ökat. Med hjälp av dessa modeller kan stora mängder data analyseras för att identifiera mönster och riskfaktorer som annars hade varit svåra att upptäcka.

Det här projektet syftar till att utveckla en modell som ska kunna förutsäga om en person löper risk för hjärt- och kärlsjukdom eller inte. Genom att analysera ett dataset med information om patienters hälsa och livsstilsfaktorer kan modellen identifiera vilka faktorer som har störst påverkan på risken för hjärtproblem. I samband med prediktionsmodellen genomförs även en mindre klusteranalys för att undersöka om patienter kan delas in i olika grupper baserat på hälsorelaterade variabler och därigenom ge en bättre förståelse för olika riskprofiler.

Frågeställningar som kommer undersökas i denna rapport är:

* **Hur säker är prediktionsmodellen?**
* **Vilka faktorer har störst påverkan på om en person har risk för hjärt- och kärlsjukdom**

# Teori

I det här projektet har olika maskininlärnings- och klustermodeller använts för att både förutse hjärt- och kärlsjukdomar samt att analysera kluster. Syftet med dessa modeller är att dels förutsäga risken för hjärt- och kärlsjukdomar samt identifiera grupperingar av personer med liknande hälsoprofiler. Detta avsnitt går igenom teorin bakom de metoder som används för de olika modellerna.

## Prediktionsmodeller

### Logistisk regression

Logistisk regression är en metod som ofta används för att uppskatta sannolikheten att en observation tillhör en viss klass, vilket gör den lämplig för binära klassificeringsuppgifter. I denna metod tilldelas observationen till den positiva klassen ("1") om sannolikheten överstiger 50 %, annars tilldelas den till den negativa klassen ("0"). Denna metod används ofta när målet är att klassificera datapunkter i en av två kategorier baserat på sannolikhetsuppskattningar.

### Random Forest Classifier (RFC)

Random Forest Classifier är en maskininlärningsmodell som bygger på en samling av beslutsträd. Den tränas oftast med en metod som kallas bagging, där varje träd tränas på ett slumpmässigt urval av träningsdatan. En viktig egenskap hos RandomForestClassifier är att den inför extra slumpmässighet när träden byggs upp genom att vid varje nod endast välja den bästa splitfunktionen från ett slumpmässigt urval av funktioner. Detta ökar variationen mellan träden, vilket leder till en mer robust modell som tenderar att generalisera bättre och minska överanpassning jämfört med enskilda beslutsträd.

### Support Vector Machine (SVM)

Support Vector Classifier (SVC) är en maskininlärningsmodell som används för att hitta den optimala gränsen (eller hyperplanet) som separerar olika klasser i datan. SVC fungerar genom att maximera marginalen mellan de närmaste datapunkterna från varje klass, vilket ger en klar separering mellan klasserna. Modellen kan anpassas till både linjära och icke-linjära separeringar med hjälp av olika kärnfunktioner (som linjär, polynomial, och RBF-kärna). SVC är särskilt effektiv när klasserna är svårseparerade och kan ge bra resultat även i komplexa och högdimensionella datamängder (många variabler för varje observation).

### XGBoost

XGBoost, som står för Extreme Gradient Boosting, är en optimerad version av gradient boosting-algoritmen, utvecklad för att vara snabbare, mer skalbar och resurseffektiv. Den är utformad för att fungera smidigt i Python och är kompatibel med Scikit-Learn, vilket gör den enkel att använda för dom som är bekanta med det ramverket.

En av dom praktiska funktionerna i XGBoost är dess förmåga att automatiskt avbryta träningen när förbättringarna av modellen avstannar (early stopping). Detta sparar både tid och datorkraft, eftersom träningen avslutas så fort ytterligare iterationer inte längre ger någon signifikant prestandaförbättring. Tack vare dessa egenskaper har XGBoost blivit ett populärt verktyg inom maskininlärning, särskilt för stora datamängder där effektivitet är avgörande.

## Klustermodeller

### K-Means

K-Means är en enkel och snabb klusteralgoritm som grupperar datapunkter baserat på deras närhet till centrumpunkter, eller "centroider". Algoritmen börjar med slumpmässiga positioner för varje centroid och justerar dem tills klustren stabiliseras. Antalet kluster (k) måste anges i förväg, algoritmen tilldelar sedan varje datapunkt till närmaste kluster. K-Means är särskilt effektiv för dataset med tydlig klusterstruktur och används ofta för sin snabbhet och enkelhet.

### K-Prototypes

K-Prototypes är en klusteralgoritm som kombinerar funktioner från både K-Means och K-Modes, vilket gör att den kan hantera dataset som innehåller både numeriska och kategoriska variabler. Algoritmen grupperar datapunkter baserat på en kombination av avståndsberäkningar för numeriska attribut och matchningsbaserade likheter för kategoriska attribut. Genom att anpassa beräkningsmetoderna för olika datatyper är K-Prototypes särskilt användbar för dataset som innehåller en blandning av numeriska och kategoriska variabler.

### HDBSCAN

HDBSCAN är en klustringsalgoritm som används för att identifiera grupper i stora, komplexa dataset. Den bygger på principen om densitetsbaserad klustring, där kluster definieras som områden med hög täthet av datapunkter som skiljs åt av glesare områden. Till skillnad från andra algoritmer, som k-means, kan HDBSCAN hitta kluster av godtyckliga former och hantera brus eller extremvärden i datan. En stor fördel med HDBSCAN är att den kan upptäcka kluster med olika tätheter genom att bygga en hierarki av kluster som sedan optimeras för att hitta de mest stabila grupperna. Detta gör HDBSCAN till ett flexibelt verktyg, särskilt användbart för data med varierande täthet och höga dimensioner.

# Metod

Detta avsnitt beskriver tillvägagångssättet för hantering, bearbetning och analysering av datan som användes för att bygga prediktionsmodellen för hjärt- och kärlsjukdomar. Datasetet kommer från en nationell undersökning och där 19 variabler har handplockats från 304 stycken.

## Datainsamling

Datan för detta projekt kommer ursprungligen från Behavioral Risk Factor Surveillance System (BRFSS), ett omfattande system i USA för telefonundersökningar som samlar in hälsoinformation om invånarnas riskbeteenden, kroniska sjukdomstillstånd och användning av förebyggande vårdtjänster. En förbehandlad version av datan fanns tillgängligt på Kaggle, där 19 av från början 304 variabler som relaterar till en persons livsstilsfaktorer och som kan bidra till risken för någon form av hjärt- och kärlsjukdom valdes ut.

## Exploratory Data Analysis (EDA)

EDA utfördes för att få en djupare förståelse för datans struktur och mönster. Följande steg genomfördes.

### Grundläggande statistiska analyser

Beskrivande statistik användes för att undersöka andelen för de olika variablerna. Här kan vi se hur det är fördelat för några av dem.

En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning En bild som visar text, Teckensnitt, skärmbild, vit

Automatiskt genererad beskrivning  
*Figur 1: Visar fördelning för hudcancer Figur 2: 10 Visar fördelning för kön*

En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning  
*Figur: 3 Visar fördelning för andra typer av cancer Figur: 4 Visar fördelning för hjärtsjukdom*

### Visualisering av numeriska variabler

Histogram och låddiagram skapades för att identifiera fördelningar och eventuella extremvärden i datan. Här nedan figurer visualiseras fördelning och spridning för olika variabler.

En bild som visar text, diagram, skärmbild, Graf

Automatiskt genererad beskrivning  
*Figur 5: Har kan vi se hur dom olika numeriska variablerna är fördelade där x-axeln visar värdet och y-axeln visar antalet observationer*

*En bild som visar skärmbild, text, diagram, linje

Automatiskt genererad beskrivning En bild som visar skärmbild, text, diagram, linje

Automatiskt genererad beskrivning  
Figur 6: Visar spridningen för BMI Figur 7: Visar spridningen för Vikt*

### Analys av kategoriska variabler

I figurerna i avsnitt 3.2.1 ”Grundläggande statistiska analyser” ser vi några figurer som visar fördelningen mellan variablerna. För detta projekt används också Stapeldiagram för kategoriska variabler för att visualisera extra.

En korrelationsmatris skapades för att identifiera samband mellan variabler, särskilt mellan riskfaktorer och förekomsten av hjärt-kärlsjukdom. Variabler med starka korrelationer identifierades.

EDA-analysen gav värdefulla insikter om datans struktur och relationer mellan variabler.  
  
En bild som visar text, skärmbild, Graf, nummer

Automatiskt genererad beskrivning  
*Figur 8: Visar förhållandet mellan olika åldersgrupper och hjärt- och kärlsjukdomar*  
  
En bild som visar text, skärmbild, kvadrat, Rektangel

Automatiskt genererad beskrivning  
*Figur 9: Visar en korrelationsmatris mellan olika variabler*

## Datatransformering och förbearbetning

För att förbereda datasetet för analys och modellträning genomfördes flera steg.

### Omvandling av kategoriska variabler

Alla kategoriska variabler i datasetet omvandlades till numeriska format med hjälp av dummy-variabler. Detta steg är avgörande för att göra data kompatibel med maskininlärningsmodeller, som ofta kräver numerisk indata. I processen skapades en binär variabel för varje kategori, vilket möjliggör korrekt hantering av de ursprungliga kategoriska värdena i analysen.

### Hantering av obalans i datan

Datasetet innehöll en tydlig obalans mellan antalet individer med och utan hjärt- och kärlsjukdomar, där friska individer var i majoritet. För att minska risken för att modellen blir partisk mot den vanligaste klassen, användes två olika tekniker för att balansera datasetet.

#### Slumpmässig underrepresentation

För att reducera antalet friska individer valdes ett slumpmässigt urval av datapunkter från den dominerande klassen (friska individer). Målantalet för denna underrepresentation sattes till 1,5 gånger antalet individer med hjärt-kärlsjukdom. Detta skapade en bättre balans i datasetet mellan friska och sjuka individer.

#### SMOTENC

Den andra metoden som testades var SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous data), en metod som skapar syntetiska exempel för den underrepresenterade klassen (individer med hjärt-kärlsjukdom). SMOTENC är särskilt anpassad för dataset som innehåller både numeriska och kategoriska variabler. Detta steg säkerställde att klasserna blev mer jämnt fördelade och skapade bättre förutsättningar för modellen att lära sig från båda klasserna.

## Träning och utvärdering av prediktionsmodeller

Efter förbearbetning tränades flera prediktionsmodeller för att uppskatta risken för hjärt- och kärlsjukdomar. Modellerna som testades var Logistisk Regression, Random Forest Classifier, Support Vector Machine (SVM) och XGBoost. Varje modell tränades på både balanserade och obalanserade data.

Modellerna utvärderades och jämfördes för att hitta den mest pålitliga modellen för prediktion av hjärtproblem.

## Utvärdering av kluster

För att identifiera och analysera grupperingar användes tre klusteralgoritmer: K-Means, K-Prototypes och HDBSCAN. Klusteranalysen syftade till att identifiera och förstå de underliggande strukturerna i datan.

Varje algoritms prestanda utvärderades med hjälp av silhouette score för att mäta klustrens sammanhållning och separering. Silhouette score användes för att bedöma kvalitén hos klustren och gav en indikation på vilken klusterstruktur som bäst representerade datasetet. Trots att vissa inställningar resulterade i högre silhouette score med endast tre kluster, valdes en annan modell med fler kluster för att kunna göra en djupare och mer nyanserad analys av grupperna.

### K-Means

För K-Means testades olika antal kluster och justerade parametern k för att få en rimlig klusterindelning som uppfyllde både kravet på tydlig separation och önskan om fler kluster. Genom att undersöka silhouette-värden för olika klusterantal hittades den bästa K-Means modellen.

### K-Prototypes

K-Prototypes testades för att hantera datasetet som innehöll både numeriska och kategoriska variabler. Genom att justera antalet kluster och testa olika inställningar kunde den bäst presterande modellen utvärderas. K-Prototypes tillåter att kombinera både numeriska och kategoriska variabler vilket passar datan.

### HDBSCAN

HDBSCAN användes för att identifiera kluster av varierande täthet, samt för att hantera extremvärden som inte passade in i de övriga klustren. Eftersom HDBSCAN inte kräver ett förbestämt antal kluster, kunde identifiering av kluster ske baserat på naturliga täthetsvariationer i datasetet. Det gjorde det möjligt att inkludera kluster av olika storlekar och former. Här användes Silhouette Score för att analysera den bäst presterande modellen.

## Streamlitapplikation

En Streamlitapplikation skapades för att underlätta visualisering, prediktion och analys av riskfaktorer för hjärt- och kärlsjukdomar samt klusterinsikter för individer med olika hälsoprofiler. Applikationen består av tre huvudsakliga sektioner som erbjuder specifika funktioner:

### Prediktion av hjärt- och kärlsjukdomar

Användare kan mata in personlig hälsoinformation för att få en prediktion om eventuell risk för hjärt- och kärlsjukdom.

### Klustring

Visualiserar förtränade klustringsresultat, vilket gör det möjligt för användare att utforska indelade hälsoprofiler och få insikter i de olika klustrens egenskaper.

### Exploratory Data Analysis (EDA):

Erbjuder ett omfattande verktyg för EDA, där datasetets struktur, statistiska sammanfattningar, korrelationsanalyser och relationer mellan kategoriska variabler och sjukdomsutfall kan undersökas.

# Resultat och Diskussion

I detta avsnitt presenteras och analyseras resultaten från de olika prediktiva modeller som användes för att förutsäga risken för hjärt- och kärlsjukdomar. Här utvärderas modellernas prestanda med hjälp av precision, recall, F1-score och ROC-kurvor för att bedöma hur väl modellerna kan skilja mellan friska och sjuka individer.

## Resultat för Prediktiva Modeller

Tabellen sammanfattar resultaten för de olika prediktiva modellerna, där "True" representerar den positiva klassen (sjuka) och "False" den negativa klassen (friska).

En bild som visar text, skärmbild, programvara, Datorikon

Automatiskt genererad beskrivning En bild som visar text, skärmbild, programvara, Datorikon

Automatiskt genererad beskrivning  
*Figur 10: Visar Classification report för RFC modellen Figur 11: Visar Classification report för XGBoost modellen*

En bild som visar text, skärmbild, programvara, Datorikon

Automatiskt genererad beskrivning En bild som visar text, skärmbild, programvara, Datorikon

Automatiskt genererad beskrivning  
*Figur 12: Visar Classification report för SVC modellen Visar Classification report för LR modellen*

### Random Forest Classifier (RFC)

Random Forest Classifier (RFC) visar sig ha stark prestanda för detta klassificeringsproblem. Modellen uppnår hög precision och recall för den negativa klassen (friska individer), vilket innebär att den klassificerar friska korrekt med stor noggrannhet. För den positiva klassen (individer med hjärt- och kärlsjukdom) har RFC en recall på 0,92, vilket innebär att den fångar majoriteten av dessa fall och minimerar risken för falskt negativa resultat. Dock är precisionen för den positiva klassen relativt låg (0,31), vilket innebär att en betydande andel av de individer som klassificeras som sjuka i själva verket är friska.

### XGBoost (XGB)

XGBoost-modellen presterar väl, men något sämre än RFC när det gäller att identifiera individer med hjärt- och kärlsjukdom. För den positiva klassen har XGBoost en recall på 0,75, vilket är lägre än RFC och innebär att modellen missar fler sjuka individer, vilket kan leda till fler falskt negativa resultat. Modellen har relativt hög precision och recall för den negativa klassen, vilket ger ett balanserat resultat för den friska gruppen.

### Support Vector Classifier (SVC)

Support Vector Classifier (SVC) presterar liknande XGBoost, med hög precision och recall för den negativa klassen, men svagare prestanda för att korrekt identifiera sjuka individer. SVC

recall för den positiva klassen är 0,72, vilket innebär att modellen riskerar att missa ett antal individer med hjärt- och kärlsjukdom och därmed producera fler falskt negativa än RFC.

### Logistic Regression

Logistic Regression-modellen visar den lägsta prestandan för den positiva klassen, med en recall på 0,68. Detta innebär att modellen har svårare att korrekt identifiera individer med hjärt- och kärlsjukdom och därmed tenderar att producera fler falskt negativa än både SVC och XGBoost. Modellen uppnår dock hög precision och recall för den negativa klassen, vilket innebär att den effektivt klassificerar friska individer korrekt.

En bild som visar text, skärmbild, linje, diagram

Automatiskt genererad beskrivning  
*Figur 15: Visar hur bra dom olika modellerna presterade*

## Resultat för Klusteranalys

Tabellen nedan sammanfattar de genomsnittliga Silhouette Score för de olika klustringsmetoderna (KMeans, KPrototypes, och HDBSCAN) och antal kluster.

|  |  |  |
| --- | --- | --- |
| Klustringsmetod | Antal Kluster | Genomsnittligt Silhouette score |
| KMeans | 3 | 0,2769 |
| KMeans | 5 | 0,2208 |
| KMeans | 10 | 0,1814 |
| KMeans | 15 | 0,1834 |
| KMeans | 20 | 0,1835 |
| KMeans | 30 | 0,1752 |
| KMeans | 35 | 0,1768 |
| KPrototypes | 3 | 0,2769 |
| KPrototypes | 5 | 0,2109 |
| KPrototypes | 10 | 0,1833 |
| KPrototypes | 15 | 0,1770 |
| KPrototypes | 20 | 0,1737 |
| KPrototypes | 30 | 0,1714 |
| KPrototypes | 35 | 0,1805 |
| HDBSCAN | 52 | 0,2749 |

### KMeans Klustring

KMeans-klustringsmetoden testades med olika antal kluster (från 3 till 35). Resultaten visar att genomsnittliga silhouette-värden är högst vid 3 kluster, med ett värde på 0,2769, vilket tyder på att datan är mest sammanhållen och väl separerad med denna klusterindelning. När antal kluster ökar minskar silhouette-värdet, vilket indikerar att klustrens sammanhållning blir svagare och separationen mellan klustren försämras.

### KPrototypes Klustring

KPrototypes-klustringsmetoden användes också med samma antal kluster (3 till 35). Resultaten liknar KMeans, med det högsta genomsnittliga silhouette-värdet vid 3 kluster 0,2769. Detta indikerar att KPrototypes också finner 3 kluster som den optimala indelningen för detta dataset. Silhouette-värdet sjunker när antalet kluster ökar, vilket tyder på att fler kluster ger mindre sammanhängande grupper.

### HDBSCAN Klustring

HDBSCAN-klustringsmetoden upptäckte 52 kluster med ett genomsnittligt silhouette-värde på 0,2749. Denna metod är särskilt användbar för att hantera brus i datan, och den identifierade 2178 datapunkter som brus. HDBSCAN presterade bäst med följande kluster baserat på deras genomsnittliga silhouette-värden:

### Val av klustermodell

De bästa modellera presterade väldigt lika, men eftersom både K-Means och K-Prototypes bara hade 3 kluster valdes HDBSCAN-modellen i stället. Anledningen till detta var för att kunna analysera flera kluster

#### Kluster för analys

Nedanför ser vi dom bäst presterande klustren. Dom som presterade bäst hade väldigt få antal datapunkter vilket kan vara anledningen till dom fick så högt silhouette score. Kluster 26 är ett av de klustren som valdes att göra en analys på, eftersom den innehåller fler datapunkter med ett relativ bra silhouette score.

Top 5 kluster baserat på genomsnittlig silhouette score:

|  |  |  |
| --- | --- | --- |
| Kluster | Silhouette score | Antal Datapunkter |
| 49 | 0,8254 | 14 |
| 42 | 0,8073 | 6 |
| 33 | 0,7780 | 6 |
| 30 | 0,7581 | 6 |
| 44 | 0,6922 | 7 |

Top 5 kluster baserat på genomsnittlig silhouette score för kluster med 20 eller fler datapunkter:

|  |  |  |
| --- | --- | --- |
| Kluster | Silhouette score | Antal Datapunkter |
| 26 | 0,4120 | 34 |
| 45 | 0,3371 | 32 |
| 40 | 0,3203 | 28 |
| 16 | 0,3122 | 51 |
| 48 | 0,2826 | 301 591 |

#### Analys av kluster 26

**ANOVA, F-test Resultat:**

* De numeriska variablerna som längd (Height\_(cm)), vikt (Weight\_(kg)), BMI, alkoholkonsumtion (Alcohol Consumption), fruktkonsumtion (Fruit Consumption), grönsakskonsumtion (Green Vegetables Consumption) och friterad potatiskonsumtion (Fried Potato Consumption) visar alla mycket höga F-statistikvärden och extremt låga p-värden.
* Dessa stora F-statistikvärden indikerar att individerna i kluster 26 har tydligt avvikande mönster när det gäller fysiska attribut och kostvanor jämfört med andra kluster. Detta kan tyda på specifika livsstils- eller näringstrender inom denna grupp, som potentiellt kan kopplas till hälsorisker eller unika beteendemönster.

**Chi-Squared Test Resultat:**

* **Hälsokontroller de senaste 2 åren:** Detta har ett p-värde på 0,0203, vilket kan tyda på en statistiskt signifikant koppling mellan frekvensen av hälsokontroller och kluster 26. Detta kan innebära att individer i klustret prioriterar regelbundna hälsokontroller, eventuellt på grund av medvetenhet om sina hälsotillstånd.
* **Artrit (Arthritis\_Yes):** Med ett p-värde på 0,0016 är artritstatusen starkt associerad med kluster 26, vilket kan indikera en högre förekomst av artrit inom denna grupp. Detta kan vara ett tecken på att vissa fysiska och livsstilsfaktorer i klustret korrelerar med artrit.
* **Rökvanor (Smoking History\_Yes):** Med ett p-värde på 0,028 finns en signifikant koppling mellan rökvanor och kluster 26. Detta innebär att rökning kan vara vanligare i detta kluster eller att det har en unik relation inom gruppen, vilket kan vara relevant för deras hälsoprofil.

**Sammanfattning:**

* Kluster 26 utmärker sig från de andra klustren genom unika karakteristika, särskilt när det gäller fysiska attribut, kostvanor och specifika hälsotillstånd eller beteenden som regelbundna hälsokontroller, artrit och rökvanor.
* Denna distinkta profil kan indikera en grupp med särskilda hälsorisker eller beteenden som påverkar deras hälsostatus. Kostvanorna och den högre förekomsten av rökning och artrit kan ha en betydande inverkan på hälsan, och detta kluster kan vara av intresse för ytterligare analys.

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning  
 *Figur 16: Visar signifikanta variabler i kluster 26*

## Diskussion

### Diskussion prediktiva modeller

Dom olika prediktionsmodellerna visar tydliga skillnader i sin förmåga att korrekt identifiera individer med hjärt- och kärlsjukdom, vilket understryker vikten av att välja modell utifrån det specifika syftet. Random Forest Classifier (RFC) utmärker sig som den starkaste modellen för att identifiera sjuka individer tack vare dess höga recall för den positiva klassen (0,92). Denna egenskap gör RFC särskilt användbar när det är viktigt att minimera antalet missade fall av hjärt- och kärlsjukdom, även om modellen tenderar att klassificera några friska individer som sjuka på grund av dess lägre precision för den positiva klassen.

XGBoost (XGB) presterar något sämre än RFC när det gäller recall för den positiva klassen, med en recall på 0,75, vilket innebär en större risk för falska negativa jämfört med RFC. Samtidigt uppvisar XGBoost en hög precision och recall för den negativa klassen, vilket ger ett mer balanserat resultat. Denna balans gör XGBoost till ett bra alternativ när både korrekt identifiering av sjuka och en minimering av falska positiva är viktiga men där maximal identifiering av sjuka individer inte är lika avgörande.

Support Vector Classifier (SVC) ger resultat som liknar XGBoost men med något lägre recall (0,72) för den positiva klassen. Detta innebär att SVC riskerar att missa fler individer med hjärt- och kärlsjukdom, vilket begränsar dess användbarhet i sammanhang där korrekt identifiering av högriskindivider är kritiskt. SVC kan dock övervägas i scenarier där en god precision och recall för friska individer efterfrågas samtidigt som balans mellan precision och recall prioriteras.

Logistic Regression uppvisar den lägsta recall för den positiva klassen (0,68), vilket innebär en större risk att missa sjuka individer än både SVC och XGBoost. Modellen har dock hög precision och recall för den negativa klassen, vilket gör den lämplig för scenarier där korrekt identifiering av friska individer är mer kritiskt än att identifiera varje enskilt fall av hjärt- och kärlsjukdom.

Sammanfattningsvis visar analysen att RFC är den mest effektiva modellen för att identifiera individer med hjärt- och kärlsjukdom, och den är därmed det bästa valet i situationer där maximal identifiering av sjuka är avgörande. XGBoost och SVC fungerar bra som alternativ i sammanhang där både precision och recall för båda klasserna är viktiga. Logistic Regression är bättre lämpad för scenarier där korrekt klassificering av friska individer prioriteras över att upptäcka alla fall av hjärt- och kärlsjukdom.

### Diskussion klusteranalys

Klusteranalysen bidrar med viktiga insikter i datauppsättningens struktur och de mönster som finns bland individer med liknande hälsoprofiler. Både KMeans och KPrototypes identifierade tre kluster som den optimala indelningen för detta dataset, vilket antyder att dessa metoder är effektiva på att fånga enhetliga grupperingar med hög intern sammanhållning och god separation. Denna treklustersstruktur kan representera olika hälsoprofiler som delas av individer med liknande riskfaktorer eller livsstil.

HDBSCAN-modellen upptäckte ett mycket större antal kluster, vilket antyder en högre grad av variation och mångfald i datasetet. Denna metod lyckades även hantera brus och identifierade 2178 datapunkter som extremvärden, vilket visar att HDBSCAN är särskilt användbar när datasetet har komplexa strukturer och kluster av varierande storlek och täthet.

Vid analys av kluster med höga Silhouette Score observerades att många av dessa grupper hade få datapunkter, vilket kan förklara deras höga värden. Dock valdes kluster med fler datapunkter, exempelvis kluster 26, för vidare analys, eftersom ett större kluster med ett acceptabelt Silhouette Score kan bidra till en bättre analys. Kluster 26 visade sig ha unika egenskaper i fråga om fysiska attribut, kostvanor och specifika hälsotillstånd som artrit och rökvanor, vilket gör denna grupp särskilt intressant för vidare studier av samband mellan livsstil och hälsoprofil.

Sammanfattningsvis visar klusteranalysen att HDBSCAN är särskilt värdefull för att hantera datauppsättningar med varierande klusterstrukturer och kan fånga upp grupper med specifika hälsoprofilmönster, vilket kan vara av särskilt intresse vid riskbedömningar och förebyggande insatser.

# Slutsatser

Projektets syfte var att utveckla en prediktionsmodell för att förutse risken för hjärt- och kärlsjukdomar baserat på ett dataset som inkluderar hälsorelaterade och livsstilsfaktorer. Genom att använda prediktionsmodeller och klusteralgoritmer blev det möjligt att både förutsäga om en person har risk för hjärt- och kärlsjukdom och identifiera grupper av individer med liknande riskprofiler. Resultaten från analyser visar att databalans är avgörande för att förbättra modellernas prestanda, särskilt när det gäller att minska risken för falskt negativa resultat. Random Forest och XGBoost presterade bäst av de testade modellerna, medan klusteranalysen med HDBSCAN gav värdefulla insikter i datans struktur och identifierade grupper med unika hälsoprofiler.

**Hur säker är prediktionsmodellen?**

Prediktionsmodellen förbättrades märkbart när datasetet balanserades. Random Forest-modellen presterade bäst med ett AUC-värde på 0,87, vilket visar att modellen effektivt kan skilja mellan friska och sjuka individer. Att balansera data bidrog till att modellen kunde identifiera fler individer med hjärt- och kärlsjukdom, vilket är viktigt i sammanhang där det är avgörande att fånga upp så många personer i riskzonen som möjligt. Random Forest framstår som en stabil och välfungerande modell för detta ändamål.

**Vilka faktorer har störst påverkan på om en person har hjärtproblem?**

Genom analysen identifierades flera faktorer som har en stark koppling till risken för hjärt- och kärlsjukdomar. De mest påverkande variablerna inkluderar:

**BMI och vikt:** Dessa variabler är direkt relaterade till individens kroppskomposition. BMI Kan dock vara missvisande, eftersom måttet bara är baserat på förhållandet mellan längd och vikt och tar inte hänsyn till om vikten kommer från exempelvis muskler eller fett. Generellt så indikerar ett högt BMI övervikt, vilket i sin tur kan ha påverkan för hjärt- och kärlsjukdom.

**Alkoholkonsumtion och kostvanor:** Variabler som alkoholkonsumtion, fruktkonsumtion och konsumtion av grönsaker och friterad potatis visade sig ha betydande inverkan. Speciellt en kost rik på frukt och grönsaker är vanligtvis associerad med lägre risk för hjärtsjukdomar, medan överdriven konsumtion av fet och friterad mat kan öka risken.

**Hälsokontroller:** Frekvensen av hälsokontroller är en viktig faktor. Individer som regelbundet genomgår hälsokontroller har ofta en högre medvetenhet om sina hälsorisker, vilket kan påverka deras förebyggande åtgärder och därmed minska risken.

**Rökvanor:** Rökning är en välkänd riskfaktor för hjärt- och kärlsjukdomar och bidrar till flera hälsoproblem som påverkar hjärtat och blodkärlen.

**Artrit:** Förekomsten av artrit kan ha en indirekt koppling till hjärtsjukdomar, då kronisk inflammation ofta är förknippad med en ökad risk för hjärtproblem (Dagens Medicin, 2008).

Sammanfattningsvis har en Random Forest Classifier utvecklats för att uppskatta risken för hjärt- och kärlsjukdomar. Även om modellen visar lovande resultat i att identifiera riskmönster, är den långt ifrån perfekt och lämnar utrymme för förbättring vad gäller noggrannhet och tillförlitlighet. Modellen är därför inte lämplig för kliniskt bruk utan tjänar snarare som ett verktyg för inlärning inom ramen för denna rapport.

Klusteranalysen kompletterade prediktionsmodellen genom att ge insikter om hur individer med liknande hälsoprofiler kan grupperas. Denna information kan potentiellt användas för att utveckla riktade förebyggande åtgärder och individanpassad rådgivning, vilket på sikt kan förbättra hälsoutfall för riskgrupper.

Denna modell och analys är dock inte avsedd för praktiskt användande inom vården utan kräver ytterligare utveckling och validering för att kunna betraktas som en pålitlig och användbar lösning.

# Självutvärdering

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.**

Att välja ett ämne att jobba med samt vilka metoder man skulle använda. Vi kom fram till att vi ville testa på klustring eftersom vi skall jobba med det på vår LIA

Projektet blev lite för omfattande med både klustring och prediktion, vi delade upp uppgifterna mellan oss.

1. Vilket betyg du anser att du skall ha och varför.

Lite osäker, Rapporten hade kunnat vara med utförlig

1. Något du vill lyfta fram till Antonio?

Lärorikt med ett annat upplägg än det vi har haft tidigare. Genom att tänka ut ett projekt själv känns det som att man utvecklas mer

# Källförteckning

Illustrerad Vetenskap. (2020). *De vanligaste dödsorsakerna i världen*. Publicerad den 06 mars 2020. Text: Redaktionen. <https://illvet.se/medicin/sjukdomar/topp-5-de-vanligaste-dodsorsakerna>

Dagens Medicin. (2008). Reumatiker löper större risk att få hjärt-kärlsjukdom. Publicerad: 29 oktober 2008, 16:01. <https://www.dagensmedicin.se/specialistomraden/hjarta-karl/reumatiker-loper-storre-risk-att-fa-hjart-karlsjukdom/>