UPPSALA
UNIVERSITET

UPTEC STS 20007
Examensarbete 30 hp
Mars 2020

Applicera maskininlärning på vägtrafikdata för att klassificera gatutyper i Stockholm

Alexander Engberg

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Apply Machine Learning on Road Traffic Data in order to Classify Street Types in Stockholm

Alexander Engberg

In this thesis, two different machine learning models have been applied on road traffic data from two large cities in Sweden: Gothenburg and Stockholm. The models have been evaluated with regard to classification of street types in urban environments. When planning and developing road traffic systems it is important that there is reliable knowledge about the traffic system. The amount of available traffic data from urban areas is growing and to gain insights about historical, current and future traffic patterns the data can be used for traffic analysis.

By training machine learning models that are able to predict what type of street a measuring location belongs to, a classification can be made based on historical data. In this thesis, the performance of two different machine learning models are presented and evaluated when street types are predicted and classified. The algorithms used for the classification were K-Nearest Neighbor and Random Forest which were applied to different combinations of attributes. This was done in order to identify which attributes that lead to the optimal classification of street types in Gothenburg. For training the algorithms the dataset consisted of traffic data collected in Gothenburg. The final model was applied on the traffic data in Stockholm and hence the prediction of street types in that area were obtained. The results of this study show that a combination of all tested attributes leads to the highest accuracy and the model that obtained these results was Random Forest. Even though there are differences between topography and size of the two cities, the study leads to relevant insights about traffic patterns in Stockholm.

Handledare: Per-Anders Staav (Norconsult Astando AB)
Ämnesgranskare: Andreas Lindholm
Examinator: Elisabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 20007

Populärvetenskaplig sammanfattning

När antalet människor som bor i de svenska städerna fortsätter att öka växer även behovet av säkra och smarta lösningar för transporter. Det ställer höga krav på beslutstagare och trafikplanerare där ett flertal faktorer som påverkar utformningen och planeringen av vägtrafiksystem måste tas hänsyn till. För att bemöta efterfrågan och behov på tillgänglighet som finns inom en stad samlas trafikdata in för att kunna göra tillförlitliga prognoser och bedömningar. Kunskapen om hur trafik flödar inom ett vägtrafiksystem ligger sedan till grund för vilka beslut som är nödvändiga och driver på så sätt utvecklingen framåt.

För att säkerställa att tillförlitliga beslut om utbyggnad och utformning av vägtrafiksystem är möjliga krävs att statistiska och matematiska beräkningar utförs på kvalitetssäkrad trafikdata och pålitlig information. Trafikprognoserna används inom en rad olika applikationer som att förutspå trafikstockning, ruttplanering, planläggning av ombyggnationer och analyser av påverkan från ekonomiska incitament. Utifrån information om trafikflöde och hastigheter kan vägars funktioner analyseras och utvärderas och slutligen bidra till ett helhetsperspektiv för det studerade vägtrafiksystemet. För att utöka möjligheterna vid trafikanalys ämnar denna studie att undersöka hur metoder inom maskininlärning kan användas för att identifiera mönster i trafikdata genom att testa olika modeller och utvärdera vilken som har bäst potential.

I takt med utvecklingen av tekniska lösningar ökar möjligheterna för tillförlitliga framtidsprognoser och specifikt i och med utvecklingen av datorbaserade program och användandet av algoritmer. Bland annat har områden som artificiell intelligens och maskininlärning öppnat upp för en rad möjligheter där potentiella användningsområden blir fler och fler. Navigeringstekniker, varningssystem och autonoma bilsystem har blivit en del av vår vardag och för att utvecklingen ska fortsätta i samma takt är det viktigt att tillgänglig kunskap om själva vägtrafiksystemet utnyttjas. Ett problem som finns idag är att mängden insamlad trafikdata, som kan ligga till grund för sådan analys, ofta är begränsad eftersom trafikmätningsutrustning i många avseenden är kostsam.

Genom att studera likheter och olikheter hos två svenska kommuner, som var för sig utfört trafikmätningar, kunde en ny trafikanalys utföras med hjälp av maskininlärning. För att identifiera trafikmönster på olika vägar och gator i Stockholms kommun användes trafikdata och kunskap om gators funktioner i Göteborgs kommun. Studien resulterade i en möjlig identifiering av vägar och gators funktionella aspekter i ett vägtrafiksystem i en stad som tidigare saknat en sådan klassificering.

Innehåll

1 Introduktion

Till följd av ökade befolkningsmängder i de svenska storstadsområdena ökar också mängden trafik. Det är en utmaning eftersom det visat sig finnas stora kapacitetsbrister i det Svenska vägtrafiksystemet, framförallt i storstadsområdena. Vägar och gator är inte tillräckligt dimensionerade för att upprätthålla en säker och pålitlig transport när fler och fler människor ska dela på samma yta. Samtidigt går utvecklingen mot att effektivisera det befintliga vägtrafiksystemet i den grad att infrastrukturåtgärder inte behöver vidtas. Det handlar istället om att ta beslut som optimerar systemet så att det kan användas på ett smartare, effektivare och mer hållbart sätt (Trafikverket, 2012a).

I handboken *Trafik för en attraktiv stad* (TRAST) nämns trafikanalys som ett viktigt instrument för att ta lämpliga beslut i trafikrelaterade frågor. Syftet med trafikanalyser är att den ska ge ett underlag som kan beskriva effekter av föreslagna och genomförda åtgärder i vägtrafiksystemet. Grunden för trafikanalyser är insamlad data som kan användas för att beskriva ett nuläge eller användas för att beskriva en framtid. Den data som samlas in är trafikuppgifter från gator och vägar som kan bestå av hur många människor som använder en viss typ av väg under en viss tidsperiod, vilket typ av trafikslag som förekommer eller fordonshastighet (Trafikverket et al., 2015a).

Inom vägtrafiksystem utvecklas ständigt nya tekniska lösningar där maskininlärning och artificiell intelligens används i en allt större utsträckning. Utvecklingen av utrustning som samlar in trafikdata har banat väg för mer exakta uppskattningar av restid, bättre prediktioner av trafikstockning och annan trafikinformation som kan användas vid utveckling och förståelse av vägtrafiksystem. Möjligheterna för ökad beräkningskraft och en växande mängd insamlad trafikdata har alltså bidragit till att de områden som maskininlärning kan appliceras på blivit fler. Trots att det redan finns en mängd lösningar inom vägtrafiksystem som bygger på maskininlärning, är problemen fortfarande många och behoven för nya lösningar fortsätter att växa samtidigt som kraven förändras (Tizghadam et al., 2019).

En begränsning i trafikanalyser är ofta tillgången till data från tillräckligt många vägar under tillräckligt lång tid. I framförallt stora städer som består av många gator är det ett problem eftersom att kostnaden för fasta mätanläggningar som mäter trafiken året om ofta är dyra. När data inte finns tillgänglig uppskattas trafiken på gator vanligtvis genom att jämföra med andra gator som anses vara liknande. Det görs bland annat genom att installera mobila mätutrustningar som mäter trafiken under kortare mätperioder (mindre än 365 dagar) och som justeras med statistiska parametrar beräknade från de fasta

mätanläggningarna för att kompensera för den korta mätperioden. En sådan jämförelse kan vara föremål för stora fel då gator med fasta mätanläggning kan uppfylla en helt annan funktion i vägnätet och därför ha mer trafik än gator där mobila mätutrustningar har installerats. Det finns därför anledning att undersöka hur trafikmätningar med korta mätperioder kan användas för att förbättra trafikanalyser i städer.

Till den här studien har trafikdata från Göteborgs och Stockholms kommun varit tillgänglig. Den är framförallt baserad på trafikmätningar utförda under korta mätperioder av bilar och tung trafik och innehåller trafikuppgifterna trafikmängd, fordonshastighet och mängden tung trafik. Data från Göteborgs kommun är dessutom klassificerad utifrån de gator där trafikmätningar utförts på. Klassificeringen är baserad på de typer av gator som Göteborgs kommun ansett vara relevanta för att tydliggöra vilken funktion olika gator har i vägnätet. Det är dels gator som är placerade i ett visst område, såsom bostadsområden, centrumområden och industriområden, och gator som fyller ett specifikt behov. Trafikdata från Stockholm har inte denna typ av klassificering, varför det skulle vara intressant att studera i vilken grad olika typer av trafikuppgifter kan kopplas till en viss typ av gata i Stockholm, mer bestämt den funktion gatan uppfyller. Det skulle inte bara ge ett mer funktionsindelat vägnät utan även en förståelse för trafiken som rör sig där.

1.1 Syfte

Syftet med den här studien är att undersöka om maskininlärningsalgoritmer kan användas för att klassificera olika gatutyper i Stockholms kommun utifrån redan klassificerad data från Göteborg. Målet med arbetet är att bidra med information som kan användas för vidare trafikanalyser och insikter om trafikmätningar med korta mätperioder.

Följande frågeställningar har formulerats för att uppnå syftet:

1. Vilken maskininlärningsmodell kan skilja på olika gatutyper med hjälp av historisk data från trafikmätningar i Göteborg?

    - För att besvara denna fråga ska maskininlärningsmodellerna k-nearest neighbor (K-NN) och random forest för klassificering undersökas genom sex hypoteser. Baserat på hur väl modellerna presterar på hypoteserna presenteras resultatet.

2. Vilka variabler som input i maskininlärningsmodellerna ger bäst resultat? Trafikmängd (antal fordon per timme över en dag), hastighet och andel tung trafik (som andel av den totala trafikmängden) är 3 vanligt förekommande storheter från tra-

fikmätningar. För att fastställa vilka variabler/storheter som ger bäst resultat i modellerna kommer de undersökas enskilt och i kombination:

Enskilda variabler:

- trafikmängd
- hastighet
- andel tung trafik

Kombinerade variabler:

- trafikmängd, hastighet
- trafikmängd, andel tung trafik
- trafikmängd, hastighet, andel tung trafik

1.2 Avgränsningar

I den här studien har följande avgränsningar gjorts:

- Den trafikdata som används i den här studien är enbart data från motorfordonstrafik.
- Fokus i studien ligger vid de trafikmätningar som utförts av de två kommunerna Göteborg och Stockholm. Det medför att all trafikdata är begränsad till de respektive trafikmätningar som gjorts inom kommunerna.
- Studien går inte in på detalj om tekniken bakom olika trafikmätningsutrustningar.
- Eftersom studien utgår från de klassificeringar av gatutyper som finns i Göteborgs kommun går det inte att göra en direkt koppling till de klassificeringar som är nödvändiga i Stockholm. Manuellt arbete krävs för att säkerhetsställa att de tilldelade gatutyperna är representativa.

1.3 Disposition

I kapitel 2 presenteras de två kommunerna, som har legat till grund för den här studien, tillsammans med de trafikmätningar som gjorts och de mätplaser där trafikdatan har samlats in. Den avslutande delen i kapitlet behandlar viktiga aspekter som bör tas i beaktning vid analys av trafik. Kapitel 3 behandlar relevant teori bakom maskininlär-

ning där de två modellerna k-NN och random forest presenteras. Efter följer de metoder som används för att utvärdera modellerna tillsammans med statistiska och matematiska mått som ligger till grund för den analys av resultaten som görs. Nästa kapitel, kapitel 4, innehåller en kort beskrivning av den trafikdata som har använts i studien. Kapitel 5 innehåller de forskningsmetoder som applicerats på den här studien. Bland annat presenteras de tekniska verktyg som använts, hur all data har bearbetats och eventuella felkällor som kan finnas. Alla resultat som erhållits i studien presenteras i kapitel 6 tillsammans med den slutgiltiga modell som presterat bäst. De två efterföljande kapitlen, kapitel 7 och 8, består av diskussion om resultaten tillsammans med de slutsatser som kan dras utifrån den här studien.

2 Bakgrund

Sveriges vägnät är i uppdelat i allmänna och enskilda vägar. De allmänna vägarna är uppdelade i statliga och kommunala vägar. Trafikverket ansvarar för drift och underhåll i det statliga och det enskilda vägnätet som får statligt bidrag medan Sveriges kommuner och regioner (SKR) ansvarar för sina respektive vägnät. Det innebär också att de två aktörerna utför trafikmätningar i sina respektive vägnät för att följa trafikutvecklingen på nationell och lokal nivå. Det sker samtidigt ett utbyte av information mellan aktörerna. Bland annat använder Göteborgs och Stockholms kommun data från trafikmätningar gjorda av Trafikverket som angränsar kommunerna för att bistå i trafikrelaterade analyser. Dessutom finns information lagrad om det Svenska vägnätet i nationell vägdatabas (NVDB). Den innehåller dels information om hur olika vägar hänger ihop och information knutna till vägar som bland annat beskriver vägarnas egenskaper och de regler som gäller för de enskilda vägarna (Trafikverket, 2018).

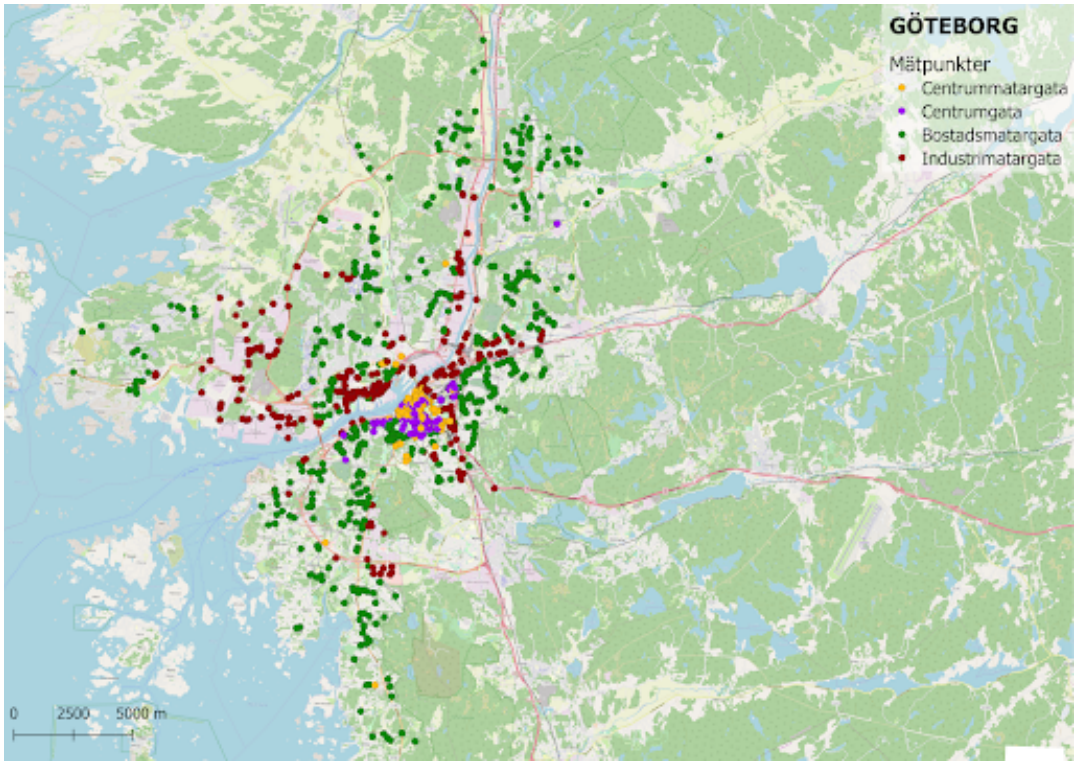
Norconsult Astando AB har utvecklat ett trafikmätningssystem åt Göteborgs och Stockholms kommun som gör det möjligt för kommunerna att importera trafikdata och beräkna trafikrelaterade parametrar (Norconsult, 2020). Kopplat till trafikmätningssystemet finns en utvecklingsmiljö med en databas med delar av kommunernas trafikmätningar som har använts i den här studien.

2.1 Göteborg

Det finns totalt 16 olika gatutyper som har tilldelats mätplatserna och är specifika för just Göteborg. Gatutyperna är följande:

- Primärled
- Primärled, in- utfartsled
- Primärled, centrumled
- Primärled, fritidsled
- Sekundärled
- Sekundärled, in- utfartsled
- Sekundärled, centrumled
- Sekundärled, fritidsled
- Sekundärled, industriled
- Centrummatargata
- Bostadsmatargata
- Industrimatargata
- Centrumgata
- Bostadsgata
- Industrigata

För den här studien har inte primärleder och sekundärleder varit intressant att ta med eftersom lederna dels är av en typ som är enklare att identifiera i vägnätet och som det inte det finns lika många av jämfört med övriga. De gatutyper som har varit av intresse är således: centrummatargata, bostadsmatargata, industrimatargata, centrumgata, bostadsgata och industrigata. Men efter granskning av tillhörande data, som beskrivs i kapitel *4. Data*, valdes gatutyperna: centrummatargata, bostadsmatargata, industrimatargata, centrumgata. Matargata syftar till en typ av gata som har funktionen att den leder in och "matar" trafik till andra gator. Till exempel är en bostadsmatargata en gata som leder in trafik till en eller flera bostadsgator. Figur 1 visar en karta över Göteborg där alla mätpunkter är utplacerade för de gatutyper som används i studien. Färgerna i figuren återspeglar vilken gatutyp som mätpunkterna har.

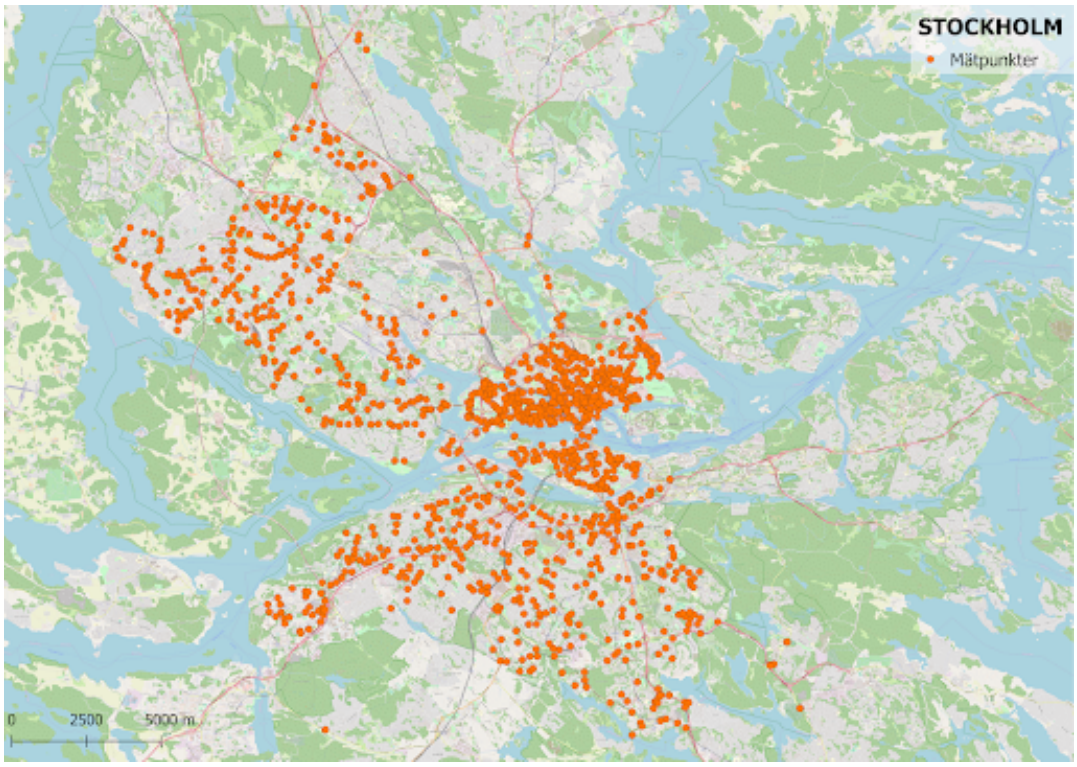


Figur 1: Karta över Göteborgs stad med utplacerade mätpunkter tillhörande gatutyperna centrummatargata, centrumgata, bostadsmatargata och industrimatargata.

Varje mätpunkt är kopplad till en specifik mätplats som innehåller information om vilket namn gatan har och vilken del av gatan som mätningen utgör. En mätplats kan ha en eller två mätpunkter kopplade till sig beroende på om det är en enkelriktad eller tvåfilig väg.

2.2 Stockholm

Som nämnt tidigare finns det i nuläget ingen liknande klassificering av gator i Stockholm som det finns i Göteborg. Det existerar olika typer av klassificeringar av vägnätet både på lokal och nationell nivå. Bland annat innehåller NVDB *funktionell vägklass* där alla vägar är tilldelade en siffra mellan 0-9 som anger hur viktig vägen är för det totala vägnätets förbindelsemöjligheter. Det är användbart vid ruttplanering och navigering eftersom den säger hur fordon ska ta sig från en väg som har funktionell vägklass 0 till en väg med funktionell vägklass 9 på snabbaste sätt (Trafikverket, 2012b). Den säger alltså inget om trafiken som går där, varför det finns anledning att hitta ett ytterligare sätt att klassificera vägar. I figur 2 nedan ses en karta över Stockholms stad där alla de mätpunkter som används i den här studien är utplacerade. Området där mätpunkterna ligger som tätast på kartan utgör de mest centrala delarna av Stockholm. På samma sätt som för mätplatserna i Göteborg kan en mätplats i Stockholm bestå av en eller två mätpunkter.



Figur 2: Karta över Stockholms stad med utplacerade mätpunkter.

2.3 Trafikvariationer

Den observerade trafiken från trafikmätningarna som har utförts på de olika mätplatserna i Göteborgs och Stockholms kommun varierar i tid och rum. Att förstå hur trafik varierar på olika platser under olika tidsperioder är en viktig del vid analys av trafik. Information

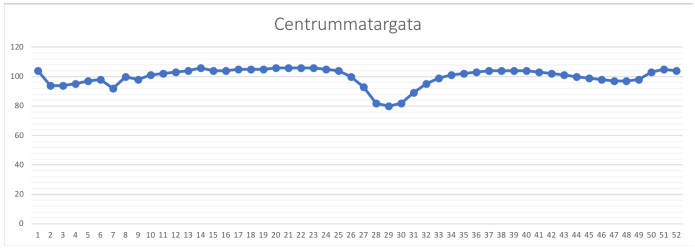
om vart och när trafiken varierar ger en förståelse för vart i trafiksystemet stora skillnader i kapacitet finns. Dessutom kan det ge en indikation om vilken funktion en gata uppfyller genom att analysera aktivitetsmönsterna som finns där.

Vanliga parametrar som används i trafikanlyser är parametrar som beskriver trafiken på en genomsnittlig dag (årsdygnstrafik: ÅDT) eller på en genomsnittlig vardag (årsmedelvardagsdygnstrafik: ÅMVD). ÅDT och ÅMVD beräknas enkelt när data från fasta mätanläggningar finns tillgänglig. ÅDT är den totala trafikmängden räknat i antal fordon som passerar en mätpunkt under ett kalenderår, dividerat med antal dagar under det givna året. ÅMVD beräknas på samma sätt som ÅDT men där helgdagar exkluderas från beräkningen. ÅDT och ÅMVD ger alltså en representation av den typiska trafikmängden på en given väg under en genomsnittlig dag eller vardag under ett givet år. För mätningar utförda under korta mätperioder kompenseras ofta mätningarna genom att mätningen justeras med en faktor för att uppskatta ÅDT eller ÅMVD. Faktorn baseras ofta på en eller flera mätplatser där permanenta mätanläggningar är installerade som mäter trafiken året om och varierar beroende på månad och dag under året. Eftersom den här studien undersöker hur trafikmätningar, utförda under korta mätperioder, kan användas i analys av trafik utan att justera mätningarna behöver variationerna studeras (Trafikverket, 2015b). Det här avsnittet förklarar dessa variationer. Först presenteras tidsmässiga variationer mellan de olika tidsskalorna tillsammans med några av de bakomliggande orsaker som kan förklara varför variationerna förekommer. Efter det presenteras rumsliga variationer och hur det är relevant för den här studien.

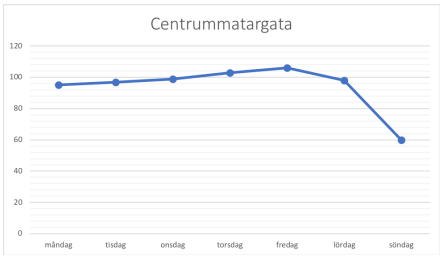
2.3.1 Tidsmässiga variationer

Trafik varierar regelbundet under året (mellan olika månader), under veckan (mellan olika dagar) och under dygnet (mellan timmar). Den drivande orsaken bakom variationerna skiljer sig mellan olika tidsskalor. I stadstrafik orsakas framförallt trafikljus de kortsiktiga variationerna. Variationer mellan timmar under dagen och mellan dagar under veckan orsakas huvudsakligen av efterfråga och variation i kapacitet i vägnätet. Under vardagar dominerar människors resebehov av resor till och från arbete eller skola, fritidsaktiviteter och andra ärenden under ett relativt förutsägbart tidsintervall (Trafikverket et al., 2015a). Under vardagar i stadsmiljö når trafiken generellt två toppar, den första under förmidagstimmarna och den andra under eftermidagstimmarna samt en period av lägre trafik mellan topparna, där toppen under eftermiddagen generellt är högre än förmiddagen (Festin, 1996). Variationer mellan dagar under veckan går att dela upp i veckodagar, helgdagar, säsongsbaserade ledigheter (t.ex jullov,sportlov, påsklov och sommarlov). Under vardagar har trafiken ungefär samma fördelning mellan måndag och torsdag men med

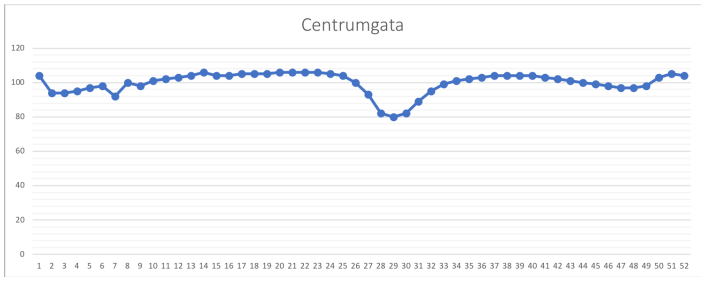
generellt högre andel trafik på fredagar. Helgdagar har ofta lägre trafik och där regelbundenheten i trafiken inte följer samma mönster som under veckodagarna (Festin, 1996). För att illustrera hur årsvariationer och veckovariationer ser ut på de gatutyper som ingår i den här studien har åtta grafer tagits fram utifrån tillgänglig trafikdata. Dessa variationer kan ses i figur 3 på nästa sida. För samtliga gatutyper går det att se att sommarperioden går ner kraftigt i trafikmängd i jämförelse med resterande delen av året. Industrimatargata är den gatutyp som når den lägsta trafikmängden jämfört med de andra gatutyperna. Alla gatutyper förutom bostadsmatargata visar en relativt regelbunden trafik över årets veckor med undantag för vecka 7, som är sportlovsvecka, och mindre förändringar som kan ses vid början och slutet av året. För veckovariationerna går det att se en tydlig skillnad mellan veckodagar och helgdagar, som också nämnts vara en genomgående trend. Dessutom kan yttre faktorer påverka variationer mellan dagar. Det är faktorer som både påverkar kapaciteten på en väg men också faktorer som påverkar attityden att välja ett trafikslag före ett annat. Vägarbeten och olyckor är exempel på faktorer som påverkar kapaciteten eftersom normalt flöde inte kan uppnås. Väderomslag (regn och snö) är också exempel som kan påverka kapaciteten men det kan också göra att människor väljer att ta bilen istället för att till exempel ta bussen (Trafikverket et al., 2015a).



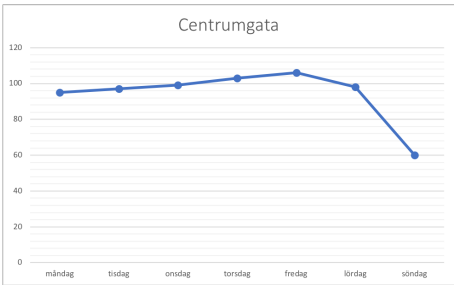
(a) Årsvariation



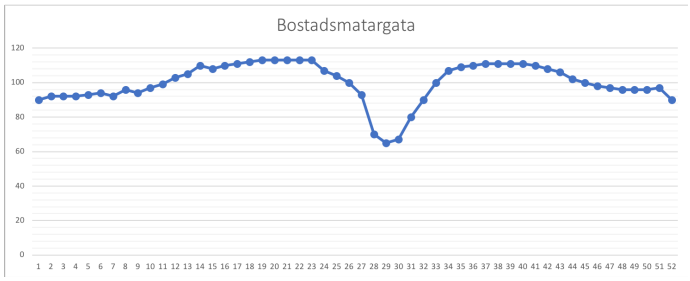
(b) Veckovariation



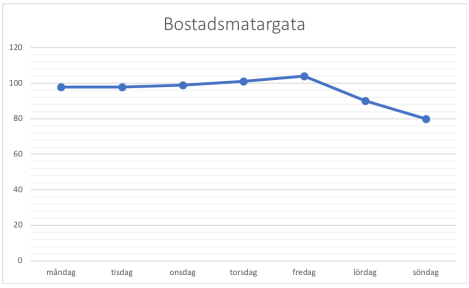
(c) Årsvariation



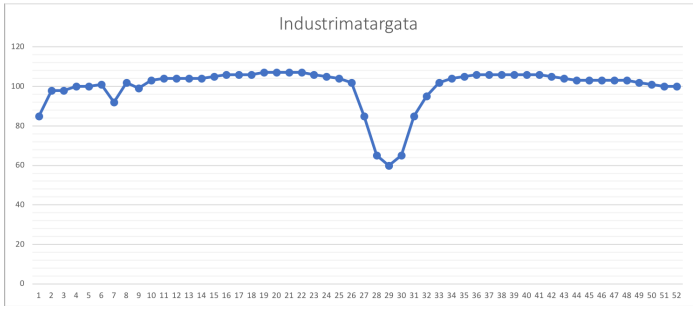
(d) Veckovariation



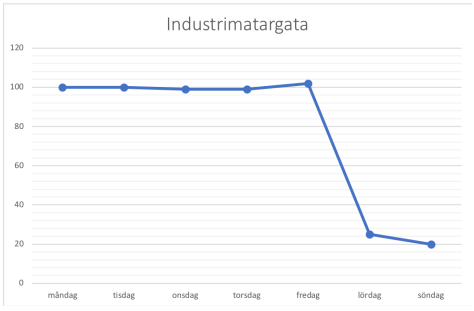
(e) Årsvariation



(f) Veckovariation



(g) Årsvariation



(h) Veckovariation

Figur 3: Års- och veckovariationer för fyra olika gatutyper. Variationerna är baserade på trafikdata från Göteborg. Y-axeln representerar antalet fordon och x-axeln är tidsskalan.

2.3.2 Rumsliga variationer

Rumsliga variationer i trafik innebär att trafiken varierar beroende på vilken plats eller område som undersöks. Om de tidsmässiga variationerna inte tas hänsyn till vid en rumslig analys blir resultatet enbart en identifiering av specifika platser med en viss trafikmängd

under ett specifikt tidsintervall. Därför är det viktigt att ta de tidsmässiga variationerna i beaktning vid analys av rumsliga variationer. Det medför att likt tidsmässiga variationer kan analyser om rumsliga variationer utföras på olika tidsskalor där tim-, dygns- och årsvariationer kan identifieras för olika områden.

Weijermars beskriver i sin avhandling *Analysis of urban traffic patterns using clustering* vikten av förståelsen för hur trafikvolymer förändras i stadsmiljöer för att förbättra trafikanalyser. Genom att analysera variationer i en specifik stadsmiljö belägen i Nederländerna klassificerar Weijermars trafikdata genom klusteranalys som baseras på så kallade *traffic flow profiles*. Trafikflödesprofilerna används för att identifiera specifika trafikmönster med avseende på tidsmässiga och rumsliga variationer inom olika grupper som vägar och områden kan delas in i. Profilerna bygger bland annat på hur trafiken varierar på en plats per timme eller dygn och används för att identifiera rumsliga kluster inom ett vägtrafiksystem (Weijermars, 2007). Den här studien ämnar inte att ta fram sådana profiler utan målet är att med hjälp av maskininlärning identifiera mönster i trafikdata genom redan klassificerade vägar. Utan att gå in på djupet på vilka variationer som finns för de olika gatutyperna undersöks möjligheterna till att identifiera vägars funktioner, precis som de platsspecifika indelningar som Weijermars gör.

Zhao och Chung (2001) utvecklade en metod för att uppskatta ÅDT på vägar där trafikmätningar inte utförs. Undersökningen gick ut på att studera vilka faktorer som påverkar ÅDT på vägar i stadsområdet Broward county, Florida. Genom att dela in oberoende variabler i fyra olika grupper: vägkaraktäristiska data, socioekonomiska data, regional tillgång till motorvägar och regional tillgång till arbete kunde olika regressionsmodeller utformas och testas. De variabler som bland annat ingick i de fyra olika grupperna var antal körfält, områdestyp, vägens funktion, andel sysselsatta, möjligheter till arbete och minimal distans till motorväg. Av de fyra regressionsmodellerna utvärderades de olika variablernas relevans vid beräkning av ÅDT. Vägens funktion och antal körfält visade sig vara de två mest betydelsefulla variablerna vid prediktion av ÅDT (Chung et al., 2001). De olika funktionerna som vägarna representerade var: huvudgata (urban principal arterial), mindre huvudgata (urban minor arterial), förbindelseväg (urban collector) och ospecificerad gata (Åkerlund, 2007). Till de variabler som inte valdes ut som mest relevanta var bland annat områdestyp, vilken var uppdelad i områden utifrån bland annat: stadskärna (central business district, CBD), bostadsområde (residential) och glesbebygd (rural area). Men trots att områdestyp inte ingick i de slutliga modellerna erhölls en relevans på 49% vid de initiala testerna och är ett intressant resultat för den här studien.

3 Maskininlärning

Konceptet maskininlärning syftar till att ge datorer förmågan att lära och anpassa sig till en uppgift utan att specifikt vara programmerad till det. Själva lärandet innebär att tillgänglig data kombineras med matematiska modeller för att i slutändan resultera i en färdig modell där, till exempel, parametrar har anpassats till ett problem. Målet med att använda maskininlärning är att den tränade modellen ska kunna dra slutsatser om ny, okänd data som inte ingått under själva träningen (Lindholm et al., 2019).

Inom maskininlärning finns tre olika områden: *supervised learning*, *unsupervised learning* och *reinforcement learning*. De tre områdena beskriver hur ett problem ska hanteras och vilken typ som lämpar sig bäst bestäms utifrån problemets natur. Om en modell ska tränas på data som kan kopplas till ett redan känt resultat används supervised learning. Modellen tränas genom att hantera varje datapunkt tillsammans med det kända resultatet för att i slutändan prediktera hur ny data bör bete sig (ta beslut eller göra prediktioner). Men om den data som ligger till grund för prediktionerna saknar ett känt resultat (utdata) lämpar sig unsupervised learning. I motsats till supervised learning tränas modellen enbart på observerad data (indata) för att finna mönster eller trender som kan beskriva datan. Utifrån de identifierade underliggande strukturer som (möjligtvis) finns i datan ska den tränade modellen prediktera ny, osedd data. Slutligen finns reinforcement learning som innebär att ett program lär sig att ta beslut som ska gagna systemet på bästa sätt. Genom återkoppling från den miljö som den så kallade aktören befinner sig i lär sig programmet att successivt ta bättre och bättre beslut (Castañón, 2019, Lindholm et al., 2019). I det här arbetet är det supervised learning som används för att analysera data och hitta underliggande strukturer. Syftet med att använda maskininlärning är att undersöka om det finns mönster i den trafikdata som används för att kunna klassificera ny trafikdata utifrån det. Eftersom utgångspunkten för studien är att hitta mönster i en redan klassificerad data för att i slutändan klassificera ny data lämpar sig supervised learning.

3.1 Supervised learning

För att det ska vara lämpligt att använda supervised learning måste problemet alltså vara på formen $\{x_i\ ,\ y_i\ \}_{i=1}^{n_i}$ där x_i är den indata och y_i den utdata som modellen tränas på. Den träningsdata som ingår vid träningsfasen är således "märkt" och visar det verkliga uppmätta resultatet till varje specifik datapunkt. Vid stora datamängder ökar komplexiteten i relationerna mellan indata och utdata vilket gör att det inte går att se mönstren med blotta ögat. Då kan en statistisk modell användas för att tränas och lära sig relationerna

mellan x och y. Den tränade modellen appliceras sedan på ny data som saknar resultat, där svaret kallas prediktering. Syftet är att datorn ska lära sig de mönster som finns för att sedan kunna replikera det (Lindholm et al., 2019).

Det finns två huvudområden där supervised learning är applicerbar: klassificering eller regressionsanalys. Vid klassificering representeras indata i vektorform med tillhörande kategori eller klass, som benämns attribut, och vid regressionsanalys representeras det tillhörande svaret i form av siffror. Kategorier eller klasser inom klassificering kan också bestå av nummer, men då representerar det numret ett specifikt attribut snarare än en siffra i decimalform. Vid regressionsanalys är målet att prediktera vilket värde en specifik indata bör få. Inom både klassificering och regressionsanalys kan indata vara på numerisk form (Lindholm et al., 2019).

Trafikdata består ofta av uppmätta hastigheter, antal fordon, geografisk information och annan relevant data. Indata kan därmed bestå av både kategoriska och numeriska värden. Målet med den här studien är att undersöka hur uppmätt trafikdata kan kopplas till en viss gatutyp och eftersom gatutyp (utdata) är av kategorisk karaktär handlar det om klassificering.

3.2 Klassificering

Vid klassificering behandlas svaret av ett problem som kvalitativt eftersom den kan bestå av namn, benämningar eller siffror (som inte följer någon naturlig ordning). Syftet med klassificering är att förutspå det kvalitativa svaret (också benämnt kategorisk) utifrån ett antal inmatningar (kvalitativa och/eller kvantitativa) (Hastie et al., 2013). Eftersom svaret är av kvalitativ karaktär måste det finnas ett bestämt antal svar som kan antas. Om uppsättningen av antal svar benämns K bestäms klasserna Y utifrån de klasser eller kategorier svaret kan anta (Lindholm et. al. 2019). Vid klassificering av trafikdata utifrån de bestämda kategorierna *gatutyp* antar Y de olika gatutyperna, till exempel (Y=14) med gatutypen {Centrumgata}. Klassificeringen kan antingen vara binär, det finns två klasser som svaret kan anta, eller bestå av flera klasser, K $>$ 2 (på engelska: *multiclass*) (Murphy, 2012).

Klassificering handlar om att förutspå ett resultat utifrån data vilket kan beskrivas genom det statistiska uttrycket *p(y|x)* som beskriver sannolikheten att y antar ett specifikt svar givet en viss inmatning x. Trots att svaret y är kvalitativ kan således svaret uttryckas kvantitativt i form av sannolikhet. Anledningen till att undersöka svaret kvantitativt är att kunna utvärdera hur väl en modell presterar där sannolikheten varierar mellan 0 och

1 (Lindholm et al., 2019).

Inom klassificering finns olika modeller som kan användas och beroende på vilket problem som undersöks passar vissa modeller mer eller mindre bra. Nedan presenteras de modeller som används i den här studien. Den första modellen är en så kallad *instance-based metod* och den andra är en trädmodell som utökas till en *ensemble metod*. Instansbaserade metoder är vanligt förekommande vid träning av modeller inom maskininlärning eftersom de ofta är enkla men effektiva. Den andra metoden, som i enklare mening är en ihopsättning av flera modeller, kan finna mer komplicerade mönster och används för att undersöka om en mer komplicerad modell är att föredra (Lindholm et al., 2019).

3.2.1 Instance based learning

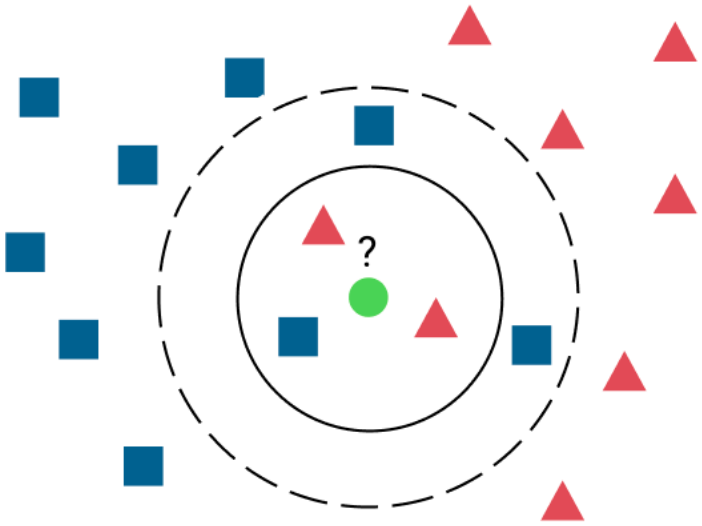
Inom maskininlärning finns en grupp av inlärningsalgoritmer som kallas *instance-based learning*. Istället för att enbart utföra explicit generalisering (ett grundläggande begrepp inom maskininlärning som innebär att en modell generaliserar data för att kunna förutspå framtiden) jämför modellen nya exempel med exempel som används vid träningsfasen. Mer specifikt innebär det att det bildas hypoteser direkt från ett träningsexempel där komplexiteten hos hypoteserna växer i och med att mängden träningsdata gör det. En fördel med att använda instance-based learning är förmågan hos algoritmen att lagra ny information om en instans och samtidigt kunna förkasta information om en gammal instans. Det medför möjligheter för algoritmen att ständigt anpassas och därmed prestera väl på ny data. Ett exempel på en sådan algoritm är *k-nearest neighbors* (k-NN) som lagrar all träningsdata för att sedan vid körning härleda ett svar utifrån den nya instansens närmaste datapunkter. I motsats till instance-based algoritmer finns bland annat beslutsträd och neurala nätverk (Keogh, 2011).

3.2.2 K-Nearest neighbor

Som namnet antyder bygger k-NN på att associera en ny datapunkt med k närmsta datapunkter som finns bland de datapunkter som använts vid träning. Sannolikheten att input x tillhör en specifik klass y ($p(y|x^*)$) grundar sig därmed i vilka klasser dess närmsta träningsdatapunkter ingår i. Värdet på k bestäms av användaren och bestämmer egenskaperna hos modellen vilket kan uttryckas som:

$$p(y = j|x_\star) = \frac{1}{k} \sum_{i \in R_*} \mathbb{I}\{y_i = j\} \tag{1}$$

där $j = 1, 2, \ldots, k$ och k är antalet närmsta datapunkter samt där $R_\star$ innehåller de k närmsta datapunkterna till $x_\star$. För att avgöra vilken klass som har den högsta sannolikheten att ny data tillhör, används majoritets votering. Det innebär att den klass som representeras flest gånger bland de k närmsta datapunkterna är den klass som får högst sannolikhet. K-NN är både en icke-parametrisk och icke-linjär modell där avståndet från en ny datapunkt beräknas gentemot tidigare punkters vektorer där det vanligaste måttet är euklidiskt avstånd (Lindholm et al., 2019). För att förstå hur värdet på k påverkar resultatet illustreras ett exempel i figur 4 nedan. Om $k = 2$ får den nya datapunkten klassificeringen röd. Om $k = 5$ blir den nya datapunkten istället klassificerad som blå.

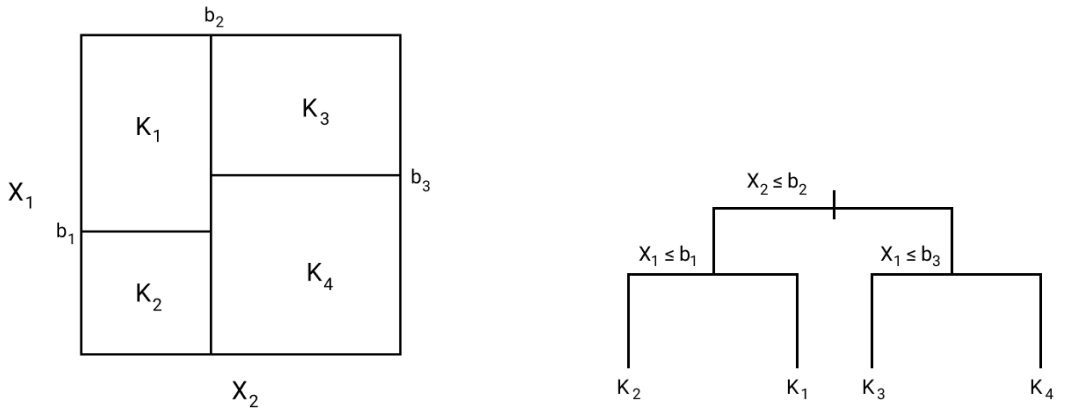


Figur 4: Ett exempel på hur värdet på k i k-NN påverkar resultatet.

Alltså har värdet på k en stor inverkar på hur väl modellen kommer att prestera. Att välja k är därför ett viktigt steg där slutmålet är att göra modellen tillräckligt generaliserad för att undvika överanpassning. Överanpassning innebär att modellen har blivit för anpassad till träningsdata vilket gör att den presterar dåligt på osedd data. Modellen börjar då diagnostisera mönster i slumpmässiga variationer som finns i data snarare än att generalisera korrekt. Att välja värdet på k är komplicerat och det är inte alltid helt självklart hur en ska gå tillväga. Ett vanligt sätt att välja k är därför att testa olika värden och sedan välja det k som genererar bäst resultat utifrån det. Ett bra mått på hur bra en modell presterar är att använda korsvalidering (Lindholm et al., 2019). En beskrivning av hur det går till presenteras i avsnitt *3.3 Utvärdering av modell* i det här kapitlet.

3.2.3 Beslutsträd

Träd-baserade modeller delar in indata i flera regioner där varje region, *p(y/x)* modelleras som den empiriska distributionen bland den träningsdata som finns i den specifika regionen. Hur indata ska delas in i dessa regioner kan summeras i ett träd, vilket gett upphov till namnet på metoden. Beslutsträd kan användas inom både klassificering och regressionsanalys. Om metoden appliceras inom klassificering representerar varje region en predikterad klass. Dessa predikterade klasser kan vara densamma, alltså kan K_1 vara samma klass som K_4 (Rokach et al., 2005). I figur 5 representeras ett beslutsträd som innehåller fyra predikterade klasser, $K_1 - K_4$, med två attribut, X_1 och X_2. Till vänster presenteras hur attributen är indelade i de olika regionerna med tillhörande gränser och till höger är själva beslutsträdet.



(a) Beslutsgränser.

(b) Beslutsträd.

Figur 5: Ett beslutsträd bestående av fyra predikterade klasser (K_1 , K_2 , K_3 , K_4), två attribut (X_1 , X_2) och tre gränser (b_1 , b_2 , b_3).

Längst ner i beslutsträdet är de klasser som regionerna delas in i och i själva trädet benämns dessa noder *löv*. Om ett beslutsträd skulle vara perfekt konstruerat skulle varje löv enbart innehålla datapunkter som tillhör den specifika klassen. De algoritmer som utgör beslutsträdet går ofta uppifrån ner vilket betyder att varje delning sker vid det gränsvärde som algoritmen anser utgöra den bästa delningen. Målet med varje delning är att få så få felklassificeringar som möjligt och utgör således kravet för var delningen görs (Rokach et al., 2005). Men även om kravet på hur den bästa delningen ska göras är entydig är verkligheten ofta mer komplex än så. Till exempel ökar komplexiteten i hur en delning ska göras vid högre dimensioner av modellen och det finns olika algoritmer som använder olika mått på hur den "bästa" delningen ser ut. Två vanliga mått för att beräkna andelen felklassificeringar är Gini index och entropi. Om vi har K klasser där i = {1, 2, ... , K} och är p_i andelen datapunkter som tillhör klass K i det området (p_1

är således andelen datapunkter som tillhör klass 1 osv) beräknas Gini index och entropi enligt:

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

$$Entropi = -\sum_{i=1}^{K} p_i log_2 p_i \tag{3}$$

Om indelningen av alla löv är felfri, det vill säga om inget löv har blivit tilldelad fel klass, kommer både Gini index och entropin att vara 0 (Lindholm et al., 2019, Hastie et al. , 2017).

3.2.4 Ensemble metoder

En teknik inom maskininlärning som skiljer sig från instance-based learning är *ensemble methods*, vilket refererar till modeller som, istället för att bestå av en enskild modell, använder flera olika metoder för att förbättra resultaten. Eftersom de består av en uppsättning av modeller kan de refereras till meta-algoritmer där idén är att den slutliga prediktionen ska bestå av ett flertal prediktioner som görs genom en ihopsättning av så kallade basmodeller (Lindholm et al., 2019).

Bagging, som är en förkortning på *bootstrap aggregating*, är ett vanligt sätt att träna ensemble modeller där poängen är att träna flera modeller parallellt. De modeller som tränas är av samma typ men tränas på olika delar i trängsdatan. Eftersom resultatet från ensemble modellerna är medelvärdet av alla prediktioner kan variansen minskas i jämförelse med att enbart träna en modell (Lindholm et.al. 2019). Att ta medelvärdet på resultaten från flera modeller betyder att det brus som skapas vid varje modell som tränas kan reduceras, varpå variansen minskar (Hastie et.al. , 2017).

3.2.5 Random forest

En kraftfull ensemble metod som använder bagging är *Random Forest*, vilken är den andra metoden som används i den här studien. Metoden bygger på beslutsträd som basmodeller för det klassificeringsproblem som ska lösas. Beslutsträd är en bra kandidat vid bagging eftersom beslutsträd i sig ofta kan hantera komplexa strukturer i data och om de tränas tillräckligt djupt kan de även minska *bias* (en hög bias innebär att modellen missar re-

levant information i datasetet) (Hastie et al., 2017). Slutligen består prediktionen av en sammansättning av *alla* de beslut som gjorts inom träden och skapar således ensemble metoden (eller skogen) (Microsoft Azure, 2019).

Det finns flera parametrar som kan ändras i random forest klassificeraren för att optimera resultaten. Bland annat kan kriteriet för delningar ändras (gini index, entropi), storleken på träden och även antalet träd som ska tränas. Processen för att träna en random forest algoritm kan beskrivas i sex olika steg (Hastie et al., 2017):

1. Bestäm antal träd.
2. Använd ett slumpat stickprov från träningsdatan för varje träd.
3. Bygg det bestämda antalet träd genom att välja vilka variabler som modellerna ska tränas på och bestäm kriterium för delning.
4. Resultatet från alla träd kan presenteras.
5. För varje träd görs en klassprediktion.
6. Den klass med flest röster (räknat från alla träd) blir den slutgiltiga prediktionen.

Ett beslutsträd i sin ensamhet kan leda till hög bias men låg varians. Genom att utöka metoden till random forest kan alltså låg varians uppnås tillsammans med låg bias. Felet reduceras genom medelvärdet från varje träd samtidigt som en slumpmässighet läggs till inom varje träd för att minska korrelationen träd emellan (Lindholm et al. , 2019).

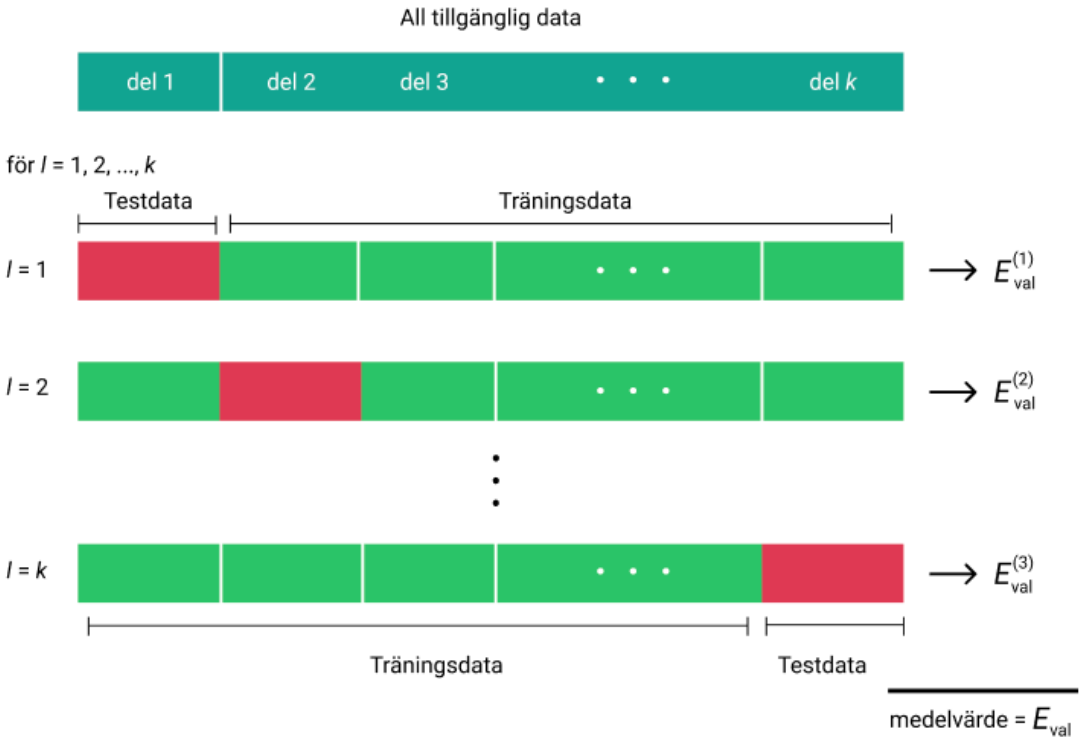
3.3 Utvärdering av modell

Att utvärdera en modell är en viktig del inom klassificering och målet är att välja den modell som bäst passar ett problem. Det finns olika sätt och mått för att utvärdera en modell och i det här arbetet används metoden *k-korsvalidering* för att jämföra hur modeller presterar i jämförelse med varandra. Andra sätt som kan användas för att utvärdera hur väl en modell presterar är genom visualisering av rätt- och felklassificeringar, vilket kan göras med en *confusion matris* (Lindholm et al. , 2019). I det här avsnittet förklaras metoderna närmare tillsammans med applicering och relevans.

3.3.1 K-korsvalidering

När väl en modell tränas på en mängd träningsdata kan det vara nödvändigt att utvärdera hur bra modellen presterar. Det finns olika metoder att utvärdera en klassificeringsmodell där en metod kallad k-korsvalidering är en välkänd och ofta använd metod. Syftet med att använda k-korsvalidering är att samla kunskap om hur *exakt* resultatet blir från en modell när den appliceras på ny data. Informationen som erhålls genom korsvalidering kan bland annat användas till att jämföra hur olika modeller har presterat, vilket möjliggör att *olika* typer av modeller kan jämföras. När en modell tränas på träningsdata finns det verkliga svaret tillgängligt eftersom all träningsdata är märkt. Men eftersom testdata inte har någon känd utdata är det svårare att få en uppskattning om hur modellen presterat i slutändan. Genom att använda korsvalidering görs en skattning utifrån den träningsdata som finns för att kunna säga något om hur modellen troligen kommer att prestera på ny data (Lindholm et al. , 2019).

Tekniken går ut på att all tillgänglig data som är märkt slumpmässigt delas in i k lika stora delar med samma antal observationer i varje del, se figur 6. Genom att varje del iterativt får agera testdata går det att vid varje träning och test beräkna hur stort fel det blir vid varje klassificering. Det innebär att vid första iterationen är det den första delen som är testdata samtidigt som modellen tränas på resterande observationer. Vid den andra körningen är det nästa del som är testdata och på så sätt fortsätter utvärderingen tills alla delar har agerat testdata. Vid varje steg beräknas felet E_{val} som i slutändan används för att beräkna ett medelvärde på felet som modellen kan förväntas ha. Ett vanligt värde på k är 10 men borde alltid bestämmas utifrån de modeller som ska utvärderas. Till exempel, om en modell är beräkningstung (det krävs många steg för att träna modellen) kan det vara fördelaktigt att inte ha ett för stort k eftersom det leder till lång beräkningstid och stor prestanda. Men om en modell istället är av enklare karaktär går det att välja större k utan att det påverkar tiden det tar att köra modellen (Lindholm et al. , 2019).



Figur 6: En illustration över hur k-korsvalidering fungerar. Träningsdata är uppdelad i k lika stora delar och för varje iteration agerar en del som testdata och resterande som träningsdata. För varje iteration beräknas felet som i slutändan utgör det uppskattade felet för modellen.

När alla steg i korsvalideringen har gjorts tränas modellen återigen på all tillgänglig, märkt data som blir den slutgiltiga modellen. Skattningen av E_{val} beräknas och resultatet blir det fel, E_{new} (förväntat fel), som modellen kan tänkas få på ny data (Lindholm et al. , 2019).

3.3.2 Confusion matris

Ett sätt att visualisera resultatet från en klassificerare som tränats och sedan testats på testdata är att göra en *confusion matris*. Om det finns två klasser som har predikterats går det att dela in testdatan i fyra olika grupper som är beroende av utdata och skattad utdata (Lindholm et al. 2019). Om testdatan kan predikteras som antingen positiv eller negativ kan matrisen se ut som följande:

Tabell 1. Confusion matris innehållande två klasser (positiv och negativ).

Klass	**Predikterade som positiva**	**Predikterade som negativa**	**Summa**
Positiv målvariabel	Antal sant positiva (SP)	Antal falskt negativa (FN)	SP + FN
Negativ målvariabel	Antal falskt positiva (FP)	Antal sant negativa (SN)	FP + SN
Summa	SP + FP	FN + SN	SP + FP + FN + SN

Antal sant positiva (SP) innebär antalet datapunkter som predikterats till positiva och faktiskt tillhör den positiva klassen. Falskt positiva (FP) är de datapunkter som predikterats till positiva med i verkligheten tillhör den negativa klassen. På samma sätt följer de falskt negativa (FN) och sant negativa (SN). Likväl som det går att konstruera matrisen utifrån två klasser går det att utöka till klassificering av flera klasser. Dimensionen av matrisen ökar då parallellt med antalet k klasser, vilket skapar matrisen $k\ x\ k$. För att illustrera hur en sådan tabell kan se ut ökar vi antalet klasser till tre: A, B och C (Lindholm et al., 2019, Webb, 2010). I tabell 2 nedan representerar den första raden antalet datapunkter som egentligen tillhör klass A, den andra raden är antalet faktiska datapunkter som tillhör klass B och slutligen antalet punkter som tillhör klass C i den sista raden. Till exempel har 4 av 5 datapunkter som tillhör klass A blivit korrekt klassificerade medan enbart 4 av 8 punkter tillhörande klass B fått rätt klassificering. Om klassificeringen vore perfekt skulle enbart diagonalen innehålla värden och resterande celler skulle vara noll. Vid en sådan klassificering skulle inga datapunkter blivit felklassificerade.

Tabell 2. Confusion matris innehållande tre klasser (A,B och C).

Klass	**Predikterade som klass A**	**Predikterade som klass B**	**Predikterade som klass C**	**Summa**
Målvariabel A	4	1	0	5
Målvariabel B	3	4	1	8
Målvariabel C	1	3	5	9
Summa	8	8	6	22

Matrisen används inte enbart för visualisering av resultatet utan kan användas för att beräkna statistiska mått som noggrannhet, precision, känslighet och F-score. Hur måtten kan användas och hur de beräknas presenteras nedan.

3.3.3 Noggrannhet

Att beräkna noggrannheten innebär att man på ett enkelt sätt kan utvärdera hur väl en klassificeringsmodell har presterat. Måttet är bland de enklare sätten att utvärdera en modell och är även en av de mer intuitiva. Noggrannheten, som beräknas enligt ekvation 4, mäts i procent genom att ta antalet sant positiva tillsammans med antalet sant negativa delat med det totala antalet datapunkter (Webb, 2010). Resultatet är ett mått på hur stor procent av modellen som är "rätt". Exempelvis, om noggrannheten är 0.805 innebär det att modellen är ungefär 80% rätt.

$$Noggrannhet = \frac{SP + SN}{SP + SN + FP + FN} \tag{4}$$

Eftersom noggrannheten inte tar hänsyn till asymmetriska dataset (antalet falskt positiva och falskt negativa är inte lika många) är det viktigt att ta hänsyn till fler parametrar än vad noggrannheten beräknar för att utvärdera modellens prestation (Webb, 2010).

I den här studien används fler än två klasser (ej binära som negativa och positiva) vilket innebär att det krävs ytterligare sätt att utvärdera en klassificeringsmodell. Hur detta hanteras förklaras sist i det här kapitlet, under avsnittet *3.3.7 Macro och micro värden.* Men eftersom måtten bygger på de som förklaras i det här och i de kommande 3 delarna är det av värde att de förklaras i dess enklaste form till en början.

3.3.4 Precision

Precision visar förhållandet mellan antalet rätt predikterade positiva observationer och den totala mängden predikterade positiva observationer. Syftet med att beräkna precisionen är att kunna besvara frågan om hur många som predikterats som positiva faktiskt tillhör den klassen (Webb, 2010). Högre precision innebär att antalet falskt positiva minskar: modellen har presterat bra.

$$Precision = \frac{SP}{SP + FP} \tag{5}$$

3.3.5 Känslighet

På engelska används termen recall som också kan associeras med ordet *sensitivity* (känslighet). Känsligheten för en modell visar relationen mellan predikterade positiva observationer och det faktiska antalet observationer som tillhör den klassen. Måttet visar på

hur stor del av de observationer som tillhör den specifika klassen som faktiskt blev rätt predikterade (Webb, 2010).

$$Känslighet = \frac{SP}{SP + FN} \tag{6}$$

3.3.6 F-score

Det sista måttet som används för att utvärdera en klassificeringsmodell i den här studien är f-score. F-score är en slags viktad sammansättning av de två tidigare måtten precision och känslighet. Det innebär att måttet både tar hänsyn till modellens precision samtidigt som antalet rätt predikterade observationer också tas i beaktning (Shung, 2018). Även för det här måttet innebär det att en högre siffra ger en bättre modell. Värdet på f-score sträcker sig mellan 0 och 1.

$$F - score = \frac{2 * (Precision * Känslighet)}{Precision + Känslighet} \tag{7}$$

Alla de ovan presenterade måtten för att utvärdera en klassificeringsmodell används på data som utgör *två* klasser. När problemet innehåller flera klasser (multiklass-baserad klassificering) går det att använda så kallade *macro* och *micro* som beräknar ett medelvärde på f-score.

3.3.7 Macro och micro medelvärden

Vid multiklass-baserad klassificering går det med hjälp av macro och micro medelvärden beräkna samma mått (f-score) fast för flera klasser. Ett macro medelvärde baseras på resultatet från varje klass utan att ta hänsyn till hur många datapunkter som varje klass innehåller. För micro medelvärden beräknas istället den totala summan av SP, FP, SN och FN. Det innebär att om klasserna innehåller olika antal datapunkter kan ett macro medelvärde ge ett svar som är missledande. Men inom problem där klasserna är balanserade är ett macro medelvärde ett bra mått på f-score (Schmueli, 2019).

Enligt ekvation 7 ovan, där f-score beräknas, behövs precision och känslighet för att beräkna värdet på f-score. Därför presenteras de ekvationer för att beräkna macro och micro

medelvärden för precision och känslighet nedan:

$$Macro\ precision = \frac{1}{k}\sum_{i=1}^{k}\frac{SP_i}{SP_i + FN_i} \tag{8}$$

$$Micro\ precision = \frac{\sum_{i=1}^{k} SP_i}{\sum_{i=1}^{k} SP_i \sum_{i=1}^{k} FN_i} \tag{9}$$

$$Macro\ känslighet = \frac{1}{k}\sum_{i=1}^{k}\frac{SP_i}{SP_i + FP_i} \tag{10}$$

$$Micro\ känslighet = \frac{\sum_{i=1}^{k} SP_i}{\sum_{i=1}^{k} SP_i \sum_{i=1}^{k} FP_i} \tag{11}$$

De olika beräkningarna för precision och känslighet kan i sin tur användas i ekvation 7 på föregående sida för att beräkna ett macro eller micro f-score.

4 Data

I det här avsnittet presenteras den data som använts i det här arbetet. Bland annat förklaras hur all data bearbetats, vilka problem som uppstått och hur det har hanterats samt vilken data som slutligen valdes till den här studien.

Den data som ligger till grund för det här arbetet är historisk och baseras på de trafikmätningar Göteborg och Stockholms kommun har utfört. Den data som används från Göteborg baseras på trafikmätningar utförda mellan åren 1987-2015 och den data från Stockholm baseras på trafikmätningar utförda mellan åren 2005-2018. Antalet trafikmätningar som har utförts på varje mätplats i Göteborgs och Stockholms kommun under ett givet år skiljer sig från mätplats till mätplats. På samma sätt skiljer det sig när på året mätningarna är utförda. Majoriteten av mätplatserna innehåller data som knappt täcker en månad, i bästa fall sträcker sig mätningarna över två månader. De flesta mätningarna är utförda under 1-3 dagar under året.

Mätningar som laddats upp i trafikmätningssystemet lagras i en SQL-databas, vilken har gjort det möjligt att generera de dataset som använts i modellerna för det här arbetet. Varje mätning som är registrerad i databasen har ett unikt mätplats id som kopplar mätningen till den plats mätningen utförts på. För varje mätplats id innehåller databasen datum för mätningen, trafikmängd, fordonshastighet, andel tung trafik och annan information som samlats in på mätplatsen. I listan nedan finns alla de attribut som finns lagrat till varje mätplats (och som är relevanta för den här studien) med tillhörande beskrivning om vad de innehåller. Attributet "AVGSPDPERCENTILE85" står för 85 percentilen av hastigheten på en gata och innebär att 85% av bilarna har en hastighet lika med eller lägre än angiven hastighet (Göteborgs Stad, 2019). Hädanefter benämns det här attribut som *hastighet*.

- **MATPLATS_ID** - Ett id kopplat till den mätplats där mätningen utfördes
- **TIMESTAMP** - Datum för när mätningen utfördes
- **FLOW_PROFILE_ID** - En främmande nyckel till flödesprofil (gatutyp)
- **TRAFFIC0001** - Antal observerade fordon mellan 24:00 - 01:00

  .
  .
  .

- **TRAFFIC2324** - Antal observerade fordon mellan 23:00 - 24:00
- **AVGSPDPERCENTILE85** - Uppmätt hastighet under dagen
- **HEAVYTRAFFICRATIO** - Andel observerad tung trafik under dagen

Som tidigare nämnt, valdes initialt gatutyperna: industrigata, industrimatargata, bostadsgata, bostadsmatargata, centrumgata och centrummatargata. Men efter utvärdering av mängden tillgänglig data togs industrigata och bostadsgata bort. Antalet trafikmätningar som tillhörde de två gatutyperna skiljer sig med mer än en tiopotens än resterande gatutyper. Eftersom mängden mätningar var så pass mycket mindre valdes endast resterande fyra gatutyper för analys. I tabell 3 visas antalet mätningar för varje gatutyp samt hur stor procent av mätningarna som innehöll värden för attributen hastighet och andel tung trafik. I den efterföljande tabellen, tabell 4, presenteras det totala medelvärdet av trafikmängd, hastighet och andel tung trafik per dygn från de mätningar som gjorts för vardera gatutyp.

Tabell 3. Antalet mätningar per gatutyp i Göteborg med hastighet och tung trafik.

Gatutyp	**Antal mätningar**	**Andel med hastighet**	**Andel med tung trafik**
Bostadsmatargata	2375	30%	30%
Centrumgata	5461	1.7%	1.7%
Centrummatargata	3913	2.3%	2.3%
Industrimatargata	5478	5.6%	5.6%

Tabell 4. Medelvärden för trafikmängd, hastighet och andel tung trafik för respektive gatutyp i Göteborg.

Gatutyp	**Medelvärde trafikmängd (antal fordon)**	**Medelhastighet (km/h)**	**Medelvärde andel tung trafik (andel tunga fordon)**
Bostadsmatargata	3078	43	6.9%
Centrumgata	5333	38	6.4%
Centrummatargata	12745	33	9.1%
Industrimatargata	5715	51	12.7%

5 Metod

I det här avsnittet presenteras de verktyg som användes för att möjliggöra projektet. Verktygen utgörs av programmeringsspråk, utvecklingsmiljöer, bibliotek och program för hantering av geografiska informationssystem. Utöver de verktyg som användes i den här studien presenteras även implementeringen av maskininlärningsmodellerna, hur modellerna utvärderades samt vilka potentiella felkällor som kan påverka resultaten. Den avslutande delen av det här avsnittet handlar om hur all data bearbetats i form av datautvinning, bearbetning av data, val av input till modellerna samt datakvalitet.

5.1 Verktyg

I den här studien har den största delen kod skrivits i Python med hjälp av Jupyter Notebook med ett flertal befintliga bibliotek i Python. För implementation av maskininlärningsmodellerna användes biblioteket Scikit-learn och för datahantering användes bland annat Pandas. All kommunikation med databasen har gjorts med PostgreSQL. Slutligen har datorprogrammet QGIS använts för hantering av geografiska informationssystem för att skapa kartor.

5.1.1 Python och Jupyter Notebook

Python är ett objektorienterat högpresterande programmeringsspråk som med dynamisk maskinskrivning kombinerat med en välutvecklad inbyggd datastruktur låter programmerare utforma komplicerade program med enkelhet. Det är ett populärt språk som har en syntax som är lätt att hantera samtidigt som språket stödjer en mängd bibliotek (Python, 2020). All kod som har skrivits för att hantera data och testa maskininlärningsmodellerna har skrivits i Python med undantag för kommunikationen med och hämtning av data från databasen.

För att smidigt konstruera det program som använts i den här studien har webbapplikationen Jupyter Notebook använts för att lättare köra, skriva och testa kod. Jupyter Notebook är skriven med öppen källkod och används främst för att skapa och dela dokument som innehåller kod som skrivs i realtid, användning av ekvationer och även för visualisering av data. Webbapplikationen stödjer över 40 olika programmeringsspråk, däribland Python (Jupyter, 2019).

Genom att använda Jupyter Notebook går det att enkelt köra delar av koden på nytt utan att påverka resterande outputs som blivit exekverade inom andra sektioner. Det medför både en tydlig bild av vilka delar i programmet som gör vad samtidigt som beräkningar

kan göras på ett smidigt sätt och därmed minska tiden för test och körningar (Jupyter, 2019).

5.1.2 Pandas

Python biblioteket Pandas (Python Data Analysis Library) är ett bibliotek som är uppbyggt av datastrukturer som är anpassade för enkel dataanalys. Syftet med biblioteket är att möjliggöra bättre analys och modellering av data (Pandas, 2019).

5.1.3 Scikit-learn

Ett bibliotek i Python som används för maskininlärning är scikit-learn. Biblioteket stödjer modeller inom klassificering, regression och klustring. Scikit-learn bygger på andra välkända bibliotek i Python som NumPy, SciPy och matplotlib. I det här projektet har modellerna k-NN, random forests och k-korsvalidering implementerats med hjälp av scikit-learn (Scikit-learn, 2020).

5.1.4 PostgreSQL

Kommunikationen med den databas som innehåller all data för det här projektet gjordes i PostgreSQL. Programmet är skriven med öppen källkod och följer den allmänna SQL-standarden i hög grad. Systemet för databasen är objekt- och relationsbaserad och erbjuder användaren att lagra olika procedurer eller triggers samtidigt som den ger stöd åt funktioner skapade av användaren i flera olika programmeringsspråk (PostgreSQL, 2019).

5.1.5 QGIS

För att hantera geografisk visualisering och placering av datapunkter används ofta geografiska informationssystem, förkortat GIS. I det här projektet används QGIS som genom ett grafiskt användargränssnitt möjliggör visualisering och utplacering av geodata på kartor. I programmet finns möjligheter att hämta upp data från databaser eller från filer som är geografiskt kodade (QGIS, 2019).

5.2 Implementering av maskininlärningsmodeller

Med hjälp av biblioteket scikit-learn har modellerna som testats i det här arbetet implementerats. Biblioteket innehåller modellerna k-NN, random forests och korsvalidering vilket möjliggjort en enkel implementering av modellerna och ingen modell har således implementerats på egen hand. En fördel med att använda befintliga modeller i biblioteket för

maskininlärning är att de har blivit grundligt testade och minskar således potentiella felkällor. Till exempel kan fel i matematiska beräkningar leda till missvisande resultat.

5.3 Utvärdering av maskininlärningsmodeller

I det här projektet jämförs två olika modeller inom maskininlärning för att uppnå det optimala resultatet vid klassificering av gatutyper. För att gör en jämförelse mellan modellerna användes korsvalidering som till en början användes för att göra en uppskattning av modellernas prestation. När korsvalidering implementeras presenteras resultatet i form av ett valideringsfel. Det är utifrån de erhållna valideringsfelen som uppskattningar gällande modellernas prestationer har gjorts.

Processen har följt en iterativ struktur utifrån det huvudsakliga syfte som projektet har. Varje nytt test följde av de hypoteser som nämndes i *1.1 Syfte*, vilka ställdes upp utifrån vilka attribut som skulle testas och varför. Vid varje nytt test där ett attribut lades till eller togs bort utvärderades modellernas prestation utifrån valideringsfelet. Av testen kunde således både attributens relevans och modellernas applicerbarhet på problemet värderas och analyseras. Måttet på valideringsfelet värderas med avseende på om värdet förbättras eller försämras.

5.4 Val av attribut

Relevant data och val av attribut är viktiga aspekter som ligger till grund för hur väl en modell kommer att prestera. Det finns flera saker att ta i beaktning vid val av attribut som i sin tur leder till vilka möjligheter som finns för att modellen tränas på rätt data. En avgörande del vid val av attribut ligger vid expertkunskap och det sammanhang som problemet är begränsat inom (Aryes, 2007). För att ta bra beslut är det därför viktigt att så mycket kunskap som möjligt har samlats in om problemets natur och vad de olika attributen innehåller. I det här arbetet valdes attributen utifrån tillgänglighet i databasen där en intuitiv utvärdering gjordes genom att varje attribut analyserades utifrån innehåll och relevans för problemet. Resultatet blev de attribut som presenterades i avsnitt *4. Data*.

5.5 Datautvinning

Den data som används för träning av maskininlärningsmodellerna är den som är insamlad i Göteborg. Eftersom den data som finns i Stockholm skiljer sig från träningsdatan har utgångspunkten för datautvinning och bearbetning varit Göteborgs data. För att i

slutändan kunna applicera modellerna på trafikdata från Stockholm där samma bearbetning upprepades som på Göteborgs data efter att alla beslut fastställdes.

Som presenterat i föregående kapitel, *4. Data*, innehåller databasen ett antal olika gatutyper som anger vilka gatutyper som finns i Göteborgs kommun. Det första steget i att hämta och ta ut data var att filtrera på gatutyperna som representerar "Centrumgata", "Bostadsmatargata", "Industrimatargata" och "Centrummatargata". Till varje gatutyp följer ett dataset som innehåller x antal trafikmätningar med tillhörande attribut, som finns presenterade i tabell 4, avsnitt *4. Data*. I databasen finns ytterligare attribut än de presenterade men med avseende på syftet är det enbart dessa som valts. De fyra dataset som filtrerades ut (med avseende på gatutyp) lagrades i separata CSV-filer för att möjliggöra en enkel hantering av data i Python, Jupyter Notebook. Det är enbart mätningar som innehåller trafikflöden som ingår i det här arbetet.

5.6 Begränsningar i data

Det finns begränsningar i den data som finns tillgänglig i den här studien, både i det dataset som kommer från Göteborg och i den data som finns i Stockholm. Det handlar delvis om dubbletter, avsaknad av värden och begränsad mängd data. För att säkerställa att all data som används vid träning av modellerna innehåller fullständiga mätningar har relevanta åtgärder vidtagits. Vissa av de åtgärder som varit nödvändiga leder bland annat till en minskad datamängd, generella värden eller ofullständig bild av verkligheten. Det som ligger till grund för vilka åtgärder som varit nödvändiga beror till största del på Trafikkontorets egna noteringar. Det handlar bland annat om noteringar om att data är bristande eller inte täcker hela dagen. Utöver Trafikkontorets noteringar har ingen större analys av data gjorts för att identifiera fall som kan ha missats. Hur datakvaliten därefter tas hänsyn till presenteras i slutet på det här avsnittet. Nedan följer förklaringar kring de åtgärder som gjordes samt vilka metoder som användes vid bearbetning av data.

5.7 Bearbetning av data

Syftet med att bearbeta data i den här studien är att filtrera datan för att öka datakvalitén samt att fylla ut bristfällig data för att öka mängden data som modellerna kan tränas på. En viktig aspekt i den här processen är att kunna identifiera vilka möjligheter och begränsningar som finns i den data som finns tillgänglig.

5.7.1 Ofullständig mätning

Mätningar där uppräkningen av passerande fordon av någon anledning inte registrerats under ett mätintervall blir i databasen markerade som en "ofullständig mätning". Det kan bero på ett tekniskt fel i mätutrustningen eller att data faller bort vid dataöverföringen till databasen. Något annat som markerar en mätning som ofullständig är om mätningen förefaller på den första eller sista dagen under mätintervallet. Anledningen är att den första och sista dagen sällan har registrerad mätdata under hela det dygnet eftersom det beror på när mätningen påbörjats och avslutats. Alla mätningar som är markerade som ofullständiga filtreras därför bort i det här arbetet.

5.7.2 Flödesriktning

De flesta mätningar baseras på trafikdata från två körfält vilket innebär att mätningarna sker i båda riktningarna. Eftersom syftet med det här arbetet är att studera den totala trafiken för varje gatutyp beräknas den totala mängden trafik i båda riktningarna för sådana mätningar. Mätningar med enbart ett körfält behålls oförändrad.

5.7.3 Anmärkningar

Vissa mätningar har en anmärkning tillsammans med en förklaring som anger om mätningen påverkats på något sätt. Det kan handla om en yttre påverkan på mätningen så som trafikarbete, trafikomläggning eller annan störning. Om mätningen registreras som "fullständig" rapporteras övrig information om den geografiska platsen. Alla de mätningar som påverkats på något sätt och således inte markerats som fullständiga har filtrerats bort.

5.7.4 Veckodag

I det här arbetet är det endast mätningar utförda på veckodagar som tagits med. Anledningen, som presenterades i avsnittet *2. Bakgrund*, är att variationer i trafiken skiljer sig mellan veckodagar och helgdagar där helgdagar inte följer samma regelbundenhet.

5.7.5 Kalenderhändelser

För att säkerhetsställa att valda mätningar med avseende på veckodag inte förefaller på röda dagar eller afton-dagar (svenska helgdagar eller högtider) skapades en tabell med alla kalenderhändelser som förefallit under alla de berörda åren. Utifrån dessa datum filtreras eventuella veckodagar som inte är vanliga arbetsdagar bort.

5.7.6 Säsong

Den säsong som visade på störst skillnad i trafiken i jämförelse med resterande år var sommarperioden juni-augusti. Mätningar utförda under dessa månader togs därmed bort.

5.7.7 Standardisera data

En nödvändig del i processen att bearbeta den trafikdata som användes i den här studien var att standardisera trafikmängden. Trafikmängd är räknad timvis med absoluta tal för respektive trafikmätning. Eftersom Stockholm är större som stad och har mer invånare än Göteborg kan mängden trafik på gator i Stockholm antas vara större. Därför har trafikmängden för varje timme standardiserats så att varje timme istället utgör en andel av den totala trafiken. Den totala trafiken är uppskattad från varje mätplats under ett givet år som ett ÅMVD värde (se kapitel *2. Bakgrund*). Beräkningen har gjorts genom att summera den totala mängden trafik för varje mätplats och för varje mättillfälle under ett givet år dividerat med antalet mätningar under det givna året. Det uppskattade ÅMVD värdet har då använts för att dividera varje timvisa mätning tillhörande mätplatsen för det givna året. För att exemplifiera: om tre trafikmätningar utförts på mätplats nr 4 under 2015 och den totala trafiken för varje mättillfälle var [1930, 2100, 2003] antal passerade fordon. Den totala mängden fordon för mätningarna blir 6033 som i sin tur divideras med 3 (antalet trafikmätningar). Resultatet blir ett ÅMVD för mätplats nr 4, vilket i det här fallet blir 2011. Om trafikmätningar utförs på mätplats nr 4 ett annat år beräknas ÅMVD på nytt. Om första mättillfället på mätplats nr 4 under 2015 hade 18 passerande fordon under första timmen under dygnet (24:00-01:00) standardiseras timmen till $18/2011 \approx 0.00895$, alltså 0.895% av trafiken passerade under dygnets första timme. Det här gör att trafiken representeras i andelar istället för absoluta tal.

5.7.8 Hantering av avsaknad data

Om mätutrustningen för en given mätplats inte haft täckning för att mäta fordonshastighet eller tung trafik tilldelas ett null-värde i databasen för den information som saknas. Saknas information om fordonshastighet innebär det även att informationen för tung trafik saknas. Anledningen är att om en mätutrustning inte har täckning att registrera mätningarna faller all potentiell trafikinformation bort, och således kan aldrig en av attributen förekomma ensamt. Andra anledningar till utesluten registrering är bland annat att slangar som mäter hastighet och andel tung trafik har lossnat eller att mätutrustningen är av äldre slag och därmed inte kan mäta hastighet alls. För att kompensera för den typen av avsaknad i data har följande metod gjorts.

Vid ett null-värde på de två attributen fordonshastighet och andel tung trafik i x antal trafikmätningar i en specifik klass (gatutyp) togs det minsta och största värde fram utifrån resterande trafikmätningar (som hade värden på de två attributen). De minsta och största värden verkade som gränserna för ett intervall som förklarar hur värdena inom klassen kan variera. För varje trafikmätning som saknade värden på attributen slumpades därefter ett värde inom intervallet och mätningen tilldelades således ett nytt värde på fordonshastighet och andel tung trafik var för sig. Anledningen till att använda slumpmässiga värden inom intervallet är att minska risken för att skapa oavsiktliga trender i data.

5.7.9 Skalning av data

För modellen k-NN är det viktigt att den data som tränas och testas på inte har för stor variation i skala sinsemellan. Eftersom modellen använder majoritets votering med avseende på närliggande datapunkter är den känslig för om distanser mellan attribut skiljer sig för mycket. I det här arbetet används euklidiskt avstånd för att beräkna närmsta datapunkter vilket innebär att den blir känslig för variationer i avstånd. Ett sätt att hantera det på och därmed undvika att särskilda attribut tappar värde är att skala om data för att minimera avstånden mellan olika attribut. I det här arbetet har attributen tung trafik och hastighet skalats om för att vara på decimalform och därmed vara jämförbara med attributet trafikmängd. Till exempel kan ett värde på hastighet vara “85 km/h” och genom att dividera alla värden på hastighet med 100 erhålls istället värdet “0.85” för samma mätning. Genom att göra den skalningen kan värdet nu jämföras med attribut som är små. Samma typ av skalning har gjorts på attributet andel tung trafik.

Skalning av data har enbart gjorts för modellen k-NN eftersom modellen random forest inte är känslig för den typen av ojämnheter.

5.8 Datakvalitet

Av de metodbeskrivningar som presenterats för bearbetning av data i föregående avsnitt framgår vilka begränsningar som finns i tillgänglig data och hur det har hanterats. Bland annat finns mätningar som tagits bort på grund av ofullständighet, anmärkningar och andra omständigheter som gör att mätningen inte kan tas med vid analys. Åtgärder som rör standardisering och hantering av avsaknade värden har vidtagits för att kompensera för dataförlusten. Allt det gör att mängden data har minskat och förändrats för att passa den problemformulering som den här studien utgör. Vid diskussion och analys av resultat är det därför viktigt att ta det i beaktning så att resultaten kan tolkas på ett så tillförlitligt sätt som möjligt. Trots att bearbetning av data förändrat den trafikdata som ligger till grund för den här studien har det varit nödvändigt och av värde att hantera och bearbeta

datan utifrån rådande förutsättningar. All data som används i den här studien presenterades i föregående avsnitt, *4. Data*, där antalet mätpunkter presenterades för vardera gatutyp.

5.9 Felkällor

I den trafikdata som används i den här studien finns det ett antal olika potentiella felkällor. Utöver de begränsningar som finns i tillgänglig data som är beskrivet ovan, kan fel uppstå på grund av systematiska eller slumpmässiga fel. När trafikdata registreras finns det utrymme för fel som beror av mänsklig karaktär: till exempel att fel data har registrerats till fel mätplats. Att en person har gjort en felregistrering av trafikdata som innebär att data kopplas till fel plats räknas det till slumpmässiga fel. Ibland uppstår andra slumpmässiga fel som inom det här området kan uppstå på grund av ombyggnationer av vägar, väderförhållanden och andra störningar som kan påverka trafiken. Ett exempel på fel som ingår under systematiska fel kan vara fel som beror på mätfel som uppstår på grund av den mätutrustning som registrerar trafikdata eller att en operatör har kopplat de slangar som mäter trafikdata till fel kontakter.

Andra felkällor som kan förekomma i det här projektet är fel som uppstår i och med egen implementation av matematiska beräkningar eller annan hantering av data från databasen. Om en uträkning, som är skriven genom kod i programmet, innehåller fel leder det till ett missvisande värde eller resultat. Till detta hör även fel som kan uppstå vid hämtning och hantering av data. Exempel på det kan vara hur data som saknar något av attributen hastighet och andel tung trafik och har tilldelats ett slumpmässigt värde som inte återspeglar verkligheten.

Den sista felkälla som nämns i det här avsnittet handlar om det tillvägagångssätt av att standardisera data som valts. Eftersom det finns begränsningar i de dataset som finns tillgängliga har det varit nödvändigt att standardisera data och om de värden på ÅMVD som används vid standardiseringen inte är representativt kan det leda till missvisande resultat.

6 Resultat

I det här avsnittet presenteras resultaten från den här studien. I de tre första delarna presenteras resultaten från de fallstudier som utförts på trafikdata från Göteborg. Den första modellen som testas är k-NN och därefter appliceras modellen random forest. Modellerna tränas och testas först på de enskilda variablerna var för sig och sedan görs samma tester på de olika kombinationerna ({trafikmängd, hastighet}, {trafikmängd, andel tung trafik}, {trafikmängd, hastighet, andel tung trafik}). Innan modellerna appliceras och utvärderas på trafikdatan väljs modellparametrarna med hjälp av korsvalidering. Tillsammans med resultaten från de olika kombinationerna presenteras även hur varje kombination klassificerat och presterat för varje gatutyp. Efter att de båda modellerna har undersökts på data från Göteborg illustreras även resultatet av den modell som presterar bäst i form av en confusion matris. Slutligen appliceras den valda modellen på trafikdata från Stockholm. I den sista delen i det här avsnittet följer en sammanfattning av de mest relevanta resultaten.

6.1 K-NN

För att välja rätt antal k, antal närmsta datapunkter som optimerar modellen, används 10-faldig korsvalidering, vilket innebär att modellen testas 10 gånger iterativt. Vid varje körning delas all data upp i två delar, en bestående av testdata och den andra av träningsdata som genereras slumpmässigt. Förhållandet är 20/80 vilket innebär att vid varje körning agerar 20% av datasetet testdata och resterande är träningsdata. Korsvalideringen gjordes för alla enskilda variabler samt för varje möjlig kombination av variabler och av resultatet kunde k = 3 väljas som det optimala antalet närmsta datapunkter för modellen. Det intervall för värdet på k som testades var 1-50 med steglängd 1. Övriga parametrar som modellen ställdes in med var utifrån de standardinställningar som finns i Scikit-learn för k-NN.

6.1.1 Enskilda variabler

I tabell 5 på nästa sida presenteras noggrannheten tillsammans med den viktade sammansättningen av precision och känslighet för varje enskild variabel. Av resultatet går det att se att variabeln trafikmängd resulterade i högst noggrannhet samt f-score vilket innebär att modellen presterade bäst för den variabeln.

Tabell 5. Resultat för de enskilda variablerna med k-NN.

Variabel	**Noggrannhet**	**F-score**
trafikmängd	89.6%	89.6%
hastighet	86.8%	86.9%
andel tung trafik	86.5%	86.6%

6.1.2 Kombinerade variabler

På samma sätt som de enskilda variablerna undersöktes, undersöktes och jämfördes de kombinerade variablerna. Av resultatet i tabell 6 nedan går det att se att modellen presterar bäst med kombinationen trafikflöde och andel tung trafik. Den kombination som resulterar i lägst noggrannhet är när alla tre variabler ingår i modellen.

Tabell 6. Resultatet för de tre kombinationerna med k-NN.

Kombination	**Noggrannhet**	**F-score**
trafikmängd, hastighet	87.2%	87.2%
trafikmängd, andel tung trafik	88.2%	88.2%
trafikmängd, hastighet, andel tung trafik	86.8%	86.9%

Eftersom de tre kombinationerna uppnådde liknande noggrannhet presenteras, i kommande tre tabeller, hur väl varje modell predikterade de fyra klasserna. Den första tabellen, tabell 7, visar hur den första kombinationen klassificerade de olika gatutyperna. För centrummatargata och bostadsmatargata kunde modellen inte klassificera gatorna lika väl som för industrimatargata och centrummatargata. Högst f-score erhölls för industrimatargata med 95.1% vilket innebär att modellen kunde identifiera datapunkter som tillhörde den klassen bäst. I tabell 8 presenteras resultaten från den andra kombinationen: trafikmängd och andel tung trafik. I jämförelse med den första kombinationen presterade modellen bättre för alla gatutyperna vilket förklarar den höga noggrannhet som presenterades i tabell 6 ovan. Likt den första kombinationen hade modellen svårighet att identifiera datapunkter som tillhör gatutyperna centrummatargata och bostadsmatargata och högst f-score erhölls av industrimatargata. Resultaten från den sista kombinationen, som bestod av alla tre variabler, se tabell 9 på nästa sida, skiljer sig från de tidigare två med avseende på gatutypen centrummatargata, där f-score blev 77.9%.

Tabell 7. Resultatet för kombination 1 för de olika gatutyperna.

Klass	**Precision**	**Känslighet**	**F-score**	**Antal datapunkter**
Centrummatargata	77.8%	84.6%	81.1%	764
Bostadsmatargata	87.8%	73.6%	80.0%	488
Industrimatargata	96.1%	94.1%	95.1%	1139
Centrumgata	85.2%	87.9%	86.5%	1138

Tabell 8. Resultatet för kombination 2 för de olika gatutyperna.

Klass	**Precision**	**Känslighet**	**F-score**	**Antal datapunkter**
Centrummatargata	79.7%	84.6%	82.1%	792
Bostadsmatargata	88.2%	78.7%	83.2%	493
Industrimatargata	97.8%	94.8%	96.3%	1171
Centrumgata	84.7%	87.9%	85.3%	1072

Tabell 9. Resultatet för kombination 3 för de olika gatutyperna.

Klass	**Precision**	**Känslighet**	**F-score**	**Antal datapunkter**
Centrummatargata	74.4%	81.7%	77.9%	785
Bostadsmatargata	91.8%	75.6%	82.9%	517
Industrimatargata	97.5%	95.2%	96.3%	1170
Centrumgata	83.3%	86.6%	85.0%	1056

6.2 Random forest

På samma sätt som det optimala antalet k valdes i k-NN, valdes det optimala antalet träd i random forest. Intervallet på antal träd som testades var 50-200 med steglängden 1. Det optimala antalet träd som erhölls för de enskilda och kombinerade variablerna utifrån korsvalidering var 100 träd. De övriga parametrar som användes i modellen var de som är förinställda i Scikit-learn för random forest.

6.2.1 Enskilda variabler

I tabell 10 på nästa sida presenteras de erhållna noggrannheterna och de viktade sammansättningarna av precision och känslighet för de enskilda variablerna som undersöktes med random forest. Av resultatet går det att se att variabeln trafikflöde resulterade i högst noggrannhet och f-score vilket innebär att modellen presterade bäst för den variabeln. Både hastighet och andel tung trafik resulterade i låga resultat där hastighet fick lägst

värden.

Tabell 10. Resultatet för de enskilda variablerna med random forest.

Variabel	**Noggrannhet**	**F-score**
trafikmängd	91.0%	90.9%
hastighet	55.0%	47.7%
andel tung trafik	61.7%	58.6%

6.2.2 Kombinerade variabler

Resultaten från de kombinerade variablerna presenteras i tabell 11 nedan. För alla tre kombinationer erhålls en hög noggrannhet tillsammans med ett högt f-score. Det innebär att alla tre kombinationerna erhöll goda klassificeringar av gatutyperna. Bäst presterade modellen på kombinationen som innehåller alla tre variabler, där en noggrannhet samt f-score blev 93.1%. En närmare inblick i hur väl de olika kombinationerna presterade på vardera gatutyp presenteras i de tre efterföljande tabellerna (12-14).

Tabell 11. Resultatet för de tre kombinationerna med random forest.

Kombination	**Noggrannhet**	**F-score**
trafikmängd, hastighet	91.6%	91.6%
trafikmängd, andel tung trafik	92.2%	92.2%
trafikmängd, hastighet, andel tung trafik	93.1%	93.1%

För kombinationen trafikmängd och hastighet presterade modellen bäst för gatutypen industrimatargata, som erhöll ett f-score på 96.2%, se tabell 12. Även gatutyperna centrummatargata och centrumgata erhöll höga värden på f-score. Sämst predikterade modellen på bostadsmatargata, med ett f-score på 86.2%. I tabell 13 på nästa sida presenteras resultaten från den andra kombinationen: trafikflöde och andel tung trafik. För gatutyperna bostadsmatargata, industrimatargata och centrumgata har värdet på f-score förbättrats i jämförelse med den första kombinationen. Den enda klassificeringen som försämrats är för cemtrummatargata, men även denna prediktion erhöll ett högt värde i sammanhanget. I den sista tabellen, tabell 14, presenteras resultaten för den kombinationen som innehåller alla tre variabler. För bostadsmatargata erhålls samma värde på f-score som för tidigare kombination. Gatutyperna cemtrummatargata och centrumgata får en procentuellt bättre andel rätt klassificerade än vid de andra två kombinationerna, en ökning på nästan två procentenheter. Industrimatargata uppnår ett högt f-score, likt vid de andra två kombinationerna, med ett värde som närmar sig 100%.

Tabell 12. Resultatet för kombination 1 för de olika gatutyperna.

Klass	**Precision**	**Känslighet**	**F-score**	**Antal datapunkter**
Centrummatargata	94.1%	85.7%	89.7%	840
Bostadsmatargata	85.8%	86.5%	86.2%	482
Industrimatargata	97.2%	95.3%	96.2%	1148
Centrumgata	86.9%	94.6%	90.6%	1055

Tabell 13. Resultatet för kombination 2 för de olika gatutyperna.

Klass	**Precision**	**Känslighet**	**F-score**	**Antal datapunkter**
Centrummatargata	92.4%	85.7%	88.9%	768
Bostadsmatargata	89.4%	87.6%	88.5%	474
Industrimatargata	98.0%	96.4%	97.2%	1182
Centrumgata	87.4%	94.3%	90.7%	1100

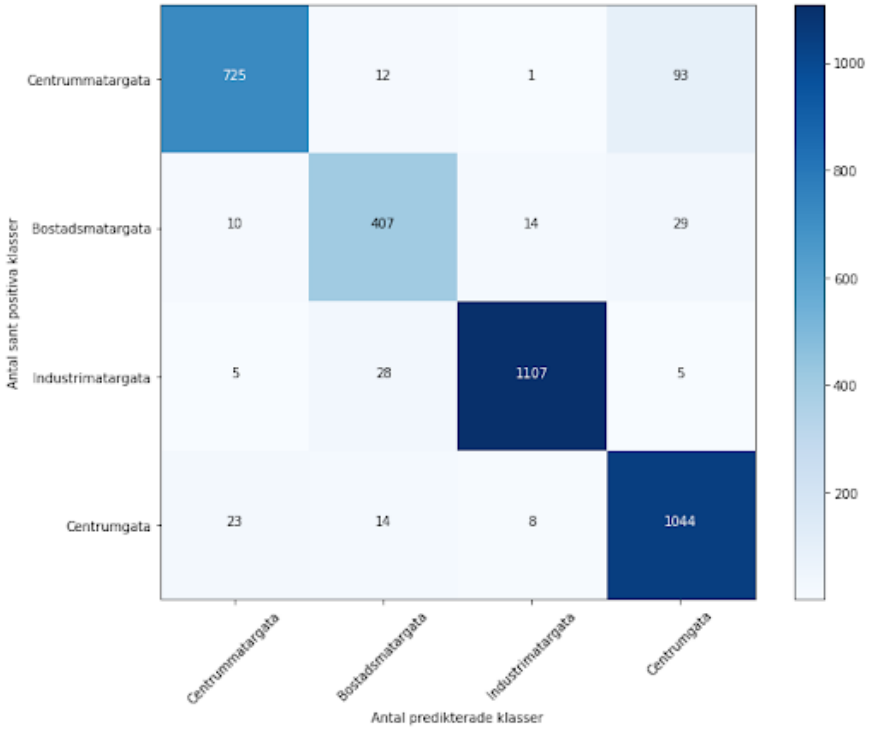
Tabell 14. Resultatet för kombination 3 för de olika gatutyperna.

Klass	**Precision**	**Känslighet**	**F-score**	**Antal datapunkter**
Centrummatargata	95.0%	87.2%	91.0%	831
Bostadsmatargata	88.3%	88.5%	88.4%	460
Industrimatargata	98.0%	96.7%	97.3%	1145
Centrumgata	89.2%	95.9%	92.4%	1089

6.3 Val av modell

Baserat på resultaten från de två undersökta modellerna för de sex hypoteserna som presenterats i tabellerna 5-14. kunde en slutgiltig modell väljas. Valet baseras på hur väl modellen predikterat de fyra klasserna utifrån de undersökta variablerna. Resultatet visar att sammansättningen av alla variabler, alltså trafikmängd, hastighet och andel tung trafik ger den bästa kombinationen med random forest. I figur 7 på nästa sida presenteras en confusion matris för att visualisera hur den valda modellen med den valda kombinationen predikterat de olika klasserna. Matrisen visar hur många datapunkter per gatutyp som klassificerades till de olika klasserna där diagonalen utgör antalet rätt klassificerade mätplatser per gatutyp. Exempelvis, om vi studerar resultatet för gatutypen centrummatargata, som utgör den första raden i matrisen, går det att utläsa att av totalt 831 mätplatser predikteras 725 som den klassen. Det innebär att av de 831 mätplatser som egentligen var centrummatargator, predikterades 725 rätt. För resterande mätplatser tillhörande centrummatargata klassificerades 12 som bostadsmatargata, 1 som industrimatargata och 93

som centrumgata. Modellen klassificerade således 93 av de 106 felklassificerade mätplatserna som centrumgata, vilken i vägtrafiksystemet får trafik från centrummatargata och indikerar således på ett bra resultat.

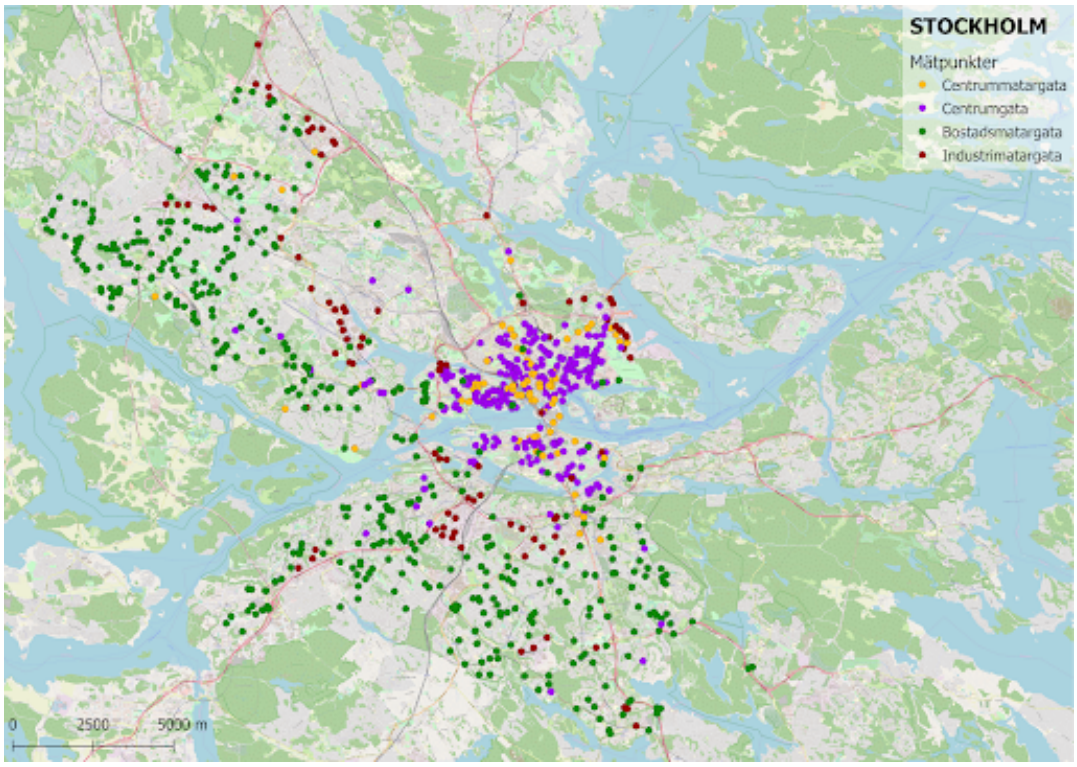


Figur 7: Confusion matris för random forest.

6.4 Applicera modellen på Stockholms trafikdata

Efter att ha valt den slutgiltiga modellen användes modellen sedan på den mätdata från de mätplatser i Stockholm som presenterades i bakgrundsavsnittet. Eftersom att arbetet med att säkerhetsställa de predikteringar som modellen genererar på Stockholms mätdata inte ingår i syftet för den här studien, så kommer resultatet istället bygga på ett antal observationer som gjorts. I ett första försök att prediktera mätplatserna med modellen extraherades och förbereddes data från databasen på samma sätt som för Göteborgs data, vilket är beskrivet i *5.7 Bearbetning av data*. Antal mätplatser som ingick i det första försöket var 1231 stycken med totalt 9383 mätningar. Resultatet av modellen gav 780 mätplatser med en entydig prediktering av någon av gatutyperna och 451 mätplatser med fler än en prediktering. Databasen i Stockholm, till skillnad från Göteborg, har en funktion som bevakar kvaliteten på varje mätning genom att jämföra det observerade antalet fordon mot ett förväntat värde. Det förväntade värdet är uppskattat från de tidigare mätningarna som gjorts på den givna mätplatsen. Överstiger det observerade värdet en godkänd nivå

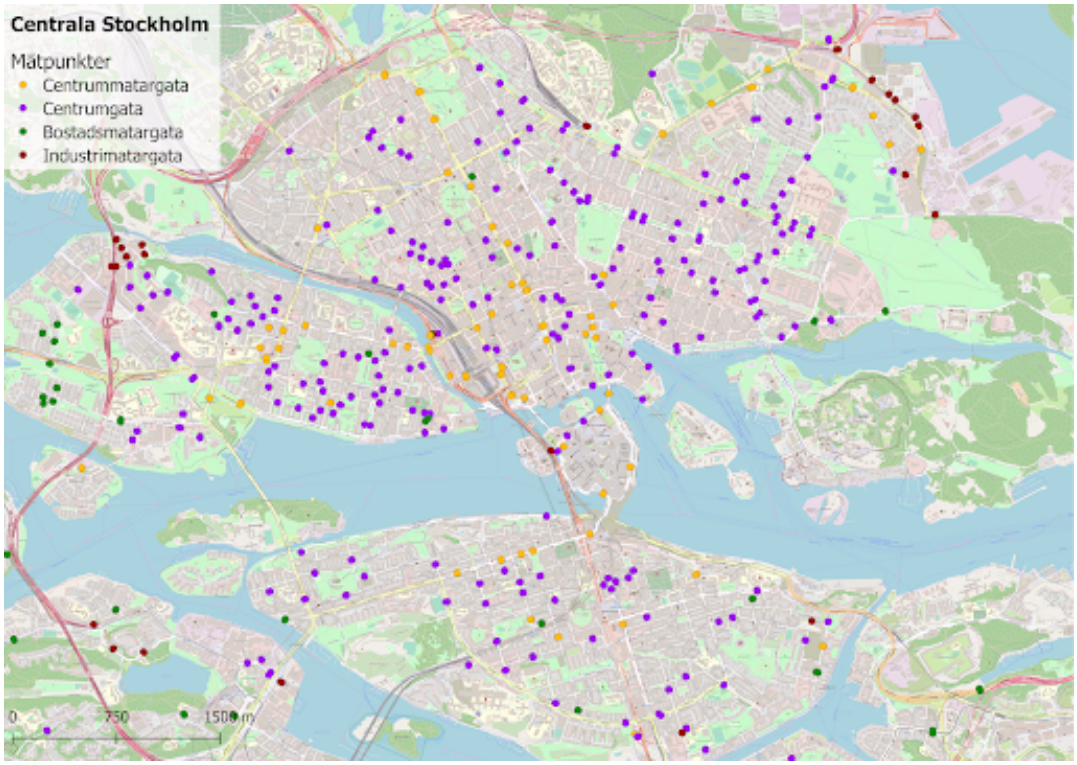
för avvikelse märks mätningen med en kvalitetsvarning. När det första försöket utfördes var den informationen inte känd så den data som användes i modellen var inte bearbetad därefter. För att undersöka om en högre andel mätplatser med entydiga predikteringar kunde uppnås utfördes ett andra försök där mätningar med en kvalitetsvarning filtrerades bort. Antalet mätplatser efter filtreringen var 1206 stycken med totalt 5980 mätningar. Resultatet gav 841 mätplatser med entydiga predikteringar och 366 mätplatser med fler än en prediktering. Att filtrera på kvalitetsvarningar gav alltså en förbättring med ungefär 7%. Resultatet för de entydigt predikterade mätplatserna med tillhörande mätpunkter presenteras i figur 8 nedan.



Figur 8: Karta över Stockholms stad med de utplacerade prediktionerna för de fyra gatutyperna. Varje punkt på kartan representerar en mätpunkt.

Genom en första anblick av kartan går det att se att majoriteten av de mätplatser som blivit klassificerade som centrumgata och centrummatargata hamnar i den mest centrala delen av Stockholms innerstad. En jämförelse med kartan i figur 1 från Göteborg (se kapitel 2) går det att se ett liknande mönster där gatutyperna centrumgata och centrummatargata är positionerade centralt. Även om majoriteten av de två gatutyperna, i kartan ovan, är lokaliserade inom ett visst område, går det att se ett antal mätplatser spridda runtom. Vad det kan bero på, eller vad det kan innebära, diskuteras i nästa avsnitt.

En förstorad bild av Stockholms mest centrala del kan ses i figur 9 på nästa sida. På kartan syns ett flertal mätplatser som klassificerats till de fyra olika gatutyperna. Majoriteten av mätplatserna har klassificerats som centrumgata eller centrummatargata, som även gick att urskilja i figur 8 ovan, och är placerade mot stadens mitt.



Figur 9: Förstorad bild över Stockholm stads mest centrala delar.

Likt kartan över Göteborg, klassificeras mätplatser som tillhör bostadsmatargator längre ifrån de mest centrala delarna av staden. Figur 10 är en förstorad karta över Stockholmsområdet Hässelby-Vällingby, som ligger i de nordvästra delarna av Stockholm. På kartan ses mätplatser som klassificerats som bostadsmatargator utifrån modellen. Området består till större delen av bostadsområden och bör således innefattas av gator av typerna bostadsgata eller bostadsmatargata. En vidare diskussion av resultaten gällande prediktioner av bostadsmatargator återkommer i nästa avsnitt.

Figur 10: Förstorad bild över Stockholmsområdet Hässelby-Vällingby.

De mätplatser som klassificerats som industrimatargator är placerade med större spridning vilket kan ses från kartan i den övergripande kartan över Stockholm, se figur 8. Detsamma gäller mätplatser av den gatutypen i Göteborg, och för att undersöka resultatet närmare har ett stickprov tagits fram. I figur 11 på nästa sida ses en karta över Västberga industriområde något sydväst i förhållande till de mest centrala delarna av Stockholm. De mätpunkter som kan ses på den förstorade delen av kartan är klassificerade som industrimatargator och av områdets karaktär bör gatorna vara av typen industrigata eller industrimatargata. En vidare diskussion av resultaten gällande prediktioner av industrimatargator återkommer i nästa avsnitt.

Figur 11: Förstorad bild över Västberga industriområde i Stockholm.

I tabell 15 presenteras medelvärden för trafikmängd, medelhastighet och andel tung trafik för varje gatutyp. Värdena kommer från de trafikmätningar som klassificerats till de fyra gatutyperna och kan användas för att analysera och diskutera resultaten ytterligare. Likt de medelvärden som togs fram för Göteborg i avsnitt *4. Data*, har industrimatargata högst medelhastighet samt högst andel tung trafik. Bostadsmatargata har en något högre medelhastighet än de vad gatutypen har i Göteborg, men för båda städerna har bostadsmatargata den näst högsta hastigheten. Medelvärdet för andel tung trafik är liknande för bostadsmatargata och centrumgata, som även var fallet i Göteborg. Centrummatargata erhåller ett medelvärde på andel tung trafik som närmar sig industrimatargata vilket gäller för båda städerna. En skillnad som kan ses är att förhållandet för medelhastigheten mellan centrumgata och centrummatargata har skiftats i Stockholm där centrummatargata istället har en högre hastighet.

Tabell 15. Medelvärden för trafikmängd, hastighet och andel tung trafik för respektive predikterad gatutyp i Stockholm.

Gatutyp	**Medelvärde trafikmängd (antal fordon)**	**Medelhastighet (km/h)**	**Medelvärde andel tung trafik (andel tunga fordon)**
Bostadsmatargata	9969	51	8.5%
Centrumgata	8582	38	8.9%
Centrummatargata	14136	45	9.4%
Industrimatargata	12007	55	11.9%

6.5 Sammanfattning

Modellen k-NN presterade liknande för alla de enskilda variablerna där en noggrannhet erhölls mellan 86.5% och 89.6%. När modellen applicerades på de tre olika kombinationerna av variablerna varierade den viktade sammansättningen av noggrannhet och känslighet (f-score) mellan 86.9% och 88.2%. Den kombination som resulterade i den högsta noggrannheten var trafikflöde tillsammans med andel tung trafik. Eftersom resultaten för kombinationerna var så pass liknande testades alla de olika kombinationerna med avseende på vilka gatutyper som klassificerats med störst f-score. Gemensamt för alla kombinationer var att gatutypen industrimatargata erhöll det högsta värdet på f-score.

För random forest presterade modellen bäst på den enskilda variabeln trafikmängd (noggrannhet 91.0%) och lägst värde fick variablen hastighet (55.0%). De tre kombinationerna av variablerna erhöll mycket snarlika resultat där f-score varierade mellan 91.6% (trafikmäng, hastighet) och 93.1% (alla tre). Eftersom liknande resultat erhölls för modellen k-NN med avseende på täthet i resultatet presenterades klassificeringarna gjorda med modellen för alla tre kombinationer. Av de tre kombinationerna presterade den sista kombinationen (alla tre variabler) bäst för alla gatutyper förutom för bostadsmatargata, där värdet på f-score var 0.1% lägre än den bäst presterande kombinationen på just den gatutypen.

Den modell som valdes för applicering på Stockholms trafikdata var random forest och att modellen presterade bra bekräftade även resultatet från confusion matrisen. Resultaten från att modellen applicerats på trafikdatan i Stockholm antyder på goda resultat och likheter mellan städerna kan ses. En djupare diskussion om resultaten från klassificeringen kommer i nästa avsnitt.

7 Diskussion

I det här avsnittet diskuteras och analyseras resultaten från den här studien. Först diskuteras val av modeller där styrkor och svagheter för det här fallet tas upp. De första två delarna behandlar val kring datahantering och bearbetning samt felkällor som kan uppstå vid den typen av hantering av data. Efter det diskuteras de resultat som erhölls när den första modellen, k-NN, applicerades på de enskilda variablerna samt de tre kombinationerna. Därefter behandlas resultaten då den andra modellen, random forest, tränades på variablerna samt kombinationerna. För att utvärdera modellernas prestation har statistiska mått som noggrannhet och f-score beräknats och valet av modell diskuteras i det efterföljande avsnittet. Utifrån resultaten valdes modellen random forest att appliceras på den trafikdata som ämnades att klassificeras i Stockholm och de erhållna resultaten diskuteras i den avslutande delen av det här kapitlet.

7.1 Val av modeller och mått för utvärdering

De två modeller som valdes för den här studien var k-NN och random forest. K-NN är en lätthanterad och enkel modell som är välanvänd inom klassificering och som inom många studier erhållit goda resultat. Modellen är enkel att implementera i Python samt erhåller resultat som vid en första anblick är enkla att tolka. Eftersom modellen k-NN utgår från distanser mellan punkter när den tränas kan den vara en bra modell att initialt testa på data där det finns okända strukturer. Det negativa med modellen är att den ibland tenderar att bli överanpassad till träningsdata eftersom den inte identifierar komplexa mönster i datan. Det leder till valet av den andra modellen som använts i det här arbetet: random forest. Random forest är en ensemble metod som genom dess sammansättning kan erhålla låg bias samtidigt som variansen är låg. Modellen bygger på beslutsträd som basmodeller och har egenskapen av att finna komplexa strukturer i data. De olika styrkorna hos modellerna gör att problemet kan ses utifrån två infallsvinklar och öka möjligheterna till goda resultat.

För att utvärdera och jämföra modeller användes confusion matris, som är en vanlig metod att använda inom maskininlärning. Syftet var att öka förståelsen för hur den valda modellen presterat på träningsdata. Utöver det, har statistiska och matematiska mått som beskriver hur väl en modell presterat på testdata använts. Måtten går att applicera inom klassificering och möjliggör att information om hur modeller presterat på specifika klasser kan tas fram. Det öppnar upp för en djupare förståelse för resultaten och kan även användas för att identifiera besvärliga klasser (klasser som innehåller data som är komplicerade eller svåra att finna strukturer i). I det här arbetet handlar det om att

kunna identifiera klasser där trafikdata inte är tillräcklig eller klasser som generellt liknar varandra.

7.2 Bearbetning och hanteringen av data

Den trafikdata som har använts kommer från två olika kommuner där insamling och mätutrustning ser olika ut. Det innebär att bearbetning av data varit nödvändig för att kunna applicera modellerna på samma sätt. Det har bland annat inneburit att viss data filtrerats samt att utfyllnad av data för vissa attribut varit nödvändig. All hantering och bearbetning av data innebär att osäkerheter och felkällor kan uppstå. Till exempel, användes en metod för standardisering av trafikmängd per timme där värdet på ÅMVD uppskattades utifrån de mätningar som fanns tillgängliga. Ingen av kommunerna innehöll mätdata för hela år vilket gör att genom beräkning av ÅMVD inte beskriver verkligheten korrekt.

I den här studien har även avsaknad av data varit nödvändig att hantera och ta i beaktning. Ibland registrerar inte en mätutrustning all data vilket medför att ett null-värde kan förekomma i dataseten. Om attributet trafikmängd saknades togs mätningen bort helt och hållet. Men om det saknas värden för attributen hastighet och andel tung trafik ersattes värdet med slumpmässigt framtaget värde. Det slumpmässiga värdet genererades på intervallet mellan det minsta och största värdet från det specifika attributet för den klassen. Anledningen till att utfyllnaden av data utfördes på det sättet var att öka andelen data som kunde tränas på och eftersom de slumpmässiga värdena var kopplade till varje enskild klass är värdena ändå relevanta.

7.3 K-NN

Som presenterat i föregående kapitel presterade k-NN modellen liknande på de tre olika attributen trafikmängd, hastighet och andel tung trafik när de testades enskilt. Modellen identifierade inte någon större skillnad mellan relevansen för de olika attributen. Något som kan förklara att modellen inte identifierar några större olikheter är att attributen skalats om för att passa modellen. K-NN är känslig mot avstickande värden eller stora skillnader mellan attribut vilket innebär att data i många fall måste skalas om, vilket även gjordes i den här studien. En indikation på att det finns olikheter mellan attributens relevans är att den andra modellen, random forest, resulterade i stora skillnader i noggrannhet mellan attributen.

När modellen testades på de tre kombinationerna blev resultatet bäst för kombinationen

trafikmängd och andel tung trafik. De andra två kombinationerna resulterade i ett f-score som var 1-1.3% lägre, vilket i sammanhanget inte någon stor försämring. Eftersom de tre kombinationerna resulterade i snarlika noggrannheter och värden på f-score presenterades mer detaljerad information om hur resultaten blivit för de olika gatutyperna. Även här var resultaten mycket lika gällande f-score och den enda uppenbara skillnaden var att den sista kombinationen inte lyckades prediktera gatutypen centrummatargata lika bra som vid de andra två. Den gatutyp som erhöll högst värde på f-score, och även för de andra två måtten precision och känslighet, var industrimatargata där resultaten aldrig var lägre än 94%. Att industrimatargata resulterade i höga värden oavsett kombination ger en indikation på att den klassen skiljer sig från de andra och således är lättare att särskilja.

Som tidigare nämnt är k-NN en relativt enkel modell som bygger på distanser mellan datapunkter. För det här problemet kan det vara så att modellen är för simpel och därför inte lyckas identifiera de bakomliggande mönster, som kan tänkas finnas, fullt ut. Trots det relativt goda resultatet (ofta högre än 85% noggrannhet) finns indikationer på att modellen inte kan identifiera alla mönster då resultaten inte är stabila. Modellen presterade generellt snarlikt för de enskilda variablerna samt för de tre kombinationerna och innebär alltså att modellen inte "lär sig mer" desto mer information som läggs till. Modellen presterar heller inte lika bra som random forest, som diskuteras i nästa del.

7.4 Random forest

Till skillnad mot k-NN, skiljde sig resultaten åt mellan de tre enskilda variablerna när random forest applicerades. För attributet trafikmängd kunde modellen prediktera rätt gatutyp med en noggrannhet på 91% medan hastighet precis översteg 55%. Även för andel tung trafik blev värdet lågt, nästan 62%, och är alltså en försämring med nästan 30% i jämförelse med resultatet från trafikmängd. Med random forest predikterade modellen alltså bättre med den information som variabeln trafikmängd innehåller. Eftersom de tre variablerna innehåller olika information om mätpunkten är det inte överraskande om variablerna utgör olika stor relevans för prediktionerna. Av de initiala resultaten med random forest kan således variabeln trafikmängd anses vara av störst värde, vilket även styrker att den variabeln bör vara med vid alla kombinationerna.

För de tre kombinationerna presterade modellen, likt k-NN, väldigt lika för alla tre kombinationer. Den kombination som erhöll lägst noggrannhet och f-score var trafikmängd tillsammans med hastighet, 91.6%. Högst fick kombinationen som innehåller alla tre variabler där både noggrannheten och f-score var 93.1%. Även om marginalerna är små,

ökar modellens prestation då variablerna trafikmängd och andel tung trafik tas med. Och eftersom marginalerna är små jämförs även resultaten mer detaljerat för prediktionerna för vardera klass inom varje kombination.

Kombinationerna som enbart innehöll två variabler resulterade i liknande resultat på f-score, där industrimatargata predikterades rätt med störst sannolikhet och bostadsmatargata med lägst. Samma resultat erhölls då modellen tillämpades på den sista kombinationen med avseende på rangordningen. Skillnaden är att resultaten på f-score för de olika klasserna blev högre för alla förutom bostadsmatargata, som var i princip samma som för de andra två kombinationerna.

Modellen random forest presterade alltså bäst för kombinationen som innehöll alla variabler där resultaten för alla gatutyper var högre än för den bästa kombinationen då k-NN användes. Gatutypen centrummatargata fick ett f-score som var 9% högre med random forest än vid k-NN, bostadsmatargata förbättrades med 5%, industrimatargata med 1% och centrumgata med nästan 6%. I nästa del diskuteras resultaten som erhölls då random forest applicerades på trafikdatan från Stockholm.

7.5 Prediktion av Stockholms trafikdata

Vid prediktion av Stockholms mätplatser finns inget tillgängligt facit att jämföra resultatet med. Genom slumpmässiga stickprov kan mätplatserna utvärderas utifrån det område som gatan finns inom. Det gör att enbart övergripande slutsatser är möjliga när modellen random forest applicerades. Inte heller går det att översätta de predikterade gatutyperna direkt till de gatutyper som finns i Stockholm utan att resultaten utvärderas. Men trots de begränsningar som finns i studien är det ändå av värde att diskutera resultaten eftersom det kan ge information om tillståndet på de olika mätplatserna.

Vid en övergripande anblick över hur mätpunkterna i Stockholm har klassificerats ser resultaten lovande ut. Genom att jämföra med de klassificerade mätpunkterna i Göteborg finns likheter med hur mätpunkterna predikterats i Stockholm. Anledningen till att det anses vara ett gott tecken är att de predikterade mätplatserna i Stockholm följer ungefär samma områdesindelningar som i Göteborg även fast topologin är olika. Majoriteten av de centrala delarna i städerna är klassificerade som centrumgata eller centrummatargata, industrimatagator är klassificerade inom industriområden och till sist bostadsmatargator är placerade i angränsning till bostadsområden, som är lokaliserade i utkanterna av städerna.

Centrumgator och centrummatargator är två gatutyper som i vägnätet befinner sig i anslutning till varandra. En centrummatargata är en gata som leder in till en centrumgata men hur trafiken förhåller sig sinsemellan är svårare att veta. Av klassificeringen av gatutyper som gjordes i Stockholm, som tidigare nämnt, placerades majoriteten av centrumgator och centrummatargator i de centrala delarna. Huruvida gatorna blivit rätt predikterade med avseende på centrumgata i jämförelse med centrummatargata är något som kräver en vidare undersökning. Det går ändå att konstatera genom en jämförelse av de medelvärden som togs fram för de respektive gatutyperna i Stockholm, se tabell 15 i *6.5 Applicera modellen på Stockholms trafikdata*, att centrummatargata generellt har högre andel tung trafik, högre medelhastighet och ett genomsnittligt högre antal fordon som passerar gatutypen per dygn. Förhållandet mellan centrumgator och centrummatargator i Göteborg liknar de i Stockholm, med undantag för att medelhastigheten är lägre för centrummatargatorna än för centrumgatorna. Att städerna har olika hastigheter, alltså att Göteborg generellt har lägre hastigheter än Stockholm, kan eventuellt förklaras av att städerna har olika kollektivtrafik. För en större stad som Stockholm borde medföra en större mängd bilar (vilket även är fallet i studien) och möjligheterna för köbildning ökar. Trots det, är medelhastigheterna högre i Stockholm vilket tyder på att så inte är fallet. En eventuell förklaring kan således vara en mer attraktiv kollektivtrafik. Men eftersom syftet med den här studien är att undersöka om det går att prediktera mätplatser i Stockholm utifrån klassificerad trafikdata i Göteborg, är resultaten för gatutyperna centrumgator och centrummatargator bra i det avseendet. De mätpunkterna som är predikterade inom de centrala delarna är således inom "rätt" område. Av resultaten ska dock tilläggas att vissa mätpunkter placerades utanför de centrala delarna. Det kan bero på att gatan antingen leder in till en plats som drar mycket folk (t.ex. fotbollsarena), att trafiken består av högre andel tung trafik (busslinjer) eller att gatan tillåter högre hastigheter.

Gator som predikterats som bostadsmatargator följer samma mönster som i Göteborg där de ligger i utkanten av den mest centrala delen av städerna, vilket är rimligt med tanke på gatans funktion. Av de framtagna medelvärdena för trafikmängd, hastighet och andel tung trafik för bostadsmatargata liknade förhållandet till övriga gatutyper de förhållanden som gick att se i Göteborg. Bostadsmatargata hade en medelhastighet som var större än både centrumgata och centrummatargata men lägre än industrimatargata. Medelvärdet för andel tung trafik var liknande värdet för centrumgata men lägre för både centrummatargata och industrimatargata, som i Göteborg. Från resultaten som ses i kartan över Stockholm klassificerades ett fåtal gator i de mest centrala delarna som bostadsmatargator. En anledning till det kan vara att vissa värden för bostadsmatargator liknar trafiken på centrumgator. Något som generellt skiljer gatutyperna åt är uppmätt hastighet, där

bostadsmatargator ofta har högre hastigheter. Att modellen predikterat vissa gator som bostadsmatargator skulle därför kunna bero på att det förekommit lägre hastigheter på de mätplatserna. En annan orsak till prediktionerna skulle kunna vara att det finns mindre områden i de mest centrala delarna i Stockholm som faktiskt utgör bostadsområden i den bemärkelsen.

Den sista gatutypen som klassificerades i Stockholm var industrimatargata. Av exemplet som studerades i slutet på resultatdelen identifierades ett industriområde där gatorna predikterats som industrimatargator. Även andra områden predikterades som industrimatargator i staden. Varje område har inte studerats närmare, men en orsak till prediktionerna kan vara att det exempelvis finns en hamn i närheten, att området faktiskt är ett industriområde, att det tillåts högre hastigheter eller att det passerar en stor andel tung trafik. De medelvärden som erhölls för mätpunkterna som klassificerades som industrimatargator liknade de värden som togs fram för gatutypen i Göteborg. Andel tung trafik var högst för den gatutypen för båda städerna, samtidigt som medelhastigheten var högst av de fyra olika gatutyperna. Modellen hade även högst antal mätningar som kom från industrimatargator som träningsdata (från Göteborg) och ger således en indikation på att modellen har stor möjlighet att identifiera mönster för den gatutypen. Det resultatet kan även stärkas från den confusion matris som gjordes för Göteborgs data: av 1089 mätningar predikterades 1044 som industrigata under träningsfasen. I de mest centrala delarna av Stockholm predikterades en viss andel gator som industrimatargator och även i utkanterna går det att finna gator som predikterats till gatutypen. En potentiell orsak till det kan vara att busslinjetrafik påverkar klassificeringen. Inom centrala områden kan det förekomma en stor andel bussar och eftersom de ingår i kategorin tung trafik kan en sådan förklaring vara trolig. Men för att kunna undersöka det vidare och kvalitetssäkra resultaten krävs en vidare undersökning.

8 Slutsatser

I den här uppsatsen har två stycken maskininlärningsmodeller använts på trafikdata från två olika städer, Göteborg och Stockholm. Uppsatsens syfte har varit att utvärdera möjligheterna att klassificera gator utifrån trafikdata i Stockholm genom att använda redan klassificerade gator i Göteborgs trafikdata. För att uppnå syftet har de två maskininlärningsmodellerna tränats på data från Göteborg och utvärderats med avseende på hur väl modellerna predikterat de olika gatutyperna. Gatutyperna som har varit intressanta i den här studien är de av Göteborgs kommun benämnda: centrummatargata, centrumgata, bostadsmatargata och industrimatargata. Modellerna k-NN och random forest utvärderades utifrån 6 olika hypoteser för att identifiera vilka attribut och kombinationer av attribut som resulterar i bäst resultat. De enskilda attributen var trafikmängd (räknat per timme under 24 timmar), medelhastighet (85 percentilen) och andel tung trafik. De tre kombinationer som undersöktes var följande: {trafikmängd, medelhastighet}, {trafikmängd, andel tung trafik} och {trafikmängd, medelhastighet, andel tung trafik}. Den modell som presterade bäst utifrån hypoteserna applicerades sedan på trafikdatan i Stockholm.

Resultatet av studien visar att den modell som presterade bäst var random forest då alla de tre attributen ingick i modellen. Med modellen kunde en noggrannhet på 93.1% uppnås där det beräknade f-score för vardera gatutyp uppgick till mellan 88-97%. Bäst presterade modellen på gatutypen industrigata och sämst för bostadsmatargata. Den tränade modellen applicerades därefter på Stockholms trafikdata och utvärderades genom att jämföra resultaten med Göteborg. Jämförelsen utgick från hur gatorna klassificerats i förhållande till hur gatutyperna är placerade i Göteborg. Bland annat studerades hur de olika mätplatserna klassificerades i olika områden, som till exempel centrumområde, bostadsområde och industriområde. De stickprov som analyserades visade på goda resultat och ger en indikation på att modellen har presterat tillfredsställande i den här studien.

Att ha klassificerade mätplatser i vägtrafiksystemet kan fungera som ett analysinstrument för att enklare förstå vilken funktion en gata uppfyller med avseende på den trafik som rör sig där. För att upptäcka trafikförändringar, som kan användas både till trafikplanering men även för att förstå hur trafiken beter sig under specifika händelser (t.ex. evenemang eller utvärdera trafikstrategiska beslut), kan klassificerade gator i vägtrafiksystemet vara ett användbart verktyg. Om en mätplats som stabilt klassificeras som en viss gatutyp men som plötsligt byter klassificering, kan det ge en indikation på att trafiken har förändrats och att gatan eventuellt har bytt funktion. Förutom det kan förändringen också ge en antydan på att hanteringen av mätningen innehåller felkällor eller att tekniska fel har uppstått i mätutrustningen.

9 Framtida forskning

Den här studien har syftat till att utreda möjligheterna till att applicera modeller inom maskininlärning på trafikdata från två olika kommuner för att i slutändan klassificera mätplatser utifrån gatutyper. Tre specifika attribut har legat till grund för vilken data som maskininlärningsmodellerna har tränats och testats på och det finns en rad olika aspekter som inte tagits med vid analysen i den här rapporten. Det här avsnittet behandlar tankar och idéer om framtida forskning som kan bygga vidare på de erhållna resultaten.

Resultaten från Stockholm visade att vissa gator lokaliserade nära de centrala delarna av staden klassificerades till bostadsmatargator. En orsak till det skulle kunna vara att det inte passerar en förväntad mängd tung trafik, som annars kan vara ett kännetecken för centrumgator och centrummatargator. Det skulle kunna bero på att det inte passerar lika stor andel bussar eller annan kollektivtrafik. Ett sätt att utreda och möjligen komplettera informationen om vägnätet vid analysen skulle vara att ta hänsyn till busslinjerna. Om det på ett smidigt sätt gick att komplettera kartan med en busslinjekarta, till exempel SL:s egen, kan det ge indikationer på varför en gata fått en specifik klassificering. Det kan i sin tur användas för att identifiera felklassificeringar eller för att öka förståelsen för hur klassificeringarna av gatutyperna går till.

En annan aspekt som skulle kunna leda till bättre resultat är att utöka de attribut som används i modellerna. I Stockholm registreras bland annat avståndet mellan bilar vid varje mätplats vilket skulle kunna användas i kombination med hastighet för att identifiera mönster i trafiken. Avståndet mellan bilar tillsammans med hastighet kan ge en indikation på när det sker köbildning och således under vilka timmar på dygnet som det sker med störst sannolikhet. Det skulle kunna bidra med ytterligare information om hur trafiken beter sig hos de olika gatutyperna.

10 Referenser

Aryes, I., *Super Crunchers: Why Thinking–By–Numbers Is The New Way To Be Smart*, Bantam Books, (2007).

Castañón, J., *10 Machine Learning Methods that Every Data Scientist Should Know*, Towards Data Science, (2019). Tillgänglig: `https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9`

Festin, M., *SUMMARY OF NATIONAL AND REGIONAL TRAVEL TRENDS: 1970-1995*, United States Department of Transportation, Federal Highway Administration, Washington, D.C (1996).

Hastie, T., Tibshirani, R., Friedman, J., *The elements of statistical learning: Data mining, inference, and prediction.* 2nd ed., corrected at 11th printing, Springer, (2017).

Hüllermeier, E., *Possibilistic instance-based learning*, Artificial Intelligence, volume 148 (2003).

Jupyter Notebook, *The Jupyter Notebook*, hämtad: 2019-12-13. Tillgänglig: `https://jupyter.org/`

Keogh, E., *Instance-Based Learning* In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA, (2011).

Lindholm, A., Wahlström, N., Lindsten, F., Schön, T., *Supervised Machine Learning, Lecture notes for the Statistical Machine Learning course*, Department of Information Technology, Uppsala University, (2019-03-12).

Microsoft Azure, *Förbättrat beslutsträd med flera klasser*, hämtad: 2020-02-19. TIllgänglig: `https://docs.microsoft.com/sv-se/azure/machine-learning/algorithm-module-reference/multiclass-boosted-decision-tree`

Murphy, K., *Machine Learning - A Probabilistic Perspective*, The MIT Press Cambridge, Massachusetts London, England, (2012).

Norconsult Astando, *System för trafikmätningar*, hämtad: 2020-03-05. Tillgänglig: `https://www.norconsultastando.se/innehall/referensprojekt/system-trafikmatning/`

Pandas, *Pandas*, hämtad: 2019-12-13. Tillgänglig: https://pandas.pydata.org/

PostgreSQL, *PostgreSQL: The World's Most Advanced Open Source Relational Database*, hämtad: 2019-12-20. Tillgänglig: https://www.postgresql.org/

Proposition 2008/09:93. Mål för framtidens resor och transporter, Näringsdepartementet.
Pyhton, *About*, hämtad: 2020-03-10. Tillgänglig: https://www.python.org/about/

QGIS, *QGIS - The Leading Open Source Desktop GIS*, hämtad: 2020-01-17. Tillgänglig: https://qgis.org/en/site/about/index.html

Rokach, L., Maimon, O., *Top-down induction of decision trees classifiers - a survey*, IEEE Transactions on Systems, Man, and Cybernetics, Part C, volym 35, utgåva 4, (2005).

Schmueli, B., *Multi-Class Metrics Made Simple, Part II: the F1-score*, Towards Data Science, (2019). Tillgänglig: https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

Scikit-learn, *Machine Learning in Python*, hämtad: 2020-01-17. Tillgänglig: https://scikit-learn.org/stable/

Shung, K. P., *Accuracy, Precision, Recall or F1?*, Towards Data Science, (2018). Tillgänglig: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Tizghadam, A., Khazaei, H., Moghaddam, M., Hassan, Y., *Machine Learning in Transportation*, Journal of Advanced Transportation, volym 2019, (2019).

Trafikverket, *Transportsystemets behov av kapacitetshöjande åtgärder – förslag på lösningar till år 2025 och utblick mot år 2050*, (2012a).

Trafikverket, *Dataproduktspecifikation – Funktionell vägklass Version 3.0*, Borlänge, (2012b).

Trafikverket, Boverket, Sveriges Kommuner och Landsting, *Trafik för en attraktiv stad*, (2015a).

Trafikverket, *Metodbeskrivning - Undersökningen av ÅDT* (2015b).

Trafikverket, *Översikt vägdata Version 4.0*, (2018).

Webb, G. I., Encyclopedia of Machine Learning, Springer US, (2010).

Weijermars, W., *Analysis of Urban Traffic Patterns using Clustering*, TRAIL Research School (2007).

Zhao, F. and Chung, S., *Contributing Factors of Annual Average Daily Traffic in a Florida County, Exploration with Geographic Information System and Regression Models,* Transportation Research Record 1769, (2001).

Åkerlund, B., *TECH-lex - Teknisk ordlista för V*, Lund (2007).