



EXAMENSARBETE INOM TEKNIK,
GRUNDNIVÅ, 15 HP
STOCKHOLM, SVERIGE 2021

Data mining inom tillverkningsindustrin

En fallstudie om möjligheten att förutspå
kvalitetsutfall i produktionslinjer

LISA JANSON

MINNA MATHISSON

DATA MINING INOM TILLVERKNINGSINDUSTRIN

Minna Mathisson och Lisa Janson

Abstract—As the adaptation towards Industry 4.0 proceeds, the possibility of using machine learning as a tool for further development of industrial production, becomes increasingly profound. In this paper, a case study has been conducted at Volvo Group in Köping, in order to investigate the wherewithals of predicting quality outcomes in the compression of hub and mainshaft. In the conduction of this study, three different machine learning models were implemented and compared amongst each other. A dataset containing data from Volvo’s production site in Köping was utilized when training and evaluating the models. However, the low evaluation scores acquired from this, indicate that the quality outcome of the compression could not be predicted given solely the variables included in that dataset. Therefore, a dataset containing three additional variables consisting of fabricated values and a known causality between two of the variables and the quality outcome, was also utilized. The purpose of this was to investigate whether the poor evaluation metrics resulted from a non-existent pattern between the included variables and the quality outcome, or from the models not being able to find the pattern. The performance of the models, when trained and evaluated on the fabricated dataset, indicate that the models were in fact able to find the pattern that was known to exist. Support vector machine was the model that performed best, given the evaluation metrics that were chosen in this study. Consequently, if the traceability of the components were to be enhanced in the future and an additional number of machines in the production line would transmit production data to a connected system, it would be possible to conduct the study again with additional variables and a larger data set. The fact that the models included in this study succeeded in finding patterns in the dataset when such patterns were known to exist, motivates the use of the same models. Furthermore, it can be concluded that with enhanced traceability of the components and a larger amount of machines transmitting production data to a connected system, there is a possibility that machine learning models could be utilized as components in larger business monitoring systems, in order to achieve efficiencies.

Sammanfattning—I detta arbete har en fallstudie utförts på Volvo Group i Köping. I takt med övergången till industri 4.0, ökar möjligheterna att använda maskininläring som ett verktyg i analysen av industriell data och vidareutvecklingen av industriproduktionen. Detta arbete syftar till att undersöka möjligheten att förutspå kvalitetsutfall vid sammanpressning av nav och huvudaxel. Metoden innefattar implementering av tre maskininlärningsmodeller samt evaluering av dess prestation i förhållande till varandra. Vid applicering av modellerna på monteringsdata från fabriken erhöles ett bristfälligt resultat, vilket indikerar att det utifrån de inkluderade variablerna inte är möjligt att förutspå kvalitetsutfallet. Orsakerna som låg till grund för resultatet granskades, och det resulterade i att det förmodligen berodde på att modellerna var oförmögna att finna samband i datan eller att det inte fanns något samband i datasetet. För att avgöra vilken av dessa två faktorer som var avgörande skapades ett fabricerat dataset där tre nya variabler introducerades. De fabricerade värdena på dessa variabler skapades på sådant sätt att det fanns syntetisk kausalitet mellan två av variablerna och kvalitetsutfallet. Vid applicering av modellerna på den fabricerade datan, lyckades samtliga modeller identifiera det syntetiska sambandet. Utifrån det drogs slutsatsen att det bristfälliga resultatet inte berodde på modellernas prestation

utan att det inte fanns något samband i datasetet bestående av verklig monteringsdata. Det här bidrog till bedömningen att om spårbarheten på komponenterna hade ökat i framtiden, i kombination med att fler maskiner i produktionslinjen genererade data till ett sammankopplat system, skulle denna studie kunna utföras igen, men med fler variabler och ett större dataset. Support vector machine var den modell som presterade bäst, givet de prestationsmått som användes i denna studie. Det faktum att modellerna som inkluderats i den här studien lyckades identifiera sambandet i datan, när det fanns vetskap om att sambandet existerade, motiverar användandet av dessa modeller i framtida studier. Avslutningsvis kan det konstateras att med förbättrad spårbarhet och en allt mer uppkopplad fabrik, finns det möjlighet att använda maskininlärningsmodeller som komponenter i större system för att kunna uppnå effektiviseringar.

Index Terms—Data mining, maskininläring, kvalitetskontroll, industriproduktion, logistisk regression, k-nearest neighbor, support vector machine

I. INLEDNING

DET svenska välståndet har varit, och är än idag, starkt sammankopplat till de svenska exportföretag vars konkurrenskraft bygger på förmågan att anpassa såväl produkter som produktion efter föränderliga marknader [28]. Ur ett initiativ för att öka den tyska industrins konkurrenskraft, växte begreppet Industri 4.0 fram och idag har konceptet bakom begreppet fått sådan slagkraft att det också benämns som den fjärde industriella revolutionen. Med industri 4.0 menas den revolution som IoT, automatisering, maskininläring och tillgång på realtidsdata kommer att medföra [19].

Övergången till industri 4.0 medför såväl utmaningar som möjligheter för industrin och teknikutvecklingen. Framförallt betonas vikten av att kunna omvandla de stora mängder data som industri 4.0 medför, till nya insikter kring alltifrån affärer till effektiviseringsmöjligheter i produktionen. Denna förmåga anses vara kritisk för industrins framtida konkurrenskraft. Den 21 januari 2016 beslutade regeringen om en nyindustrialiseringsstrategi för Sverige, benämnd som Smart Industri, där Industri 4.0 uttrycks som ett av fyra fokusområden. Där uttrycks det att målsättningen är att svenska industriföretags digitala utveckling ska vara ledande, liksom deras förmåga att utnyttja de möjligheter som digitaliseringen introducerar [28]. Detta poängterar vikten av digitaliseringen inom industrin och markerar området som detta arbete kommer att utföras inom.

A. Intressenter

Ett intressant område, inom vilket verktygen som industri 4.0 medför skulle kunna nyttjas, är kvalitetskontroller i produktionslinjer. Detta arbete undersöker dessa möjligheter, genom en fallstudie på Volvo Group i Köping, där möjligheten att förutspå kvalitetsutfall vid sammanpressning av nav och huvudaxel utreds. Att kunna förutspå huruvida en komponent

kommer att bli godkänd eller ej, skulle medföra möjlighet att hinna göra justeringar hos komponenten som gör att den kan bli godkänd i kontrollen. Detta skulle i sin tur kunna leda till besparingar i form av såväl pengar, som tid, samt ge bättre inblick i vilka variabler i produktionen som är relevanta för kvalitetsutfallet. Ett företag som påverkas i allra högsta grad av omställningen till smarta fabriker är Volvo Group, som kommer att vara extern samarbetspartner. På Volvo Group Trucks i Köping tillverkas växellådor till företagets lastbilar, bussar och dumprar, samt marina drev åt Volvo Penta. I en specifik del av tillverkningen erhålls fler icke-godkända kvalitetsutfall än önskat för en viss typ av komponent. Därmed finns ett intresse för att undersöka huruvida implementation av maskininlärningsmodeller, i kombination med utnyttjande av monteringsdata, hade kunnat ge nya insikter kring vilka variabler som påverkar kvalitetsutfallet för komponenten. I denna aspekt sammanförs därmed Volvo Groups intresse med syftet i detta arbete, eftersom det som undersöks skulle kunna vara ett potentiellt användningsområde för den produktionsdata som i nuläget enbart sparas utan vidare bearbetning.

Utöver Volvo Group, bedöms arbetet även generellt vara intressant för andra tillverkningsföretag som är intresserade av att nyttja data från den egna produktionen i syfte att identifiera möjliga effektiviseringar. Arbetet kommer att kunna tjäna som ett exempel på hur den ökade datatillgången, vilken Industri 4.0 medför, kan utnyttjas.

Arbetet kan också vara av intresse på grund av att det kan kopplas till ett av målen i regeringens nyindustrialiseringsstrategi, närmare bestämt att svenska industriföretag ska vara ledande gällande förmågan att utnyttja de möjligheter som digitaliseringen introducerar. Det skulle också kunna kopplas till det nionde globala utvecklingsmålet:

“Bygga upp en motståndskraftig infrastruktur, verka för en inkluderande och hållbar industrialisering och främja innovation” [35]

Dels på grund av arbetets fokus på de nya teknologier och modeller som framtidens smarta industrier eventuellt kan möjliggöra, men också på grund av att arbetets resultat skulle kunna användas för att identifiera möjliga effektiviseringar inom produktionslinjer i industrin, vilket i sin tur möjliggör resurssparande. Etiska aspekter anses irrelevanta, då arbetet enbart behandlar data från industriproduktion och därmed ingen persondata.

B. Problemformulering

Projektet syftar till att besvara följande frågeställning:

Vilken av modellerna Support Vector Machine, logistisk regression och K-nearest-neighbor presterar bäst vad gäller att förutspå kvalitetskontroller i industriproduktion?

Dessa tre modeller har valts på grund av att de har åtskilda tillvägagångssätt gällande att klassificera datapunkter, vilket är önskvärt då detta arbete innefattar en jämförelse av olika maskininlärningsmodellers klassificeringsförmåga. Uppgiften kommer innebära att maskininläring tillämpas på data som genererats från en avgränsad del i monteringen av växellådor

på Volvo Group i Köping. Med hjälp av historisk data som genererats under sammanpressning av nav och huvudaxel, kommer maskininläring tillämpas i hopp om att kunna förutspå hur olika värden på navets dimensioner, respektive huvudaxelns dimensioner, medför att den sammanpressade komponentens kvalitetsutfall blir godkänt eller ej. Med dimension avses kulmättet. Ett godkänt kvalitetsutfall kommer karaktäriseras av att presskraften som krävs för att skapa den sammanpressade komponenten hamnar inom ett förutbestämt intervall, vilket har såväl nedre som övre gräns. När “godkänt kvalitetsutfall”, respektive “icke-godkänt kvalitetsutfall” diskuteras i denna rapport, syftas det alltså på huruvida presskraften varit inom det godkända intervallet vid sammanpressningen av navet och huvudaxeln.

Det finns många variabler i produktionslinjen som skulle vara intressanta att inkludera i datasetet som modellerna får träna på, såsom exempelvis temperatur på komponenterna. I dagsläget finns dock inte verklig mätdata för detta. Därav kommer även ett fabricerat dataset att framställas, vilket kommer att innehålla tre ytterligare variabler. Dessa ytterligare variabler kommer att utgöras av “temperatur på navet”, “antal dagar på lager” och “temperatur i pressningsmaskinen”. De två första variabelerna är sådana som tagits upp under diskussioner med handledarna på Volvo. Under diskussionerna har det konstaterats att dessa två variabler hade varit intressanta att inkludera i studien om det hade funnits mätvärden att tillgå för dessa. Eftersom så inte är fallet, kommer det att skapas fiktiva värden för dem i det fabricerade datasetet. Gällande den tredje variabeln, “temperatur i pressningsmaskinen”, anses den vara relevant eftersom pressmaskinen är den del av produktionslinjen där pressningen som avgör kvalitetsutfallet sker och där fokus koncentreras i detta arbete. Därav, hade det i ett verkligt scenario varit rimligt att inkludera temperaturen i pressmaskinen, då detta skulle kunna vara en variabel som påverkar exempelvis maskinens prestanda.

Värdena i det fabricerade datasetet kommer att framställas på sådant sätt att det finns kausalitet mellan en del av variabelerna och huruvida presskraften är godkänd eller ej. Vetskapen om att det finns ett samband i datasetet, gör att modellernas förmåga att hitta detta samband kan utvärderas. Detta kan sedan sättas i relation till hur samma modeller presterar på datasetet som innehåller verklig data över enbart dimensioner på nav respektive huvudaxel. När datasetet innehållandes verklig monteringsdata diskuteras i rapporten, kommer detta dataset benämnas “verkligt dataset”.

Gällande arbetets tekniska relevans, anses det bidra med insikter om relevanta metoder och modeller i diskussionen kring Smart Industri, samtidigt som det bidrar till kunskaperna kring vilka möjligheter som följer från den Smarta Industrins ökade tillgång på data.

II. TEORI

I denna del redogörs teoretisk bakgrund till de modeller och koncept som är relevanta för studien. Avsnittet om data mining presenterar en förklaring av det område som detta arbete utförs inom. Arbetets övergripande metodologi beskrivs i stycket som innefattar CRISP-DM. De teoriavsnitt som behandlar

support vector machine, logistisk regression respektive K-nearest neighbor beskriver de modeller som jämförs i arbetet. Utöver teorin om modellerna presenteras också teori om hur dessa kan utvärderas, i styckena Prestationsmått, Precision-Recall-kurvor samt ROC-kurvor. Grundläggande teori om uppbyggnaden och hanteringen av dataset vid maskininläring förmedlas i teoriavsnitten Obalanserade dataset och K-fold cross validation. Teoridelarna Decision Support Systems och Business Activity Monitoring bidrar med kunskap om hur arbetet, och modellerna det innefattar, kan användas i praktiken inom industriproduktion.

A. Data Mining

Grundstenen inom data mining är att omvandla rådata till användbar information [23]. Det kan åstadkommas genom att, utifrån stora mängder data, upptäcka gömd information, vilket i sin tur uppnås med hjälp av att kombinera olika teknologier. Dessa kan exempelvis vara maskininläring, grafvisualisering och mönsteranalys. [37]

B. CRISP-DM

CRISP-DM är en metodologi för utformning av data-mining-projekt, bestående av följande sex faser:

1) Förståelse för verksamheten

En djupare förståelse för verksamheten där projektet genomförs ska erhållas i denna fasen. Vad önskas uppnås och varför? Syftet med denna fas är att kunna planera projektet och utforma tydliga mål.

2) Förståelse för datan

I denna fas ska initial data utforskas. Vilken data finns det tillgång till? Hur fås åtkomst till denna och hur är datan formaterad? Det bestäms vilken data som är önskvärd att erhålla och faktorer som påverkar kvalitén av denna data bör undersökas.

3) Dataförberedelser

För att kunna utforma modeller, krävs efterforskning i hur datan bör organiseras och formateras för att kunna användas på önskvärd sätt.

4) Modellering

I denna fas kan olika modeller testas för att undersöka vad som bäst passar datan och motiven.

5) Evaluering

Hur bra presterar modellen? Hur tillförlitliga är resultaten? Detta är frågor som bör undersökas i denna fas.

6) Spridning

Avslutningsvis behöver projektets resultat sammanställas på sådant sätt att det kan utnyttjas av intressenter.

Ofta krävs att projektet går tillbaka till tidigare steg upprepade gånger, vilket gör att den interna ordningen mellan de sex faserna kan vara relativt flytande. [10]

C. Obalanserade klasser

I dataset där mängden observationer i varje klass är oproportionerlig, är klasserna obalanserade. Eftersom många maskininlärningsalgoritmer försöker maximera accuracy, uppstår problem när antalet datapunkter i varje klass inte är av ungefär

samma storlek. Vid hantering av obalanserade klasser kan accuracy som utvärderingsmått därmed vara mycket missvisande och andra alternativ, såsom exempelvis F-measure, bör övervägas. En vanlig teknik som kan användas för att justera obalanserade klasser, är att utöka datasetet genom att framställa syntetiska datapunkter. En metod som kan användas för detta är exempelvis Synthetic Minority Oversampling Technique (SMOTE). Metoden använder en nearest-neighbors-algoritm, vilken gör att syntetiska datapunkter kan tas fram för att utnyttjas i träningsfasen. Vid användning av denna teknik, är det viktigt att påpeka att utökningen av datapunkter i minoritetsklassen enbart bör göras i träningsdatan, för att motverka överfitting. [9]

D. Support Vector Machine

Det finns många olika maskininlärningsalgoritmer, och en som ofta används för klassificeringsproblem är support vector machine (SVM). Det är en icke-linjär algoritm som är användbar vid både binärklassificering och multiklassificering. Dessutom är metoden lämplig om det förekommer mycket överlappning mellan klassernas data. SVM går ut på att datan formateras om till en högre dimension och ett hyperplan som separerar de två klasserna på bästa sätt genereras. Hyperplanet tas fram med hjälp av stödvektorer, det vill säga datapunkter från två olika klasser som är nära hyperplanet och därmed påverkar dess position och orientering, och som ger maximala marginaler. Vid klassificering, klassificeras en datapunkt utifrån vilken sida av hyperplanet som den befinner sig på. Klasstillhörigheten för punkten blir alltså densamma som klasstillhörigheten hos de punkter som befinner sig på den sidan [16].

Klassificeringsproblem där klasserna inte är linjärt separerbara kan lösas med hjälp av det så kallade Kernel-tricket. En kernel-funktion uttrycker ett förhållande mellan två vektorer i ett vektorrum av viss dimensionalitet. Det kernel-tricket innebär, är att olika kernel-funktioner kan appliceras så att data kan transformeras till högre dimensioner på effektivt sätt, utan att beräkningar blir alltför kostsamma. Genom att applicera en kernel-funktion utökas dimensionen i det ursprungliga rummet, så att klasserna kan separeras linjärt i den erhållna högre dimensionen. En bild som förtydligar detta kan ses i bilaga 1. Det finns flera olika kernel-funktioner, exempelvis linjär, polynomial eller gaussian-radial-basis [13]. Djupare teori kring kernel-funktionen gaussian-radial-basis finns att tillgå i bilaga 2.

E. K-Nearest Neighbor

K-nearest neighbor (KNN) är en övervakad maskininlärningsalgoritm som kan användas för att lösa både klassificerings- samt regressionsproblem [18]. Vid övervakad maskininläring tränar modellen på data som innehåller den korrekta klassificeringen, tillsammans med övriga parametrar. KNN-algoritmen har som utgångspunkt att liknande datapunkter existerar nära varandra. För att klassificera ny data baserar modellen sin bedömning på datapunktens k närmaste grannar. Antalet grannar som modellen tar hänsyn till beror på vilket värde k initierats till. KNN-algoritmen mäter avståndet till

grannarna, och den klassificering som är högst förekommande bland dessa datapunkter används för att klassificera den aktuella datapunkten. Vid mätningen av avstånd är det vanligast att det euklidiska avståndet används, vilket ges av formeln i bilaga 3. [20]

F. Binär logistisk regression

Regressionsanalys är en metod vars syfte är att analysera sambandet mellan två eller fler variabler, där en variabel är en responsvariabel som beror av resterande variabler [26]. Binär logistisk regression är en typ av regressionsanalys där responsvariabeln är binär och därmed endast har två utfall bestående av 0, alternativt 1 [25]. Det som ligger till grund för klassificeringen är att vardera datapunkt representeras av en särdragsvektor, bestående av dess värden på de olika variablerna. Utöver detta viktas även variablerna baserat på dess importans, och dessa vikter representeras av en vektor theta (θ). Med hjälp av maskininlärning kan dessa vikter bestämmas automatiskt och utan att det behövs assistans från domänexperter, vilket tidigare varit fallet. Givet datapunkten x , söks den klass y som maximerar följande uttryck:

$$\arg_y \max P(y|x)$$

Om datapunkten x representeras av särdragsvektorn med variablerna w_1, w_2, \dots, w_n gäller att uttrycket ovan kan skrivas om enligt nedan:

$$\arg_y \max P(y|w_1, \dots, w_n)$$

Det utförs vektormultiplikation av dessa två vektorer, x och θ , för vardera datapunkt. Detta resulterar i ett tal som ska representera sannolikheten att datapunkten ska klassificeras som den positiva klassen. Detta tal kan dock både vara negativt, och större än ett, så för att erhålla en sannolikhet mellan 0 och 1 används den logistiska funktionen, vilken visas nedan. [5]

$$\sigma(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

Den logistiska funktionen är en sigmoid-funktion och det returnerade värdet tolkas som sannolikheten att datapunkten tillhör den positiva klassen. [6]

G. K-fold cross validation

Korsvalidering används för att dela upp ett dataset i träningsdata samt testdata [24]. Det kan dessutom göras i flera olika kombinationer för att vidare testa modellen. Vid användning av k-fold cross validation delas datan in i k antal mindre delar, så kallade folds. Därefter testas modellen på varje scenario där en fold av datan i taget utgör testdata, medan resterande utgör träningsdata. Vilket värde k bör sättas till, kan bestämmas genom analyser över vad som medför bäst resultat, men om sådana analyser inte utförs är det standard att sätta k till 10 [29].

H. Prestationsmått

Ett vanligt sätt att evaluera hur väl maskininlärningsmodeller har presterat är att utgå från en confusion matrix, och utifrån den använda sig av standardiserade prestationsmått. Dessa mått kan exempelvis vara Accuracy, Average Accuracy, Precision, Recall samt F-measure. I bilaga 4 beskrivs de olika måtten samt hur de förhåller sig till varandra mer ingående. [4]

I. Precision-Recall-kurvor

En precision-recall-kurva har precision på y-axeln och recall på x-axeln. Olika tröskelvärden på sannolikheten att en datapunkt tillhör klass 1 används för att skapa grafen, där precision och recall markeras för dessa tröskelvärden. Precision-Recall-kurvor är passande i situationer där det finns hög andel av den negativa klassen och låg andel av den positiva. Baseline i en precision-recall-kurva beror på andelen datapunkter som tillhör den positiva klassen. Eftersom baseline utgörs av en klassificerare som inte kan urskilja mellan klasserna och som därmed klassificerar slumpmässigt, eller enbart utifrån en klass i samtliga fall, utgörs baseline av en horisontell linje. En perfekt klassificerares precision-recall-kurva representeras av en graf som löper från punkten (0, 1) till (1,1) och avslutas i (1, 0). En klassificerare som kan anses ha någon form av klassificeringsförmåga representeras av en graf som befinner sig ovanför baseline och som böjer av mot den perfekta klassificerarens precision-recall-kurva.[7]

J. ROC-kurvor

Receiver Operating Characteristics-kurvor (ROC-kurvor) kan användas för att evaluera en klassificerares prestation [34]. En ROC-kurvas y-axel utgörs av True Positive Rate (TPR). TPR beskriver hur många av den faktiskt positiva klassen som har lyckats klassificeras korrekt, uttryckt i procent. X-axeln däremot utgörs av False Positive Rate (FPR). FPR visar, procentuellt, hur många av den negativa klassen som modellen har lyckats klassificera korrekt. En ROC-kurva byggs upp av punkter som består av ett värde på FPR respektive TPR. [34]. Dessa punkter framträder utifrån olika värden på den tröskel som avgör vid vilken sannolikhet en viss datapunkt ska klassificeras som antingen negativ eller positiv. Genom att sedan koppla samman dessa punkter uppkommer ROC-kurvan. En perfekt ROC-kurva karaktäriseras av att den utgörs av sammankoppling av endast tre punkter. Dessa punkter är (0, 0), (0, 1) samt (1, 1). En ROC-kurva som genererats av en slumpmässig klassificerare däremot, är på formen $f(x) = x$, från punkten (0, 0) till (1, 1).

Ett av värdena som ROC-kurvan ger upphov till och som kan användas för att jämföra hur olika klassificerare har presterat, är arean under grafen (AUC). Ett AUC-värde på 1 är det optimala, som den perfekta ROC-kurvan besitter. [33]

K. Decision Support System

Decision support systems (DSS) är datorbaserade system som används för att förenkla processen vid beslutsfattning. Systemen blir som en knutpunkt för information som inhämtas

från olika platser, tillsammans med diverse relevanta modeller och dess förutsättningar. Dessa i kombination med varandra ger beslutsfattare en möjlighet att få stöd i processen för att i slutändan kunna fatta gynnsamma beslut. Detta är delvis önskvärt då det i många fall handlar om stora mängder ostrukturerad data, som för en enskild beslutsfattare är svårt att ta till sig på ett användbart sätt. Ett DSS kan till exempel hjälpa till att få åtkomst till data, utveckla passande modeller samt utvärdera modellens resultat. [30]

L. Business Activity Monitoring

Business Activity Monitoring (BAM) handlar om att använda data tillsammans med datatekniska verktyg för att definiera och analysera kritiska KPI:er i realtid för att förbättra såväl effektivitet, som lönsamhet. Målet med BAM är alltså att förse organisationen med realtidsinformation gällande status för exempelvis utvalda processer. Tre huvudsteg kan anses utgöra basen i en effektiv implementation av BAM. För att meningsfulla resultat ska kunna erhållas, krävs att det först säkerställs att det finns tillgång på tillräckligt mycket relevant data. Nästa steg utgörs av att bearbeta datan för att identifiera vilka faktorer som är relevanta i de aspekter som anses vara intressanta. Därefter ska datan analyseras och resultaten ska delges genom ett användarvänligt gränssnitt, exempelvis en dashboard där utvalda KPI:er visas tillsammans med realtidsresultatens inverkan på dessa. Genom att resultaten förmedlas på detta sätt, underlättas beslutsfattandet. [21]

III. TIDIGARE STUDIER

I en studie där data mining inom tillverkningsindustrin utvärderas konstateras det att CRISP-DM är en av de mest etablerade metodikerna inom projekt där data mining involveras [17]. Detta stöds av ytterligare en studie, som konstaterar att metodiken ger detaljerade riktlinjer för projektets utformning, samt att den resulterande processen är mycket användbar i verklighetsbaserade tillämpningar [1]. Användning av logistisk regression i syfte att kunna utföra binär klassificering på datapunkter, beskrivs i en studie som undersökte möjligheten att förutspå sannolikheten att uppnå uppsatta mål gällande vägsäkerhet. Författarna i studien använder sig dessutom av klusteralays, för att först erhålla förståelse för det dataset som utgjorde grunden för projektet. Att initialt utföra en klusteranalys för att få förståelse för datan går hand i hand med metodiken i CRISP-DM [32]. Även om författarna i studien inte specifikt behandlar kvalitetskontroller, anses den ändå vara av intresse, då den visar hur binär logistisk regression kan tillämpas för att utföra binär klassificering. Användning av logistisk regression för binär klassificering återfinns även i en annan studie, som beskriver hur relationen mellan data som genererats under valsning av stålstänger och dess binära kvalitetsutfall (godkänt eller icke godkänt), undersökts. Med hjälp av binär logistisk regression erhöles en modell som ansågs ha så låg felfrekvens att den hade varit möjlig att använda för verkliga kvalitetskontroller vid valsning. [22]. Vikten av att arbeta utefter en väldefinierad metod belyses i en artikel där Business Activity Monitoring (BAM)

integreras i studiens metod, tillsammans med viktiga aspekter i prediktion [31]. Idén är att prediktioner ska kunna göras på realtidsdata som genereras i Business Process Management Systems (BPMS) i syfte att minska fördröjningen mellan datainsamling och analys. Metoden baseras på Six Sigma och syftar därmed till att uppnå effektivisering. Utifrån de fem stegen i Six Sigma (DMAIC), togs följande fem steg fram i artikeln: prediction preparation, predictors modeling, prediction model definition, prediction model application och prediction model controlling. Den framtagna metoden demonstreras genom ett exempel där den efterföljs vid reparation av en telefon och det konstateras att metoden sannolikt medför att reaktionstiden hos beslutsfattare kan minskas, samt att kombinationen av BAM och prediktiva analyser av data från BPMS därmed ökar effektiviteten.

En viktig aspekt som konstateras i en studie som syftar till att framställa en datateknisk verktygslåda för analys av industriella processer, är att just kvalitetskontroller ofta medför obalanserade dataset. I studien används ett dataset från industriell tillverkning i en av Boscchs fabriker. I datasetet kan det ses att den del av datan som utgörs av godkända kvalitetskontroller når upp till cirka 99%, vilket innebär att endast drygt 1% av datan beskriver den negativa klassen [14]. Detta är något som även uppmärksammats i en studie gällande prediktionsmodeller i monteringen under tillverkningsprocesser [15]. Vidare, belyses problemet också i en studie som handlar om att med hjälp av SVM-basmodeller träna på data som endast består av datapunkter av den positiva klassen, samt omärkt data. Författarna konstaterar där att "Robust Ensemble of SVMs" är bäst lämpat för att lösa problemet [12]. Ytterligare en studie som tar upp problemet förklarar att SVM oftast utvärderas med hjälp av accuracy i det binära fallet, men att det vid obalanserade klasser är mer lämpligt att använda sig av F-measure [11].

IV. BEGRÄNSNINGAR

På grund av att full spårbarhet för vardera komponent inte finns i dagsläget på Volvo, skapas en begränsning för arbetet i form av att exakt dimensionsmått för varje enskild komponent inte kan tas fram. Istället behöver stickprover från varje batch användas och dessa stickprover generaliseras för komponenterna i just den batchen. Eftersom full spårbarhet inte finns på fabriken idag, görs alltså inte mätningar på varje komponent, vilket är anledningen till att detta varit en nödvändig åtgärd i arbetsprocessen. Däremot anses det inte vara en alltför stor begränsning, eftersom arbetet i sig syftar till att visa potentialen i hur data från monteringen kan nyttjas och inte nödvändigtvis till att fokusera på resultaten som just dessa dataset ger upphov till. Istället är förhoppningen att arbetet ska kunna bidra med insikter om vilka möjligheter som kan realiseras gällande att nyttja data för att få nya insikter om den egna verksamheten, om resurser läggs på att förbättra komponenternas spårbarhet. Med andra ord kan det liknas vid att syntetisk spårbarhet försöker framställas på ett så verklighetstroget sätt som möjligt, för att i sin tur belysa möjliga metoder och modeller som kan appliceras på den

erhållna datan, om spårbarheten realiseras i eventuella framtida produktionslinjer.

V. METOD

De tidigare studiernas positiva syn på CRISP-DM har medfört att metodiken i denna studie genomgående följt CRISP-DMs sex faser. Initialt utfördes ett besök på Volvo Groups fabrik i Köping. Besöket inleddes med en kort företagspresentation och följdes av en rundtur där produktionslinjens olika delar visades upp och förklarades. Fokus var på sammanpressningen av navet och huvudaxeln. En genomgång av hur pressmaskinen fungerar, följt av enskild analys och iakttagelse av de två komponenterna som ingår i sammanpressningen, gav en helhetsbild av processen och vad den innefattar. Det klargjordes vilka intervall som är godkända för vardera komponents dimension, samt vilket som är det godkända intervallet på presskraften vid sammanpressning.

A. Sammanställning verkligt dataset

Det påbörjades därefter manuell loggning i fabriken vid byte av den batch nav respektive huvudaxel som användes vid pressningen, där batchnummer och tidpunkt noterades. Det verkliga datasetet togs fram genom att manuellt sammanställa data från monteringen. Detta gjordes genom att den antecknade tidpunkten användes för att matcha ihop en viss presskraft med batchnummer på huvudaxel respektive nav. Med hjälp av batchnumret kunde därefter presskraften sammankopplas med mätvärden på dimensionerna för vardera komponent, utifrån data som samlats in från stickprovskontroller. Efter att denna data sammanställdes i form av ett Excelark med fyra kolumner bestående av vilken typ av de två typerna av huvudaxel som använts, presskraft, dimension för huvudaxel samt dimension för nav, lades ytterligare en kolumn till med den korrekta klassificeringen. En presskraftsmätning utanför det godkända intervallet resulterade i en nolla, och godkända pressningar representerades av en etta i den tillagda kolumnen.

B. Implementation av modeller

Efter att det verkliga datasetet var färdigställt, togs tre modeller fram för att testas i kombination med datan. Dessa modeller var Support Vector Machine (SVM), k-nearest neighbor (KNN) samt logistisk regression. Vid tillämpandet av dessa modeller användes biblioteket scikit-learn för python, mer precist funktionerna SVM, KNeighborsClassifier respektive LogisticRegression. Vid skapandet av SVM-modellen användes Gaussian radial basis function (RBF) som kernel-funktion, eftersom det förekom mycket överlappning mellan datapunkterna i de två klasserna. Vid framtagandet av KNN-modellen genomfördes en analys för att bestämma det mest lämpliga värdet på k . Varje tal mellan 1 till 31 testades som värden på k , och det k som resulterade i högst f -measure ansågs ha presterat bäst. Utfallet av analysen var att det mest satisfierande resultatet genererades vid betraktelse av de tre närmaste grannarna, och därmed nyttjades värdet tre på k i KNN-modellen. Gemensamt för samtliga modeller var att det utfördes en k -fold cross validation (KFCV) på datasetet samt

att SMOTE användes för att generera syntetiska datapunkter, då det endast fanns en begränsad mängd data. Den KFCV som användes var funktionen KFold som Python-biblioteket scikit-learn innefattar, där värdet på k sattes till tio. De syntetiska datapunkterna som SMOTE gav upphov till genererades i vardera fold, med hjälp av biblioteket imblearns funktion SMOTENC. Detta var på grund av att det var där träningsdata samt testdata för varje iteration delades upp. Anledningen till att SMOTENC användes, var att variabeln "typ av huvudaxel" är kategorisk, vilket SMOTENC kunde ta hänsyn till vid syntetiseringen av nya datapunkter. För det verkliga datasetet gjordes en upsampling av minoritetsklassen med hjälp av SMOTE så att antalet punkter i den positiva klassen var 60% av antalet negativa. Detta förhållande valdes på grund av att det låga antalet datapunkter tillhörande den positiva klassen i det ursprungliga datasetet gjorde att en upsampling till ett 1:1 förhållande inte ansågs rimligt.

C. Utvärdering av modeller

När modellerna färdigställdes, användes de för att göra klassificeringar på datasetet som innehöll verklig monteringsdata. Efter att klassificeringarna utförts, sammanställdes resultatet och modellerna utvärderades. Detta gjordes genom att prestationsmåttens average accuracy, recall, precision och F-measure beräknades. Alla dessa prestationsmått beräknades för alla tio folds och därefter beräknades medelvärdet av dem för att ge de slutgiltiga värdena. Dessutom framställdes ROC-kurvor och precision-recall-kurvor för varje fold för att ge ytterligare insikt i modellernas klassificeringsförmågor. Kurvorna framställdes med hjälp av biblioteket scikit-learn. Ur detta bibliotek användes funktionen `plot_precision_recall_curve` för att skapa precision-recall-kurvor, medan `roc_curve` användes för ROC-kurvor. Utifrån ROC-kurvorna kunde värdet på arean under grafen (AUC) beräknas, och med hjälp av precision-recall-kurvorna kunde average precision (AP) beräknas. Dessa beräkningar gjordes med hjälp av funktionerna `roc_auc_score`, respektive `average_precision_score`, där även dessa funktioner importerats från scikit-learn. För att kunna få en sammanfattad bild av modellernas prestation i dessa avseenden, beräknades medelvärdet av samtliga folds värden på AUC respektive AP. Med hjälp av erhållna medelvärden på AUC och AP kunde modellerna jämföras sinsemellan på ett enklare sätt, jämfört med om bara graferna hade tagits fram. Efter detta skapades även inlärningskurvor för att kunna utvärdera modellernas prestation i förhållande till mängden träningsexempel. Närmare bestämt framställdes dessa med hjälp av biblioteket scikit-learn, från vilket funktionen `learning_curve` importerades. Som score användes accuracy.

D. Fabricerat dataset

Efter att resultatet sammanställdes för klassificeringar utförda på datasetet innehållandes verklig monteringsdata, framställdes ett fabricerat dataset. I det fabricerade datasetet inkluderades ytterligare tre variabler. De nya variablerna utgjordes av *temperatur i pressningsmaskinen*, *temperatur på navet* och *antal dagar på lager*, varav den sista syftade på antalet dagar som navet befunnit sig inomhus i lagerlokalen.

De fiktiva värdena på de tre variablerna valdes på sådant sätt att det framställdes ett syntetiskt samband mellan två av variablerna och icke-godkända presskrafter. Dessa två variabler var temperaturen på navet och temperaturen i pressmaskinen. Navtemperaturer på 18-20 grader Celsius och pressmaskinstemperaturer på 59-60 grader Celsius parades, i de flesta fall, ihop med icke-godkända presskrafter. Detta gjordes dock på sådant sätt att datapunkter med godkända presskrafter ett fåtal gånger hade samma värden på dessa två variabler som de värden som kunde identifieras hos datapunkter med icke-godkända presskrafter. Syftet med detta var att göra det fabricerade datasetet så verklighetstroget som möjligt. Det ansågs vara realistiskt att även om exempelvis en navtemperatur på 19 grader i de allra flesta fall medför en icke-godkänd presskraft, kommer det rimligtvis finnas fall där det faktiskt medfört en godkänd presskraft. Ett antal sådana undantagsfall finns därmed representerade i det fabricerade datasetet.

När det fabricerade datasetet färdigställts, kunde samma tre modeller som tidigare appliceras för att utföra klassificeringar. Innan KNN användes, gjordes en ny analys för att hitta mest lämpligt k , vilket resulterade i att k bestämdes till sju. På samma sätt som för det verkliga datasetet användes 10-fold-cross-validation i kombination med SMOTE, så att det enbart skapades syntetiska datapunkter i träningsdata och inte i testdata.

Samma prestationsmått och utvärderingskurvor som för det verkliga datasetet sammanställdes för vidare jämförelser mellan modellerna.

När detta färdigställts, kunde modellernas prestationer givet de två olika dataseten, slutligen sättas i relation till varandra.

VI. RESULTAT

Resultatet inleds med det resulterande förhållandet mellan antal datapunkter i de två klasserna innan, respektive efter, upsampling skett med hjälp av SMOTE. Därefter delas resultatet in i två delar, där den ena delen behandlar resultaten för det verkliga datasetet och den andra behandlar resultaten från det fabricerade. I båda delar redovisas resultat från de tre modeller som använts.

A. Upsampling

Det går att avläsa från bilaga 5 att antalet datapunkter av den positiva klassen, bestående av icke godkända presskrafter, gick från 53 innan SMOTE till 164 efter att upsampling ägt rum. Antalet nollor, det vill säga datapunkter av den negativa klassen, var 274 både innan och efter att SMOTE applicerats. Denna jämförelse av förhållandet mellan klasserna innan och efter upsampling har utförts för vardera fold, och i bilaga 5 kan det mest representativa sambandsdiagrammet ses.

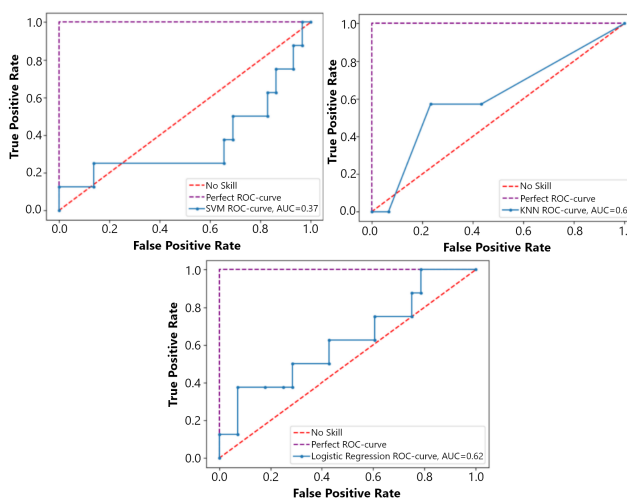
B. Verkligt dataset

Nedan presenteras resultatet för det verkliga datasetet.

B.1 Prestationsmått verkligt dataset

Bilaga 6 beskriver hur modellerna SVM, KNN, respektive Logistisk regression har presterat enligt medelvärden på dess average accuracy, recall, precision, F-measure, AUC samt AP. Ur bilagan kan det avläsas att värdena på average accuracy, recall, precision och F-measure är samma för både SVM och Logistisk Regression. Utöver det visas att KNN hade lägre värden på dessa mått, förutom recall, där KNN presterade bäst. KNN gav upphov till ett värde på 0.412 i recall, till skillnad från de andra modellernas värde på 0.343. Även medelvärden på precision och F-measure låg under 0.5 för samtliga modeller. Average accuracy låg runt 0.58 för vardera modell. När det gäller AUC kan det avläsas att SVM hade det lägsta värdet på 0.373, medan KNN erhöll det högsta på 0.624, där Logistisk Regression låg närmare KNNs värde på 0.589. Bilagan beskriver även modellernas värde på AP, som låg högst för logistisk regression, med ett värde på 0.311 och lägst för SVM, med värdet 0.182.

B.2 ROC-kurvor verkligt dataset



Figur 1. Figuren visar ROC-kurvor för det verkliga datasetet vars AUC-värde är närmast medelvärdet som tagits fram utifrån k -fold cross validation för respektive modell.

För varje fold från k -fold cross validation, har vardera modell skapat en ROC-kurva, med ett tillhörande värde på AUC. De ROC-kurvor vars AUC-värde är närmast medelvärdet som tagits fram utifrån k -fold cross validation för respektive modell har valts ut och visas i figur 1. Figuren visar att både KNN och Logistisk regression uppnådde ett AUC-värde över 0.60, medan ROC-kurvan för SVM endast uppnådde ett AUC-värde på 0.37. Det går dessutom att avläsa att alla de tre modellernas grafer var närmare baseline-kurvan än kurvan som beskriver en perfekt klassificering. Utöver detta tydliggörs det att ROC-kurvan för logistisk regression är den enda av de tre som inte vid något tillfälle är under baseline-kurvan. Till skillnad från detta är majoriteten av kurvan för SVM under baseline-kurvan.

B.3 Precision-Recall-Kurvor verkligt dataset

Utifrån bilaga 7 går det att se att en större del av SVMs Precision-Recall-kurva ligger under baseline, i förhållande till logistisk regression som endast är ovanför baseline, samt KNN vars kurva till en klar majoritet ligger ovanför. Det går dessutom att urskilja att logistisk regression hade högst medelvärde på AP, vilket gör det till den graf som till störst del låg ovanför baseline-kurvan. Det går även att se att alla grafer över modellernas Precision-Recall-kurva är närmare baseline-kurvan än kurvan som utgörs av en perfekt klassificering.

B.4 Analys av variabelers inverkan på det verkliga datasetet

Bilaga 8 visar att varken modellen SVM eller logistisk regression fann att någon av dimensionerna hade inverkan på kvalitetsutfallet, men däremot att variabeln över vilken typ av huvudaxel som var med i pressningen hade ett marginellt värde på cirka 0.10 på importans. Till skillnad från dessa två modeller fann KNN att det fanns en starkare inverkan från de två dimensionsvariablerna än typen av huvudaxel. Importansvärdena för de två dimensionsvariablerna var cirka 0.30.

C. Fabricerat Dataset

Nedan återfinns resultatet för det fabricerade datasetet.

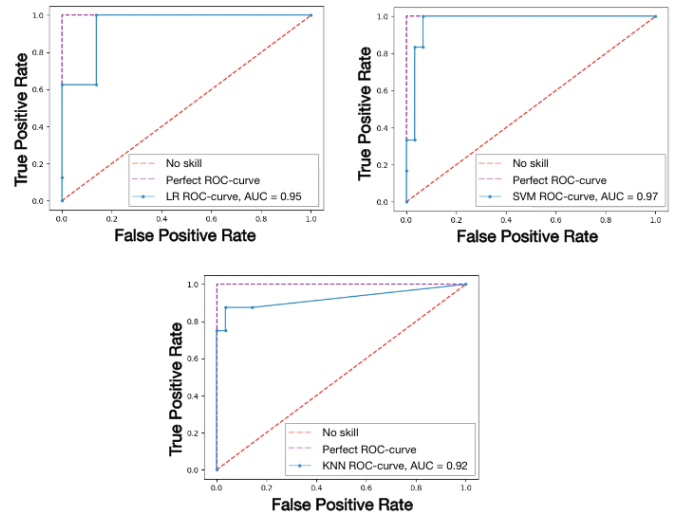
C.1 Prestationsmått fabricerat dataset

Ur bilaga 9 kan det utläsas att den modell som erhållit högst värde på recall, närmare bestämt 0.926, är logistisk regression. Värdet på precision för samma modell är 0.608, vilket är markant lägre än de andra modellerna. Värdena för KNN, respektive SVM, är 0.745 respektive 0.801. Gällande F-measure, uppvisar SVM värdet 0.827, vilket är högst bland de tre modellerna. Detta följs av värdet 0.810, som erhålls genom KNN, samt 0.729 som erhålls genom logistisk regression. För värdena på AUC, konstateras det att högst uppnådda värde är 0.958 och ges av SVM, medan det lägsta är 0.944 och uppnås av KNN. Även för AP är det SVM som uppvisar högst värde, i form av 0.896, medan det lägsta värdet är 0.828 och erhålls av KNN.

C.2 ROC-kurvor fabricerat dataset

I figur 2 visas den trade-off som råder mellan andelen korrekt positivt klassificerade och andelen falskt positivt klassificerade för de tre modellerna. Det kan utläsas att alla modellerna presterar bättre än baseline. För SVM kan det ses att dess ROC-kurva följer formen av en perfekt klassificerades ROC-kurva från och med att false positive rate är cirka 0.1. För KNN avviker ROC-kurvan något mer från en perfekt klassificerades, jämfört med SVM. Avvikelsen sker

från och med att true positive rate är cirka 0.75. Vad gäller ROC-kurvan för logistisk regression, avviker denna något mindre från en perfekt klassificerare än KNN, men aningen mer än SVM.



Figur 2. Grafen visar ROC-kurvorna vars AUC-värde är närmast medelvärdet som tagits fram utifrån k-fold cross validation för respektive modell, genererat utifrån det fabricerade datasetet.

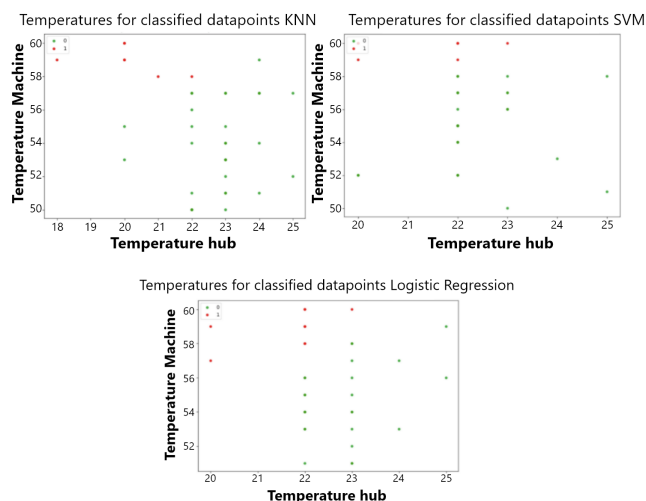
C.3 Precision-Recall-Kurvor fabricerat dataset

Det kan utläsas ur bilaga 10 att för både SVM och logistisk regression, krävs en stor uppoffring i precision för att erhålla ett högre värde på recall än 0.8. För KNN, avviker precision-recall-kurvan från 1.0 i precision redan i början, men uppoffringen i precision för att öka recall är markant mindre än för SVM och logistisk regression. Då recall är cirka 0.7, sker ytterligare en liten uppoffring i värdet på precision för KNN.

C.4 Analys av variabelers inverkan fabricerat dataset

Det går att avläsa från bilaga 11 att samtliga modeller resulterade i att den variabel som har störst inverkan på kvalitetsutfallet är den temperatur som infinder sig i maskinen vid pressning. Modellerna SVM samt KNN mätte båda att dess inverkan sträcker sig till ett importans-värde runt 0.5, i kontrast till logistisk regression som endast nådde ett värde av importans på drygt 0.4 för variabeln. Logistisk regression presenterar däremot det högsta värdet på importans av variabeln som beskriver navets temperatur. Samtliga modeller visar att navets temperatur är den variabel som har näst högst inverkan. SVM, samt logistisk regression, uppvisar att övriga variabler inte har någon inverkan på kvalitetsutfallet. Till skillnad från detta visar KNN att navets dimension samt antalet dagar på lager har en marginell inverkan med ett importans-värde på cirka 0.01.

Navtemperaturer och temperaturer i pressmaskinen för såväl positivt klassificerade som negativt klassificerade datapunkter presenteras i figur 3. I denna figur kan datapunkternas klassificering sättas i relation till vilka värden punkterna hade på temperaturen på navet, respektive i pressmaskinen vid pressningen. Generellt kan det konstateras att negativt klassificerade datapunkter framför allt återfinns i den övre vänstra delen av respektive scatterplot. Detta korresponderar till höga temperaturer i pressmaskinen, samt låga temperaturer på navet.



Figur 3. I figuren presenteras sambandsdiagram med de två variablerna som visat sig vara mest betydelsefulla i variabelanalysen för det fabricerade datasetet. Figuren visar värden för den mest representativa folden för vardera modell. På X-axeln ses temperaturen på navet och på Y-axeln ses temperaturen i pressmaskinen vid pressningen.

VII. DISKUSSION

När enbart komponenternas dimensioner inkluderades som variabler i modellerna, erhöles generellt låga värden på alla de prestationsmått som undersöktes. Det var också ett faktum att både ROC-kurva och precision-recall-kurva var markant närmare baseline än perfect-skill för samtliga modeller. Dessa två konstateranden tyder, i kombination, på att modellerna inte kunnat identifiera samband mellan värden på dimensioner för huvudaxel, respektive nav och huruvida sammanpressningen av dessa krävt en godkänd presskraft eller ej.

Att ett samband inte har kunnat påvisas i just detta dataset med hjälp av just dessa tre modeller, innebär dock inte att det kan uteslutas att sambandet existerar. Till exempel har det faktum att datasetet framställdes manuellt, genom matchning av tidpunkter från när de olika komponenterna befunnit sig på olika stationer i produktionslinjen, introducerat felkällor i datasetet. En annan möjlighet är att det finns andra modeller än de tre som använts, vilka kanske hade passat bättre för just denna data. Dessutom kan det vara så att dimensionerna, i kombination med andra variabler, påverkar vilken presskraft som krävs och därmed huruvida kvalitetsutfallet blir godkänt

eller ej. Dimensionsvariablerna kanske alltså samverkar med andra faktorer, vilka inte inkluderats i detta dataset.

Det är viktigt att ta samtliga mått i beaktning vid utvärdering av resultatet, detta på grund av att enskilda värden kan vara överdrivet optimistiska eller pessimistiska och därmed inte återspeglar det faktiska förhållandet mellan parametrarna. Det visar sig tydligt i att KNN, vid analysen över variabelers inverkan på det verkliga datasetet, resulterade i att de två dimensionsparametrarna hade viss inverkan på kvalitetsutfallet, vilket skiljer sig från de två övriga modellernas resultat. Genom att dessutom sätta det i relation till resterande resultat såsom exempelvis F-measure, ROC-kurvor samt precision-recall-kurvor, blir det tydligt att även om variabelanalysen för KNN tyder på ett samband mellan dimensionsvariablerna och kvalitetsutfallet, så verkar inte rätt samband mellan dessa ha identifierats. Resultatet KNN erhöles vid variabelanalys är därmed missvisande och bör ej fokuseras på enskilt, vilket belyser vikten av att se till helheten vid utvärdering.

Samtliga modeller har upptäckt ett samband givet det fabricerade datasetet, vilket därmed har resulterat i fler korrekta klassificeringar och i sin tur högre värden på samtliga mått, i förhållande till värdena som det verkliga datasetet medförde. För att kunna dra någon slutsats utifrån dessa värden måste en bedömning göras gällande vilket av prestationsmått som bör värderas högst. Vilket prestationsmått som är av störst vikt när det gäller en klassificerare som syftar till att klassificera kvalitetsutfall i monteringen, beror på situationen. Om fallet är sådant att de komponenter som inte blir godkända måste kasseras, kan det vara önskvärt att fokusera på att erhålla högt recall. Då kommer maskinen att varna om att en viss pressning inte kommer att lyckas i en del fall när pressningen kanske faktiskt hade godkänts. I den här situationen lyder logiken med andra ord att komponenterna hellre justeras en gång extra, än en gång för lite, för att materialsvinnet ska minskas. Om det däremot exempelvis är mer kostsamt på sikt med den tid som krävs för att byta ut komponenterna när det varnas för att pressningen inte kommer att fungera, är precision viktigare att maximera. Då är det prioriterat att minimera antalet gånger det varnas i onödan angående en pressning. Vad gäller produktionslinjen på Volvo i Köping, plockas komponenterna bort för vidare inspektion i mättrummet i de situationer där flera komponenter efter varandra resulterar i ett icke godkänt kvalitetsutfall. Det är inte önskvärt att detta sker i onödan. Det bör dock också konstateras att det inte heller är önskvärt att det missas att varnas på alltför många komponenter vars kvalitetsutfall sedan inte blir godkänt, eftersom det blir tidskrävande i längden att först försöka göra en sammanpressning, för att sedan plocka ur komponenterna, samt byta ut dem. Det skulle alltså kunna konstateras att precision och recall är av ungefär samma vikt i Volvos situation, vilket medför att F-measure anses vara det prestationsmått som bör fokuseras på i jämförelsen mellan modellerna.

SVM är den modell som erhöll högst värde på F-measure, följt av KNN för det fabricerade datasetet och tillsammans med logistisk regression för det verkliga datasetet. Utöver det hade SVM den ROC-kurva som mest efterliknar en perfekt sådan, med högst värde på AUC. SVM är dessutom en av de modeller som presterade bäst när det gäller precision-recall-kurvor. Att SVM visar på bra värden och grafer för både ROC samt Precision-Recall är avgörande. Detta på grund av att den upsampling som utförs på det fabricerade datasetet inte har ett förhållande på formen 1:1, utan att antalet punkter av den positiva klassen endast är 60% av antalet negativa punkter. Då ROC-kurvor är bäst anpassade för balanserade dataset kan det medföra att dessa kurvor i detta fall ger en något optimistisk bild. Kombinationen av att dessa tre olika mätvärden alla ger en bild av att SVM är den modell som har presterat bäst, tyder på att det stämmer. Det är dock viktigt att poängtera att resterande modeller inte har presterat dåligt, och därmed inte bör förbises endast på grund av att SVM genererat bättre resultat i dessa mått.

På det fabricerade datasetet utfördes även analys av variabelernas inverkan på beslutet gällande att klassificera datapunkten som godkänd eller icke-godkänd. Resultatet från analysen kan ses i bilaga 11, utifrån vilken det konstaterades att temperaturen i pressningsmaskinen var överlägset mest betydelsefull i samtliga modeller. Eftersom datasetet är fabricerat, bör det poängteras att det som önskas belysas genom variabelanalysen inte är resultatet i sig, eftersom detta inte baseras på verklig data. Det är redan känt att det finns ett samband mellan variabelerna temperatur på navet och temperatur i pressmaskinen och det som undersöks är istället hurvida modellerna lyckas identifiera detta samband eller ej. Analysen syftar med andra ord till att visa att det faktiskt är möjligt att utnyttja monteringsdata för att bättre förstå vilka punkter i produktionslinjen som medför extra kritiska variabler, samt vilka tröskelvärden på dessa variabler som i sin tur ger inverkan på kvalitetsutfallet. I framtiden kan samma metod användas för att göra analys på verklig data. Sådan analys skulle eventuellt kunna erhålla resultat som kan möjliggöra att nödvändiga förändringar identifieras i produktionslinjen, vilka i sin tur minskar antalet icke-godkända komponenter, genom att de mest betydelsefulla variabelerna för kvalitetsutfallet lyckas identifieras. Detta hade kunnat ge upphov till besparingar i form av såväl tid som material, vilka båda kan medföra bättre lönsamhet på sikt. I figur 3 ses en figur där navtemperaturer och temperaturer i pressmaskinen för såväl positivt klassificerade som negativt klassificerade datapunkter i den mest representativa folden kan ses. Utifrån dessa är det tydligt att det är höga temperaturer i pressningsmaskinen, samt låga temperaturer för navet som resulterar i icke-godkända klassificeringar. Hade det fabricerade datasetet istället bestått av verklig data, hade en möjlighet för Volvo kunnat vara att fokusera sina åtgärder utefter detta. Exempelvis genom att vidare undersöka orsaken till de höga temperaturerna som, enligt det fabricerade datasetet, ibland förekommer i pressningsmaskinen. För att öka temperaturen på navet hade en möjlig åtgärd kunnat vara

att undersöka möjligheter att flytta lagerhållningen av dessa till en del av fabriken där temperaturen är något högre. Det som vill belysas med dessa exempel, är att modellerna öppnar för möjligheten att nyttja data från den egna verksamheten för att identifiera eventuella förbättringar.

De två dataseten som använts har båda genomgått upsampling med hjälp av SMOTE, där syntetiska datapunkter genererats, vilket med största sannolikhet har påverkat resultatet. I enlighet med vad som konstaterats i de tidigare studier som tagits upp gällande obalanserade klasser, har användningen av SMOTE förmodligen haft en positiv inverkan på så sätt att den har bidragit till en balans mellan de två klasserna, vilket lett till att modellen haft mer data att träna på, vilket i sin tur kan ha medfört bättre resultat. Det är dock inte säkert att det endast haft en positiv inverkan, då det är syntetiska datapunkter som inte alltid blir likvärdiga de faktiska datapunkterna. Vid analys av exempelvis sambandsdiagram som beskriver fördelningen mellan de två klasserna före och efter upsampling med hjälp av SMOTE, blir det synligt att en del av den syntetiska datan som ska representera den negativa klassen får samma värden på komponenternas dimensioner, som datapunkter av den positiva klassen. Detta kan i sin tur leda till att det blir svårt för modellen att skilja på klasserna, och därmed även att klassificera dem. Sett till resultatet från det verkliga datasetet har inget samband mellan komponenternas dimensioner och kvalitetsutfallet identifierats av modellerna, vilket hade kunnat vara på grund av användningen av SMOTE. Det skulle dock kunna argumenteras för att upsampling med SMOTE inte haft alltför stor inverkan, eftersom samtliga modeller lyckas fånga sambandet i det fabricerade datasetet. Om det i framtiden skulle samlas in tillräckligt mycket data för att kunna bygga balanserade dataset utan SMOTE, skulle denna möjliga felkälla, helt försvinna. Det är därmed en önskvärd åtgärd i framtiden.

Om spårbarheten för komponenterna förbättras i monteringen på Volvo, kan det alltså i framtiden vara möjligt att framställa ett dataset, vilket liknar det fabricerade som använts i denna studie, men som istället innehåller verklig data. I denna studie har dimensionerna på komponenterna som nämnt bestämts utifrån de stickprov som tagits på varje batch, eftersom det i dagsläget inte finns spårbarhet för vardera enskild komponent på Volvo. Att förbättra spårbarheten hade möjliggjort fler liknande studier angående andra variabelers inverkan på kvalitetsutfallet och dessutom ökat tillförlitligheten i denna typ av studier. Med förbättrad spårbarhet blir det dessutom lättare att lagra datan på ett sätt som förenklar möjligheten att välja ut och sammanställa värden för vardera komponent över variabler av intresse från produktionslinjen. I ett sådant scenario skulle manuell sammanställning av datasetet alltså kunna undvikas, vilket i sin tur minskar antalet felkällor som introduceras. I kombination med att SMOTE, som nämnt, inte skulle behöva tillämpas finns alltså potential att skapa ett verkligt dataset med hög tillförlitlighet inför att studien eventuellt görs om i framtiden. Om så skulle ske, är jämförelsen mellan modellernas prestation på det verkliga och det fabricerade datasetet något som kan motivera

användningen av just dessa modeller även i framtida studier. Det faktum att modellerna faktiskt presterar bra på det dataset där vi vet att det finns samband mellan två av variablerna och kvalitetsutfallet, tyder på en existerande förmåga hos modellerna att hitta mönster.

Det har alltså konstaterats att det i framtiden eventuellt kan vara möjligt att konstruera modeller utifrån verkliga monteringsdata för att kunna förutspå kvalitetsutfall. Men hur hade denna möjlighet kunnat nyttjas och integreras i produktionslinjen och vad skulle det kunna medföra?

Ett möjligt användningsområde för de modeller som tagits fram, är att implementera Business Activity Monitoring (BAM) och integrera en av modellerna i systemet. Detta gjordes i den tidigare nämnda studien A Method and Tool for Predictive Event-Driven Process Analytics, där en föreslagen metod även presenterades. Som nämnt, baserades metoden på de fem stegen i Six Sigma (Define, Measure, Analyze, Improve, Control) och utgjordes av prediction preparation, predictors modeling, prediction model definition, prediction model application och prediction model controlling. Att metoden bygger på Six Sigma, tyder på ett effektiviseringsfokus, vilket är önskvärt i tillverkningsindustrin. Deras metod demonstrerades även genom ett exempel, där modellerna konstruerades och prediktioner gjordes gällande reparationstid och huruvida reparationen behövde göras om eller ej. Prediktionsresultaten förmedlades till arbetarna genom en visualiserings-frontend, vilket gjorde att de kunde vidta proaktiva åtgärder om det till exempel visade sig att en reparation tog längre tid än normalt.

Liknande implementation skulle kunna ske på Volvo. En dashboard för visualisering skulle kunna tas fram, vilken kan placeras vid pressmaskinen. På Volvo Group AB finns tillgång till program som exempelvis Power BI samt ThingWorx, vilka båda är kompatibla med realtidsdata och eventuellt hade kunnat nyttjas i framställningen av dashboards för visualisering. Relevanta KPI:er för den aktuella produktionslinjen skulle behöva bestämmas, för att inkluderas på dashboarden. Ett exempel skulle kunna vara Quality Ratio hos pressmaskinen, vilket visar hur stor andel komponenter som är godkända för en viss maskin. Ett annat exempel är Utilization Efficiency, vilket visar andelen upptagen tid som är värdeskapande tid hos maskinen. Utöver dessa KPI:er skulle det även vara intressant att visa variablerna som inkluderats i modellen och som visat sig vara mest betydelsefulla för kvalitetsutfallet. I fallet med modellerna som använts på det fabricerade datasetet, skulle dessa alltså vara temperaturen i pressmaskinen och temperaturen på navet. När en pressning ska ske visas temperaturen på navet, temperaturen i pressmaskinen och huruvida kvalitetsutfallet prediceras bli godkänt eller ej, på dashboarden vid pressmaskinen. Dessutom visas kvalitetsutfallets inverkan på de KPI:er som valts ut. Om en viss pressning prediceras bli icke-godkänd skulle det alltså i detta exempel gå att se hur det icke-godkända kvalitetsutfallet skulle påverka Quality Ratio, respektive Utilization Efficiency

för pressmaskinen, vilket i sin tur enkelt kan sättas i relation till de mål som definierats för dessa KPI:er. Genom att det även går att se temperaturen i pressmaskinen, respektive temperaturen på navet kan det beslutas om proaktiva åtgärder för att minska defekter och därmed ledtider. Visar det sig till exempel att temperaturen i pressmaskinen är för hög under flera pressningar, kanske det kan vara mer tidseffektivt att pausa pressning för att åtgärda temperaturen, jämfört med att få fortsatt många defekter på rad, vilka måste pressas om. Skulle det visa sig att temperaturen på navet är för låg under ett antal pressningar på rad, skulle det vara rimligt att dessa nav kanske placerats i en del av fabriken där temperaturen varit något lägre. Genom att identifiera detta, skulle det kunna bytas så att pressning sker med nav från en annan pall, i väntan på att de kalla naven värms upp.

Ytterligare ett sätt att utnyttja både realtidsdata och komponenters spårbarhet är att implementera ett decision support system. Det skulle exempelvis vara relevant ur syftet att för vardera dag i monteringen ta fram den optimala mixen av produkter som ska framställas, då Volvo Group AB i Köping jobbar mot beställningar. Genom att samla in en stor mängd data över bland annat tid det tar för vardera produkt samt dess olika komponenter att ta sig igenom produktionslinjen och vilka delar av produktionen som montering av dessa produkter/komponenter kräver, skulle en DSS kunna lära sig av detta. I sin tur hade då detta system kunnat fatta egna beslut vad gäller vilka produkter som ska produceras under dagen för att optimera användningen av exempelvis maskiner. Detta skulle ge upphov till effektivisering inom monteringen och skulle vara ett relevant och intressant område för vidare studier.

I diskussionen kring hur modellerna kan nyttjas, bör det återigen kommenteras att variablerna temperatur i pressmaskin, respektive temperatur på nav haft fiktiva värden och att sambandet med kvalitetsutfallet därmed är fabricerat. Om samma studie görs om i framtiden, med ett verkligt dataset av satisfierande storlek och innehållandes fler variabler, kommer det med stor sannolikhet visa sig vara andra variabler som är avgörande för kvalitetsutfallet. Förhoppningen är dock att de exempel som tagits upp kan generaliseras trots andra variabler och att samma tankesätt kan appliceras, eftersom syftet kommer att vara detsamma - det vill säga minska mängden defekter, förkorta ledtider och uppnå ökad effektivitet i produktionen.

VIII. SLUTSATS

Diskussionen mynnar sammanfattningsvis ut i att det, med genomgående spårbarhet och en uppkopplad fabrik, finns möjlighet att implementera maskininlärningsmodeller som kan nyttjas praktiskt i fabriken för att uppnå effektiviseringar. Av de tre modellerna som jämförts i detta arbete, presterade SVM bäst givet de prestationsmått som ansågs vara mest relevanta. För att dra slutsatser kring huruvida det hade varit lönsamt att implementera förutsägelser av kvalitetsutfall i produktionslinjen, krävs först att en investeringskalkyl görs.

Till en sådan diskussion är det också relevant att undersöka om det är vinstgivande på sikt att implementera de system som diskuterats enbart i eventuella framtida produktionslinjer eller om förändringar även bör göras i den befintliga. Med de möjligheter och krav som Industri 4.0 introducerar, kan det dock argumenteras för att arbetet med att nyttja tillgången på data är högst relevant. De metoder som föreslagits i detta arbete hade därmed kunnat vara ett första steg mot omställningen till Industri 4.0, vilken kommer att behöva ske i sinom tid, för att upprätthålla konkurrenskraft.

TACKSÄGELSE

Vi känner tacksamhet gentemot alla som bidragit med kunskaper som förbättrat detta arbete. Ett extra stort tack riktas till Marcus Ek, Tobias Högfeldt och Robby Kloos, vilka agerat handledare på Volvo Group och vars vägledning varit mycket hjälpsam. Dessutom vill vi tacka Jonas Beskow och Mattias Wiggberg från KTH för deras handledning och de råd som de givit oss under utformningen av arbetet.

FÖRFATTARPRESANTATION

Minna Mathisson

Minna Mathisson studerar Industriell Ekonomi på Kungliga Tekniska Högskolan i Stockholm, med teknikinriktningen Datateknik och Kommunikation.

Lisa Janson

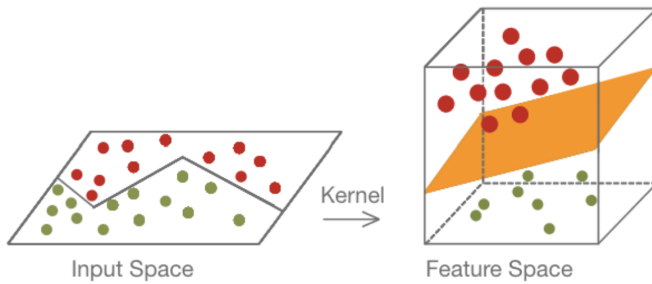
Lisa Janson studerar tredje året på Industriell Ekonomi vid Kungliga Tekniska Högskolan i Stockholm. Som teknikinriktning studerar hon Datateknik och Kommunikation.

REFERENSER

- [1] Bekar, Ebru Turanoglu; Nyqvist, Per & Skoogh, Anders. 2020. An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. *Advances in Mechanical Engineering*. DOI: 10.1177/1687814020919207 (Hämtad 2021-02-25).
- [2] Belkin, Mikhail; Hsu, Daniel; Ma, Siyuan & Mandal, Soumik. 2019. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences* 116 (32) 15849-15854. <https://doi.org/10.1073/pnas.1903070116> (Hämtad 2021-02-26).
- [3] Boye, Johan. DD1418, föreläsning 7, del 2. *The bag-of-words representation k-nearest neighbors*. 2020-11-09.
- [4] Boye, Johan. DD1418, föreläsning 7, del 4. *Evaluation of text classification*. 2020-11-09.
- [5] Boye, Johan. DD1418, föreläsning 8, del 1. *More on Classification*. 2020-11-12.
- [6] Boye, Johan. DD1418, föreläsning 8, del 2. *Logistic regression*. 2020-11-12.
- [7] Brownlee, Jason. 2018. How to Use ROC Curves and Precision-Recall Curves for Classification in Python. *Machine Learning Mastery*. [Link to Machine Learning Mastery](https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/) (Hämtad: 2021-04-28)
- [8] Brownlee, Jason. 2020. Cost-Sensitive SVM for Imbalanced Classification. *Machine Learning Mastery*. [Link to Machine Learning Mastery](https://machinelearningmastery.com/cost-sensitive-svm-for-imbalanced-classification/) (Hämtad: 2021-03-01)
- [9] Boyle, Tara. Dealing with Imbalanced Data. *Towards data science*. [Link to article](https://towardsdatascience.com/dealing-with-imbalanced-data/) (Hämtad 2021-03-01)
- [10] Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin & Wirth, Rüdiger. 2000. CRISP-DM 1.0. CRISP-DM consortium. [Link to article](https://www.crisp-dm.org/) (Hämtad: 2021-03-01)
- [11] Cheng, Fan; Zhou, Yuan; Gao, Jian & Zheng, Shuangqi. Efficient Optimization with Cost-Sensitive SVM. *Mathematical Problems in Engineering*. Vol. 2016, 11 pages. <https://doi.org/10.1155/2016/5873769> (Hämtad 2021-02-23).
- [12] Claesen, Marc; De Smet, Frank; Suykens, Johan A.K. & De Moor, Bart. 2015. A robust ensemble approach to learn from positive and unlabeled data using SVM base models. *Neurocomputing*. Vol. 160, 73-84. <https://doi.org/10.1016/j.neucom.2014.10.081> (Hämtad 2021-02-23).
- [13] Dengler, Sebastian; Lahiri, Said; Trunzer, Emanuel & Vogel-Heuser, Birgit. 2021. Applied machine learning for a zero defect tolerance system in the automated assembly of pharmaceutical devices. *Decision Support Systems*, 113540.
- [14] Flath, Christoph M. & Stein, Nikolai. 2018. Towards a data science toolbox for industrial analytics applications. *Computers in Industry*. Vol. 94, 16-25. <https://doi.org/10.1016/j.compind.2017.09.003> (Hämtad 2021-02-23).
- [15] Galetto, Maurizio; Verna, Elisa & Genta, Gianfranco. 2020. Accurate estimation of prediction models for operator-induced defects in assembly manufacturing processes. *Quality Engineering*, 32:4, 595-613, DOI: 10.1080/08982112.2019.1700274 (Hämtad 2021-02-23).
- [16] Gandhi, Rihith. 2018. Support Vector Machine - Introduction to Machine Learning Algorithms. *Towards data science*. [Link to Towards data science](https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms/) (Hämtad 2021-02-27).
- [17] Harding, J. A.; Shahbaz, M.; Srinivas & Kusiak, A. 2005. Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering* 128(4): 969-976. <https://doi.org/10.1115/1.2194554> (Hämtad 2021-02-25).
- [18] Harrison, Onel. 2018. Machine Learning Basics with the K-Nearest Neighbors Algorithm. *Towards data science*. [Link to article](https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm/) (Hämtad: 2021-05-02)
- [19] Hofmann, Erik & Rüscher, Marco. 2017. Industry 4.0 and the current status as well as future prospects on logistics. *Computers in Industry*. Volume 89: 23-24. [Link to article](https://www.sciencedirect.com/journal/computers-in-industry/volume/89) (Hämtad: 2021-03-01)
- [20] IBM Cloud Education. 2020. Supervised Learning. IBM. <https://www.ibm.com/cloud/learn/supervised-learning#toc-unsupervised-Fo3jDcmY> (Hämtad: 2021-05-02)
- [21] Kang, Jin Gu & Han, Kwan Hee. (2008). A Business Activity Monitoring System Supporting Real-Time Business Performance Management. 2008 Third International Conference on Convergence and Hybrid Information Technology, 1, 473-478.
- [22] Jin, Ran; Li, Jing & Shi, Jianjun. 2007. Quality prediction and control in rolling processes using logistic regression. *University of Michigan*. [Link to article](https://www.umich.edu/~enginerep/papers/rolling_processes_using_logistic_regression.pdf) (Hämtad 2021-02-25).
- [23] Nationalencyklopedin, data mining. <https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/data-mining> (Hämtad 2021-03-01).
- [24] Nationalencyklopedin, korsvalidering. <https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/korsvalidering> (Hämtad 2021-03-01).
- [25] Nationalencyklopedin, logistisk regression. <https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/logistisk-regression> (Hämtad 2021-03-01).
- [26] Nationalencyklopedin, regressionsanalys. <https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/regressionsanalys> (Hämtad 2021-03-01).
- [27] Ranjan, Chitta. 2019. Understanding the Kernel Trick with fundamentals. *Towards data science*. [Link to Towards data science](https://towardsdatascience.com/understanding-the-kernel-trick-with-fundamentals/) (Hämtad: 2021-04-28)
- [28] Regeringskansliet. 2018. En politik för tillväxt och utveckling i svensk industri. Regeringens skrivelse 2017/18:202. <https://www.regeringen.se/rattsliga-dokument/skrivelse/2018/03/skr-201718202/> (Hämtad: 2021-02-25).
- [29] Sanjay, M. 2018. Why and how to Cross Validate a Model? *Towards data science*. [Link to Towards data science](https://towardsdatascience.com/why-and-how-to-cross-validate-a-model/) (Hämtad 2021-02-27).
- [30] Sauter, Vicki. 2010. *Decision Support Systems for Business Intelligence*. 2. uppl. New Jersey: John Wiley Sons. [Link to article](https://www.wiley.com/9781118202142) (Hämtad 2021-02-27).
- [31] Schwegmann, Bernd; Matzner, Martin; & Janiesch, Christian, Ä Method and Tool for Predictive Event-Driven Process Analytics" (2013). *Wirtschaftsinformatik Proceedings 2013*. Paper 46. [Link to article](https://www.wirtschaftsinformatik.de/proceedings/2013/paper46/) (Hämtad 2021-04-29).
- [32] Sze, N. N.; Wong, S. C. & Lee, C. Y. 2014. The likelihood of achieving quantified road safety targets: a binary logistic regression model for possible factors. *Accident; analysis and prevention*, 73, 242-251. <https://doi.org/10.1016/j.aap.2014.09.012> (Hämtad 2021-02-25).
- [33] The scikit-learn developers. 2020. Receiver Operating Characteristic (ROC). [Link to scikit-learn](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html) (Hämtad 2021-04-28).
- [34] Toshniwal, Ruchi. 2020. Demystifying ROC Curves. *Towards data science*. [Link to Towards data science](https://towardsdatascience.com/demystifying-roc-curves/) (Hämtad 2021-04-28).
- [35] Unicef. De Globala Målen. Ansvarig utgivare: Lee, Åsa. <https://unicef.se/vad-vi-gor/de-nya-globala-utvecklingsmalen> (Hämtad: 2021-02-25)

BILAGOR

A. Bilaga 1



När SVM modelleras, används en kernel-funktion. Figuren visar hur en kernel-funktion används för att utöka dimensionen i det ursprungliga rummet, så att klasserna kan separeras linjärt i den erhållna högre dimensionen.

B. Bilaga 2

Kernel-funktionen RBF är populär att använda som kernel-funktion i SVM och definieras enligt följande formel:

$$K(x, y) = e^{-\gamma \|x-y\|^2}, \gamma > 0$$

Funktionsvärdet beror alltså på avståndet mellan vektorerna x och y i originaldimensionen och representerar relationen mellan de två punkterna i den högre dimensionen. Ju högre funktionsvärde som erhålls, desto närmare är de två vektorerna. Ett sätt att hantera överlappning mellan datapunkterna från olika klasser, vilka inte är linjärt separabla, är att använda SVM med RBF som kernel-funktion. I grunden innebär en RBF en oändlig summering av polynoma kernel-funktioner, vilket i sin tur medför att vektorerna (det vill säga x och y) projiceras i ett vektorrum av oändlig dimension. Gamma representerar hur stor inverkan vardera datapunkt i träningsmängden har när nya datapunkter ska klassificeras. Om x och y är två olika datapunkter, går det att se i formeln ovan att avståndet mellan de två datapunkterna kvadreras. Utifrån detta kan det konstateras att den påverkan en viss datapunkt har på en annan datapunkt, är beroende av värdet på det kvadrerade avståndet mellan dem. Med hjälp av gamma, är det möjligt att påverka hur mycket omgivande datapunkter ska kunna påverka. Exempelvis medför ett högre värde på gamma att andra datapunkter i träningsmängden behöver vara närmare för att påverka. [27]

C. Bilaga 3

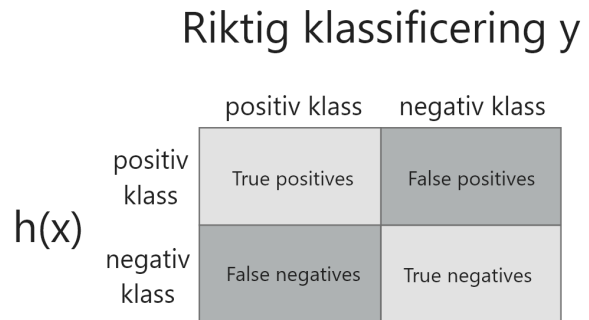
$$\|x - y\|_2 = \sqrt{\sum_{n=i} (x_i - y_i)^2}$$

Formeln visar hur det euklidiska avståndet beräknas [3].

D. Bilaga 4

En confusion matrix (CM) utgörs av den korrekta klassificeringen y på ena axeln och modellens klassificering, $h(x)$, på den andra. Gemensamt för båda dessa axlar är att de dessutom

delas in i positiv respektive negativ klass, vilket förtydligas på bilden nedan:



Figuren illustrerar en binär confusion matrix, vars olika beståndsdelar namnges.

De $h(x)$ som stämmer överens med dess associerade y kallas för true, och de datapunkter där $h(x)$ och y skiljer sig från varandra benämns false. Detta resulterar, i det binära fallet, i fyra olika klasser av datapunkter; True Positives (TP), False Positives (FP), True Negatives (TN) samt False Negatives (FN). Utifrån dessa klasser kan prestationsmått som exempelvis Accuracy, Average Accuracy, Precision, Recall och F-measure beräknas.

Accuracy visar, procentuellt, hur många exempel som har klassificerats korrekt av modellen.

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

Utöver Accuracy kan även Average Accuracy beräknas, där accuracy för vardera klass beräknas genom att ta korrekt klassificerade av klassen dividerat med totalt antal exempel i klassen. Därefter tas medelvärdet av dessa värden för att få ut Average Accuracy.

Precision beskriver procentuellt hur många av de exempel som modellen klassificerade som positiva, som faktiskt var positiva.

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

Recall är ett mått som visar hur många procent av de exempel som faktiskt var av den positiva klassen som modellen lyckades klassificera korrekt.

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

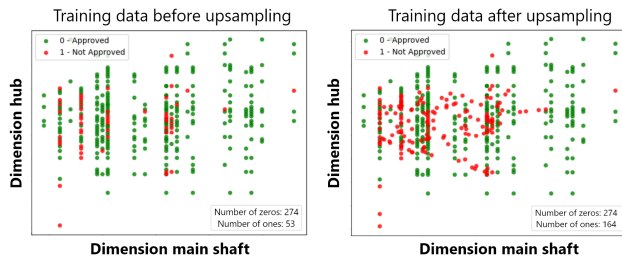
F-measure däremot är ett mått som sammankopplar både precision och recall, då det är det geometriska medelvärdet av de två måtten. Den visar hur väl modellen har presterat med både precision och recall tagna i beaktning.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Det finns en trade-off mellan precision och recall. Det värde som bör fokuseras på, beror på vad syftet med klassificeringen är. I situationer där det är viktigt att minimera mängden

false negatives, exempelvis vid klassificering av tumörer som benigna eller maligna, bör fokus vara att uppnå högt värde på recall. Situationen är däremot annorlunda för en klassificerare som ska klassificera mail som spam eller icke spam. I en sådan situation är det inte lika viktigt att minimera false negatives, utan istället viktigt att minimera false positives, eftersom det inte är önskvärt att missa viktiga mail. Detta medför att det är viktigt att uppnå högt värde på precision. Det kan dock även uppstå situationer där de två måtten bör värderas lika högt. Vid sådana tillfällen är det av godo att använda sig av F-measure för att få en helhetsbild över hur modellen har presterat på både precision samt recall. [4]

E. Bilaga 5



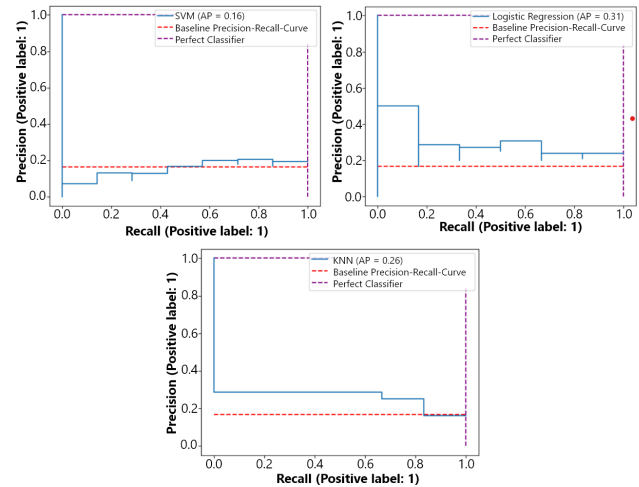
Figuren presenterar scatterplots över ett exempel på hur fördelningen av datapunkter i träningsdatan såg ut för en fold innan och efter SMOTE applicerats, vilket visas till vänster respektive höger. Y-axeln beskriver dimension på nav, medan x-axeln beskriver dimension på huvudaxel.

F. Bilaga 6

Prestationsmått verkligt dataset			
Medelvärde	SVM	KNN	LR
Average Accuracy	0.584	0.572	0.584
Recall	0.343	0.412	0.343
Precision	0.291	0.242	0.291
F-measure	0.305	0.298	0.305
AUC	0.373	0.624	0.589
AP	0.182	0.275	0.311

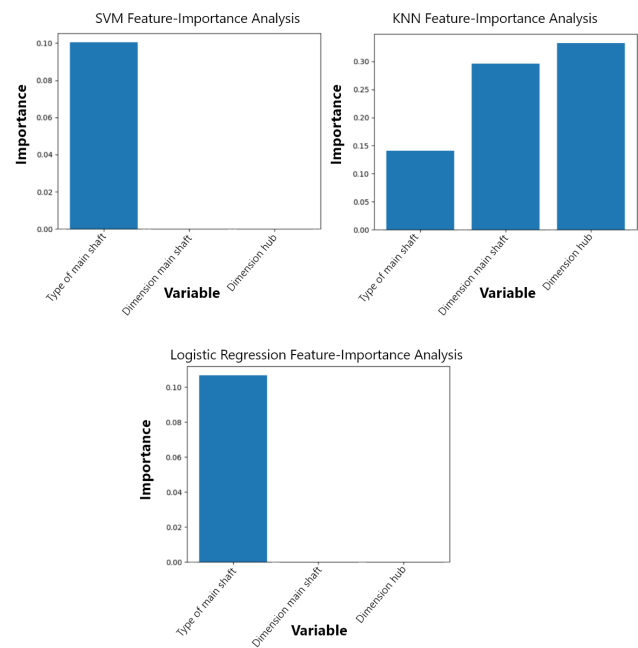
Tabell I: Tabellen visar medelvärden på de inkluderade prestationsmått för modellerna SVM, KNN och Logistisk Regression, givet det verkliga datasetet. Medelvärdena bygger på de värden som respektive fold har gett upphov till.

G. Bilaga 7



Figuren presenterar, för det verkliga datasetet, Precision-Recall-kurvan för SVM, KNN och Logistisk regression vars värde på AP var närmast det medelvärde på AP som k-fold cross validation resulterade i. Varje modells Precision-Recall-kurva sätts i relation till en perfekt klassificerarens kurva, samt en baseline som utgör lägsta nivån.

H. Bilaga 8



Figuren presenterar de olika variabelernas inverkan på kvalitetsutfallet, givet det verkliga datasetet, utifrån den fold som anses vara mest representativ för vardera modell.

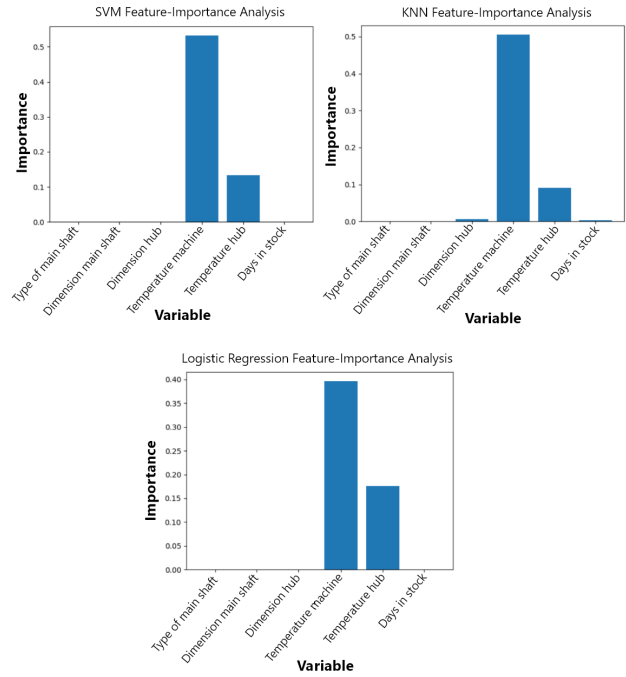
I. Bilaga 9

Prestationsmått fabricerat dataset			
Medelvärde	SVM	KNN	LR
Average Accuracy	0.920	0.924	0.903
Recall	0.892	0.913	0.926
Precision	0.801	0.745	0.608
F-measure	0.827	0.810	0.729
AUC	0.958	0.944	0.955
AP	0.896	0.828	0.875

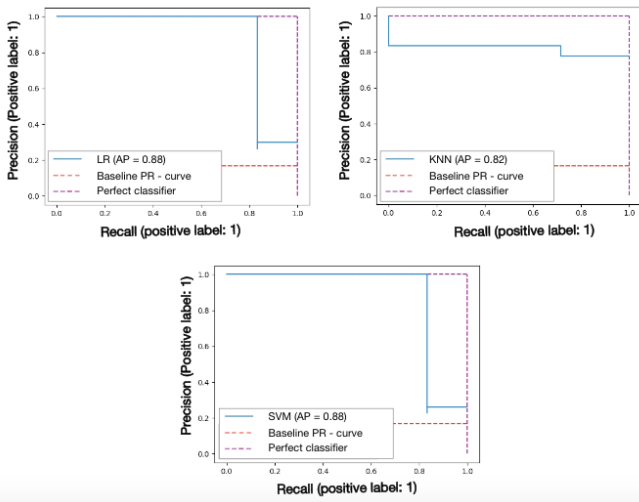
Tabell II: I tabellen kan medelvärden på inkluderade prestationsmått, givet det fabricerade datasetet, utläsas. Dessa medelvärden bygger på de värden som respektive fold har gett upphov till.

J. Bilaga 10

K. Bilaga 11



Figuren presenterar de olika variabelernas inverkan på kvalitetsutfallet utifrån den fold som anses vara mest representativ för vardera modell, givet det fabricerade datasetet.



För vardera modell har den precision-recall-kurva som genererats utifrån det fabricerade datasetet, vars värde på AP var närmast det medelvärde på AP som k-fold cross validation resulterade i, valts ut. Dessa visas i figuren. Modellernas prestationer kan sättas i relation till såväl en baseline, som en perfekt klassificerades precision-recall-kurva, i vardera graf.

TRITA-EECS-EX-2021:364