P1 - Klassificering af Kardiovaskulær sygdom

Daniel Kristensen[1], Lanja Mozafar Khorshid[2], Mathilde Vibæk[3], Mikkel Stouby Holm[4], Salar Mahmoud[5], and Tobias Reinholt[6]

[1]**dkrist21@student.aau.dk**
[2]**lkhors24@student.aau.dk**
[3]**mvibak24@student.aau.dk**
[4]**mholm24@student.aau.dk**
[5]**smahmo24@student.aau.dk**
[6]**treinh24@student.aau.dk**

ABSTRACT

This project aims to implement and test different machine learning methods for the development of a model to classify whether patients have or don't have cardiovascular disease (CVD). The intent is to combine the model with a software product for general practitioners to use. The purpose of this product is to reduce mortality and morbidity by recommending certain patients for further assessment. It was found that certain risk factors are used to evaluate the risk of CVD. This was important in deciding which features from the dataset to include in the model. Initially three classifiers where chosen through a Quick and Dirty process where a total of six classifiers were tested, with special attention to recall and F1. SVM was implemented and tested in this project along with Random Forest and Logistic Regression. They were initially tested separately to find their optimal hyperparameters before being combined in an ensemble model. Soft voting was used and the hyperparameters were once again tuned to maximize the collective performance for the ensemble. The training showed, that the ensemble model performed the best with an accuracy of 81%, precision of 84%, recall of 84% and F1 of 83%. The optimal decision threshold was then estimated for the ensemble model at 0.376 resulting in an accuracy of 82%, precision of 77%, recall of 96% and F1 of 85%. Due to lack of time, this project did not succeed in creating a user interface for receiving user input in the form of medical data. The project fulfilled the requirement of a minimum recall of 95%, but did not achieve the required minimum accuracy of 90%. Future work on the project would focus on improving accuracy and the creation of a software product to receive user input.

Keywords:

1 PREFACE

Denne rapport er udarbejdet i forbindelse med semesterprojekt 1 på uddannelsen Design og Anvendelse af Kunstig Intelligens på Aalborg Universitet. Følgende projekt har til formål at reducere mortalitet og morbiditet. Dette undersøges, ved anvendelsen af machine learning metoder, som kan klassificere, hvorvidt individer har kardiovaskulære sygdomme (CVD). Til dette vil der blive udviklet et softwareprodukt, der ved brug i almen praksis, kan forudsige CVD hos patienter og derved reducere dødelighed (mortalitet) og sygdomsgrad (morbiditet). Til dette formål anvendes supervised learning-algortimer på et datasæt fra Kaggle.

2 INTRODUKTION

Kardiovaskulære sygdomme udgør en betydelig sundhedsbyrde både på globalt plan og i Danmark og er kendetegnet ved høj dødelighed og omfattende økonomiske konsekvenser. Selvom mange af disse sygdomme kan forbygges, forbliver det en af de førende dødsårsager, hvilket understreger behovet for forbedrede værktøjer til tidlig opsporing og diagnosticering for at mindske CVDs indvirkning. Den øgede tilgængelighed af sundhedsdata, kombineret med fremskridt inden for kunstig intelligens, introducerer nye muligheder for at forbedre prædiktion og behandling af sygdomme som CVD.

Rapporten udforsker både de kliniske aspekter af CVD, såsom risikofaktorer og eksisterende metoder til vurdering af CVD-risiko.

Formålet med dette projekt er at reducere mortalitet og morbiditet ved at muliggøre mere præcis og rettidig diagnosticering af CVD. Dette undersøges gennem anvendelsen af machine learning til forbedret diagnosticering samt udvikling af en robust og fleksibel softwareløsning til brug i almen praksis. Rapporten adresserer samtidig de udfordringer og begrænsninger, der skal overvindes for at kunne anbefale, om patienter bør henvises til yderligere undersøgelse.

3 PROBLEMANALYSE

For at forstå, hvordan machine learning kan forbedre diagnosticering af kardiovaskulære sygdomme, er det vigtigt først at afdække omfanget af problemet. I det følgende afsnit analyseres den aktuelle situation, herunder CVD's globale og nationale betydning, økonomiske konsekvenser, sygdommens definition og de væsentligste risikofaktorer, som ligger til grund for udviklingen af CVD.

3.1 Baggrunds research

Kardiovaskulære sygdomme er en af de førende dødsårsager på verdensplan. Næsten hver tredje person dør som følge af CVD på globalt plan, hvilket overstiger dødeligheden for kræft og lungesygdomme. I Danmark dør ca. 20% af CVD. I 2021 døde over 20 millioner mennesker af en CVD globalt, herunder ca. 933.000 (150,8 pr. 100.000) og 13.300 (99,1 pr. 100.000) i hhv. USA og Danmark[5],[13] [20].

Ligeledes er den økonomiske byrde for CVD i sundhedsvæsenet stor, med udgifter opdelt i direkte (indlæggelser og medicin) og indirekte omkostninger (tabt produktivitet, uarbejdsdygtighed, handicap og for tidlig død). Ifølge The American Heart Association var de direkte omkostninger forbundet med CVD i sundhedsvæsnet i 2020 omkring $393 milliarder i USA. Desuden opsøgte 1 ud af 3 voksne amerikanere sundhedsvæsnet på grund af enten CVD eller kendte risikofaktorer hertil (se afsnit Risikofaktorer), hvor det også blev oplyst, at udgiften til behandling af patienter for risikofaktorer udgjorde $400 milliarder. Produktivitetstabet for USA, forårsaget af CVD, var i 2020 på $234 milliarder. The American Heart Association estimerer at udgifterne til CVD i 2050 vil være næsten firdoblet og produktivitetstabet øget med 54% [5]. I Danmark, hvor ca. 65.700 årligt rammes af CVD, er de økonomiske konsekvenser for samfundet også betydelige. Ifølge beregninger lavet for Hjerteforeningen i 2017 kostede indlæggelser på grund af CVD den danske stat næsten 5,3 milliarder kr. og medicinudgiften til CVD-patienter beløb sig til 1,81 milliarder kr. Danmarks tabte produktivitet som følge af CVD var i 2017 estimeret til 1,84 milliarder kr. [20].

3.1.1 Definition af kardiovaskulær sygdom

CVD dækker over en bred gruppe af sygdomme, der påvirker hjertet og blodkarrene. Blandt de mest almindelige er koronar-, cerebrovaskulær- og perifer arteriel sygdom, som forårsager problemer i blodkarrene til hjertet, hjernen og andre organer. Reumatisk hjertesygdom og medfødte hjertefejl udgør også betydelige typer af CVD, som skyldes hhv. reumatisk feber og medfødte misdannelser i hjertets anatomi. En væsentlig faktor i udviklingen af disse CVD er åreforkalkning, hvor aflejringer ophobes i arterievæggene og kan føre til alvorlige hændelser som hjerteanfald og slagtilfælde. Desuden kan tilstande som dyb venetrombose og lungeemboli opstå, når blodpropper fra benene bevæger sig til hjertet eller lungerne [32].

Andre væsentlige CVD-tilstande omfatter hjerteinsufficiens, hvor hjertet har nedsat pumpefunktion, samt hjerterytmeforstyrrelser, hvor hjertets elektriske aktivitet er forstyrret. Hjerteklapsygdomme påvirker desuden blodstrømmen, når klapperne ikke åbner og lukker korrekt. Hjerteanfald og slagtilfælde er for mange de første synlige tegn på CVD og skyldes pludselige blokeringer i blodtilførslen til hjerte eller hjerne, som kræver øjeblikkelig behandling. Et iskæmisk slagtilfælde opstår, når et blodkar til hjernen blokeres af en blodprop, hvilket kan føre til tab af funktioner som gang eller tale. Et intracerebralt hæmoragisk slagtilfælde sker, når et blodkar i hjernen brister, ofte som følge af forhøjet blodtryk [4].

3.1.2 Risikofaktorer

Risikofaktorer for udvikling af CVD omfatter en række påvirkelige og upåvirkelige faktorer. I dette afsnit gennemgåes nogle af de mest væsentlige. Blodtryk spiller en central rolle, især når det systoliske tryk overstiger 140 mmHg, da det kan skade karvæggene og hjertemusklen. Kolesterol, særligt LDL-kolesterol, bidrager også ved at danne plak i blodkarrene, hvilket kan hæmme blodgennemstrømningen og øge risikoen for blodpropper [31]. Rygning forværrer disse effekter ved at beskadige karvægge og øge blodets

klæbrighed, hvilket fremmer blodpropdannelse. Alle de faktorer som er nævnt ovenfor er påvirkelige, hvilket vil sige at livsstilsændringer og medicin kan have indflydelse.

Alder og køn er upåvirkelige faktorer, hvor risikoen stiger med alderen og er generelt højere blandt mænd. Når flere risikofaktorer som højt blodtryk, højt kolesterol og rygning kombineres, forstærker de hinanden og øger risikoen yderligere. Andre væsentlige risikofaktorer for udvikling af CVD inkluderer overvægt, genetisk disposition, bopæl, manglende motion, kost og komorbiditet, hvor tilstedeværelse af en sygdom kan øge riskoen for udvikle en anden sygdom [10]. Forebyggelse af CVD handler derfor om at reducere disse risikofaktorer gennem livsstilsændringer og om nødvendigt medicinsk behandling.

3.1.3 Forebyggelse af præmatur mortalitet og morbiditet ved CVD

Organisation for Economic Co-operation and Devolopment (OECD), har lavet en oversigt over præmatur mortalitet for kredsløbssygdomme, hvor det er angivet om de inkluderede sygdomme kan forebygges og behandles. Her er præmatur mortalitet defineret som dødsfald inden det 75. år. I tabel 1 fra OECD kan mortaliteten for alle de pågældende sygdomme sænkes med behandling [29]. Ved tidlig opsporing af CVD kan behandlingen være med til at forebygge dødeligheden for sygdommene.

Table 1. Preventable and treatable causes of mortality

Circulatory Diseases	Preventable	Treatable	
Aortic aneurysm	x (50%)	x (50%)	
Hypertensive diseases	x (50%)	x (50%)	
Ischaemic heart diseases	x (50%)	x (50%)	Her markerer x om man overvejende
Cerebrovascular diseases	x (50%)	x (50%)	
Other atherosclerosis	x (50%)	x (50%)	
Rheumatic and other heart disease		x	
Venous thromboembolism		x	

kan sænke præmatur mortalitet med forbyggende eller behandlende tiltag. 50% er brugt når der ikke er evidens for hvilken metode der er mest effektiv.

I [6] gennemgåes en række forebyggende og behandlende tiltag, der kan reducere præmatur mortalitet samt morbiditet som følge af CVD. Her understreges også vigtigheden af tidlig opsporing af sygdom for at kunne implementere tiltag som farmakologisk behandling, monitorering af kardiovaskulære forhold og livsstilsændringer, hvormed man kan nedsætte risikoen for forværring og nye CVD events, der skal føre til mortalitet. Livsstilsændringerne beskæftiger sig med de samme påvirkelige risikofaktorer, der er gennemgået i afsnittet Risikofaktorer, hvorfor disse kan ændres både i forebyggende øjenmed, men også som en foranstaltning, der kan være med til at reducere yderligere morbiditet og mortalitet som følge af CVD efter diagnosticering.

Læger bruger SCORE-modellen til at identificere højrisikopersoner og tilpasse forebyggelsesindsatser, hvilket kan reducere antallet af hjertekarsygdomme og død som følge heraf.[9]

3.2 State of the art

Nogle af de eksisterende metoder til vurdering af, om et individ er i risiko for kardiovaskulær sygdom, beskrives her. Mere specifikt anvendes skeamet i figur 1 til at vurdere risikoen for død som følge af CVD.

3.2.1 Moderne metoder

FIGUR 3

Det Europæiske Hypertensionsselskabs risikostratificeringsskema.

	Blood pressure (mmHg)				
Other risk factors, OD or disease	Normal SBP 120-129 or DBP 80-84	High normal SBP 130-139 or DBP 85-89	Grade 1 HT SBP 140-149 or DBP 90-99	Grade 2 HT SBP 160-179 or DBP 100-109	Grade 3 HT SBP ≥ 180 or DBP ≥ 110
No other risk factors	Average risk	Average risk	Low added risk	Moderate added risk	High added risk
1-2 risk factors	Low added risk	Low added risk	Moderate added risk	Moderate added risk	Very high added risk
3 or more risk factors MS, OD or diabetes	Moderate added risk	High added risk	High added risk	High added risk	Very high added risk
Established CV or renal disease	Very high added risk	Very high added risk	Very high added risk	Very high added risk	Very high added risk

© 2007 ESC

Figure 1. Det Europæiske Hypertensions risikostratificeringskema

ESH's Risikostratificeringsskema (Figur 1), som er blevet udviklet ud fra det amerikanske Framingham Risk Score, vurdere hvor høj risikoen er for kardiovaskulær mortalitet og morbiditet [31]. Skemaet er opbygget ud fra blodtryk ud af x-aksen, og andre risiko faktorer ud af y-aksen. Blodtyk er en faktor der altid bliver vurderet, fordi risikoen for CVD stiger og falder kontinuert med blodtrykket. Derudover vurdere man mængden af andre risikofaktorer og eventuelt om de allerede har en etableret CVD eller nyresygdom.

Et andet skema system, som bruges til at beregne risikoen ved personer uden erkendt CVD, er SCORE systemet, der vises i figur 2. Dette system findes både i skemaform og i form af software, HeartScore. Score systemet laver forskellige kategoriske inddelinger af sandsynligheden for død som følge af CVD inden for 10 år. Lav risiko: $< 1\%$, Moderat risiko: 1–4%, Øget risiko 5–9%, Høj risiko $\geq 10\%$ Denne stratificering er med til at sætte behandlingsmålene for patienten, men der er ikke en universel anvendelig tærskel for hvornår medicinsk behandling bør igangsættes. Personer i høj risiko vil have samme behandlingsmål, som personer med allerede konstateret CVD. Behandlingsmål omfatter det interval hvori en patients blodtryk skal ligge ,patientens tilstræbte kolesteroltal m.m. SCORE systemet vurdere risikoen for CV-død inden for 10 år ud fra en række forskellige risikofaktorer. Den sammenligner Blodtryk, Alder, Ryger/Ikke ryger og kolesteroltal. Der er lavet to versioner af dette skema. En for lande med generel lav risiko for CVD og en for lande med høj risiko for CVD. Danmark hører til blandt kategorien lavrisiko lande.

3.2.2 Forskning

Selvom eksisterende metoder som SCORE-systemet og ESH's risikostratificering er klinisk nyttige, har de begrænsninger i forhold til at håndtere komplekse mønstre i data. For at forbedre prædiktionen af kardiovaskulære hændelser undersøger nyere forskning, hvordan machine learning kan anvendes til at skabe mere præcise og fleksible modeller.

I et studie [25] , har man forsøgt at forbedre CVD risiko stratificering ved at bygge en machine learning model der kan beregne risikoen for aterosklerotisk kardiovaskulær sygdom indenfor 10 år. Modellen benytter sig af de samme risikofaktorer som den allerede eksisterende risiko beregner, American College of Cardiology/American Heart Association (ACC/AHA) Pooled Cohort Equations Risk Calculator. I alt blev modellen trænet på 9 risiko faktorer (variabler): alder, køn, etnicitet, diabetes historie, rygestatus, Systolisk blodtryk, tidligere behandling for hypertension, HDL kolesterol og total kolesterol.

Modellen havde til formål at klassificere om et individ havde "lav risiko" eller "høj risiko" for at opleve en hændelse af aterosklerotisk CVD. Hertil delte de aterosklerotisk CVD hændelser op i to versioner. Hård CVD og alt CVD og trænede én ML model på hver af dem. Hård CVD inkluderer myokardieinfarkt,

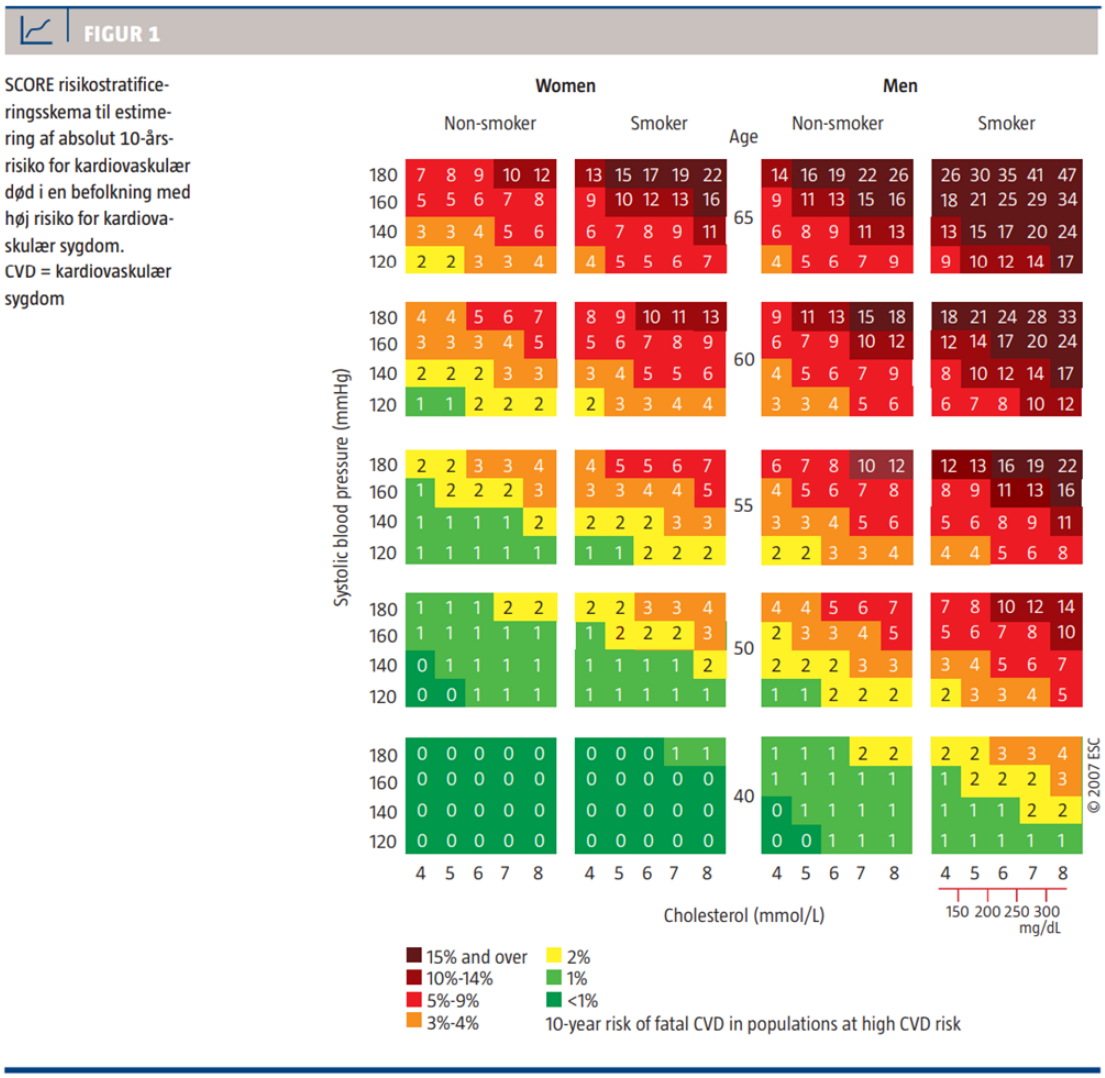

Figure 2. SCORE skema

fatal koronar hjertesygdom, slagtilfælde og fatal slagtilfælde. Alt CVD henviser til alt der indgår i Hård CVD, men også hændelser, som venstresidig hjerteinsufficiens, forbigående iskæmisk anfald, genoplivet hjertestop, andre former for CVD død mm.

Da traditionelle statistiske prædiktions metoder har en tendens til at lave antagelser om linearitet har man valgt at benytte sig af en Support Vector Machine algoritme i dette forsøg(SVM). Denne algoritme blev valgt da den læner sig op af mønstre i datasæt frem for at lave uhensigtsmæssige antagelser om linearitet. For at sikre sig en vis robusthed og generaliserbarhed af modellen blev 2-fold cross validation brugt til tilfældigt at opdele datasættet i to lige store dele. Ét træningssæt til at træne modellen og et testsæt til at evaluerer den.

For at illustrere forholdet mellem sensitivitet og specificitet for ACC/AHA risiko beregneren og ML risiko beregneren er der lavet 4 *Receiver operating characteristic* (ROC) kurver for de to beregningsmod-eller ift. hård CVD og alt CVD i Figur 3.

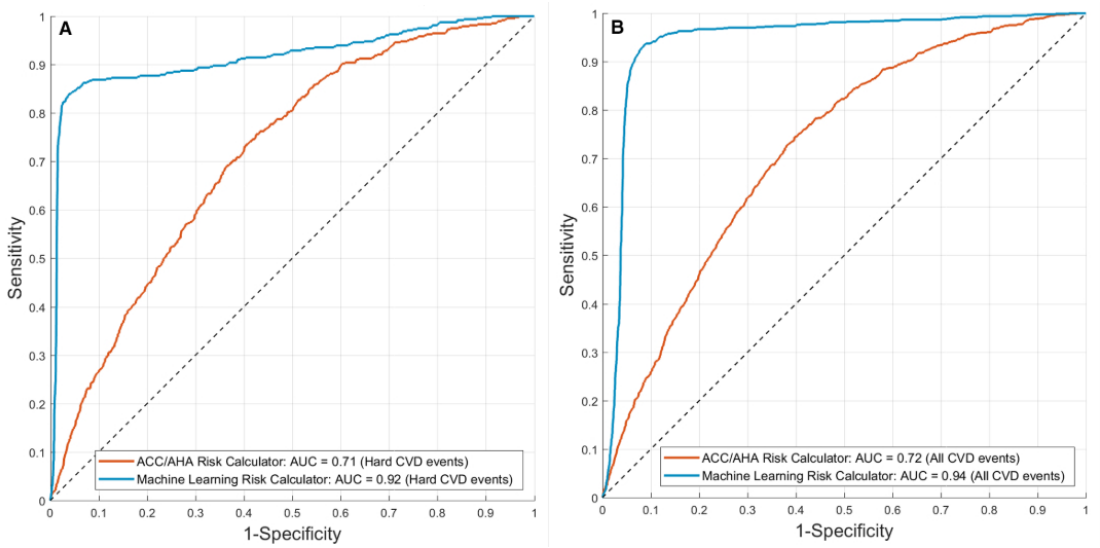

Figure 3. Sammenligning af ROC kurver for SVM machine learning risiko beregneren (blå linje) og ACC/AHA risiko beregneren (orange linje). Billede A illustrerer modellernes evne til at prædiktere "Hård CVD" og billede B illustrerer "Alt CVD".

På både A- og B grafen illustreres det tydeligt at ML risiko beregneren genererer bedre resultater end ACC/AHA risiko beregneren. Hvor arealet under kurven (AUC) for ML risiko beregneren er omkring 0.3 højere end ACC/AHA risiko beregneren i begge tilfælde. Tabel 2 går mere i dybden med resultaterne. Her er AUC værdierne fra ROC kurverne plottet ind, samt værdier fra risiko beregnernes Sensitivitet, Specificitet og Accuracy. Uden undtagelse, er ML risiko beregneren bedre end ACC/AHA risiko beregneren på samtlige parametre. Med en relativt høj difference i alle tilfælde.

	AUC	**Sensitivitet**	**Specificitet**	**Accuracy**
Hård CVD: ACC/AHA	0.71	0.76	0.56	0.58
Hård CVD: ML	0.92	0.86	0.95	0.94
Alt CVD: ACC/AHA	0.72	0.75	0.59	0.62
Alt CVD: ML	0.94	0.96	0.87	0.89

Table 2. Sammenligning af AUC, Sensitivitet, Specificitet og Accuracy mellem ACC/AHA og ML modeller for CVD.

Alt i alt har studiet benyttet sig af de samme parametre, som ACC/AHA risiko beregneren, til at træne en ML model til at beregne samme risiko for hændelser af CVD, og opnået betydeligt bedre resultater. Dette studie har fungeret som inspiration til projektet, hvor SVM blandt andre classifiere er inddraget i processen om valg af model. Derudover har ESH's Risikostratificeringsskema og SCORE systemet givet en forståelse for vægten af bestemte features herunder blodtryk, kolesterol, alder, køn og rygning. Det er disse features SCORE systemet baserer sin risikostratificering på og de er essentielle for diagnosticering og klassificering af CVD. Denne forståelse har været vigtig ift. beslutning af hvilke features der skulle inddrages til en ML model og hvilke features der mangler i det benyttede datasæt. Rygning er en risikofaktor af væsentlig betydning, men indgår ikke i datasættet, hvilket kan have en negativ indflydelse på generaliserbarheden af ML modellen. Dette er en vigtig refleksion at inddrage i den endelige vurdering af modellen.

3.3 Problemformulering

På baggrund af problemanalysen er følgende problemformulering blevet valgt:

Vil vi undersøge om man ved brug af machine learning metoder kan klassificere om individer har hjertekarsygdom, så vi kan bygge et software produkt der ved brug i almen praksis ***kan forebygge mortalitet og morbiditet ved at*** *anbefale om en patient bør sendes til yderligere undersøgelse.*

4 KRAVSPECIFIKATION

For at kunne lave et software produkt, der kan inddele patienter i grupperne hjertesygdom (1) og (0), udformes en kravspecifikation med funktionelle - og ikke funktionelle krav til systemet. Målgruppen for produktet er læger i almen praksis, som kravene er defineret ud fra. De funktionelle krav beskriver, hvad systemet skal kunne gøre for at fungere og de ikke funktionelle krav beskriver hvilke klassifikationsmetrikker produktet skal opfylde.

Traditionelle metoder som SCORE-systemet og ESH's risikostratificering bygger på lineære sammenhænge og faste kategorier, hvilket begrænser deres evne til at identificere komplekse mønstre i data. Machine learning (ML) løser dette ved at analysere store datasæt og opdage skjulte relationer mellem flere risikofaktorer, som ikke kan fanges med traditionelle metoder. I studier som det nævnte, hvor ML anvender algoritmer som kan opnå mere præcise og fleksible prædiktioner, hvilket gør det muligt at målrette forebyggelse bedre og reducere mortalitet og morbiditet i forbindelse med kardiovaskulære sygdomme.

4.1 Funktionelle krav:

- 1. Systemet skal kunne anvende en trænet machine learning model til at klassificere om patient har eller ikke har kardiovaskulær sygdom baseret på brugerinput

4.2 Data håndtering

- 2. Systemet skal kunne tage imod patientdata som input, som er relevante for forudsigelse af kardiovaskulær sygdom.
- 3. Systemet skal kunne validere inputdata for at sikre, at de korrekte værdier bliver indtastet
- 4. Systemet skal kunne indsamle og lagre patientdata der inputtes, så det kan bruges til at forbedre modellen løbende.
- 5. Systemet skal kunne anonymisere patientdata.
- 6. Systemet skal kunne forhindre brug af model ved manglende data

4.3 Ikke funktionelle krav:

Løsning

- 7. Systemet skal kunne klassificere patienten på under 10 sekunder

Brugervenlighed

- 8. Systemet skal være intuitivt og kræve minimal træning af sundhedspersonalet.
- 9. Det skal være nemt og simpelt at indtaste brugerinput samt forstå og evaluere output.

Klassifikationsmetrikker

- 10. Modellen skal have en accuracy på minimum 90%
- 11. Recall skal være $> 95\%$
- 12. F1 score skal være så høj som muligt med et passende trade off mellem recall og precision
- 13. Precision skal vægtes lavere end recall, da det er vigtigere at fange alle sygdomstilfælde

Tilgængelighed

- 14. Systemet skal være tilgængeligt døgnet rundt, med minimal nedetid (maksimalt 5%)
- 15. Systemet skal være tilgængeligt som en web-løsning. Den skal være tilgængelig fra moderne web-broswere (Chrome, Firefox, Internet explore, Safari etc.)
- 16. Systemet skal kunne oversættes til flere sprog hvis den skal kunne anvendes i andre lande

4.4 Vedligeholdelse af model

- 17. Modellen skal monitoreres regelmæssigt da den skal sikre høj accuracy og pålidelighed
- 18. Modellen skal opdateres og gentrænes hver 6. måned med ny data

4.5 Skalerbarhed

- 19. Systemet skal kunne håndtere store mængder patientdata effektivt
- 20. Systemet skal kunne tilpasses til andre lande, da befolkningsgrupper kan have varierende risiko for at udvikle CVD
- 21. Systemet skal kunne håndtere mange samtidige brugere uden det påvirker ydeevne

4.6 Afgrænsning

Idet dette projekt er underlagt en begrænset tidsramme, fokuseres der på kernefunktionaliteten bag systemet. Fokus med projektet er derved at udvikle og evaluere en machine learning model, der kan klassificere om en patient har eller ikke har kardiovaskulær sygdom. Punkterne i kravspecifikationen, der bliver behandlet i denne rapport er: 1, 10, 11, 12, og 13

5 DATA

For at opfylde kravene i kravspecifikationen og udvikle en model, der lever op til de opstillede mål for præcis diagnosticering, er det nødvendigt at arbejde med relevante og pålidelige data. I det følgende afsnit beskrives det anvendte datasæt.

Datasættet 'Heart Failure Prediction' fra Kaggle består af 1190 observationer, hvoraf 272 er dublikerede, hvilket resulterer i i alt 918 unikke observationer. Dataene er indsamlet fra forskellige kilder, herunder Cleveland (303 observationer), Ungarn (294 observationer), Schweiz (123 observationer), Long Beach VA (200 observationer) samt 270 observationer fra ukendte lokationer. Sidstnævnte er hentet fra UCI Machine Learning Repository.

Herunder findes en gennemgang af de features, der er inkluderet i datasættet:

- Age: Angiver patientens alder i år ved dataindsamling
  Numerisk værdi fra 28 til 77 år
- Sex: Angivet i M og F og repræsenterer hhv. mænd og kvinder. Fordelingen er ulige idet 21 procent af de inkluderede personer er kvinder og 79 procent er mænd
- ChestPainType: Typen af brystsmerter, hvilket kan indikere hjertesygdom
- TA: Typical angina - opfylder følgende kriterier:
  - i: Indsnævrende ubehag foran på brystet, i halsen, kæberne, skulderen eller armen
  - ii:Udløst af fysisk anstrengelse
  - iii:Bliver forbedret ved hvile eller nitrater i løbet af få minutter
- -ATA: Atypical angina - opfylder to af ovenstående kriterier
- -NAP: Non anginal pain - opfylder en eller ingen af ovenstående kriterier.
- -ASY - Asymptomatic - intet abnormt[11]
- RestingBP: Numerisk værdi målt i mmHg i hvile - Værdier fra 0 - 200 mmHg. Det er ikke angivet om værdierne er det systoliske, diastoliske eller middelarterietryk.
- Cholesterol: Numerisk værdi, der angiver serum kolesterol i mm/dl. Værdier fra 60,3 - 603 mm/dl. Normalt målt som enten mg/dl eller mm/L. Bliver i resten af rapporten omtalt som mg/dl. Se forklaring i afsnit Dataindsamling og Bias.[**?**]
- FastingBS: Fastende blodsukker

  - 1: over 120 mg/dl
  - 0: alt derunder

- RestingECG: EKG målt i hvile, hvilket måler hjertets elektriske aktivitet
  - Normal: Ingen abnormitet
  - ST: Abnormiteter i ST-segmentet i EKG'et f.eks. elevation eller depression som kan indikere iskæmi
  - LVH: Hypertrofi af venstre ventrikel vurderet ud fra Estes kriterier [19][28]

- MaxHR: Maksimalt opnået puls for datasættets deltagere. Numerisk værdi fra 60 til 202 slag/minut
- ExerciseAngina: Træningsudløste brystsmerter. Kan indikere nedsat blodgennemstrømning i koronararterierne under arbejde.
  - Y: Yes
  - N: No

- Oldpeak: Beskriver graden af ST-depression i forhold til baseline målt i hvile. Oldpeak bliver målt under belastning eller umiddelbart efter, hvor et højere tal indikerer iskæmi i myokardiet. [**?**] Numerisk værdi målt i depression fra - 1,72 til 4,44
- ST-slope: Beskriver ST-segmentets udseende under peak-belastning:
  - Up: Upsloping, indikerer at hjertets iltbehov er opfyldt under arbejde
  - Flat: Kan være grundlag for yderligere test, da det kan indikere begyndende eller mild iskæmi
  - Down: Downsloping, betragtes som en indikator for alvorlig iskæmi under belastning. [28]

- Heart Disease: Tilstedeværelse eller fravær af hjertekarsygdom
  - 1: Heart Disease
  - 0: No Heart Disease

5.1 Databehandling

For at sikre, at det beskrevne datasæt kan anvendes effektivt til at træne og evaluere en machine learning-model, er det nødvendigt at behandle og forberede dataene korrekt. Databehandling er en central proces, hvor data renses og tilpasses, så de sikrer pålidelige resultater. I de følgende afsnit beskrives de specifikke trin og metoder, der er anvendt i databehandlingen, herunder håndtering af manglende værdier, normalisering af data og valg af relevante variabler

5.2 Dataindsamling og Bias

Formålet med databehandlingen er at forbedre datakvaliteten og sikre et robust grundlag for machine learning-modellen. Dette opnås gennem håndtering af manglende værdier, fjernelse af outliers og justering af datastrukturen, så datasættet bliver velegnet til præcis klassificering af kardiovaskulære sygdomme.

I datasættet er det ikke oplyst, hvornår eller hvordan data er indsamlet. Man kan altså ikke vide om de personer, der er inkluderet i dataen er blevet testet tilfældigt, på grund af symptomer eller i forbindelse med deres diagnosticering. Dette kan introducere selektionsbias, hvilket kan påvirke klassifikationsmetrikkerne og generaliserbarheden. Ydermere ved vi heller ikke om målingerne er foretaget på hospitalet eller hos egen praktiserende læge og her kunne forskellige måleinstrumenter også påvirke resultaterne, hvilket kan medføre teknisk bias. I datasættet er værdierne for kolesterol angivet i mm/dl, hvilket ikke almindeligvis er en målenhed man anvender[12]. Dette kan også introducere teknisk bias. I et af de mindre datasæt, som dette datasæt er sammensat af, var måleenheden for kolesterol angivet som mg/dl, hvilket også

er det vi vil benytte i resten af denne rapport. Desuden er det heller ikke angivet for kolesterol om der er tale om total, hdl eller ldl kolesterol. Idet over 50 procent af de inkluderede i datasættet har en hjertekarsygdom kunne det godt tyde på, at de personer, der indgår i datasættet er udvalgte og at man har ønsket en ligevægt mellem grupperne HeartDisease 0 og 1. I datasættet, hvor fordelingen mellem kønnene er skæv, kan det have betydning for machine learning modellens skalerbarhed til kvinder, som her er det underrepræsenterede køn. Det at datasættet er fundet før problemet vi skal behandle, har den indvirkning at features og den måde de er indsamlet på er bestemt på forhånd. Hvis vi selv skulle have udvalgt features kunne andre kendte risikofaktorer for hjertekarsygdomme som f.eks. rygning, kost, motion, genetisk disposition og BMI have været relevante at inddrage. Desuden kunne man også oplyse om, hvor de enkelte patienter stammede fra, da dette har betydning i kraft af at risikoen for udvikling af hjertekarsygdomme er forskellige alt efter, hvor man bor. En anden relevant oplysning, der ikke er tilgængelig i vores datasæt, er om patienterne har andre sygdomme eller sundhedstilstande og om de tager medicin. Alle de ovennævnte faktorer kan føre til bias, hvilket kan have den konsekvens at modellen ikke præsterer lige så godt på nye datasæt specielt, hvis disse repræsenterer en mere varieret population.

5.3 Visualisering af data

For at få et dybere indblik i datasættets struktur blev der udført en række visualiseringer. Dette trin er essentielt i dataanalyse for at identificere mønstre, outliers og skævheder i data, for derefter at kunne bruge disse informationer til at fjerne manglende værdier, håndtere outliers og tilpasse machine learning modellen til dataens struktur.

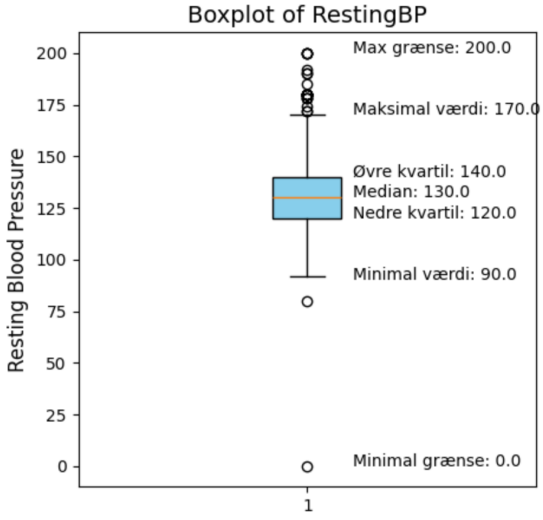

Figure 4. RestingBP boxplot

Den første visualisering blev gennemført ved hjælp af boxplots, der er særligt velegnede til at identificere centrale tendenser som medianen, og fordelingen af værdier indenfor det interkvartile interval. Her ses det at der er potentielle outliers da et *RestingBP* på 0 ikke er muligt Se Figur 4

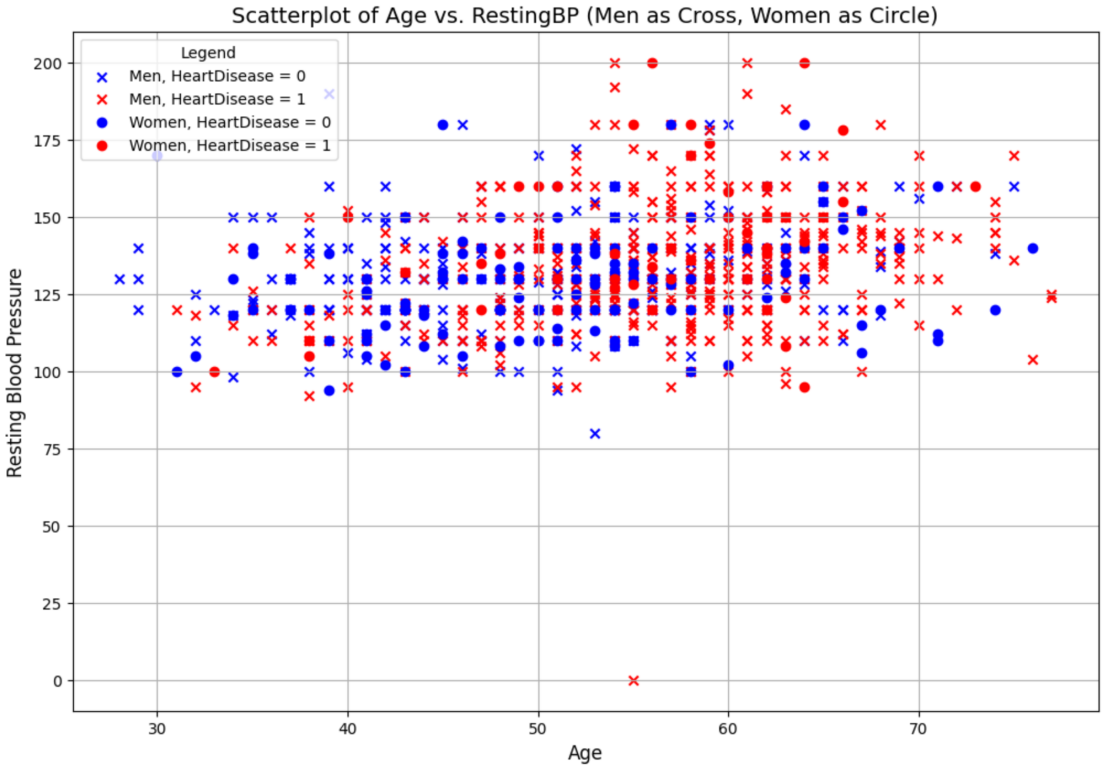

Figure 5. RestingBP Scatterplot

I figur 5, der viser sammenhængen mellem alder, køn og *RestingBP*, havde én person et hvileblodtryk på 0, og her kan det også bemærkes, at der er fire observationer med en måling på 200. Dette repræsenterer den maksimale værdi i datasættet. Det er dog muligt, at den faktiske maksimalværdi er højere, men at der var begrænsninger i enten måleudstyret eller i rammerne for datasættet.

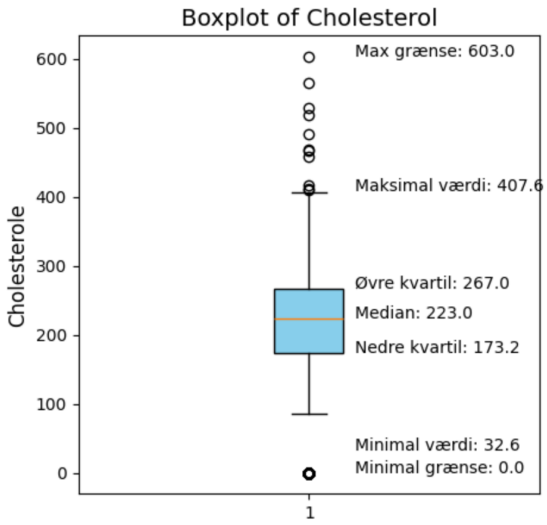

Figure 6. Cholesterol boxplot

I boxplottet i figur 6 for variablen *Cholesterol* sås også outliers med værdier, der spænder fra 0 til 603. Den observerede minimalværdi på 36,2 understreger en markant skævhed i datafordelingen. Scatterplots for *Cholesterol* viste en høj koncentration af outliers på tværs af aldersgrupper og køn, men primært blandt mænd. I *cholesterol* findes også 137 værdier på 0, hvilket enten skyldes fejl i dataregistreringen

eller ufuldstændige målinger og derfor betragtes som manglende værdier.

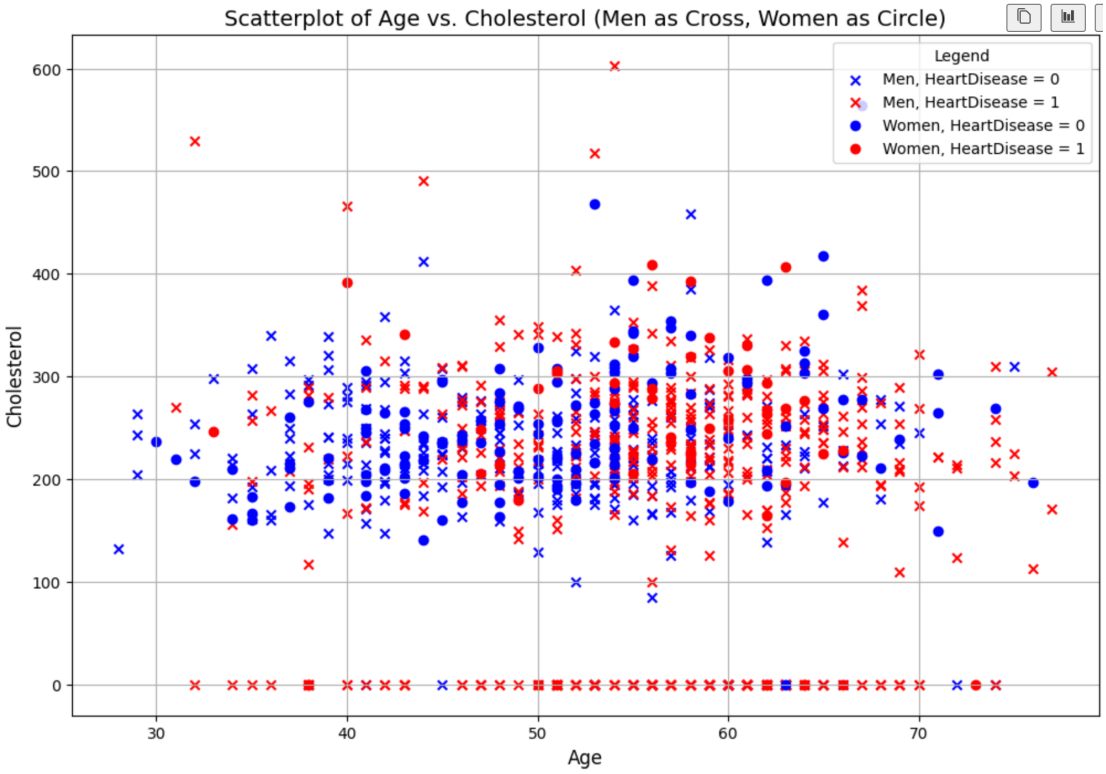

Figure 7. Cholesterol scatterplot

for at understøtte analysen blev der lavet en visualisering i figur 8 af datasættets fordeling mellem mænd og kvinder, hvilket gav et klart billede af kønsforskelle i datasættet og potentielle biases.

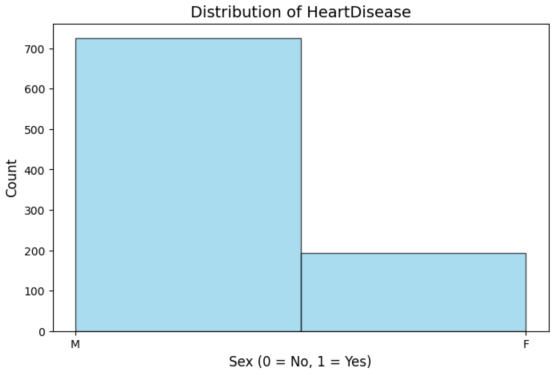

Figure 8. Sex boxplot

5.4 Datahåndtering

Efter visualisering af data, hvor 137 af datapunkterne i 'cholesterol' havde værdien 0, må dette betragtes som manglende data, da et kolesteroltal på 0 mg/dl ikke er foreneligt med liv. Af disse havde 120 hjertekarsygdom. Herunder i figur 9 er et billede af hvordan spredningen i de numeriske værdier i vores datasæt fordeler sig inden behandling af data:

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	HeartDisease
count	734.000000	734.000000	734.000000	734.000000	734.000000	734.000000
mean	53.480926	132.847411	200.797003	0.231608	137.299728	0.554496
std	9.398502	18.797220	110.817665	0.422147	25.514249	0.497360
min	28.000000	0.000000	0.000000	0.000000	60.000000	0.000000
25%	47.000000	120.000000	173.500000	0.000000	120.000000	0.000000
50%	54.000000	130.000000	224.000000	0.000000	139.500000	1.000000
75%	60.000000	140.750000	268.000000	0.000000	156.000000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	1.000000

Figure 9. Train Data før imputing og fjernelse af outliers

Da formålet med modellen er, at den skal benyttes i almen praksis fjernes features Oldpeak og ST-slope, der ikke kan måles i denne sammenhæng. Disse måles normalt på mere avanceret udstyr end der almindeligvis er tilgængeligt i almen praksis og indgår derfor ikke i datahåndteringen. I figur 9 fremgår det at også 'RestingBP' har 0 som den laveste værdi, hvilket også må betragtes som en manglende værdi.

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	HeartDisease
count	597.000000	597.000000	597.000000	597.000000	597.000000	597.000000
mean	52.902848	133.661642	246.876047	0.169179	140.402010	0.480737
std	9.500827	17.375137	60.897742	0.375225	24.825077	0.500048
min	28.000000	92.000000	100.000000	0.000000	69.000000	0.000000
25%	46.000000	120.000000	209.000000	0.000000	122.000000	0.000000
50%	54.000000	130.000000	237.000000	0.000000	140.000000	0.000000
75%	59.000000	141.000000	277.000000	0.000000	160.000000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	1.000000

Figure 10. Train Data med Cholesterol kolonner = 0 fjernet

Der er først forsøgt med fjernelse af kolonnerne med kolesterolværdier = 0 i figur10. Her er gennemsnittet i kolesterol = 246,88 mg/dl imod de 200,8 mg/dl der var inden kolonnerne blev fjernet i figur 9. 137 entries blev fjernet. For at undgå at miste en relativ stor del af datasættet er der forsøgt med Scikit learns KNN-imputing af manglende kolesterolværdier, da vi ønskede at bevare så meget data som muligt. Første trin var at erstatte de manglende kolesterolværdier med NaN: $train_data['Cholesterol'] = train_data['Cholesterol'].replace(0, np.nan)$

Derefter anvendtes KNN-imputer: $knn_imputer = KNNImputer(n_neighbors = 5) train_data_imputed = pd.DataFrame(knn_imputer.fit_transform(train_data_imputed), columns = train_data_imputed.columns$

Denne metode beregner gennemsnittet af naboværdiernes 'Cholesterol'-værdier, hvor alle naboer (her fem) bliver vægtet lige. Ved KNN - imputering bliver naboer udvalgt ved at finde de andre entries med værdier tættest på dem selv. Her sammenlignes de værdier som ingen af de pågældende entries mangler[33]. Vi forsøgte os både med flere og færre naboer og med en KNN-imputer, der tog hensyn til, hvor langt de respektive naboer befandt sig fra den manglende værdi. Det der gav det bedste resultat var, KNN-imputation med fem naboer hvor alle blev vægtet lige. I figur 11 ses værdierne for 'cholesterol' efter KNN-imputering. Her nærmer den nye gennemsnitsværdi på 247,37 mg/dl sig den der blev opnået ved at fjerne de manglende kolesterolværdier

	Age	RestingBP	Cholesterol	FastingBS	MaxHR
count	734.000000	734.000000	734.000000	734.000000	734.000000
mean	53.480926	132.847411	247.370572	0.231608	137.299728
std	9.398502	18.797220	56.342045	0.422147	25.514249
min	28.000000	0.000000	100.000000	0.000000	60.000000
25%	47.000000	120.000000	213.450000	0.000000	120.000000
50%	54.000000	130.000000	240.000000	0.000000	139.500000
75%	60.000000	140.750000	274.750000	0.000000	156.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000

Figure 11. Train Data with KNN-imputation i 'Cholesterol'

For at undgå at ekstreme værdier introducerer bias i modellen, blev outliers fjernet. Til dette har vi benyttet Z-score, der beregnes med formlen

$$Z = \frac{X - \mu}{\sigma},$$

hvor μ er gennemsnittet, som repræsenterer midten af normalfordelingen, og σ er standardafvigelsen, der beskriver spredningen og definerer intervallerne i bell curven (Se Figur 12). Når en given X-værdi indsættes i formlen, beregnes en Z-score. Denne Z-score kan bruges i en Z-score-tabel til at finde sandsynligheden for, at en observation falder inden for et bestemt interval.

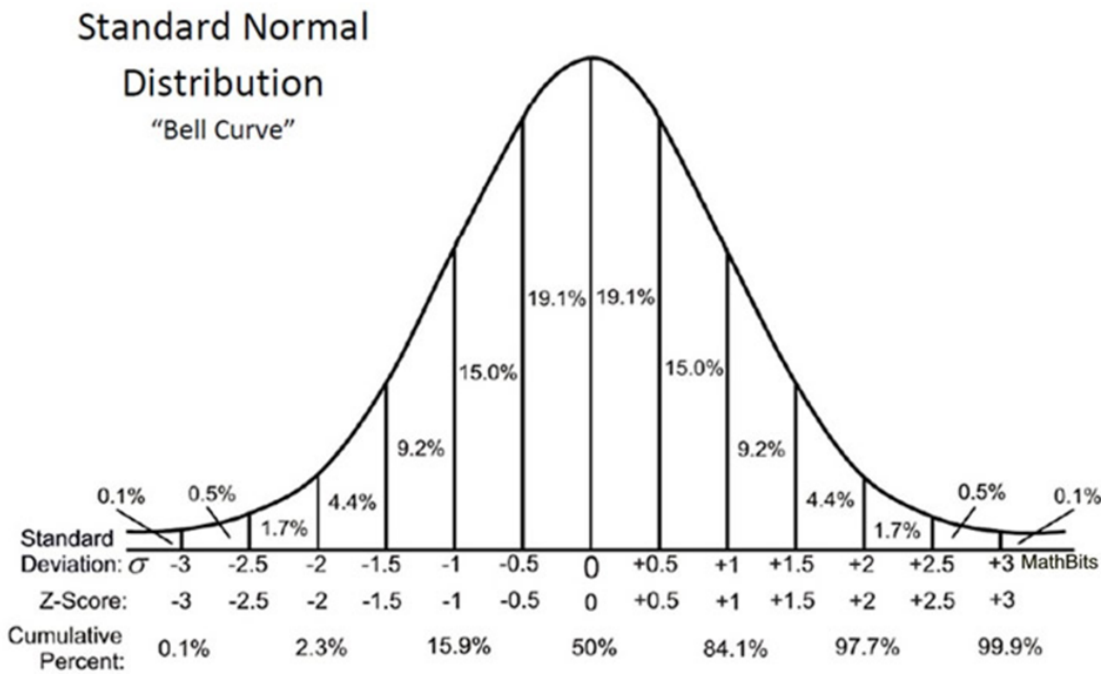

Figure 12. Z-Score

Ved fjernelse af outliers fjernede vi alle entries med en Z-score udenfor intervallet [-3, 3]. Først blev de outliers der skulle fjernes printet i terminalen, så vi kunne inspicere om det ville introducere andre problemer med vores data. Den mest sjældne ChestpainType typical angina blev flagget som en outlier. Dette syntes problematisk, da dette er en kategorisk variabel. For at imødekomme dette fjernede vi 'ChestpainType' fra vores datasæt inden vi fjernede outliers og tilføjede den efter fjernelse af outliers. Dette resulterede i at 18 outliers blev detekteret, heriblandt RestingBP på nul. Efter fjernelse af disse opnåede vi en 0,1 procent højere accuracy ved cross validation med en SVM-model.

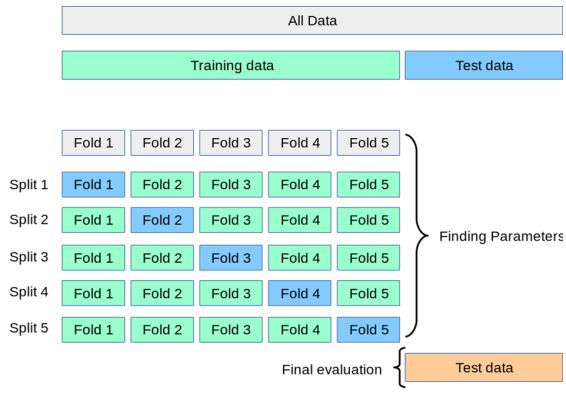

Figure 13. Cross Validation

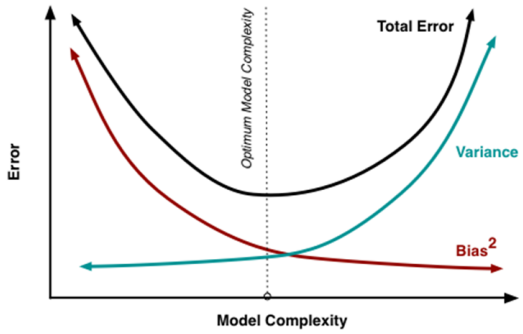

Figure 14. Bias og varians

6 VALG AF MODEL

For at udvælge hvilke modeller der egner sig bedst til dette klassificeringsproblem, er en række modeller med standard indstilling af deres hyperparametre blevet afprøvet (quick and dirty proces). Her blev den gennemsnitlige accuracy ved brug af cross validation benyttet som målestok for deres præstation. De tre, der havde den højeste gennemsnitlige accuracy blev udvalgt som dem, der skulle videreudvikles på. De afprøvede modeller og deres gennemsnitlige accuracy er: logistisk regression (0,7943), decision trees (0,69), random forest (0,7834), support vector machine(SVM) (0,8156), KNN (0,7710) og Naive Bayes (0,7711). Derfor arbejdes der videre med logistisk regression, random forest og SVM. Herunder bliver disse tre modeller og cross validation præsenteret.

6.1 Cross Validation

Cross-validation er en metode til at evaluere og validere maskinlæringsmodeller ved at opdele træningsdataene i flere undergrupper (\textit{folds}). Under denne proces bliver hvert fold på skift anvendt som testdata, mens de resterende folds bruges til at træne modellen. Ved at gentage denne procedure for alle folds sikres det, at hver observation i datasættet anvendes som testdata præcis én gang, samtidig med at modellen trænes på de resterende data. Denne tilgang kan anvendes udelukkende på træningsdataene, hvilket betyder, at man eksempelvis i et 80/20 split stadig vil reservere de sidste 20 af datasættet til en endelig uafhængig test. Cross-validation forbedrer modelvurderingens robusthed og giver en bedre indikation af modellens generelle ydeevne, da den reducerer risikoen for overfitting og afhængighed af en specifik opdeling af datasættet13.

Når en model bliver mere kompleks, vil variansen typisk stige. Dette medfører risikoen for overfitting, hvor modellen tilpasser sig træningsdataene i en sådan grad, at den fejler i at generalisere til det underliggende mønster i data. Omvendt gælder det for bias: En høj bias-værdi indikerer, at modellen er for simpel og dermed ikke i stand til at fange det underliggende mønster i data, hvilket fører til underfitting. Det er derfor afgørende at finde en balance i modellens kompleksitet for at minimere den samlede fejl og opnå optimal ydeevne. Dette kaldes ofte \emph{bias-varians-afvejningen} og er en central udfordring

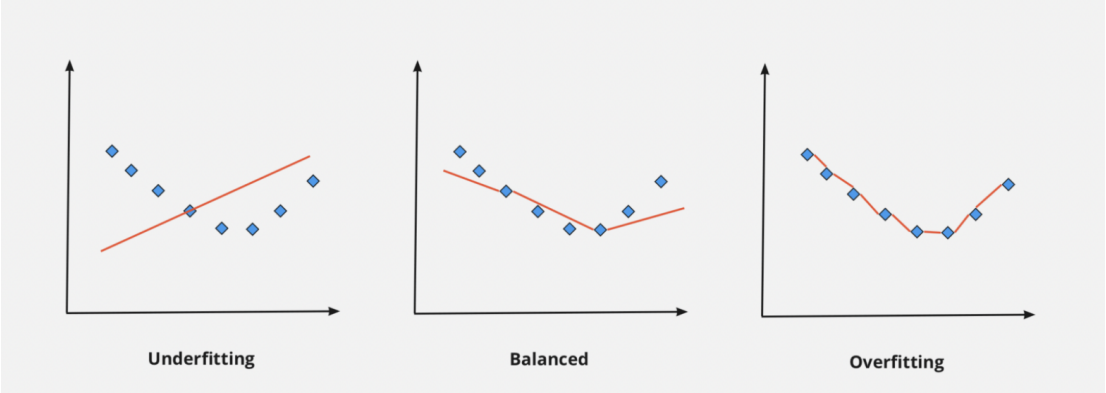

Figure 15. Illustration af underfitting, balanced og overfitting

inden for maskinlæring14.

Figur 15 viser tre forskellige scenarier for, hvordan en model kan passe til data: underfitting, balanced og overfitting.

6.2 Logistisk Regression

Den første af de udvalgte modeller fra Quick and dirty er logistisk regression, som er en statistisk teknik, der ofte bliver brugt i machine learning til forudsigelse og binære klassifikationsproblemer og den hører under 'supervised learning'. Lineær regression bruges derimod til at definere lineære forhold mellem afhængige og uafhængige variabler og bruger en lineær funktion til at finde den regressionslinje, der passer bedst på den givne data. Det kunne f.eks. bruges til at prædiktere menneskers vægt hvis vi kendte deres højde[41]. Logistisk regression kan bruge både kontinuer, diskret og kategorisk data til at prædiktere et binært resultat. Logistisk regression transformerer det kontinuere værdi-resultat som funktionen for lineær regression giver, til et kategorisk værdi-resultat af enten 0 eller 1. Resultatet logistisk regression giver, er en sandsynlighed repræsenteret som et tal mellem 0 og 1. Til binær klassifikation vil en værdi mindre end 0.5 prædiktere 0 og en værdi større end eller lig med 0.5 prædiktere 1 [22]. I stedet for blot at kigge på om denne talværdi er over eller under 0.5, kunne den også bruges til at udtrykke risici eller hvor sikker modellen er i sin prædiktion. En værdi tæt på 1 ville f.eks. udtrykke en stor risiko eller sikkerhed for det givne udfald.

6.2.1 Den Logistiske Funktion

Logistisk regression bruger Sigmoid funktionen: $Sigmoid = \frac{1}{1+e^{-z}}$, hvor z er formlen for en lineær regressionslinje, $a*x+b$, som ved logistisk regression ofte skrives B_0+B_1*x. Den logistiske funktion kan altså skrives: $P = \frac{1}{1+e^{-(B_0+B_1x)}}$ I Sigmoid funktionen er z den prædikterede værdi, som er modellens rå og ubearbejdede resultat. Logistisk regression beregner en lineær kombination af input features, $z = B_1x_1 + B_2x_2 + \ldots + B_nx_n + B_0$ hvor B_n er parametre (også kaldet weights), x_n er input features og B_0 er en konstant også kaldet intercept. Sigmoid funktionen konverterer z, den prædikterede værdi, til en sandsynlighedsværdi mellem 0 og 1 [16]

Det er denne logistiske funktion, $P = \frac{1}{1+e^{-(B_0+B_1x)}}$, som skaber S-kurven (Se Figur 16), der så bliver 'fittet' på den givne data. S-kurven bliver fittet på dataene med brug af metoden 'Maximum Likelihood Estimation' (MLE).

6.2.2 Maximum Likelihood Estimation

MLE finder de parametre (B_n) i modellen, som maksimerer sandsynligheden for at observere de givne data. I praksis med machine learning bliver den negative log-likelihood funktion brugt her som en tabsfunktion, som bruges til at beregne modellens afvigelse fra den sande værdi. Det flere forkerte prædiktioner en model laver, det større er tabsfunktionen. Den negative log-likelihood funktion bliver optimeret af gradient descent, som er en algoritme der er designet til at minimere funktioner. Dette er derfor gavnligt, da en mindre tabsfunktion betyder færre forkerte prædiktioner [14]

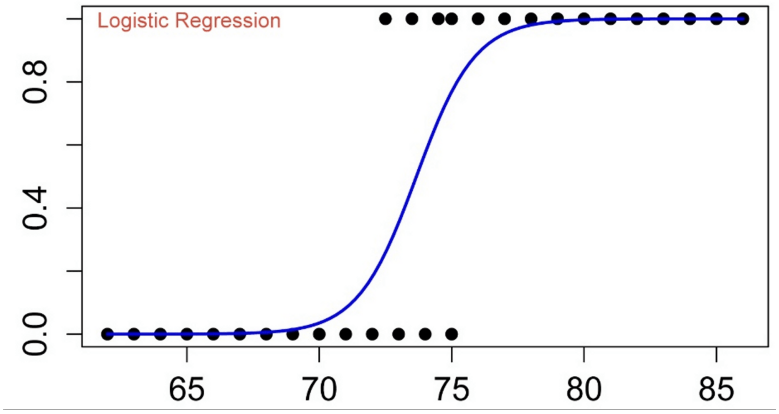

Figure 16. S-kurven for den logistiske funktion

6.2.3 Regularisering

Logistisk regression kan have tendens til overfitting, især når der er mange features i modellen. Regularisering bliver ofte brugt til at straffe store parameter-koefficienter når der er mange features. Logistisk regression har to hyperparametre, 'penalty' og 'C', som begge er regulariserings hyperparametre. Der er tre typer af penalty: L1, L2 og elastic net. L1, som også er kaldet Lasso, er en regulariserings metode, der tilføjer et strafled til log-likelihood funktionen. Dette strafled er lig med den absolutte værdi af koefficienterne (parametrene), hvilket gør at mange af koefficienterne bliver præcis lig nul. L2, også kaldet Ridge, tilføjer et strafled til log-likelihood funktionen, som er lig med kvadratet af koefficienterne. Dette gør at alle koefficienterne bliver tæt på nul, men ikke nødvendigvis lig med nul. L2 er mindre tilbøjelig til overfitting end L1, da L2 har en lavere varians og højere bias, og L2 er ofte brugt som standard valg. Elastic net kombinerer L1 og L2 og tilføjer et strafled til log-likelihood funktionen, som er en kombination af den absolutte værdi og kvadratet af koefficienterne. Dette gør at nogle af koefficienterne er lig med nul og andre er tæt på nul. Elastic net er særligt nyttigt til hvis der er korrelerede features i ens datasæt. C afgør styrken af regulariseringen. C er en invers værdi, hvilket betyder at jo lavere C-værdien er, jo større er regulariserings styrken, hvilket dermed fører til mindre overfitting[36].

6.3 Random Forest

Den anden af de tre modeller som havde den højeste gennemsnitlige accuracy før, justering af hyperparametre, er Random forest.

Random forest er en ensemble-metode inden for machine learning, der består af flere beslutningstræer for at forbedre modellens klassifikationsmetrikker og reducere overfitting. Random forest er en supervised learning metode og bruges til klassifikation og regression. Metoden bygger på ideen om at anvende flere svagt korrelerede beslutningstræer og kombinere deres forudsigelse for at blive mere præcis og generaliserbar.

Beslutningstræer som random forest fungerer ved at opdele data i mindre, mere homogene grupper ved at stille spørgsmål i hver node, som f.eks. "Er alderen ¿ 50?". Hver beslutning i træet deler dataene i to grene afhængigt af spøgsmålets udfald. Denne proces fortsætter, indtil træet når bladene/leaf nodes, som indeholder de endelige forudsigelser. For at vurdere kvaliteten af en opdeling bruges typisk gini Impurity og entropi. Formlen for gini-impurity er:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{1}$$

G repræsenterer Gini-impuriteten, som er et mål for urenheden eller graden af blanding af klasser i en node. Jo lavere Gini-værdien er, desto renere er noderne (dvs. jo mere ensartede er de data, der ligger i noden).

- p(i)i er andelen af datapunkter i klasse i i noden.

- N er antallet af klasser i datasættet

  Entropi måler graden af usikkerhed eller uorden i et datasæt. Entropien bruges til at vurdere, hvor blandet dataene er i en node. Jo højere entropi, desto mere uorden er der i dataene, og jo lavere entropi, desto mere homogen er noden. Formlen for entropi er:

$$H(S) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \tag{2}$$

- p(i) er andelen af datapunkter i klasse i i en given node

- N er antallet af klasser i datasættet

Random Forests funktionalitet er baseret på bagging (som bliver gennemgåes senere i Ensemble afsnittet) , hvor flere træer trænes på tilfældige delmængder af træningsdata og derefter kombinerer deres resultater for at træffe en beslutning. Denne tilgang hjælper med at reducere varians og forbedre generaliserbarheden af modellen [39], [1]. Når det kommer til forudsigelser, stemmer træerne om den korrekte klasse (hard voting) eller beregner gennemsnittet af forudsigelserne (soft voting), som igen uddybes i Ensemble afsnittet [1].

En af de stærke funktioner ved Random Forest er dens evne til at bestemme *feature importance*, hvilket betyder, at modellen kan identificere, hvilke features der har størst indflydelse på forudsigelserne. Dette gør modellen særlig nyttig til analyse af komplekse datasæt. [39], [1].

Random Forest er robust over for både outliers og overfitting, især når antallet af træer er stort. [39], [1]. En ulempe ved Random Forest er dens høje beregningsmæssige krav, især når datasættene er store og træerne er dybe. Træning af et stort antal træer kan være tidskrævende og kræve meget hukommelse, hvilket kan være en ulempe ved brug af Random Forest på meget store datasæt [39].

Selvom Random Forest desuden giver indsigt i *feature importance*, kan den samlede beslutningsproces være svær at forstå, hvilket gør modellen mindre gennemsigtig end enkelte beslutningstræer. Dette kan være problematisk i situationer, hvor det er nødvendigt at forstå, hvorfor en model træffer en bestemt beslutning, f.eks. i medicinsk beslutningstagning [1].

6.4 SVM RBF Kernel

Den sidste model, der opnåede en høj gennemsnitlig accuracy før tuning af hyperparameter, er Support Vector Machine (SVM) med en RBF-kernel. I det følgende afsnit gennemgås, hvordan SVM fungerer, og hvorfor RBF-kernelen er velegnet til at håndtere komplekse og ikke-lineære datasæt.

I support vector machine modellen anvendes scikit-learns standardindstilling[27] for SVM, som benytter en Radial Basis Function (RBF), også kaldet en Gaussisk kernel. Den matematiske formel for denne kernel er vist her:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{3}$$

Her beregner kernel-funktionen $K(x_i, x_j)$ similariteten mellem de individuelle observationer, hvor $\|x_i - x_j\|^2$ repræsenterer den numeriske værdi af den euklidiske (Pythagoras-læresætningen) afstand mellem to datapunkter. Parameteren γ (gamma) styrer rækkevidden af påvirkningen fra hvert datapunkt[23].

SVM er velegnet til datasæt, der omhandler sundhedsdata, da den effektivt kan håndtere ikke-lineære relationer og operere i højdimensionelle rum. RBF-kernelen tilpasses data ved hjælp af gamma-parameteren, hvilket gør det muligt at finde en balance mellem underfitting og overfitting[23].

Dog har SVM en udfordring med gennemsigtighed, da modellen kan være vanskelig at fortolke. Dette gør det problematisk i sundhedsdata, hvor det ofte er nødvendigt at kunne forklare og dokumentere beslutninger, f.eks. i medicinsk beslutningstagning og godkendelsesprocesser[24].

RBF kan god tilpasse beslutningsgrænsen efter de vigtigste mønstre i dataen uden nødvendigvis og tage hensyn til alle punkter. I scikit-learn anvendes parameteren C til at kontrollere, hvor hårdt modellen skal straffe fejlklassifikationer. En lav C -værdi kan føre til underfitting, da modellen tillader flere fejlklassifikationer, mens en høj C -værdi kan føre til overfitting, da modellen bliver for kompleks og følsom over for støj[17].

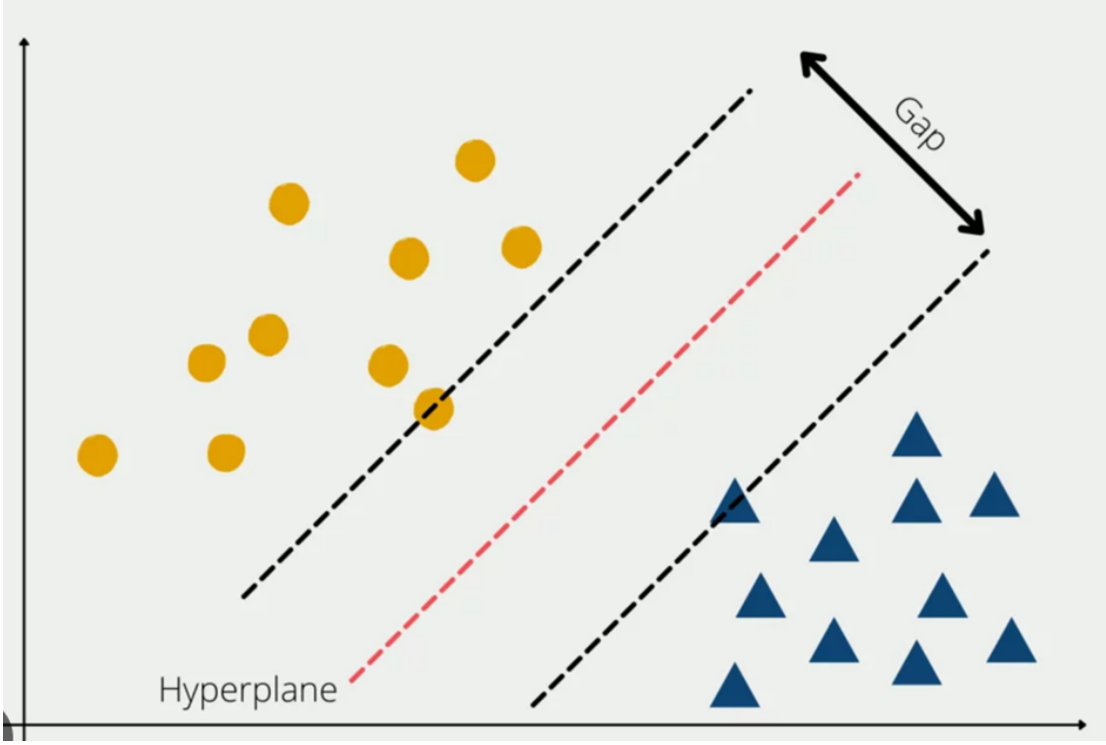

Figure 17. SVM

SVM i to dimensioner anvender en hyperplan, som i dette tilfælde er en linje, til at opdele datasættet i forskellige klasser (se Figur 17). Formålet er at maksimere marginen, som er afstanden mellem hyperplanet og de nærmeste datapunkter fra hver klasse. Disse nærmeste datapunkter kaldes supportvektorer, og de spiller en afgørende rolle i fastlæggelsen af hyperplanets position og orientering. I to dimensioner er det dog en linje, hvilket betyder, at løsningen kun bliver lineær.

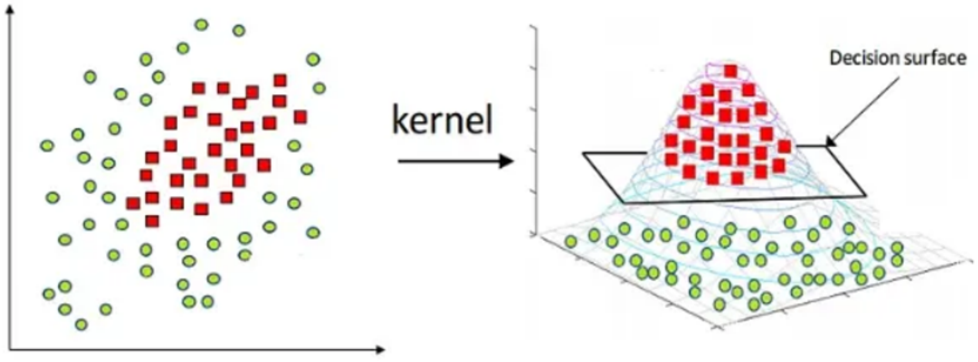

Figure 18. RBF

RBF-kernel transformerer datasættet implicit til et højere dimensionsrum, hvor det bliver lettere at klassificere data, fordi klasserne kan adskilles i dette rum (se Figur 18). RBF-kernens transformation muliggør oprettelsen af et hyperplan, der adskiller klasserne. Parameteren γ i RBF-kernen justerer, hvor meget hver datapunkt påvirker klassifikationen, hvilket indirekte styrer, hvor ”kurvet” eller ”glat” hyperplanet bliver[38].

6.5 Ensemble

For at forbedre accuracy og robusthed kombinerer ensemble learning flere modeller. Her gennemgås metoden samt forskellen mellem soft og hard voting.

En populær måde at optimere sin model er ved en teknik der kaldes *Ensemble learning.* En *ensemble* er en gruppe af modeller. *Ensemble learning* bygger på en teori kaldet *wisdom of the crowd*, som i sin essens handler om at, hvis en stor mængde af individer bliver bedt om at komme med et forslag til en løsning på et problem, da gruppens sammenlagte løsningsforslag vil være bedre end løsningsforslaget fra en enkelt ekspert. Relationen til ensemble learning er, at det sammenlagte svar til en gruppe af modeller potentielt kan levere bedre resultater end den bedste enkeltstående af disse modeller [18].

Et eksempel på, hvorfor dette fungere er visualiseret på billedet med mønt kast (Se Figur 26). I dette eksempel er mønten, der kastes med 51% partisk imod krone. Lige akkurat bedre end helt tilfældig.

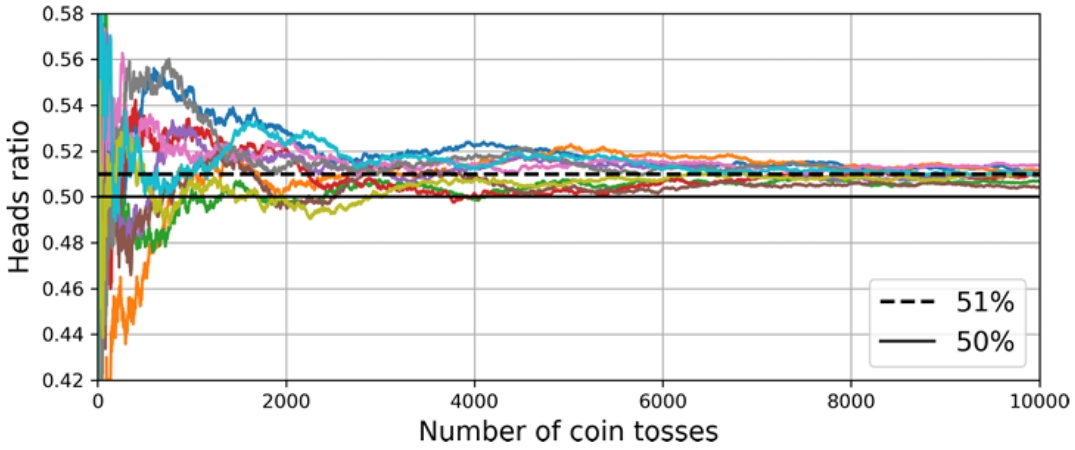

Figure 19. Visualisering af møntkast med 51% sandsynlighed for krone

Hvis mønten kastes x antal gange og man gætter på at krone fremkommer oftere end plat, så stiger sandsynligheden for at man har ret det højere x er. Som vist på billedet i Figur 26, vil der være en reel risiko for at plat vil forekomme oftest selv efter 1000 kast. Hvis mønten derimod kastes 10.000 gange, så nærmer ratioen af krone sig sandsynligheden (51%) for at få krone. Dermed stiger også sandsynligheden for, at man har ret, hvis man gætter på størst repræsentation af krone, jo flere gange mønten er kastet.

Ligeledes øger det sandsynligheden for at klassificere data korrekt med en ensemble af modeller. Selvom hver model kun har en accuracy på 51%, så vil et højt antal af disse modeller øge ensemblets samlede accuracy.

6.5.1 Soft and hard voting

Som beskrevet ovenfor baseres ensemble-metoden sig på stemmeafgivelse fra individuelle modeller. Hvis modellernes stemmer vægtes ens og klasse beslutningen træffes ud fra majoritetsstemmen, kaldes det hard voting (211). I den allerede beskrevne figur 19 , benyttes hard voting, hvilket demonstrerer, at majoritet afstemningen alene kan være aldeles betydningsfuld såfremt modellerne, der indgår i ensemblet er uafhængige (dvs. begår fejl og har ret på forskellige tidspunkter), selvom de hver for sig ikke bidrager med megen prædiktiv værdi. En anden metode til stemmeafgivelse inden for ensemble er soft voting. Her er det ikke majoritetsstemmen, der bestemmer resultatet, men derimod den samlede sandsynlighed for en given klasse, hvor den gennemsnitlige sandsynlighed udregnes på tværs af modellerne. Den endelige klasse kan derfor godt afvige fra den majoriteten af modellerne prædikterede, hvis tilstrækkeligt mange af de resterende modeller er mere 'sikre' i deres prædiktion. Soft voting har ofte en bedre prædiktiv værdi end hard voting fordi de mere sikre modeller bliver tillagt mere vægt. (215) For at anvende soft voting kræves det, at alle de individuelle modeller skal have adgang til klasse sandsynligheder. Hvis dette er tilfældet kan ensemblets hyperparameter indstilles til soft, hvorefter der foretages soft voting.

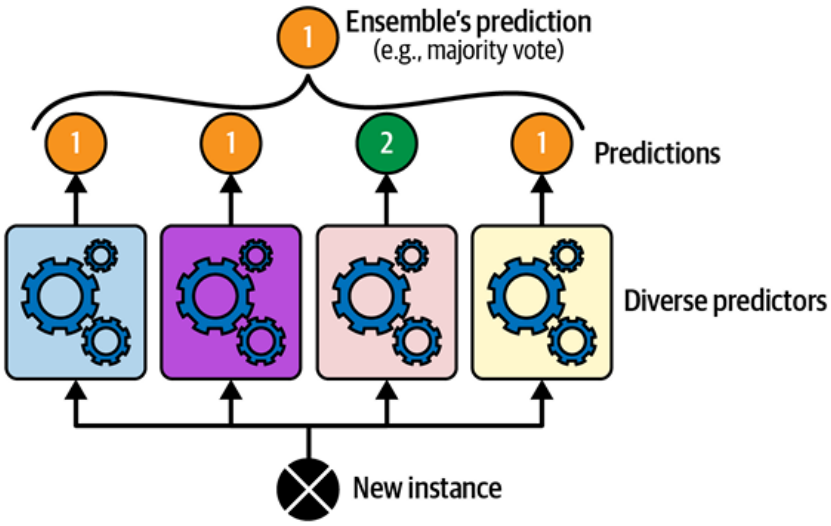

Figure 20. Illustration af hard voting

6.5.2 Bagging og Pasting

Det er vigtigt at have et diverst set af modeller til sin ensemble. Det er der flere måder at opnå det på. Én måde er at træne hver model i ensemblet på forskellige algoritmer. Ved at træne forskellige algoritmer øges sandsynligheden for at hver model laver en række af forskellige fejl, hvilket til gengæld øger den samlede accuracy for ensemblet.

En anden måde at opnå et diverst set af modeller er ved at træne hver model på en tilfældig delmængde af det originale træningsset. Denne metode kan udføres både med og uden udskiftning af datapunkter. Hvis datapunkterne ikke udskiftes, dvs., at alle datapunkterne i hver delmængde er unikke, kaldes det *Pasting*. Hvis det derimod er åbnet op for udskiftning så har hver delmængde chance for at indeholde datapunkter der også eksisterer i andre delmængder. Denne version kaldes *Bagging*, kort for *bootstrap aggregating*. Ved brug af bagging er der også sandsynlighed for at hvert datapunkt kan optræde flere gange i samme delmængde, samt risiko for at det ikke optræder i nogle af dem.

Hver enkelt model vil uanset om der bruges bagging eller pasting have en højere bias end, hvis de var trænet på det originale datasæt. Når ensemblet aggregere over modellernes resultater for at opnå den samlede prædiktion reduceres både bias og varians. Det har den effekt at ensemblet har omtrent den samme bias, men lavere varians end, hvis én model var trænet på hele datasættet.

7 IMPLEMENTATION

I dette afsnit beskrives hvordan modellernes hyperparametre fra Quick and Dirty processen er blevet tunet for at levere de bedst mulige resultater. Herudover beskrives hvordan en ensemble model er bygget baseret på valgte modeller, samt hvordan de forskellige modeller præsterede under træningen.

7.1 Tuning af hyperparametre

For at forbedre evalueringsmetrikkerne for de udvalgte modeller er der anvendt grid search til at finde de optimale hyperparametre. Grid Search er en algoritme som bliver anvendt til hyperparametertuning, og som systematisk afprøver en række kombinationer af de angivne hyperparameterværdier for at identificere dem der, der præsterer bedst. Dette gøres ved at opbygge et gitter af hyperparameterværdier og evaluere modellens ydeevne for hver kombination [18]. Efter de optimale kombinationer er fundet for hver af de tre modeller er de individuelt blevet trænet på træningsdataen.

Målestok

For at kunne evaluere modellernes præstation blev F1-score, Accuracy, Precission og Recall brugt som evalueringsmetrikker. Accuracy er en målestok for, hvor ofte en model klassificerer korrekt. Det beregnes som forholdet mellem korrekt klassificerede observationer (både positive og negative) og det samlede antal forudsigelser:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Her står:

- **TP (True Positives)**: Antallet af korrekt forudsagte positive observationer.
- **TN (True Negatives)**: Antallet af korrekt forudsagte negative observationer.
- **FP (False Positives)**: Antallet af forkerte positive forudsigelser.
- **FN (False Negatives)**: Antallet af forkerte negative forudsigelser.

Accuracy er velegnet som en generel indikator for modellens præstation, men kan være misvisende ved datasæt med skævt fordelte klasser.[18] [30]

Precision måler, hvor præcise de positive forudsigelser er, dvs. hvor mange af de forudsagte positive observationer, der rent faktisk er korrekte. Det beregnes som:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision er især vigtig, når falske positive har alvorlige konsekvenser, f.eks. i sygdomsdiagnostik, hvor en falsk positiv kan føre til unødvendige behandlinger eller bekymringer.[30] [18]

Recall, også kendt som sensitivitet eller *True Positive Rate*, måler modellens evne til at identificere alle faktiske positive observationer. Det beregnes som:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall er særligt vigtigt, når det er afgørende at minimere falske negative, som kan være tilfældet i screeningsprogrammer for alvorlige sygdomme. En model med høj recall sikrer, at få faktiske positive tilfælde overses, men det kan komme på bekostning af flere falske positive.[30] [18]

F1 score er en evalueringsmetrik som kombinerer precission og recall. Formålet med scoren er at evaluere hvor god balancen er mellem precision og recall i en model. Hvis F1 værdien er lav, betyder det at modellen laver mange falske positive klassifikationer og/eller mange falske negative klassifikationer. Hvis værdien derimod der høj (nærmer sig 1), er det et tegn på at modellen laver få fejl. Den højest mulige F1 score opnås der hvor der er bedst balance mellem precission og recall.

F1 beregnes som:

$$\text{F1} = 2 * \frac{\text{Precission} * \text{Recall}}{\text{Precission} + \text{Recall}}$$

Ensemble model

Efter de tre modeller: Random forest, Logistisk Regression og SVM havde fået tunet deres hyperparametre , blev en ensemble konstrueret. Ensemblet var baseret på disse og tilhørende optimerede hyperparametre. Herefter blev hyperparametere til ensemblet manuelt tunet for at maksimere den samlede præstation. Figure 21 illustrerer den overordnede struktur for ensemble modellen.

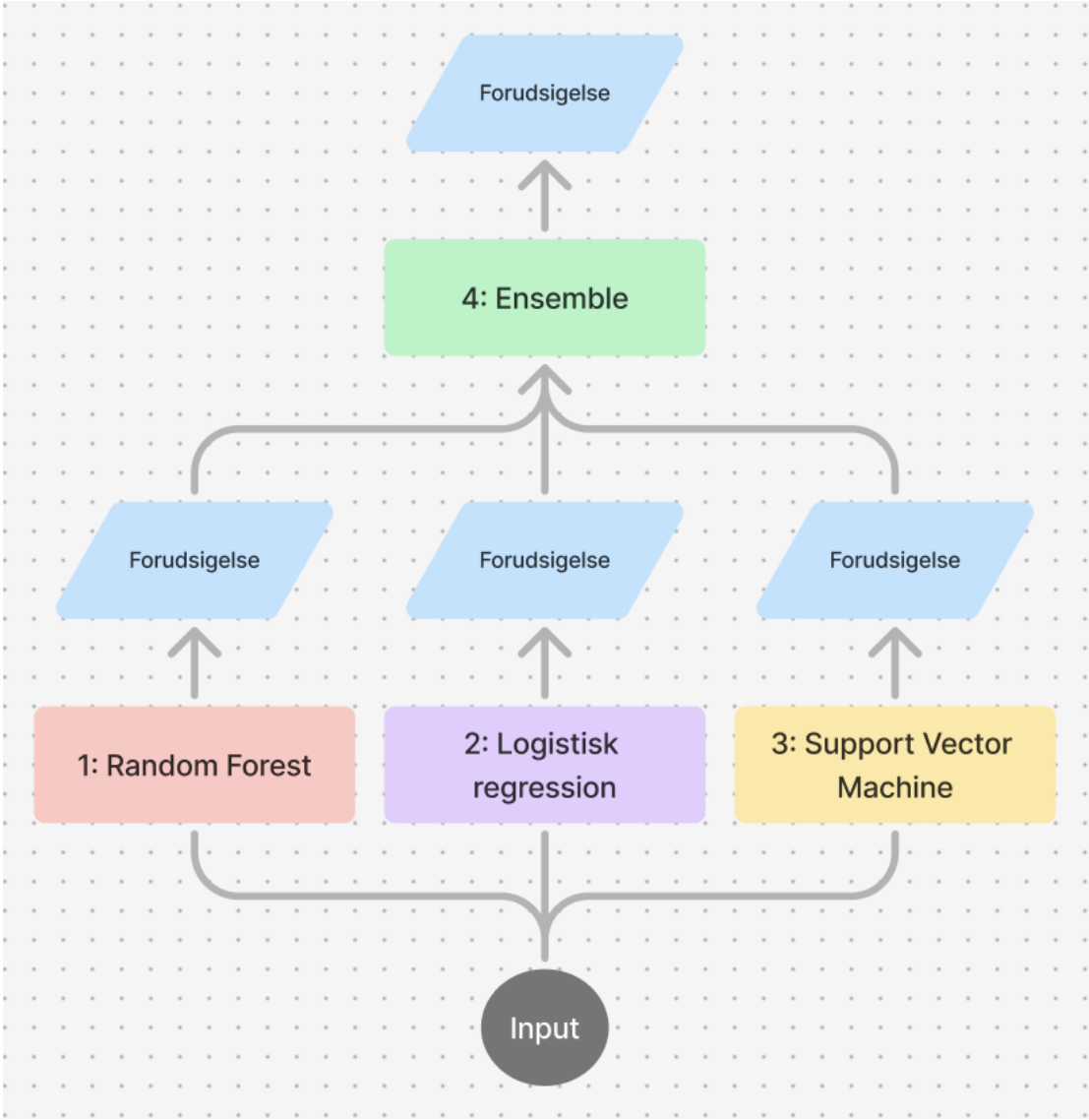

Figure 21. Ensemble flowchart

Figurene 22, 23 og 24 er kode udklip fra træningen af ensemble modellerne blok 1, 2 og 3 i figur 21 med deres opdaterede hyperparametre. Under træningen af 2 og 3 blev der bygget en pipeline med formålet at skalere datapunkterne, så de blev mere standardiserede, da de fleste machine learning modeller har bedst præstation når de forskellige features befinder sig indenfor samme skala [18]. For Random Forest, blok 1, var det ikke nødvendigt at skalere dataen, da den i forvejen er god til at håndtere features på forskellige skalaer.

```
rf = RandomForestClassifier(n_estimators=200,max_depth=20, random_state=42)
```

Figure 22. Scikit-learn kode brugt til at træne blok 1 i Figur 21: Logistisk Regression modellen.

```
log_reg =  Pipeline([
    ('scaler', StandardScaler()),
    ('classifier', LogisticRegression(max_iter=500, solver="liblinear", random_state=42))])
```

Figure 23. Scikit-learn kode brugt til at træne blok 2 i Figur 21: Logistisk Regression modellen.

```
svm = Pipeline([
    ('scaler', StandardScaler()),
    ('classifier', SVC(probability=True, kernel="rbf", random_state=13))
])
```

Figure 24. Scikit-learn kode brugt til at træne blok 3 i Figur 21: Support Vector Machine modellen.

Herefter blev ensemblet blev bygget ud fra *scikit-learns* VotingClassifier (Figure 25) metode, som tager en række modeller og kan benytte sig både hard- og soft voting. Både hard og soft voting blev eksperimenteret med, ved konstruktionen af ensemblet, hvor det viste sig at soft voting genererede de bedste resultater [40].

```
voting_model = VotingClassifier(
    estimators=[
        ('log_reg', log_reg),
        ('svm', svm),
        ('rf', rf)
    ],
    voting='soft'
)
```

Figure 25. Scikit-learn kode brugt til at træne ensemblet i blok 4 i Figur 21 ud fra modellerne i blok 1,2 og 3.

Det blev også forsøgt at udvide mængden af modeller i ensemblet ved at bruge scikit-learns Bagging-Classifier til at træne flere versioner af logistisk regression og SVM på delmængder af træningssettet. Bagging metoden viste sig ikke at bidrage med bedre resultater, end den originale ensemble med kun én af hver af de tre machine learning modeller. Alle modellerne, Random forest, Logistisk Regression, SVM og ensemblet, gennemgik herefter en crossvalidation for at øge generaliserbarheden til modellerne, samt reducere risikoen for overfitting. Nedenfor i Tabel 3 kan det ses at at ensemble modellen i hver kategori performer bedre end samtlige af de andre modeller. Det viste sig at kombinationen af enkeltsående modeller til én samlet model i dette tilfælde skaber den bedste classifier.

	F1	**Accuracy**	**Precission**	**Recall**
Random Forest	0.81	0.79	0.82	0.81
Logistic Regression	0.81	0.80	0.82	0.82
SVM	0.82	0.80	0.82	0.83
Ensemble	**0.83**	**0.81**	**0.84**	**0.84**

Table 3. Oversigt over træningsresultater fra Random Forest, Logistisk Regression, SVM og ensemble model

7.2 Argument for hvorfor recall vægtes højest

Fordi vores model prædikterer om en person har hjertekarsygdom eller ej, er det vigtigere at fange de positive instances end de negative. Hvis vores model fejlagtigt prædikterer, at en person ikke har hjertekarsygdom, kan det betyde at personen ikke bliver sendt i nødvendig behandling og kan føre til at personen bliver alvorlig syg eller afgår ved døden. Hvis modellen derimod fejlagtigt prædikterer, at en person har hjertekarsygdom, kan det betyde at personen bliver sendt til behandling uden grund og at der dermed er tale om en falsk alarm. Denne falske alarm er en betydelig mindre konsekvens end hvis et menneske får alvorlig sygdom eller mister livet, som følge af en fejlagtig prædiktion fra vores model. Det er altså vigtigere at fange de positive instances end de negative og derfor vægtes recall højest.

7.3 Threshold

En visualisering af de enkelte features i ensemblemodellen blev udarbejdet på baggrund af træningen. Formålet med denne analyse var at sikre, at ingen enkelte features korrelerede direkte med labels, hvilket potentielt kunne føre til, at en positiv klassifikation alene blev forudsagt på baggrund af en enkeltstående feature. Her ses det at det højest gennemsnit af Contribution to threshold er $Chestpaintype_ATA$ men da værdien ikke er tæt på threshold værdien 0.376 er der ikke fuldstændig korrelation mellem featuren og lablet

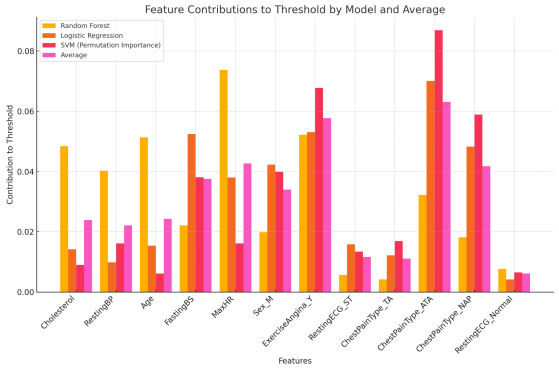

Figure 26. Feature weight

Som en del af den videregående analyse af resultaterne er der blevet udarbejdet en Precision-Recall-kurve. Denne kurve illustrerer, hvordan variation i threshold-værdier kan bruges til at opnå en ønsket recall, samtidig med at en tilstrækkelig høj precision bevares.

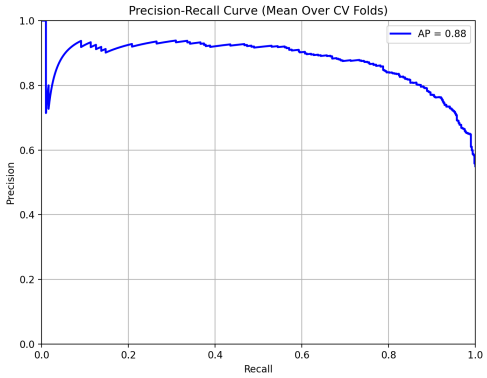

Figure 27. Precision and Recall curve

Ud fra trænings resultaterne blev der lavet en precision-recall-kurve som kan ses i Figur 27. Derefter blev threshold-værdierne justeret for at finde den bedste balance mellem høj recall og en tilfredsstillende precision. Efter nogle manuelle justeringer fandt man frem til, at den bedst egnede threshold-værdi var 0.376 den mest egnede threshold værdi. I tabel 5 ses de opdaterede træningsresultater

	F1	Accuracy	Precission	Recall
Ensemble med updatereThreshold-træning	0.82	0.79	0.76	0.89

Table 4

7.4 Binær klassifikation og sandsynlighedsscore

Binær klassifikation indebærer tildeling af en sandsynlighedsscore, mellem 0 og 1, til hver observation baseret på en kombination af datasættets features[3]. Sandsynlighedsscoren repræsenterer sandsynligheden for, at observationen tilhører den positive klasse. Som standard anvendes ofte en threshold på 0,5, hvor observationer med en score over 0,5 klassificeres som positive, og observationer med en score under 0,5 klassificeres som negative. Threshold-værdien kan dog justeres for at optimere balancen mellem recall og precision afhængigt af modellens anvendelsesområde og krav29[8].

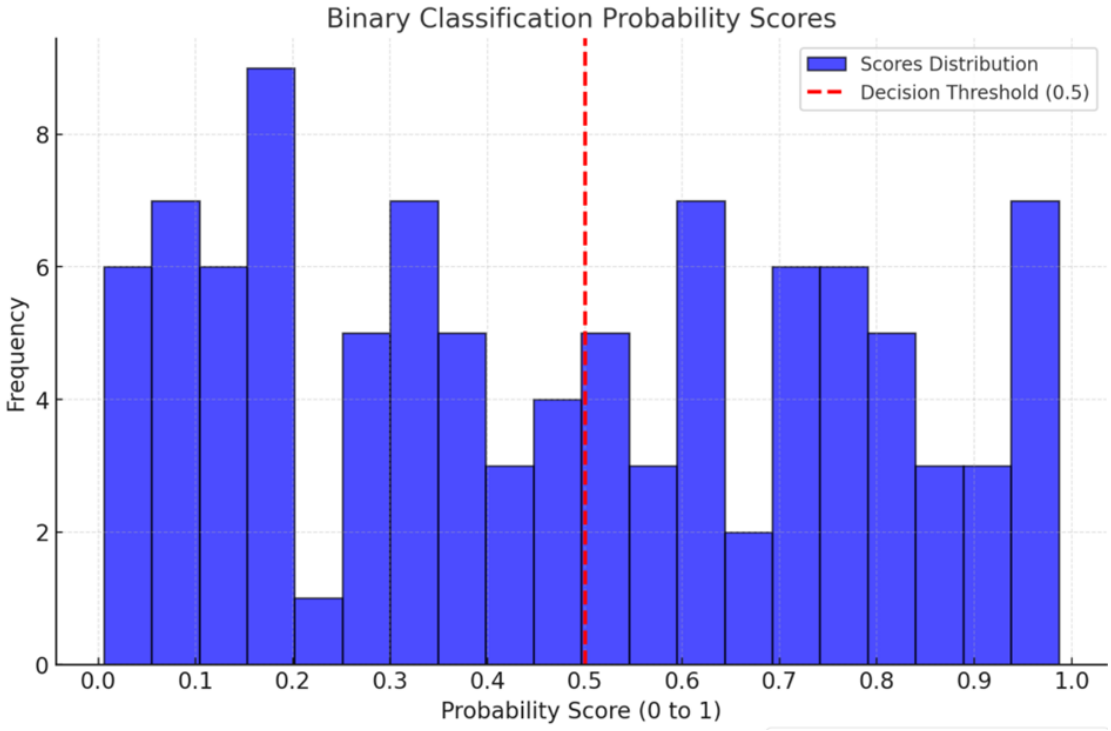

Figure 28. Binær klassifikation

8 TEST

For at vurdere modellens kapacitet er de sidste 20% af datasættet, som ikke har , været anvendt i træningsfasen, blev benyttet til test. Testen fokuserer på analysere precision, recall og balancen mellem korrekte forudsigelser og falske positive ved justering af tresholds.

Analyse af modelresultater og threshold-justering

Testen viser lovende resultater, og med en opdatering af thresholds til 0.376 viser det sig, at testen lever op til kravspecifikationen med kun fire falske positive og en recall på 95%

	F1	Accuracy	Precission	Recall
Ensemble med updateret Threshold-Test	0.85	0.82	0.77	0.96

Table 5

8.1 Falske negative

Ved afslutningen af testen blev der udført en analyse af de fire falske negative prædiktioner, da disse resultater anses for at være særligt vigtige. Imidlertid afslørede den tilhørende graf ingen tydelige mønstre blandt de falske negative observationer, som kunne forklare modellens fejlklassificeringer. Dette kan potentielt skyldes begrænsninger i datamængden, idet et større og mere omfattende datasæt muligvis ville kunne give yderligere indsigt og afdække underliggende mønstre (se Figur30).

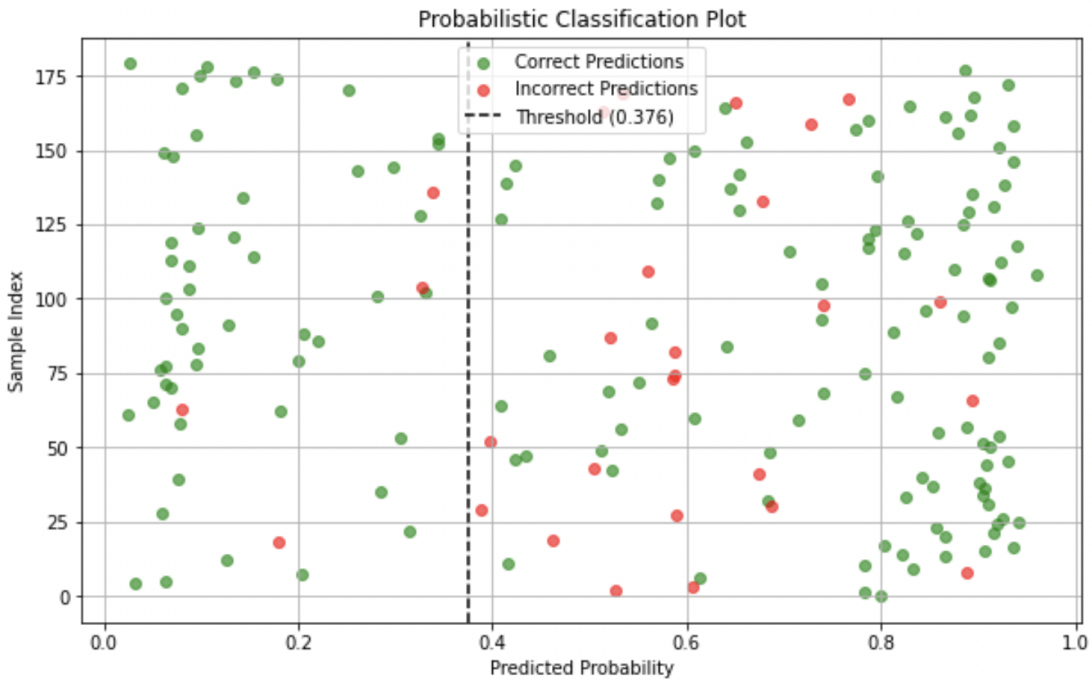

Figure 29. probability score

Instance	Probability	Age	Sex_M	Chestpain Type	ExerciseAngina yes	MaxHR	Fasting BS
18	0.18	47	0	NAP	0	170	1
63	0.08	48	1	0	0	168	0
104	0.33	58	1	ATA	0	160	0
136	0.34	40	0	0	0	130	0

Falske negative prædiktioner

Figure 30. Falske Negative Klassifikationer

8.2 Etiske aspekter i forbindelse med projektet

Ifølge Etisk Råd handler etik om, hvad man bør gøre både som enkeltindivid og samfund [35]. Etiske retninger som pligt-, nytte- og commonsense etik har forskellige tilgange til de problematikker, der kan opstå i brugen af kunstig intelligens (AI)

Den hurtige udbredelse og udvikling inden for AI introducerer ofte nye etiske spørgsmål omkring anvendelse og hvad der er forsvarligt. I artiklen Ethics of Artificial Intelligence in Medicine beskrives, hvordan skævhed i data kan resultere i diskrimination på baggrund af f.eks. køn, etnicitet og socioøkonomiske forhold[37]. Ved begrænset data for specifikke befolkningsgrupper kan prædiktioner for disse, blive dårligere selvom modellen ellers præsterer på niveau med eksperter. Her er det også væsentligt at bemærke, at en stor del af den historiske tilgængelige sundhedsdata har en overrepræsentation af hvide medvirkende, hvilke kan bidrage yderligere til diskrimination.

I datasættet 'Heart Failure Predictions' kan lignende problematikker gøre sig gældende. Kvinder er underrepræsenterede og derfor kan modellens generaliserbarhed til kvinder være begrænset. Desuden kan den manglende information om etnicitet og socioøkonomiske forhold også tænkes at påvirke modellens generaliserbarhed og retfærdighed.

Et andet etisk aspekt, der bør overvejes ved brug af AI, er, at disse metoder ikke undergår de samme strenge krav for evaluering som der ellers er et krav om i forbindelse med implementering af nye sundhedsteknologier[37]. Desuden er modellernes beslutningsgrundlag ikke altid lige så transparent som eksisterende metoder indenfor sundhedsteknologi (f.eks. Score), hvilket kan udfordre patienters autonomi og mulighed for at fortage beslutninger på et informeret grundlag.

Etiske retninger ville have forskellige tilgange til brug af AI. Pligtetik, hvor handlinger i sig selv skal være etisk forsvarlige [35] ville kunne skabe konflikter med brugen af AI fordi man uvilkårligt ville komme til at træffe nogle valg, der ville få negative konsekvenser for nogen. I ensemblet der er bygget i denne rapport, er recall - precision trade off et eksempel på et valg man ikke ville kunne træffe udfra et pligtetisk grundlag. I nytteetik, der beskæftiger sig med handlingers konsekvenser og handler om at maksimere den samlede gode effekt ville brugen af AI kunne være fordelagtig. Dette skyldes, at man blandt andet ville være i stand til at behandle meget data, ofte opnå gode resultater, der kan være med til at forbedre behandlingen for flest mulige mennesker. Hvis nytten af en handling derimod er skævt fordelt, ville det ikke leve op til de nyttetiske principper[35]. Commonsense etik er baseret på følgende principper: ikke skadevolden, fremme velvære, autonomi og retfærdighed. I forhold til denne etiske retning ville der også opstå potentielle problemer ved brug af AI. Blandt andet ville de mulige uligheder for effekten i forskellige befolkningsgrupper ikke leve op til retfærdighedsprincippet. [Kilde 4]

Ved brug af AI bør man også overveje, hvor ansvaret for de fejl, der laves skal placeres. Man kunne forestille sig situationer, hvor man misser nogle sygdomstilfælde, der ville være blevet opdaget af en ekspert, hvis eksperten for eksempel deducerer yderligere end der er tilgængeligt i data i deres møde med patienten. Sådanne tilfælde vil rejse et etisk dilemma omkring om man med brug af AI kommer til at forvolde mere skade end gavn, specielt hvis den type af fejl, der laves er systematisk og for eksempel laves på specifikke aldersgrupper, etniciteter eller køn. Dette kunne være tilfældet i denne model, hvor kvinder er underrepræsenterede og manglende oplysninger kunne forårsage diskrimination. Her ville man komme til at forringe lige behandling af alle i sundhedsvæsnet selvom man potentielt kan behandle flere mennesker i alt.

9 DISKUSSION

I det følgende afsnit vil der blive diskuteret, hvordan machine learning-modeller kan anvendes til at klassificere kardiovaskulær sygdom og anbefale yderligere undersøgelser med henblik på at reducere mortalitet og morbiditet. Der vil blive diskuteret valg af modeller, justering af thresholds, inddragelse af relevante faktorer samt etiske overvejelser for at sikre effektivitet og ansvarlig anvendelse i klinisk praksis. Derudover vil der blive diskuteret, hvordan udeladelse af visse features kan føre til bias og diskrimination, hvilket kan påvirke modellernes precision og retfærdighed.

9.1 Konsekvenser af valg af model

Ud af de tre classifiers der blev afprøvet, præsterede SVM bedst på Recall og F1.SVM ville derfor være det bedste valg udover ensemble modellen. Man kunne derfor have valgt SVM i stedet for ensemble modellen med argumentet, at den simpleste løsning som regel er den bedste. Simplicitet gavner både i forhold til krav af computerevne, strøm og ressourcer, men også gennemsigtighed, som særligt er vigtigt

i sundhedssektoren, hvor sensitive personoplysninger behandles og menneskers helbred vurderes. Når man bruger en machine learning model til at vurdere helbredsstatus, er det særligt vigtigt at vide hvordan modellen laver denne vurdering. Dette er bl.a. for at undgå bias, og at modellen kan komme til at få en negativ betyning for helbredet . Ensemblemodeller er som udgangspunkt mindre gennemsigtige end en enkeltstående classifier. Da den endelige beslutning og prædiktion er lavet af ensemblemodellen, vil der være kombinationer af flere forskellige modeller. Afhængigt af hvor mange modeller der er kombineret, og dermed hvor kompleks ensemblemodellen er, kan det være svært at forstå, hvorfor den prædiktere som den gør [7] [21]. SVM er særlig effektiv i håndtering af høj dimensionalitet i data, men jo højere dimensionaliteten er, jo mere kompleks bliver SVM, hvilket gør den mindre gennemsigtig i beslutningstagningen [15] [34]. SVM tilbyder dermed ikke en tilstrækkelig gennemsigtighed, som argument for at benytte den i stedet for ensemblemodellen. Ensemblemodellen, der blev brugt som det endelige valg til test, bruger SVM, Logistisk Regression og Random Forest. Logistisk Regression tilbyder høj gennemsigtighed [2], men Random Forest, som er en ensemble læringsmetode, lider af samme manglende gennemsigtighed som SVM [26]. Ensemblemodellen benytter dermed to classifiere med lav gennemsigtighed, hvilket kan være problematisk at forsvare til brug i sundhedssektoren på trods af at den performer bedst på alle klassifikations metrikker.

Justeringen af threshold-værdien, der er foretaget på ensemblemodellen, kunne også være blevet afprøvet på de individuelle classifiers og en sammenligning kunne laves mellem dem og ensemblemodellen. Her kunne det så afgøres om en enkeltstående classifier kunne bruges som den endelige model, baseret på særligt dens recall score. Der blev dog ikke lavet afprøvninger af classifierne med brug af threshold justering grundet begrænset tid i projektet.

9.2 Perspektivering til kravsspecifikation

Gennem projektet har der været fokus på at opfylde både de funktionelle og ikke-funktionelle krav i kravspecifikationen. I forhold til funktionelle krav er der arbejdet på enkelte mål, særligt krav 1 og 2, men disse er endnu ikke opfyldt, da modellen ikke kan tage imod brugerinput og patientdata.

For de ikke-funktionelle krav er der gjort fremskridt med krav 10, 11, 12 og 13, som sikrer en robust model i forhold til klassifikation af kardiovaskulær sygdom. Med længere tid og flere ressourcer kunne yderligere arbejde på disse krav havde været fuldført.

Fremadrettet vil der prioriteres udvikling af funktionalitet til brugerinput (krav 1) for at forbedre systemets anvendelighed. Samtidig anbefales det at afsætte flere ressourcer til at opfylde krav 10 fuldstændigt, da dette er essentielt for modellens præstation. En løbende evaluering af delmål og systemkrav vil være nødvendig for at sikre fremdrift og identificere udfordringer tidligt.

Funktionelle krav:	**I mål**	**Arbejdet på**
1. Systemet skal kunne anvende en trænet machine learning model til at klassificere om patient har eller ikke har kardiovaskulær sygdom baseret på brugerinput	✗	✓
2. Systemet skal kunne tage imod patientdata som input, som er relevant for forudsigelse af kardiovaskulær sygdom.	✗	✓
3. Systemet skal kunne validere inputdata for at sikre, at de korrekte værdier bliver indtastet	✗	✗
4. Systemet skal kunne indsamle og lagre patientdata der inputtes, så det kan bruges til at forbedre modellen løbende.	✗	✗
5. Systemet skal kunne anonymisere patientdata.	✗	✗
6. Systemet skal kunne forhindre brug af model ved manglende data	✗	✗

Table 6. Oversigt over status på de **funktionelle krav** fra kravsspecifikationen

Ikke funktionelle krav:	**I mål**	**Arbejdet på**
7. Systemet skal kunne beregne patientens status på under 10 sekunder	✗	✗
8. Systemet skal være intuitivt og kræve minimal træning af sundhedspersonalet.	✗	✗
9. Det skal være nemt og simpelt at indtaste brugerinput samt forstå og evaluere output.	✗	✗
10. Modellen skal have en accuracy på minimum 90%	✗	✓
11. Recall skal være $> 95\%$	✓	✓
12. F1 score skal være så høj som muligt med et passende trade off mellem recall og precision.	✓	✓
13. Precision skal vægtes lavere end recall, da det er vigtigere at fange alle sygdomstilfælde.	✓	✓
14. Systemet skal være tilgængeligt døgnet rundt, med minimal nedetid (maksimalt 5%)	✗	✗
15. Systemet skal være tilgængeligt som en web-løsning. Den skal være tilgængelig fra moderne web-broswere (Chrome, Firefox, Internet explore, Safari etc.)	✗	✗
16. Systemet skal kunne oversættes til flere sprog hvis den skal kunne anvendes i andre lande.	✗	✗
17. Modellen skal monitoreres regelmæssigt da den skal sikre høj accuracy og pålidelighed.	✗	✗
18. Modellen skal opdateres og gentrænes hver 6. måned med nye data.	✗	✗
19. Systemet skal kunne håndtere store mængder patientdata effektivt	✗	✗
20. Systemet skal kunne tilpasses andre lande, da befolkningsgrupper kan have varierende risiko for at udvikle CVD.	✗	✗
21. Systemet skal kunne håndtere mange samtidige brugere uden det påvirker ydeevne	✗	✗

Table 7. Oversigt over status på de **ikke funktionelle krav** fra kravsspecifikationen.

9.3 Diskussion af udvalgte features i datasættet

De features der indgår i datasættet er hovedsagligt kendte risikofaktorer for hjertekarsygdom. Dette stemmer godt overens med at AI i sundhedsvæsnet gerne skal have en transparent fremgangsmetode, hvor der er en kausal sammenhæng mellem risikoprofil og udfald. Derudover indeholder datasættet information om køn og alder. Information om etnicitet, land og socioøkonomiske forhold indgår ikke og det kan skabe bias, hvis der er forskel på hvordan risikoprofilen og udviklingen af hjertekarsygdomme ser ud for disse grupper. Hvis inklusionen i data er ulige for nogle af disse grupper som det for eksempel er for kvinder i dette datasæt, kan vores model komme til at diskriminere på baggrund af disse faktorer. Det sker fordi nogle risikofaktorer som måske er specifikke for nogle af de nævnte grupper, kunne tænkes at blive tillagt en mindre værdi end de reelt har i den pågældende gruppe. Udelukkelsen af disse features kan have negative konsekvenser for modellens præsentationsmål. Etnicitet og land kan indirekte afspejle genetiske eller miljømæssige faktorer, som påvirker sundhed, såsom forskelle i kostvaner, livsstil eller risiko for arvelige sygdomme. Ved at udelade disse variabler riskerer modellerne at overse vigtige forskelle, hvilket kan føre til mindre præcise risikovurderinger og diskrimination, hvis det påvirker bestemte grupper.

En anden udfordring er, at udeladelsen af socioøkonomiske faktorer kan begrænse modellens evne til at identificere risici for sårbare grupper. For eksempel kan lav socioøkonomisk status indikere begrænset adgang til sundhedsvæsnet og øget risiko for sygdom. Uden denne information kan modellen undervurdere risici for disse grupper, hvilket kan resultere i diskrimination.

For at reducere diskrimination mellem forskellige grupper kunne en mulighed være, at anvende

forskellig vægtning af falske positive og falske negative i modellen afhængigt af gruppen. For eksempel kan modellen prioritere at reducere falske negative for de underrepræsenterede grupper ved at sætte lavere krav til hvornår de placeres i kategorien syg. Herved kan risikofaktorer for disse grupper blive tillagt mere værdi. Dette kan også have som konsekvens at man får et større antal falske positive i disse grupper, men være med til at reducere diskrimination forårsaget af data, hvor repræsentationen på tværs af grupperne ikke er ens.

Ved manglen af features BMI, rygning, kost og motion i modellen risikerer man, at der mangler information om afgørende elementer af sygdomsudviklingen. Eksempelvis er rygning en kendt risikofaktor for udvikling af blodpropper, og forhøjet BMI kan ledsages af forhøjet blodtryk og forhøjet kolesterol. Når man udelader disse features fra en model om kardiovaskulær sygdom, kan det føre til undervurdering eller oversete risici hos personer med for eksempel overvægt eller et høj tobaksforbrug. Konsekvensen er en mindre præcis vurdering af den samlede risiko for patienten, hvilket kan lede til fejl eller mangelfulde behandlingsanbefalinger. Dette kan også lede til diskrimination, hvis de udeladte features forbundet med forhøjet risiko er mest fremtrædende hos nogle bestemte grupper.

9.4 Diskussion om resultater

Modellen opnåede positive resultater med et særligt fokus på høj recall, da omkostningerne ved falske negative er kritiske og potentielt livstruende. De fire identificerede falske negative observationer bestod af to mænd og to kvinder. En nærmere analyse af testresultaterne indikerer, at det ikke er muligt at opnå en recal på100 uden at kompromittere præcisionen væsentligt. Dette skyldes, at fejlagtige forudsigelser ofte er karakteriseret ved lave sandsynlighedsscorer, hvoraf nogle observerede værdier var helt ned til 0.08. Testen blev udført på et relativt begrænset datasæt med kun 170 observationer. Selvom resultaterne er lovende, begrænser den lille testmængde modellens evne til at identificere og generalisere mønstre for fejlklassifikationer. Derudover er fraværet af vigtige features i datasættet en betydelig udfordring. Dette resulterer i, at modellen har svært ved at forudsige visse observationer nøjagtigt, da afgørende variabler, der potentielt kunne forklare positive tilfælde med lave sandsynlighedsscorer, ikke er tilgængelige. For at forbedre modellens ydeevne fremadrettet er det afgørende at udvide datasættet både i antallet af observationer og i antallet af relevante features. Desuden kunne det være hensigtsmæssigt at udvikle separate modeller, der tager højde for køns- og regionsspecifikke forskelle, da prævalensen af positive og negative tilfælde kan variere betydeligt mellem disse grupper. En sådan tilgang vil sandsynligvis kunne forbedre modellens præcision og robusthed.

9.5 Modellens værdi for forbyggelse af mortalitet og morbiditet

Dette projekts model beskæftiger sig med data, hvor ground truth er labeled som tilstedeværelse af CVD eller ej. I forhold til forebyggelse af mortalitet og morbiditet som konsekvenser af kardiovaskulær sygdom misser man med denne model, de tilfælde, hvor man helt ville kunne forhindre udviklingen af CVD og derved have sparet både individet og samfundet for tabte leveår og økonomiske omkostninger i sundhedsvæsnet. På trods af, at denne model ikke kan bruges til at prædiktere hvem, der er i risiko for udvikling af CVD, kan der alligevel opnås forebyggelse af præmature mortalitet indenfor hjertekarsygdomme ifølge tabellen fra 1 fra OECD [29]. Hvis man implementerede en model som denne i almen praksis, hvor den indgik i screeninger af patienter når der forelå EKG og blodprøver, kunne man forestille sig, at det kunne have en besparende effekt fordi man derfor hurtigt kunne prædiktere, hvem der ville have mest behov for viderebehandling og anbefalinger om livsstilsændringer.På den anden side ville man risikere at komme til at overbehandle udrede patienter, der ikke har brug for det. Det kan både være problematisk ifht. belastning af sundhedsvæsnet og for dem, der skal udredes. I forhold til de direkte omkostninger forbundet med CVD ville man potentielt kunne opnå besparelser på de direkte omkostninger for indlæggelser fordi tidlig diagnosticering af CVD ville kunne forebygge alvorlig og mere behandlingskrævende hjertekarsygdom. Også de indirekte omkostninger for CVD ville kunne påvirkes fordi forebyggelse af mere alvorlig sygdom og præmatur mortalitet, potenitelt ville resultere i et mindre produktivitetstab som følge af uarbejdsdygtighed, handicap og for tidlig død.

9.6 Opsummering af diskussion

Her vil man kunne referere tilbage til problemformuleringen, som undersøger, om machine learning (ML) metoder kan klassificere tilstedeværelsen af hjertekarsygdomme og udvikle et softwareprodukt, der forebygger mortalitet og morbiditet ved at anbefale yderligere undersøgelse. Diskussionen viser, at ML-modeller ville kunne have potentiale til at opfylde dette mål.

Af de testede metoder præsterede ensemblemodellen bedst med hensyn til recall og F1-score. Ensemblemodellen blev derfor valgt som den endelige løsning, da en høj recall er kritisk for at minimere falske negative i en klinisk kontekst, hvor oversete tilfælde kan have alvorlige konsekvenser. Denne egenskab understøtter direkte problemformuleringens mål om at identificere patienter, der kræver yderligere undersøgelse, og dermed bidrage til forebyggelse af mortalitet og morbiditet. Samtidig viste diskussionen, at en alternativ tilgang med SVM kunne have været en relevant overvejelse. SVM har lavere krav til ressourcer, men tilbyder dog ikke større gennemsigtighed hvilket er vigtigt i sundhedssektoren, hvor forståelsen af modellens beslutningsproces er essentiel. Ensemblemodellen er mindre gennemskuelig, hvilket kan skabe udfordringer i forhold til patienternes og sundhedspersonalets tillid.

Justering af thresholds i de enkelte klassifikationsmodeller kunne potentielt opnå lignende resultater som ensemblemodellen, men som nævnt tidligere blev dette ikke udforsket grundet tidsbegrænsninger. Desuden kan manglende inddragelse af faktorer som etnicitet, socioøkonomiske forhold, BMI, rygning og kost føre til bias og lavere accuracy i modellen, hvilket er problematisk i en klinisk kontekst, hvor retfærdighed og lige behandling er vigtige. Etiske overvejelser omkring brugen af ML i sundhedssektoren omfatter behovet for gennemsigtighed, ansvarlighed og undgåelse af diskrimination.

10 KONKLUSION

Formålet med dette projekt var at undersøge, om man ved brug af machine learning metoder kunne klassificere, om individer har hjertekarsygdom for dermed at bygge et softwareprodukt, der ved brug i almen praksis kan forebygge mortalitet og morbiditet ved at anbefale, om en patient bør sendes til yderligere undersøgelse. Problemanalysen redegør for studiet, der sammenligner ACC/AHA med SVM, og fandt at SVM præsterede betydeligt bedre som risiko beregner af CVD. Dette gav inspiration til at teste forskellige machine learning modeller i en såkaldt 'Quick and Dirty' og derefter udvælge de tre bedste. Seks machine learning modeller blev testet og de tre bedste blev udvalgt: SVM, Random Forest og Logistisk Regression. De optimale hyperparametre for hver af de tre classifere blev fundet hvorefter de blev kombineret til en ensemblemodel.

Ensemble modellen viste sig at præstere bedre en samtlige af de enkeltstående modeller. Eftersom det blev vurderet at konsekvenserne ved en falsk alarm, falske positive, er mindre alvorlige end konsekvenserne ved at vurdere en syg person som rask, falsk negativ, så blev thesholden for ensensemblen flyttet tættere mod nul med henblik på at øge modellens recall, og derved fange flere af de syge personer. Efter flytningen af threshold blev ensemblet testet på testsættet og leverede en præstation på F1: 0.85, Accuracy: 0.82, Precission: 0.77 og Recall: 0.96. Derved har machine learning modellen opnået målet fra kravsspeficikationen om en recall på over 0.95. Den har ikke opnået målet om en accuracy på minimum 0.90. Projektet har ikke fuldført målet om at bygge et softwareprodukt der kan tage imod patientdata som input til forudsigelse af kardiovaskulær sygdom. Fremadrettet ville det være relevant at fortsætte arbejdet på modellens accuracy, samt at begynde på udviklingen af modellens evne til at tage imod brugerinput og softwaren omkring.

11 BRUG AF GENERATIV AI

I dette projekt har gruppen blandt andet benyttet ChatGPT, hvilket primært har været anvendt til grammatiske rettelser og forslag til omstrukturering af tekst. Derudover har vi anvendt ChatGPT til debugging for hurtigere og mere effektivt at kunne løse problemer og fejl i vores kode. Dette har været særlig nyttigt i læringsprocessen omkring kodeskrivning. Desuden har vi i gruppen anvendt ChatGPT som understøttende læringsværktøjer, hvilket har været et gavnligt supplement til at forstå og diskutere kildemateriale og metoder. Udover ChatGPT blev der anvendt Co-pilot som en anden ressource i kodningsprocessen, herunder til opsætning og implementering af en logistisk regression model. Co-pilot blev anvendt til at generere kode fungerende som inspiration samt et støttende hjælpemiddel til at sikre en korrekt og struktureret opbygning af kode. Alt i alt har ChatGPT og Co-pilot været en god støtte og hjælpemiddel i dette projekt ved at give forslag til forbedring af både det skriftlige produkt samt forståelse af metoder og kode.

12 BIBLIOGRAFI

REFERENCES

[1] (2021). What is random forest and how it works. *Towards Machine Learning*.

[2] 2OS (n.d.). Logistic regression in the credit risk industry. Tilgået: 2024-12-12.

[3] AI, P. (n.d.). Understanding probability score. Accessed: 2024-12-18.

[4] Association, A. H. (2024a). American heart association homepage. https://www.heart.org/en/. Tilgået: 2024-11-18.

[5] Association, A. H. (2024b). Article from circulation: Details available online. *Circulation*, n.d.(n.d.):n.d. Tilgået: 2024-12-06.

[6] Author(s) (Year). Addressing the global burden of cardiovascular diseases; need for scalable and sustainable frameworks. *PubMed Central (PMC)*.

[7] Bhat, S. (n.d.). Machine learning ensemble methods with chatgpt open ai. Tilgået: 2024-12-12.

[8] contributors, W. (n.d.). Probabilistic classification — wikipedia. Tilgået: 2024-12-12.

[9] for Almen Medicin (DSAM), D. S. (2024). Risikovurdering – hjerte-kar-sygdomme. Tilgået: 2024-12-17.

[10] for Disease Control, C. and (CDC), P. (2024). Heart disease risk factors. Tilgået: 2024-12-17.

[11] Ford, T. J. and Berry, C. (2020). Angina: contemporary diagnosis and management. *Heart*, 106(5):387–398.

[12] Forfatter(e) (Årstal). Artikelens titel. *Navn på tidsskrift*, Volumen(Nummer):Sidenumre. Tilgået: 2024-12-12.

[13] Foundation, B. H. (n.d.). Bhf cvd statistics: Uk factsheet. Tilgået: 2024-12-06.

[14] GeeksforGeeks (n.d.a). Probability density estimation - maximum likelihood estimation. Tilgået: 2024-12-06.

[15] GeeksforGeeks (n.d.b). Support vector machine (svm) algorithm. Tilgået: 2024-12-12.

[16] GeeksforGeeks (n.d.c). Understanding logistic regression. Tilgået: 2024-12-06.

[17] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., Sebastopol, CA, second edition. Tilgået: 2024-12-06.

[18] Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Sebastopol, CA, 3rd edition. Kapitel2, Kapitel 3, Kapitel 7.

[19] Hamed, M., Dasari, G., Casale, J. A., Kaur, N., and Karl, M. (2022). The use of romhilt-estes criteria in the presumptive electrocardiographic diagnosis of left ventricular hypertrophy in comparison to voltage-based criteria. *Cureus*, 14(8):e28003.

[20] Hjerteforeningen (n.d.). Nøgletal om hjerte-kar-sygdomme. Tilgået: 2024-12-06.

[21] Hunters, D. H. (n.d.). Ensemble learning vs single models: Maximizing predictive performance. Tilgået: 2024-12-12.

[22] IBM (n.d.). What is logistic regression? Tilgået: 2024-12-06.

[23] Jain, A. (n.d.). Svm kernels and its type. Tilgået: 2024-12-06.

[24] Jang, D. (2024). Challenges of kernel selection in support vector machines: An in-depth exploration. Tilgået: 2024-06-17.

[25] Kakadiaris, I. A., Vrigkas, M., Yen, A. A., Kuznetsova, T., Budoff, M., and Naghavi, M. (2018). Machine learning outperforms acc/aha cvd risk calculator in mesa. *Journal of the American Heart Association*, 7(22):e009476.

[26] Kalra, K. (n.d.). Random forest algorithm. Tilgået: 2024-12-12.

[27] learn Developers, S. (n.d.). sklearn.metrics.pairwise.rbf$_k ernel|scikit - learn1.5documentation.Accessed : 2024 - 12 - 18$.

[28] Manuals, C. (2018). The st segment: physiology, normal appearance, st depression & st elevation. Accessed: 2024-12-18.

[29] OECD and Eurostat (2019). Avoidable mortality: 2019 joint oecd/eurostat list of preventable and treatable causes of death. Technical report, Organisation for Economic Co-operation and Development (OECD) and Eurostat. Tilgået: 2024-06-17.

[30] Ohiri, E. (2024). Accuracy, precision and recall in deep learning. Tilgået: 2024-10-12.

[31] Olsen, M. H., Sehestedt, T., and Prescott, E. (2009). Den kardiovaskulære risikoprofil og antihypertensiv behandling. *Ugeskrift for Læger*. Tilgået: 2024-12-18.

[32] Organization, W. H. (2024). Cardiovascular diseases (cvds). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Tilgået: 2024-12-18.

[33] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011). Knnimputer - scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html. Tilgået: 2024-06-17.
[34] Roy, S. (n.d.). A closer look at svm and k-nearest neighbors for data classification. Tilgået: 2024-12-12.
[35] Råd, D. E. (n.d.). Hvad er etik? Tilgået: 2024-12-12.
[36] Sanga, R. P. (n.d.). Logistic regression and regularization: Avoiding overfitting and improving generalization. Tilgået: 2024-12-06.
[37] Savulescu, J., Giubilini, A., Vandersluis, R., and Mishra, A. (2024). Ethics of artificial intelligence in medicine. *Singapore Medical Journal*, 65(3):150–158. Tilgået: 2024-12-12.
[38] Science, T. D. (n.d.). Support vector machine explained. Tilgået: 2024-12-12.
[39] Scikit-learn contributors (2020). Randomforestclassifier. *Scikit-learn documentation*.
[40] Scikit-learn Developers (2024). Scikit-learn documentation: sklearn.ensemble. https://scikit-learn.org/1.5/api/sklearn.ensemble.html. Tilgået: 2024-12-18.
[41] Spiceworks (n.d.). Linear regression vs logistic regression. Tilgået: 2024-12-06.