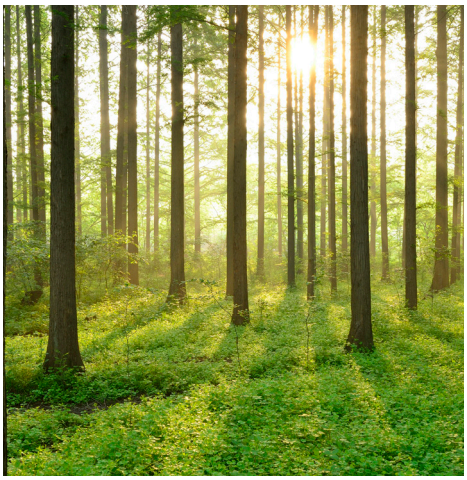
ELECTRICITY CONSUMER CLASSIFICATION USING SUPERVISED MACHINE LEARNING

RAPPORT 2021:729



ELNÄTENS DIGITALISERING
OCH IT-SÄKERHET

Energiforsk

Electricity Consumer Classification Using Supervised Machine Learning

KRISTOFFER FÜRST

ISBN 978-91-7673-729-3 | © Energiforsk February 2021

Energiforsk AB | Phone: 08-677 25 30 | E-mail: kontakt@energiforsk.se | www.energiforsk.se

Foreword

Projektet *Digitaliseringsbaserad konsumentkaraktärisering för intelligent distributionsplanering* ingår i programmet Elnätens digitalisering och IT-säkerhet och det tittar på prognoser av elnätets topplast för att kunna utvärdera behovet av att uppgradera distributionsnätet för att tillgodose fler konsumenter samt förnybara produktionskällor.

Noggrannheten beror på kunskapen om konsumenternas elektriska karaktäristik. Med tillgång till timvis elförbrukning samt data från meteorologiska- och fastighetsmyndigheter, har projektet utvecklat en mer noggrann modell för att kategorisera konsumenternas elektriska karaktäristik för en kostnadseffektiv nätplanering och dimensionering av mikronät genom att beakta mönsterigenkännings- och maskininlärningsmetoder.

Kristoffer Fürst på Chalmers tekniska högskola är projektledare och han har arbetat tillsammans med docent Peiyuan Chen på Chalmers tekniska högskola.

Ett extra stort tack till referensgruppen, som på ett mycket givande sätt har bidragit till projektet:

- Arne Berlin, Vattenfall Eldistribution
- Ferruccio Vuinovich, Göteborg Energi
- Anders Mannikoff, Herrljunga Energi
- Björn Jansson, Kungälven Energi
- Per Norberg, Vattenfall Eldistribution
- Irene Yu-Hua Gu, Chalmers

Programmets programstyrelse, som initierat, följt upp och godkänt projektet, består av följande ledamöter:

- Kristina Nilsson, Ellevio AB (ordförande)
- Arne Berlin, Vattenfall Eldistribution
- Hampus Bergquist, Svk
- Ferruccio Vuinovich, Göteborg Energi
- Torbjörn Solver, Mälarenergi AB (vice ordförande)
- Magnus Sjunnesson, Öresundskraft AB
- Peter Ols, Tekniska verken i Linköping AB
- Teddy Hjelm, Gävle Energi AB
- Claes Wedén, Hitachi ABB Power Grids AB
- Katarina Porath, ABB AB
- Björn Ållebrand, Trafikverket
- Adam Nilsson, Jämtkraft AB
- Magnus Brodin, Skellefteå Kraft AB
- Johan Örnberg, Umeå Energi Elnät
- Patrik Björnström, Sveriges Ingenjörer, MF
- Peter Addicksson, HEM AB
- Jesper Bjärvall, Karlskoga Energi AB
- Malin Wallberg, VB Energi AB
- Matz Tapper, Energiföretagen Sverige (adjungerad)

Stort tack också till de företag som har varit engagerade i programmet Elnätens digitalisering och IT-säkerhet:

- Ellevio
- Vattenfall Eldistribution
- Svenska kraftnät
- Göteborg Energi
- Mälarenergi Elnät
- Öresundskraft Elnät
- Tekniska Verken i Linköping
- Skellefteå Kraft Elnät
- Umeå Energi Elnät
- Jämtkraft Elnät
- Eskilstuna Strängnäs Energi & Miljö
- Karlstads El- och Stadsnät, Borås Elnät
- Halmstad Energi och Miljö Nät
- Luleå Energi Elnät, Borlänge Energi
- Nacka Energi
- Västerbergslagens Elnät
- PiteEnergi
- Södra Hallands Kraftförening
- Karlskoga Elnät
- Sveriges ingenjörer (Miljöfonden)
- Hitachi ABB Power Grid
- ABB
- Trafikverket
- Forumet Swedish Smartgrid
- Teknikföretagen
- Huawei Sverige
- Exeri
- Evado
- Elinorr ekonomisk förening; Bergs Tingslags Elektriska, Blåsjön Nät, Dala Energi Elnät, Elektra Nät, Gävle Energi, Hamra Besparingsskog, Hofors Elverk, Härjeåns Nät, Härnösand Elnät, Ljusdal Elnät, Malungs Elnät, Sandviken Energi Nät, Sundsvall Elnät, Söderhamn Elnät, Åsele Elnät, Årsunda Kraft & Belysningsförening och Övik Energi Nät.

Stockholm december 2020
Energiforsk AB

Susanne Stjernfeldt
Forskningsområde Elnät, Vindkraft och Solel

Sammanfattning

Baserat på diskussioner med de lokala nätoperatörerna saknas generellt noggranna modeller av den elektriska karaktäristiken hos olika typer av konsumenter. En anledning till detta är att det inte finns någon skyldighet för konsumenterna att meddela nätoperatören om energieffektivitetsåtgärder eller vilken typ av värmesystem som används. Byggnadsinformationen som är registrerad i energideklarationen kan också förändras under åren utan att det rapporteras till Boverket. Denna rapport syftar till att klassificera konsumenters uppvärmningstyper genom att använda maskininlärningsmetoder för analys av data från smarta elmätare, meteorologiska observationer och byggnadsdata. Tidsserierna analyserades vid olika tidsupplösningar, där motsvarande medel-, bas-, toppeffekt samt standardavvikelse av elförbrukningen extraherades som attribut (features) för klassificeringen. Denna rapport fokuserar på byggnader med endast ett värmesystem, som antingen är fjärrvärme, frånluftvärmepump eller direktverkande el.

I denna rapport har klassificeringsmodellen för att skilja mellan tre olika elkonsumenttyper framgångsrikt utvecklats med hjälp av support vector machine. Ett datadrivet tillvägagångssätt har använts för att klassificera enfamiljehushålls huvudsakliga uppvärmningstyp, där uppvärmningstypen samlades in från byggnadens energideklaration och egenskaperna hos elkonsumenterna extraherades från smart elmätare.

Resultatet visar att ökad datatidsupplösning ökar prestandan av klassificeringen, där prestandan är baserad på konsumenter med okänt (av modellen) uppvärmningssystem. Undantaget är när alla timvärden används som attribut, vilket minskar prestandan avsevärt på grund av överanpassning av klassificeringsmodellen. Specifikt ger timvariationer för varje månad av året den bästa prestandan, där den genomsnittliga noggrannhet ± standardavvikelsens för den femfaldiga korsvalideringen var 97.1%±0.4% för att klassificera konsumenter med fjärrvärme från konsumenter med elbaserad uppvärmningskällor; medan den genomsnittliga noggrannheten minskas till 92.4%±1.4% om modellen ska särskilja fjärrvärme, frånluftvärmepump eller direktverkande el. Dessutom visade sig lutningen av linjär regression mellan daglig medeltemperatur och daglig elförbrukning som enda attribut att ha en bra prestanda med en noggrannhet på 95.8%±0.8% när man klassificerar konsumenter med fjärrvärme från konsumenter med elbaserade värmekällor.

Analys av felklassificeringarna visar också att energideklarationerna kan vara föråldrade och att modellen kan indikera förändringen i uppvärmningsmetod, även om felmärkta exempel ingår vid träningen av klassificeringsmodellen. Ett exempel ges för ett område där 9 av 10 konsumenter med fjärrvärme klassificerades felaktigt och istället klassificerades som fjärrvärmepump enligt klassificeringsmodellen. En manuell undersökning visar att konsumenterna har ändrat sin uppvärmningstyp från fjärrvärme till frånluftsvärmepump år 2015.

Summary

Based on the discussions with the local grid operators, there is a general lack of accurate models of the electrical characteristics of different types of consumers. One reason for this is that there is no obligation for the consumers to notify the grid operator about energy efficiency measures or which type of heating system is used. The consumer information may also change over the years without being reported to the building authority either. This report aims to classify consumer heating types by applying machine learning methods for analyzing smart meter measurements, meteorological observations, and building data. The smart meter time series was analyzed at different time-resolutions, where the corresponding average, base, peak, and standard deviation of the electricity consumption was extracted as features for the classifier. This report focuses on buildings with only one heating system, which is either district heating, exhaust air heat pump, or direct electricity.

In this report, the classification model (classifier) to distinguish three different electricity consumer types has been successfully developed by using the support vector machine (SVM) algorithm. A data-driven approach has been used to classify one family household's *main* heating type, where the heating type was collected from the building's energy declaration, and the characteristics of the electricity consumers were extracted from smart meter data.

The result shows that increasing the data time resolution increases the generalization performance of the classifier. The exception is when using all the hourly measurements as features, which will reduce the performance substantially due to model overfitting. Specifically, hourly variations for each month of the year gives the best generalization performance, where the average$\pm$standard deviation accuracy of the 5-fold cross-validation was 97.1%$\pm$0.4% for classifying consumers with district heating from consumers with electricity-based heating sources; whereas the average accuracy is reduced to 92.4%$\pm$1.4% if the classifier is to tell further if the consumer uses exhaust air heat pump or direct electricity. Furthermore, using linear regression slope between temperature and power as a single feature showed to have a good performance with an accuracy of 95.8%$\pm$0.8% when classifying consumer with district heating from consumers with electricity-based heating sources.

The analysis of the misclassification also shows that the energy declarations can be outdated and that the model is able to indicate the change in the heating system, even though wrongly labeled samples are included in the training of the classifier. An example is given for an area where 9 out of 10 consumers with district heating were misclassified and instead classified as exhaust air heat pump by the classification model. A manual investigation shows that the consumers have changed their heating type from district heating to exhaust air heat pumps in 2015.

List of content

1 Introduction

1.1 BACKGROUND AND MOTIVATION

Based on the discussions with the local grid operators, there is a general lack of accurate models of the electrical characteristics of different types of consumers. This is important for several reasons, including peak load and demand flexibility estimation for dimensioning and operational purposes. One reason for this is that there is no obligation for the consumers to notify the grid operator about energy efficiency measures or which type of heating system is used. This leads to low customer knowledge, and together with a too rough customer categorization, it adds uncertainty for grid planning decisions. On the other hand, the large-scale rolling out of smart meters for consumers together with publicly available data provides a great opportunity to develop methods to characterize the end-users and their electrical characteristics, in which the method of pattern recognition and machine learning is considered to have a great potential to be applied.

From the DSO's point of view, it is important to estimate the peak load of a consumer, and it's a contribution to the peak demand in the local and upper-stream grid. With the *known* characteristics of different types of consumers, the data can be used to estimate the peak demand of new loads and their contribution to the system peak, even though no smart meter data yet exists. Similarly, if consumers change their characteristic, for example, changing from non-electric based heating to an electric base heating, the peak characteristic would change. Therefore, it is important to capture such a change before potential congestions occurs in the grid.

1.2 RELATED WORK

To increase the end-user knowledge, smart meter data is used to categorize the end-users by utilizing machine learning methods in [1], [2], [3], [4], [5], [6]. Commonly, a supervised classification or and unsupervised clustering is used. The type of method depends inter alia on the availability of data and the purpose of the categorization. In classification models (classifier), the label/category of the end-users in the dataset is known a-priori. The classifier predicts the category for a sample which have not been seen by the classifier before. In an unsupervised clustering, there is no labels/category. The aim could be to cluster consumers which have similar load patterns. An unseen sample is assigned to one of the clusters.

In [1], [2], the aim is to classify (separately) different household properties, including building properties, such as the number of bedrooms, age of the building, area of the building, and family characteristics such as family size and retirement status. The known categories are based on survey data. A set of predefined features are selected, where [1] includes 22 statistical features, and [2] includes the same features and is extended with 66 other features, including consumption, ratios, statistical and temporal characteristics. The papers review different machine learning classification methods, including support vector

machine (SVM) and k-Nearest Neighbors (k-NN). The results indicate that SVM is one of the top methods evaluated. However, the default hyperparameters (model parameter) of the classifiers defined in the machine learning library was used. Though, tuning of the hyperparameters for the given dataset is an important step in machine learning as it defines the complexity of the classifier. The classifier is not able to capture all the information in the data if the complexity is too low (underfitting), whereas the classifier captures most of the information if the complexity is too high, but it generalizes poorly on data that have not been seen by the classifier before (overfitting). Furthermore, [1], [2], did not consider weather-dependency or seasonal behavior, which can be important when classifying the heating types of different electricity consumers.

In [3], [4], [5], [6] one of the aims is to find end-users that are similar to each other, where among others the unsupervised K-means clustering was used. In [3], the features include different end-user key performance indicators, such as load factor, temperature sensitivity, and the correlation between electricity consumption and electricity spot price. In [4], [5], [6], the average/typical load profile of a day is used as features to find consumption patterns that are similar to each other. In [7], a typical load curve is defined by the average and the standard deviation of the electricity consumption for different seasons, temperatures, and hours of a day. A post-clustering analysis is performed in [5] and [6], where [6] include the type of building and [5] the heating type, household size, number of teenagers, and number of kids in the households. The household characteristics in [5] are based on a telephone survey.

Based on the literature review, we suppose that the end-user electricity consumption characteristics can be described by load curves. To extend the analysis, the consumption is analyzed at different time-resolutions, and also including average, base, peak, and standard deviation of the electricity consumption. Moreover, one of the drawbacks of unsupervised clustering is that the number of clusters is unknown. As the heating type, i.e. the class label, is known in this work, supervised learning will be used instead. Also, survey data is often expensive and time-consuming where instead a data-driven approach is used.

1.3 AIM OF THE REPORT

This report aims to classify consumer types by applying supervised machine learning methods for analyzing smart meter measurements, meteorological observations, and building data. Specifically, the focus of this work is to categorize the *main* heating source commonly used by 1-2 family households including district heating, exhaust air heat pump, and direct electricity.

1.4 DELIMITATIONS

The sequential time series from the smart meters and metrological observations is for hourly average measurements, hence sequential data with faster sampling frequencies are not considered. Data such as occupant information and detailed end-user behavior are not considered. The project does not investigate secondary

or even tertiary heating sources used by consumers. This project does not investigate reactive power-consumption by the consumers either.

1.5 BENEFITS TO NEED OWNERS

Distribution system operators (DSOs): by increasing the customer knowledge such that a more accurate statistical model on consumer's electrical characteristics can be provided. This helps to improve the accuracy in peak load prognosis of the grid and thus assists decision-making on grid upgrade, expansion, and operation in an energy-efficient, economical, and reliable way.

Local electricity grid users: grid users' bill to the grid operator can be cut down as the grid planning is carried out with a more accurate knowledge of the consumers

Flexibility aggregator: by providing a machine learning model to distinguish consumers using different heating technologies, from which the demand flexibility can be further estimated.

Keywords

Classification, machine learning, heating types, energy declaration, smart meter data, outdoor air temperature data, consumer characteristics, customer awareness

2 Description of building characteristics and its energy usage

2.1 DATA SOURCES

The main data used for the classification in this project is electricity consumption from smart meters, outdoor air temperature, and building characteristics. For the case study, smart meter data were collected from the Swedish distribution system operator (DSO) Göteborg Energi Nät AB (GENAB) [8], which has the area concession in Gothenburg municipality with 270.000 connected customers, see Figure 2.1. The historical weather observations were collected from the Swedish Metrological and Hydrological Institute (SMHI) [9]. Lastly, the building characteristic was collected from The National Board of Housing, Building, and Planning (Boverket), which is the authority that supervises and manages the register of the building's energy declarations [10].

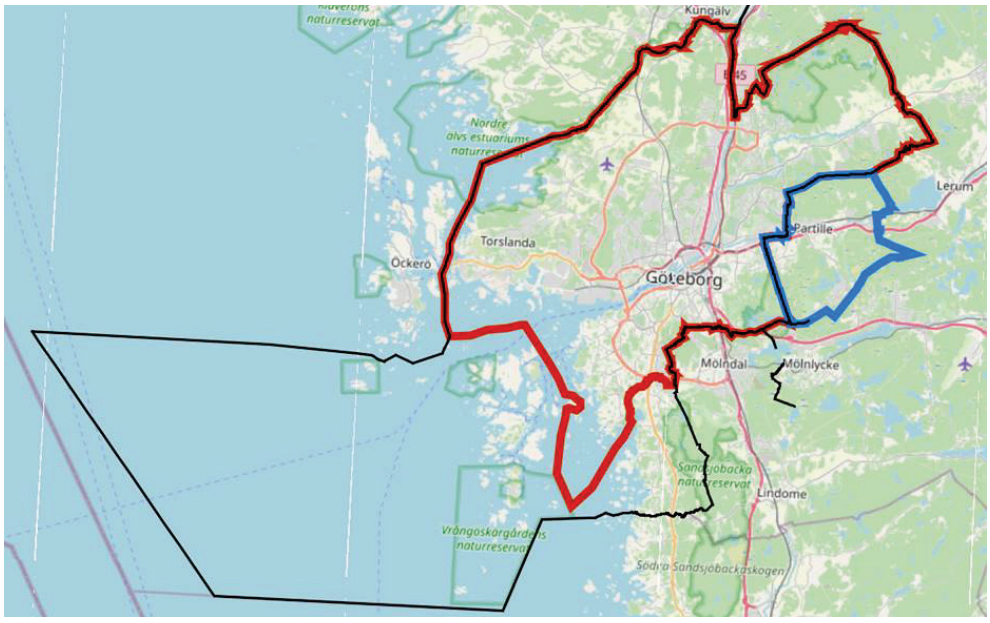


Figure 2.1 Map of GENAB's concession area. The black line shows the border of Gothenburg municipality and red the concession border of GENAB. The blue line shows Partille municipality and does not belong to GENAB's concession area. Source: [8]

2.2 ENERGY DECLARATION

The energy declaration shows the energy performance of the building and is performed by an independent and certified energy expert [11]. From the energy declaration, various well informative features can be extracted about the buildings heating characteristics, which, among others, include:

- the type of building,
- heated area (m^2 heated above 10ºC),
- the share of the heated area used for different end-use purposes, e.g. residential, offices, hotel, hospital, etc.

- measured energy usage for space and tap water heating specified for different heating sources
- ventilation system

The energy expert can also leave proposed suggestions on how to reduce energy consumption [10]. This can indicate if a consumer could reduce their energy consumption in a foreseeable time.

In general, the buildings that are required to have a valid energy declaration are [11]:

- all buildings that are larger than 250 m² and that are often visited by the public, e.g. hospitals, libraries, museums, schools, etc.
- buildings with the right of use also need an energy declaration, e.g. rental apartments, rental offices, etc.
- when buildings are to be sold, including 1-2 family households
- newly built buildings, where an energy declaration should be performed within two years after it has been put into use

The energy declaration is valid for ten years. After that, a new energy declaration needs to be conducted if the building falls under any of the requirements above [11]. This dataset gives a very good foundation for consumer characterization and increased customer knowledge. However, the energy declaration is valid for ten years, for which under this time a lot of changes can occur. Also, for 1-2 family households, it mainly *only* includes newly built houses or houses that have been sold in the last 10 years.

2.3 SPACE AND TAP WATER HEATING SYSTEMS

Tap water heating demand in a residential building is end-user specific. In other words, the demand is mainly due to end-user's behavior, e.g. the usage of showers. The space heating demand, on the other hand, is affected by multiple factors. Table 2.1 summarizes a non-comprehensive list of factors that can affect the space heating demand. The factors that are marked with bold font are data that can be found in the energy declaration and weather observations [9], [10], whereas the other factors are to the authors unknown.

2.3.1 Heating type utilization

The space and tap water heating in a dwelling can come from various heating types. Figure 2.2 shows, based on the energy declarations, the number of 1-2 family households in Sweden using different heating types/sources. Note that a building can have more than one heating type, including the use for comfort heating.

Table 2.1 The building's space heating demand. The marked factors are factors that can be found in [9] [10], whereas the other factors are unknown to the authors.

Increase in demand (positive correlation)	Decrease in demand (negative correlation)
+ Heated area of the building	- Outdoor temperature
+ Indoor comfort temperature	- Solar irradiance
+ Ventilation	- Building isolation
+ Wind	- Heat losses from electrical appliances
	- Heat from people indoors

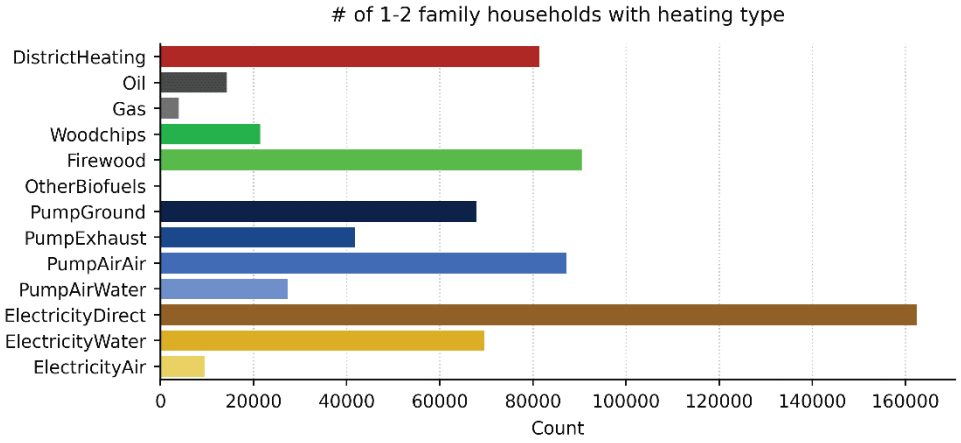


Figure 2.2 Number of 1-2 family households in Sweden with a specific heating type. Data are based on approved energy declarations between 2010-2019. Data source: [10].

Figure 2.3 shows the utilization of the different heating types based on the annual energy consumption registered in the energy declaration. The heating type utilization for a customer/building is defined as $E_i / \sum_i E_i$, where E_i is the annual energy for heating type i. With a utilization of 100% for a given heating type, only one type of heating is used. District heating, oil, gas, and heat pumps of type ground source, exhaust air, and air-to-water is used as the sole heating source, for more than 50% of the buildings that have any of these heating types. In contrast, the air-to-air heat pump and firewood are almost never used as the sole source of space heating.

2.3.2 Heat type characteristics

In Table 2.2, the heating source and heating distribution for different heating types are presented. The characteristics of the heating types that are analyzed in the case study are briefly explained below.

District heating

District heating is by far the most common *non-electricity-based* heating type that is used as a *primary* heating source in 1-2 family households in Sweden, see Figure 2.2 and Figure 2.3. Note that the firewood is more common, but as can be seen in Figure 2.3, the utilization is low, and it is more used as a complementary or comfort heating. Around 75% of the buildings with district heating does not complement their heating system with other heating types.

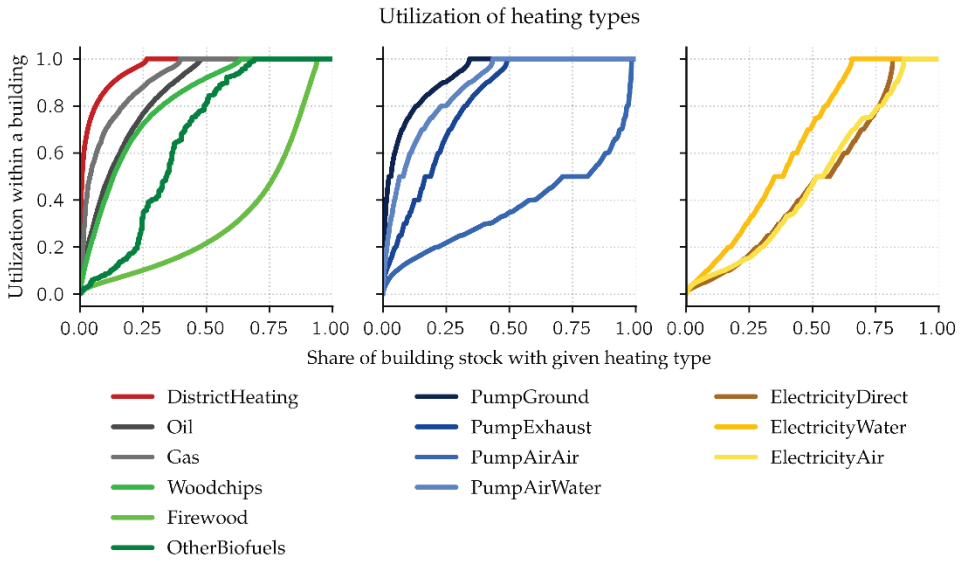


Figure 2.3 Utilization of different heating types for 1-2 family households in Sweden. Data are based on approved energy declarations between 2010-2019. Data source: [10].

Table 2.2 Heat source and heat distribution for different heating types

	District heating	Oil burner	Gas burner	Woodchip burner	Firewood	Ground source heat pump	Exhaust air heat pump	Air-to-air heat pump	Air-to-water heat pump	Direct electricity	Electricity-to-water	Electricity to air
Heat source												
District heating	●											
Fuel		●	●	●	●							
Electricity						●	●	●	●	●	●	●
Ground/rock/lake						●						
Exhaust air							●					
Outdoor air								●	●			
Heat distribution												
Tap water heating	●	●	●	●	●	●	●		●		●	
Radiator/floor heating (water for space heating)	●	●	●	●	●	●	●		●		●	
Air heating					●			●		●		●

Instead of each building producing their heat, district heating centralizes the heat production, interconnecting entire, or parts, of cities with a common pipe network. The heat is then transferred to the building, heating a waterborne heating system for space and tap water heating.

Direct electricity

The electric heaters for heating in buildings can be divided into three types: direct electric heating, electricity-to-water, and electricity-to-air heaters. Direct electricity heating is the most common heating system for 1-2 family households in Sweden, see Figure 2.2. However, in Figure 2.3 it can be seen that only around 20% of the buildings that have direct electric heating are using it as the only heating type. The direct electricity distributes the heat in the house by electric radiators or through floor/roof heating. The efficiency of the direct electricity is around 100%, that is that all electricity is converted to heat.

Exhaust air heat pumps

A more energy-efficient way of heating the house compared to full-electric heaters is by using a heat pump. A heat pump is a device that takes heat from a source, such as the air, the ground, or the water, to provide heating to a building. The heat can be transferred to a waterborne system and/or to the indoor air. The efficiency of the heat pump is dependent on the inlet temperature from the heating source. In the energy declaration, four types of heat pumps are distinguished: air-to-air heat pump, air-to-water heat pump, exhaust air heat pump, and ground source heat pump. Around 50% of the buildings with an exhaust air heat pump uses it as the only type of heating, see Figure 2.3.

Exhaust air heat pumps recover the heat from the exhaust air in the ventilation system of the building. Hence, it is limited by the heat in the exhaust ventilation air. The heat pump is connected to the water-based system that heats the indoor air and/or the warm water. As it uses mechanical exhaust ventilation, electricity is also used to drive the ventilation system. If no heat recovery is used in the inlet air, there is also a risk that cold inlet air could increase the heating demand. The benefit of an exhaust air heat pump is that it is less dependent on the outdoor temperature, compared to an air-to-air heat pump which loses its power-to-heat efficiency as the outdoor temperature decreases. A study showed that the coefficient of performance (COP) was around 2.9-3.4 during wintertime and around 3 in the summer for an exhaust air heat pump, where the COP is mainly affected by the supply temperature of the domestic hot water for a given heat pump [12]. The heat pump can also be supplied with an immersion heater which is used when the heat pump cannot cover the entire heating demand.

2.4 SMART METER DATA

The collected smart meter data from GENAB is hourly energy measurements. In general, electricity consumption and production are behind the meter for residential customers. Hence, detailed information about appliance usage, power-to-heat, etc. is not available to the authors. Neither is the load part or the

generation part of a prosumer Table 2.3 shows a non-comprehensive list of factors that affect the electricity usage of a consumer in a residential building.

Table 2.3 Electricity demand for a consumer in a residential building. The list is non-comprehensive. The building electricity is defined as the electricity that is used in common spaces, basements, outdoor electricity, etc. [13].

Demand	Supply
Household electricity (appliances)	Solar PV
Space and tap water heating	Other types of electricity production
Comfort cooling	
Building electricity	
Electric vehicles	

2.5 CASE STUDY – 1-2 FAMILY HOUSEHOLDS IN GOTHENBURG

For the case study in this report, 1-2 family households in the region of Gothenburg was used. Different characteristics associated with these households are summarized in this section, including heating type, fuse size, building age, heated area, and the corresponding outdoor temperature.

2.5.1 Heating type

The following presents some characteristics of the consumers/buildings in this region for which an energy declaration is valid. Figure 2.4 shows, based on the energy declarations, the number of 1-2 family households in Gothenburg with a specific heating type. Note that a building can use more than one heating type. Compared to entire Sweden, see Figure 2.2, the share of buildings with the fuel-based heating sources oil, gas, woodchips, and firewood are considerably less. There are also fewer buildings, relatively, with ground source and air-to-air heat pumps. District heating and electricity-to-water appear to be more common in Gothenburg compared to entire Sweden.

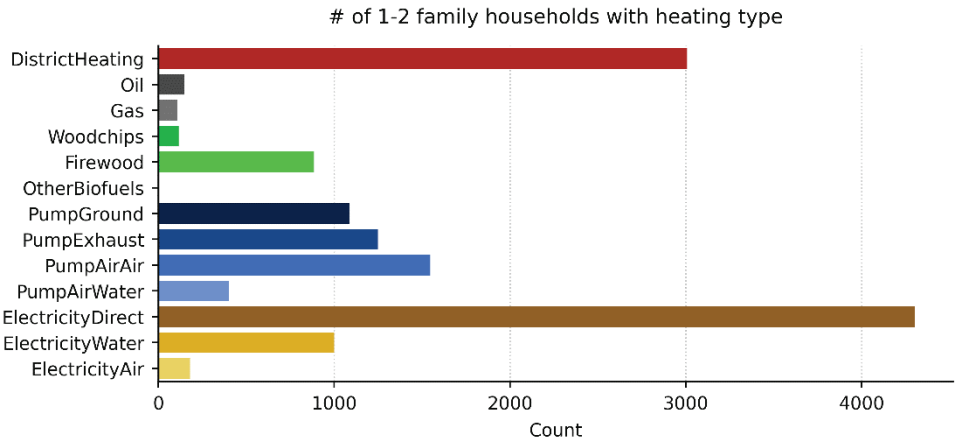


Figure 2.4 Number of 1-2 family households in Gothenburg with a specific heating type. Data are based on approved energy declarations between 2010-2019. Data source: [10].

Figure 2.5 shows the utilization of the different heating types based on the annual energy consumption registered in the energy declaration. The utilization of the heating types in Gothenburg is comparable to the entire Sweden, see Figure 2.3.

However, for buildings with firewood or electricity-to-air, it contributes less to the building's heating supply compared to entire Sweden. There are only a few buildings with Other biofuels, hence the stepwise distribution.

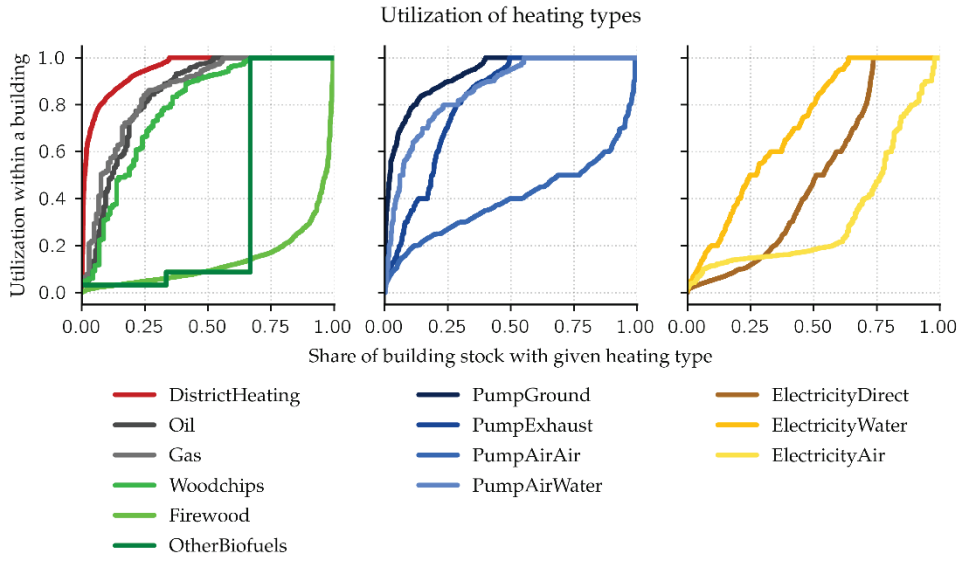


Figure 2.5 Utilization of different heating types for 1-2 family households in Gothenburg. Data are based on approved energy declarations between 2010-2019. Data source: [10].

2.5.2 Heated area

The share of the primary heating types as a function of the heated area can be seen in Figure 2.6. The primary heating type is here defined as the heating type that consumed the most energy in a year based on the measurements in the energy declaration. As the building size increase, the share of full-electric based heating {direct electricity, electricity-to-water, electricity-to-air} as the primary heating type is reduced, where instead exhaust air heat pumps are more common.

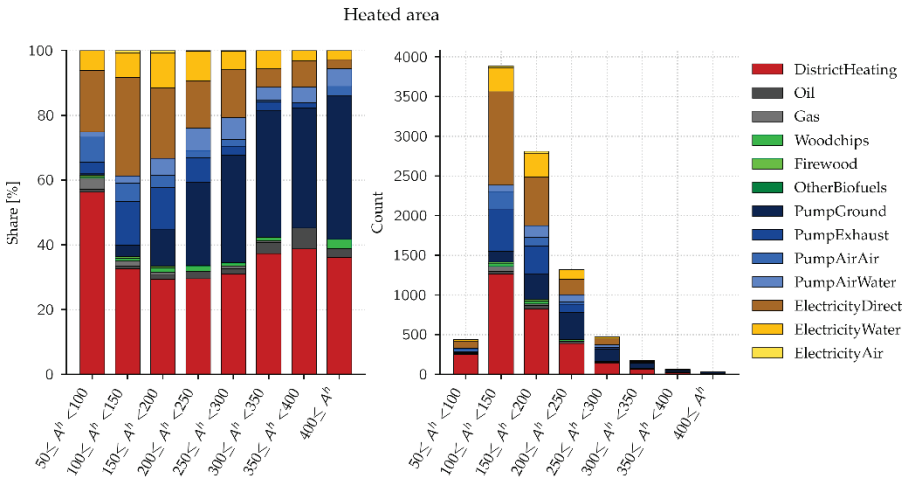


Figure 2.6 Heated area differentiated by the type of heating. Data based on approved energy declaration from 2010-2019 for 1-2 family households in Gothenburg. The colors represent the primary heating type used in the building. Data source: [10].

2.5.3 Building age

Figure 2.7 shows the number of 1-2 family households with a specific heating type given the decade when the building was built. From the energy declaration, the annual energy used for each heating type is specified. Based on that, the primary, secondary, and tertiary heating types can be separated. Note, however, that the primary is not necessarily the same as the base heating type. An example of that is the combination of direct electricity and air-to-air heat pump, where the heat pump should be operated as much as possible due to a higher power-to-heat ratio, and the direct electricity covers the remaining heat deficit.

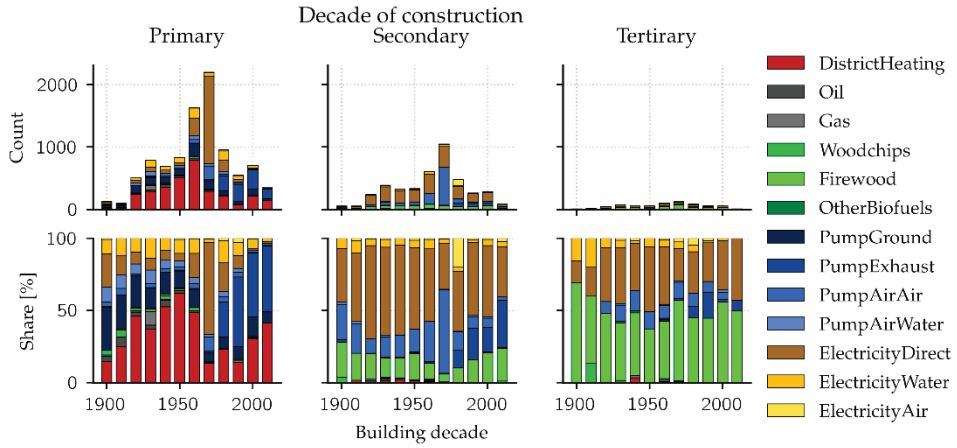


Figure 2.7 Count of 1-2 family households with primary, secondary, and tertiary heating type as a function of the building decade. Data based on approved energy declaration from 2010-2019 for 1-2 family households in Gothenburg. Data source: [10].

During the 70's oil-crises and the increase of nuclear power electricity generation, full-electricity-based heating sources became more and more popular. This trend can still be seen today, where 1-2 family households built in the '70s in Gothenburg are today dominated by direct electricity heating sources as the primary heating source, see Figure 2.7. From the primary and secondary heating types in the figure, it can also be seen that for the buildings built in the '70s, air-to-air heat pumps and direct electricity are often combined. The share of direct electricity as the primary heating source is reduced for buildings built after the '70s, and for villas built in the '90s and onward, direct electricity is seldom used as the main heating source today. Except for the poor power-to-heat ratio, there is also a limitation today on the installed capacity of the electrical appliances for space and tap water heating, including heat pumps driven by electricity [13].

2.5.4 Outdoor air temperature

The annual outdoor air temperature profile for Gothenburg can be seen in Figure 2.8.

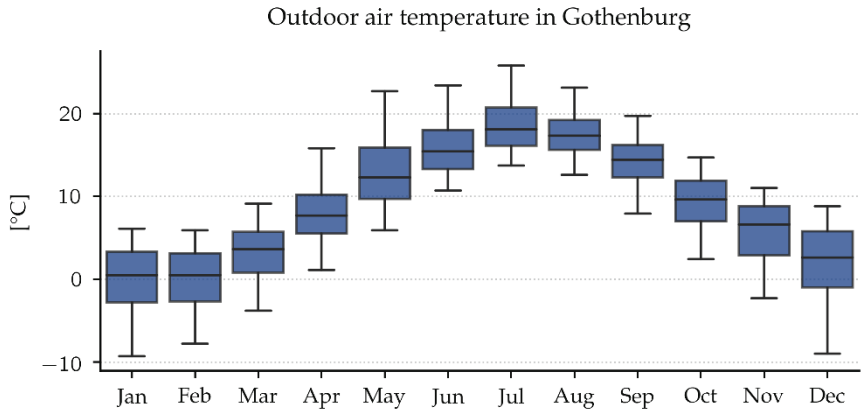


Figure 2.8 Boxplot of outdoor air temperature in Gothenburg, including the years 2009 to 2018. The boxplot shows the 25th, 50th, and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data source: [9]

2.5.5 Fuse size

Figure 2.9 shows the acquired fuse size for 1-2 family households with a valid energy declaration in Gothenburg. Most of the households (~97%) in the data set have a fuse size connection between 16 and 35 amperes. At GENAB, customers with a fuse size less than 63 amperes have the same grid tariff; whereas customers with a fuse size more than, including, 63 amperes have another grid tariff [8].

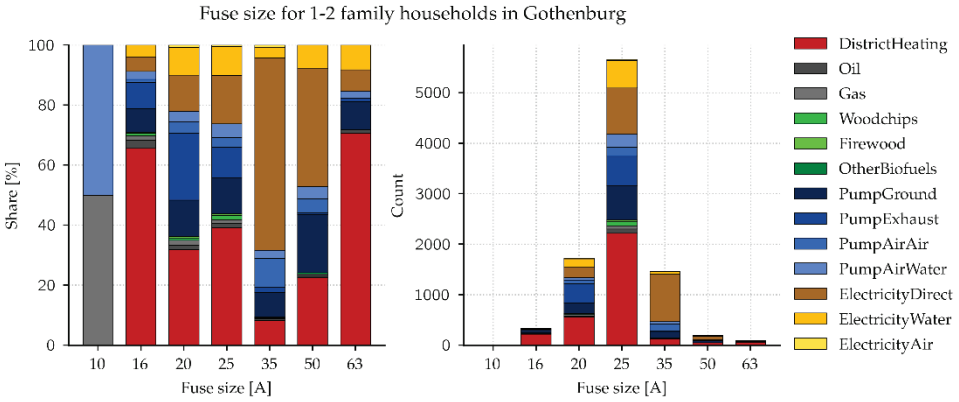


Figure 2.9 Acquired fuse size for 1-2 family households with a valid energy declaration in Gothenburg. The colors represent the main heating type used in the building. Data source: [8]

3 Classification model framework

"A computer program is said to learn from experience E with respect to some class of task T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [14]

This is a widely cited formal definition of machine learning. In supervised machine learning, the task T is to map an input X to an output Y, where the input $X \in \mathbb{R}^d$ is a d-dimensional feature vector and Y is referred to as label. Classification, which is a type of supervised machine learning, deals with categorical output, e.g. assigning a given load/building with the label district heating, exhaust air heat pump, or direct electricity as the main heating source, see Figure 3.1. Figure 3.2 shows the general framework used in this report to classify the electricity consumer's heating system by using a machine learning method.

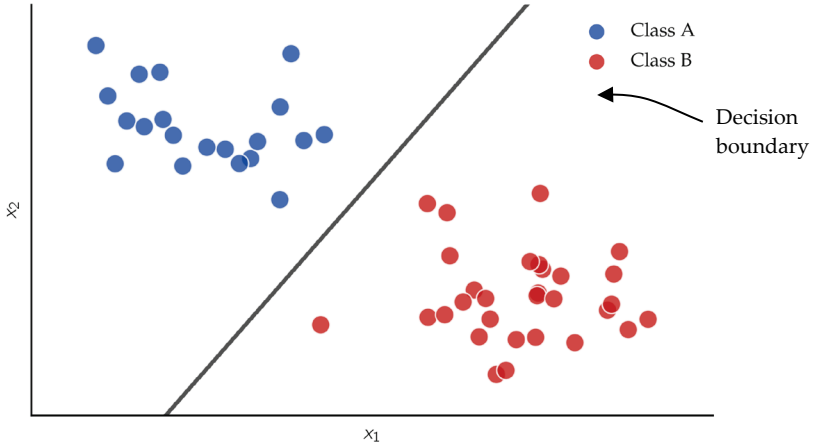


Figure 3.1 Classification of two classes, Class A and Class B, in a 2-dimensional feature space, given the two features x_1 and x_2. For a new unseen sample $x^{(i)}$, it is classified as $y'^{(i)}$ = Class A if to the left of the decision boundary, and $y'^{(i)}$ = Class B if to the right.

The input data to the classification model (classifier) is in time-domain, i.e. smart meter and outdoor air temperature time series. After removing non-representative observations/customers, features $\boldsymbol{x}^{(i)}$ are extracted from the time series, where $\boldsymbol{x}^{(i)}$ is a d-dimensional feature vector of consumer i. Define a set of N input-output pairs $\{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}),\ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})\}$, where $y^{(i)}$ is the corresponding heating type (label) of the i^{th} consumer. The input-output pairs are split into a training S_k^{train} and test set S_k^{test}, where K-fold cross-validation (resampling) is used to evaluate the classifier. k represents the split at the k^{th} fold. The machine learning method seeks a function that maps the feature vector $\boldsymbol{x}^{(i)}$ to the given label $y^{(i)}$ from the training experience E, i.e. the training set. For an unseen sample $\boldsymbol{x}^{(i)}$, the model predicts the output $y'^{(i)}$. From the test set, we know the true label $y^{(i)}$ which is used to compare to the predicted output $y'^{(i)}$. The performance on the unseen data, i.e. if $y^{(i)} = y'^{(i)}$, indicated the generalization properties of the model.

The model parameters of the machine learning method, also called hyperparameters, are first tuned with a grid-search and an L-fold cross-validation.

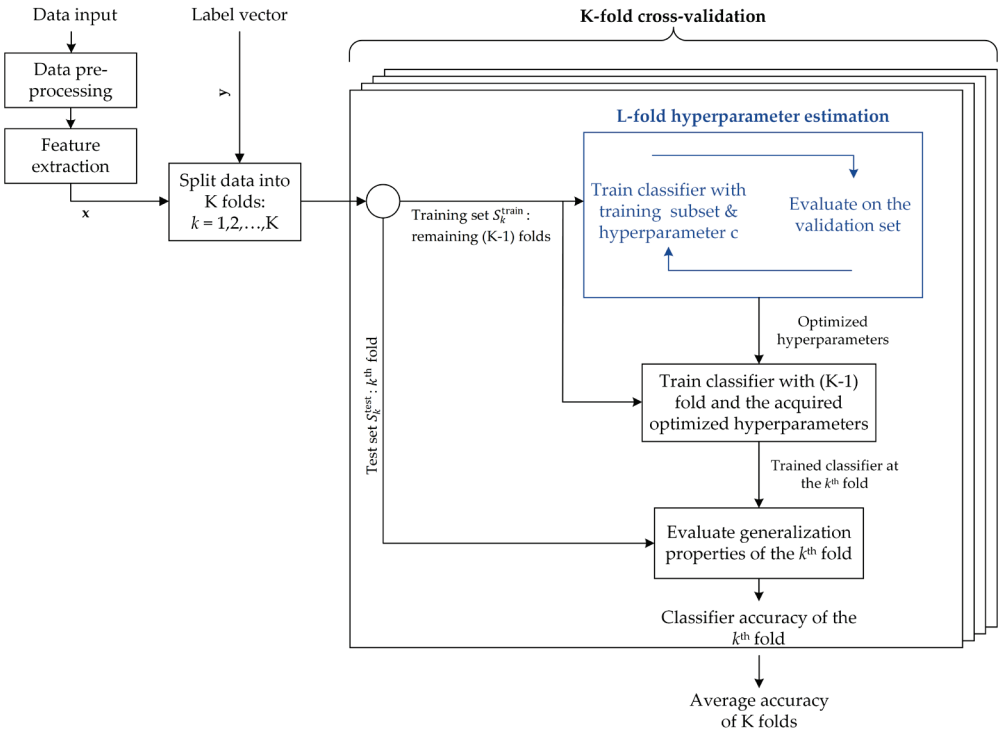


Figure 3.2 Framework for classifying electric consumer's heating type.

approach. With the optimized hyperparameters, the classifier is trained with the complete training set before being evaluated on the test set. Before the training and hyperparameter tuning, each feature is normalized with the sample mean and standard deviation of a given feature in the training set.

The data can also as be transformed to other domains, e.g. frequency domain, as a pre-processing step before the machine learning classifier. The model can also be further developed by including optimization of extracted features, e.g. by feature selection. These two are however not further analyzed in this work.

3.1 DATA PRE-PROCESSING

This step aims to correct or remove data that is incorrect or in other ways not representative of an active load, for example, outliers and missing values. Two types of characteristics that are reoccurring in the smart meter data are a change of sampling frequency and trailing zeroes. In Figure 3.3, two examples are given of trailing zero, or close-to-zero power consumption. To the left, there is a close-to-zero power consumption over a long period, which indicates that all/most of the electrical appliances in the household are shut down, or that the meter is inactive/faulty. Such data are not representative of an active load and would therefore influence the accurate modeling of the classifier. To the right: zero power consumption can also indicate a negative net consumption if the consumer has for example solar PV, and where the electricity production is behind the meter. If the smart meter has two different recordings, one for net consumption and one for net production, the net consumption would appear zero during periods of net production. Behind-the-meter electricity production would influence the

classification of loads. For this analysis, prosumers are excluded, but for the classification of all consumers, one could model the load part.

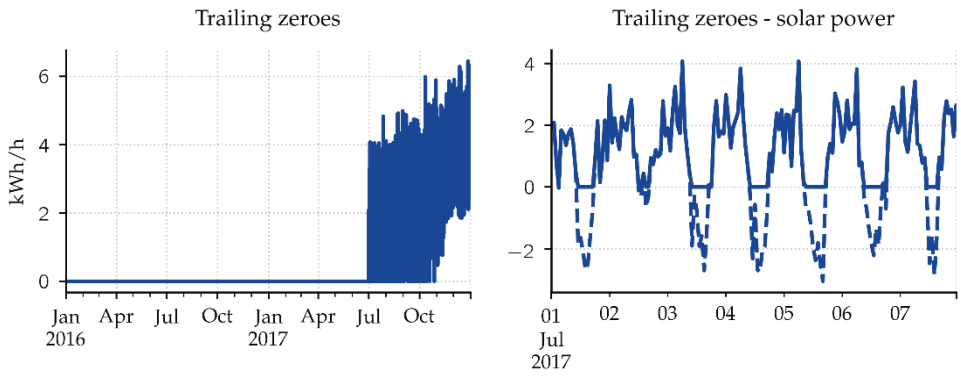


Figure 3.3 Examples of trailing zero power consumption. Left: for a period of time, Right: daily reoccurring pattern

An example of a change in sampling frequency can be seen in Figure 3.4. This could occur if the smart meter/automatic communication system is faulty and the data is downloaded from the smart meters manually, e.g. every 24 hours. The trend in the average consumption is still captured, however, information regarding peaks and intra-day variations are lost. This could for example be modeled or simply be removed if the period is for a short time. As this is out of the scope in this report, the data are removed from our analysis.

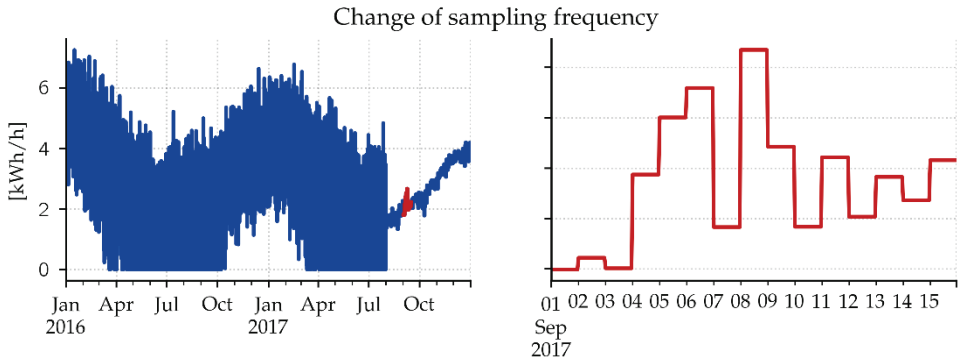


Figure 3.4 Change of sampling frequency due to faulty smart meter/automatic communication system

3.2 FEATURE EXTRACTION

As machine learning is data-driven, the key to a successful result for any machine learning task lies in the data. Ideally, only features[1] that are useful and that can improve the classification model, i.e. the classifier, to predict the correct class is used. With an overrepresentation of features, it can increase the complexity of the classifier, increase the computational cost/time, and/or it can cause a phenomenon called the curse of dimensionality. That is, the model starts overfitting the training data, and the performance on the test data is reducing. Some machine learning algorithms are more prone to the curse of dimensionality than others, for example, k-nearest neighbors (k-NN) [15]. Extracting the *key features* that describe the

[1] Feature – an individual measurable attribute

different classes is therefore essential. To extract key features from a time series, statistical measures and domain knowledge or automatic tools can be used

In this report, a feature-based representation of a time series is used where the data are analyzed at different time resolutions of an annual timescale. An annual timescale is selected where the idea is to see each year if the consumers have changed their consumer class. The effect of complexity and curse of dimensionality are analyzed further in the results

3.3 CROSS-VALIDATION AND HYPERPARMAETER TUNING

For supervised machine learning, the data is split into a training and a test set. The training set is used to train the classifier and represents the known samples. The test set represents the unknown sample and has not been seen by the classifier before. The performance of the classifier is evaluated on the test set, which gives an unbiased estimation of how well the model generalizes on the unseen data. Cross-validation is an approach to resample the train/test dataset to get a more stable estimation of the model's performance, which reduces the impact of one individual train/test dataset split. A common approach is K-fold cross-validation, where the data is split randomly into K equal-sized, and non-overlapping, subsamples [15], see Figure 3.5. Each fold/subsample is used as a test set exactly once. From the cross-validation, the average and variance of the classifier performance are obtained. Note that for each fold, the classifier is re-trained with the corresponding training set and optimized hyperparameters. In this way, the corresponding test set has not been seen by the classifier before and it has not been used for the tuning of the hyperparameters.

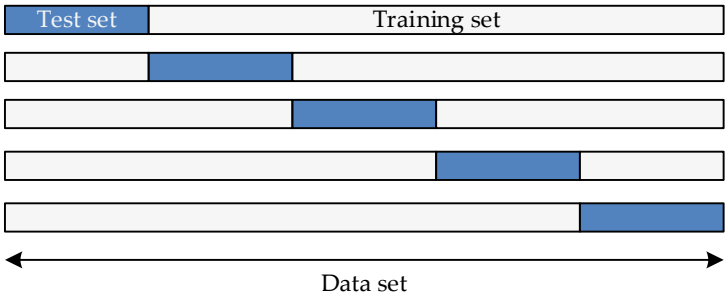


Figure 3.5 Schematic over a k-fold train/test split with $k = 5$ folds.

In machine learning, there are often so-called hyperparameters to be defined for the classification model before the actual training, such as the degree of the polynomial in polynomial regression. In other words, the hyperparameters are part of the model selection task. In this report, a simple grid-search approach is used. That is, all combinations for a finite set of hyperparameter values are evaluated. Note that with a coarse grid, the optimal value can be missed. On the other hand, with a finer grid, the calculation time/costs increase with it. However, using the entire training set for the grid search can cause a bias in the model. L-fold cross-validation (same principal as K-fold cross-validation) is used to reduce this bias in model development. The training set is further split into a training subset and a validation set, see Figure 3.6. The validation set is a holdout set that is not used for training the classifier within the model parameter estimation. The parameter that minimizes the validation set error is selected.

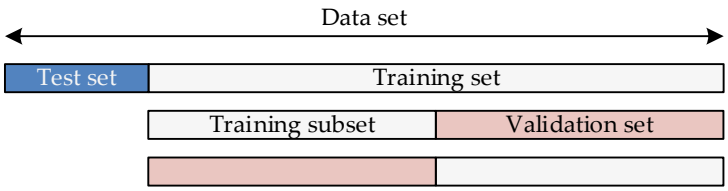


Figure 3.6 Schematic over a L-fold training subset/validation split of the data with $L = 2$ folds. The training set is split into a 50%/50% split where the training subset is used to train the classifier to estimate the optimal model hyperparameters, and the validation tests the performance of the selected hyperparameters.

When the hyperparameters have been selected, a final classifier is retrained on the entire training set with the optimized. Note that the optimal hyperparameter search is performed for each feature component and each k-fold, hence the optimized hyperparameter values are not necessarily the same for each k-fold. The pseudo-code for the K × L-fold cross-validation with hyperparameter tuning can be seen in Algorithm 1.

Algorithm 1: K × L-fold cross-validation with hyperparameter tuning

Input:
  Feature input-output pairs S: $\{(\boldsymbol{x}^1, y^1), (\boldsymbol{x}^2, y^2)$,..., $(\boldsymbol{x}^{\mathrm{N}}, y^{\mathrm{N}})\}$
  Hyperparameter combinations C: $\{\boldsymbol{c}^1, \boldsymbol{c}^2, \dots, \boldsymbol{c}^N\}$
Output:
  The average classification performance of using K-fold cross-validation approach
Algorithm:
  Split S randomly into K equal folds, $k = 1,2, \dots, \mathrm{K}$
  for each fold k in K (*outer loop*) do:
    Define fold k as the test set S_k^{test} and remaining $\mathrm{K} - 1$ folds as the training set S_k^{train}
    Split S_k^{train} randomly into L equal folds, $\ell = 1,2, \dots, \mathrm{L}$
    for each parameter combination c in C do:
      for each fold ℓ in L ($inner\ loop$) do:
        Define fold ℓ as the validation set
        Train the classifier for hyperparameter tuning on remaining $\mathrm{L} - 1$ folds given c
        Evaluate the performance of the classifier on ℓ^{th} fold
      end
      Calculate the average performance of hyperparameter tuning using L-fold cross-validation approach given c
    end
    Train the classifier with S_k^{train} with the c which shows the highest performance in the inner loop
    Evaluate the performance of the classifier on the k^{th} fold, S_k^{test}
  end
  Calculate the average performance of the classifier using K-fold cross-validation

3.4 CLASSIFICATION MACHINE LEARNING METHOD – SUPPORT VECTOR MACHINE

There are numerous machine learning methods used to develop classifiers. Which classifier is the best depends on the task and the given data. In this report, a support vector machine (SVM) will be used to develop the classifier for identifying consumer heating types. The support vector machine (SVM) is a supervised learning method that often shows good performance in various classification tasks [16], and is often considered to be one of the best off-the-shelf methods to develop classifiers [15].

3.4.1 Support vector machine

The SVM-based classifier is a binary classifier, where the aim is to find a hyperplane that best separates the two classes. More specifically, it seeks the hyperplane that gives the largest margin between the two classes, where the distance of the margin is $\frac{2}{\|\boldsymbol{w}\|}$. The hyperplane is defined as

$$b + w_1 z_1 + w_2 z_2 + \cdots = 0$$

$$\boldsymbol{w}^T \boldsymbol{z} + b = 0,$$

where $\boldsymbol{w}$ is a normal vector to the hyperplane, $\boldsymbol{z}$ a set of points, and b a constant. For a 2-dimensional feature space, the hyperplane is a straight line. Compared to a maximum margin classifier, which only allows linearly separable classes, SVM is relaxed by allowing some of the training data to violate the margin. The samples that violate the margin are penalized by $C \cdot \xi_i$, where C is a hyperparameter and ξ_i a slack variable. By allowing errors in the training set, it is more robust against individual training samples. The objective function of the SVM can be defined as [17],

$$\min_{w,b,\xi} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{\mathrm{N}} \xi_i$$

subject to

$$y^{(i)} \cdot \ (\boldsymbol{w} \cdot \boldsymbol{x}^{(i)} + b) \geq 1 - \xi^{(i)}$$

for $i = 1, \ldots, \mathrm{N}$ and $\xi_i \geq 0$. N is the number of training samples.

If $\boldsymbol{x}^{(i)}$ violates the margin $\xi_i \geq 0$, else $\xi_i = 0$. The value of ξ_i increases as the $\boldsymbol{x}^{(i)}$ is further away from the "right side" of the hyperplane. Thus, it is penalized harder than a sample that is close to the hyperplane. The regularization parameter C trades off a wide margin and training accuracy. With a small C, a large margin is encouraged, which leads to a simpler decision function. A large C allows for a high training accuracy with a more complex decision boundary. With a too small value of C, there is a risk for underfitting, whereas, with a too large value, there is a risk for overfitting. Both under and over fitting cause generalization issues for unseen data.

For non-linear separable classes, the feature vector $\boldsymbol{x}^{(i)}$ is mapped from the feature space to a higher-dimensional space. The SVM seeks a hyperplane that best

separates the classes in the higher dimensional space. In this report, a radial basis function (RBF) kernel is used to map the feature vector into a higher dimensional space. The kernel is a Gaussian function and is given as [17] [18]

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

where $\|x - x'\|$ is the Euclidean distance between two points,

$$\gamma = \frac{1}{2\sigma^2},$$

and where σ is the radius of the Gaussian function. The RBF is used as a similarity measure to compare points in the feature space. With $\|\boldsymbol{x} - \boldsymbol{x}'\|$, where $\boldsymbol{x} \neq \boldsymbol{x}'$, the points are considered to be closer to each other if a small γ (large radius σ) is used compared to a large γ. With a large γ (a small radius), the points need to be close to each other in order to be considered to be similar.

An example of the hyperparameters γ and C impact on the decision boundary can be seen in Appendix A. For further detail about the algorithm and implementation, the reader is referred to [17], [18].

3.4.2 SVM with multiclass classification

The SVM classifier is designed for binary (two-class) classification problems. In order to solve a multi-class classification problem, multiple binary SVM classifiers are often combined. Two popular approaches are one-versus-rest (OvR) and one-versus-one (OvO). In an OvO approach, all pairs of classes are evaluated one-by-one, where the number of pairs evaluated is $\binom{\mathrm{M}}{2}$, where M is the number of classes. The most frequent assigned class in the pairwise test is assigned to an unseen sample (consumer) $\boldsymbol{x}^{(i)}$ [15].

In an OvR approach, each class in M is compared against the remaining $\mathrm{M} - 1$ classes. The number of binary classifiers is then M if $\mathrm{M} > 2$. The prediction of an unseen sample (consumer) $\boldsymbol{x}^{(i)}$ is classified according to the classifier that gives the highest probability score for that sample belonging to the j^{th} class [15].

The choice of approach for multiclass classification is a part of the model selection. In this report, an OvR method was chosen.

3.4.3 Unbalanced classes

Unbalanced classes are when the number of samples (customers) for each class are different. Depending on the data set and the performance metric, it can have an impact on the generalization performance for each class. For example, if the classes are very skewed, classifying all samples as the most dominant class could give the best performance, however, the generalization of the other classes would be oblivion. To change the importance of class j, a weight factor w_j can be added to the cost parameter c to give a higher/lower cost for a given class, where the cost C_j for class j can be defined as [18]

$$C_j = w_j \cdot C,$$

and where $w_j > 0$.

Thus, a sample of class j that violate the margin are penalized by $C \cdot w_j \cdot \xi_j$. By increasing the weight factor for the j^{th} class, and thus the cost constant C_j for that class, the class is given higher importance. For balanced importance between the classes,

$$w_j = \frac{\mathrm{N}}{\mathrm{MN}_j}$$

can be used, which is inversely proportional to the number of samples of a given class j. The number of samples in the training set is denoted with N, the number of samples in class j is denoted as N_j, and M is the number of classes. Note that if the classes are balanced, $\mathrm{N}_1 = \mathrm{N}_2 = \cdots = \mathrm{N}_M$, the weight becomes one for all classes, which is the same as using no weighting factor.

3.4.4 Scaler/normalization

Before training the classifier, the data is scaled such that each feature (in the training set) have zero mean $\bar{\boldsymbol{x}}$ and unit standard deviation $\boldsymbol{s}$

$$\boldsymbol{x}' = \frac{\boldsymbol{x} - \bar{\boldsymbol{x}}_{train}}{\boldsymbol{s}_{train}}$$

For an unseen sample $\boldsymbol{x}^{(i)}$ from the test set, the sample is normalized with the values obtained from the training set.

3.4.5 Evaluation

Performance metrics are an important aspect of evaluating, comparing, and selecting suitable classifiers, including the choice of machine learning methods, hyperparameters tuning, etc. The choice of performance metric depends on the aim of the classification. In this report, a single number accuracy metric is used for comparison, i.e. the total accuracy, as the metric works for multi-class problems and it does not rate the importance of the different classes. The accuracy is described as

$$accuracy = \frac{correctly\ classified\ samples}{total\ number\ of\ samples}.$$

4 Case study results and model assessment

For the analysis in this project, the following selection for the case study was made:

- 1-2 family households with only *one* smart meter
- Buildings with only *one* heating type
- Customers with a fuse size between 16-35 amperes.

Moreover, three types of heating types are considered: district heating, exhaust air heat pump, and direct electricity, with 1811, 613, 1070 consumers/buildings respectively. The classification is evaluated on two respectively three classes. That is, for two classes, consumers with district heating and electricity-based heating sources are classified, where the electricity-based heating source includes exhaust air heat pump and direct electricity. For three classes, district heating, exhaust air heat pump, and direct electricity are classified. For the aggregation of other heating types, e.g. different heat pumps, the different types might for example have different efficiencies, installed capacity, utilization when multiple heating types are used, etc. Some heating types are also more often combined with other heating types as seen in Figure 2.5, e.g. air-to-air heat pump with direct electricity. The choice of the level of detail of the consumer categories also depends on the end-use purpose.

Extracting the key features from the smart meter measurements that discriminate the different classes is key for successful classification. In this project, we will analyze the smart meter data at different time resolutions to see the effect on classification accuracy. The analysis can be used to further develop the features or to extract the key information from the features that have the largest impact on classification accuracy. The analyzed features include electricity consumption, with and without scaling to the heated area of the building, and secondly, the outdoor air temperature effect on the electricity consumption is considered. In the end, the error of the classifier is analyzed in more detail.

4.1 ANALYSIS OF SMART METER DATA

In Sweden, for those buildings that have a valid energy declaration, the annual energy usage of different heating types is specified, see Section 2.2. However, detailed information is unknown such as the technology/brand/model of the heating system, if it is an old or a new system, and how it's operated. Furthermore, a consumer may change its heating system to a different class without informing the authorities or the DSO. The heating class of the consumer directly affects its electric power consumption, which is recorded by a smart meter. Hence, power measurement data from smart meters will be used to develop a classification model that aims to classify the heating system of a consumer.

The analysis will start with a few simple features and then increase the complexity and the number of features. This is to show to what extent the classification can be improved as the complexity increases. By increasing the complexity, it can increase the model performance. However, it comes at a cost of reduced interpretability and increased computational cost.

Four properties from the smart meter data are considered: average electricity consumption, the standard deviation of electricity consumption, base electricity consumption, and peak electricity consumption. These features are considered for different time perspectives, capturing the variation of the consumption in time. First, the variation of the year is considered, that is annual, seasonal, monthly, weekly, daily, and hourly variations. Second, the variation between different weekdays, and the hours of the day are considered. Note that these features are in this report treated as static features by the classifier. That is, the time sequence of the time-dependent features is not considered. In Section 4.4, the classification result is presented for a different level of time resolutions is considered.

Other features have been analyzed as the difference and ratio between consumption during winter and summertime, temperature correlation. However, it showed similar or worse results as the analyzed features.

4.1.1 Average consumption

The average electricity consumption $\bar{P}$ is given as

$$\bar{P} = \frac{1}{T}\sum_{t=1}^{T} P_t$$

where P_t the measured consumption at hour t, and T is the number of observations for the given time window, where the time window could be a year, a month, a day, etc.

In Figure 4.1, the average consumption for monthly, day of the week, and hour of the day variations are presented for three different types of heating systems: district heating, exhaust air heat pump, and direct electricity. For visualization, the monthly trend is removed for the day of the week variations, and the monthly and day of the week trend is removed trend for the hour of the day variations, hence the negative values.

The monthly trend shows that the two electricity-based heating sources (heat pump and direct electricity) have a clear seasonal trend, whereas district heating shows only a small seasonal trend. This indicates that district heating and electric-based heating sources are distinguishable by considering the trend of the year, especially in the winter months and early spring/late autumn. However, consumers with exhaust air heat pumps are not distinguishable from consumers with direct electricity heating. For the day of the week variations, all consumer categories are in the same range. Hence, including the day of the week variation is not likely to add value to the classifier. However, the hour of the day variations shows different profiles between the three classes, though overlapping. The largest difference between the classes can be seen in the early morning of the day. Hence, the hour of the day could improve the discrimination of the different consumer classes.

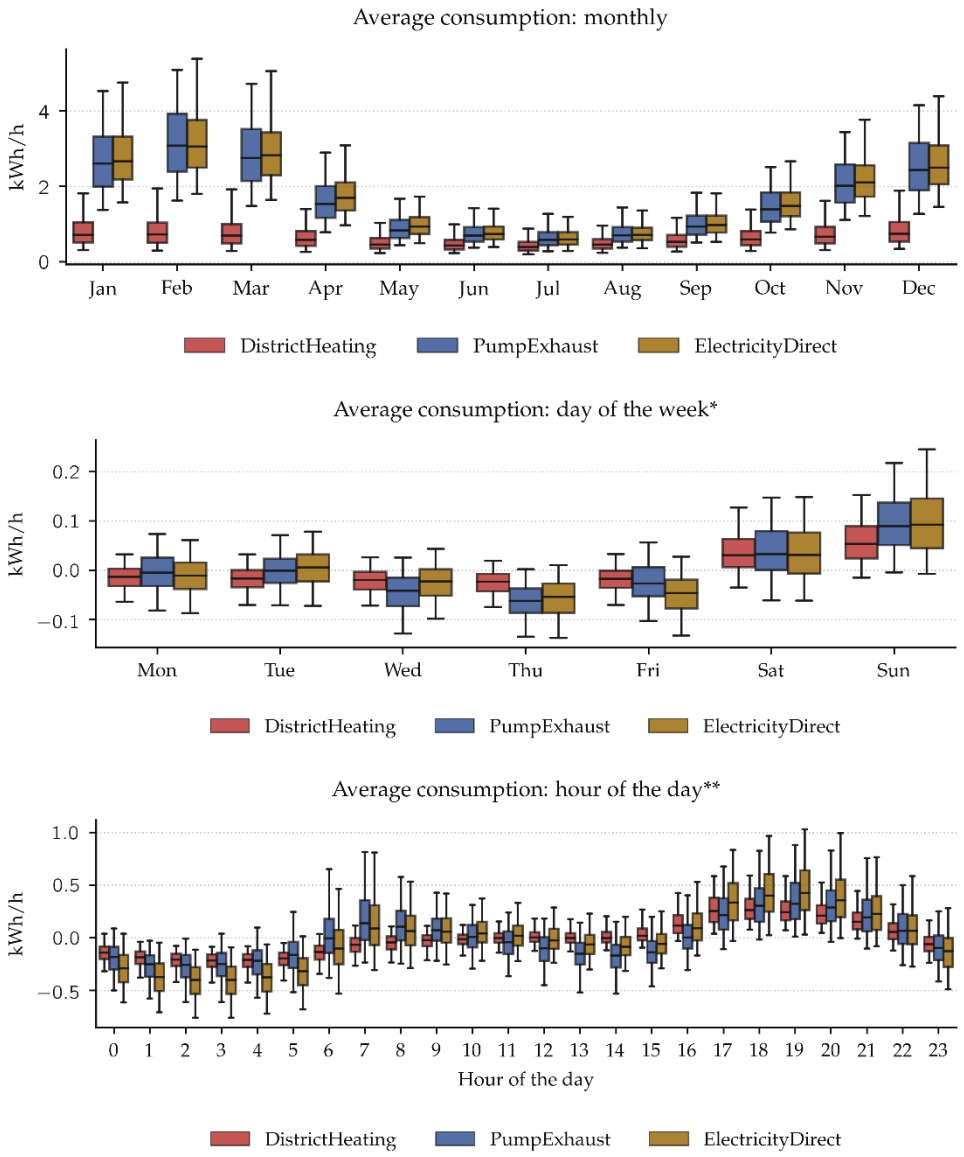


Figure 4.1 The average consumption for one family households in Gothenburg the year 2018 for Upper: monthly variation, Middle*: day of the week variation after *removing* the monthly trend, Lower: hour of the day variation after removing monthly and day of the week trends. The colors represent the different heating sources used in the buildings, where buildings with only one heating system are included. The boxplot shows the 25th, 50th, and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data sources: [8], [10].**

4.1.2 The standard deviation of the consumption

The standard deviation of the electricity consumption shows how much the consumption is changing to the mean, where the sample standard deviation s is

$$s = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(P_t - \bar{P})^2},$$

and where T is the number of observations for the given time window, $\bar{P}$ the mean consumption, P_t the measured consumption at hour t. With an electric load that is

correlated with the outdoor temperature, the standard deviation of the consumption could be increased if there is a temperature shift within the analyzed time window. The standard deviation of the consumption could also indicate that the electric heating systems do not have a constant output throughout the day.

Figure 4.2 shows the average consumption for monthly, day of the week, and hour of the day variations. Similarly, the monthly trend is removed for the day of the week variations, and the monthly and day of the week trend is removed for the hour of the day variations. The district heating shows, in general, a lower variation of the electricity consumption for all considered time resolutions, compared to electric-based heating sources. There is also a clear seasonal trend for the two electricity-based heating systems, where the exhaust air heat pump shows a higher variation during the winter months, and a lower one during the summer month, as compared to direct electricity. This difference between direct electricity and the exhaust air heat pump could, for example, be due to the variations of power-to-heat ratios in the heat pump, or that it is operated differently. The day of the week, however, does not improve the separability between the two electric-based heating source classes. For the hourly variation of the day, it shows a small difference between the classes in the early morning.

4.1.3 Base consumption

The baseload is the electricity that is typically always required for the period. It is here represented as the percentile of the electricity consumption samples for a given time window, where the 5th percentile represents the baseload P_{base}.

$$P_{\text{base}} = p_{.05}(P_t), \quad \forall\, t \in \mathrm{T}$$

where P_t is the measured consumption at hour t, $p_{.05}(P_t)$ the 5th percentile of the electricity consumption, and T the observations for the given time window

In Figure 4.3, the average consumption for monthly, day of the week, and hour of the day variations are presented. As previously, the monthly trend is removed for the day of the week variations, and the monthly and day of the week trend is removed trend for the hour of the day variations. It is mainly the seasonal trend of the base consumption that differs from the previous feature components. That is, the average base consumption of consumers with direct electricity is higher than the average base consumption of consumers with exhaust air heat pumps. Elsewise, the base consumption does not appear to contribute to further discriminate the classes.

Figure 4.3, shows the corresponding results for the base consumption for monthly variations, day of the week variations, and hour of the day variations. In the case of baseload, it is mainly the seasonal trend of the base consumption that differs from the previous feature components, i.e., the average base consumption of consumers with direct electricity is higher than the average base consumption of consumers with exhaust air heat pumps. Otherwise, the base consumption does not contribute to further discriminate the classes.

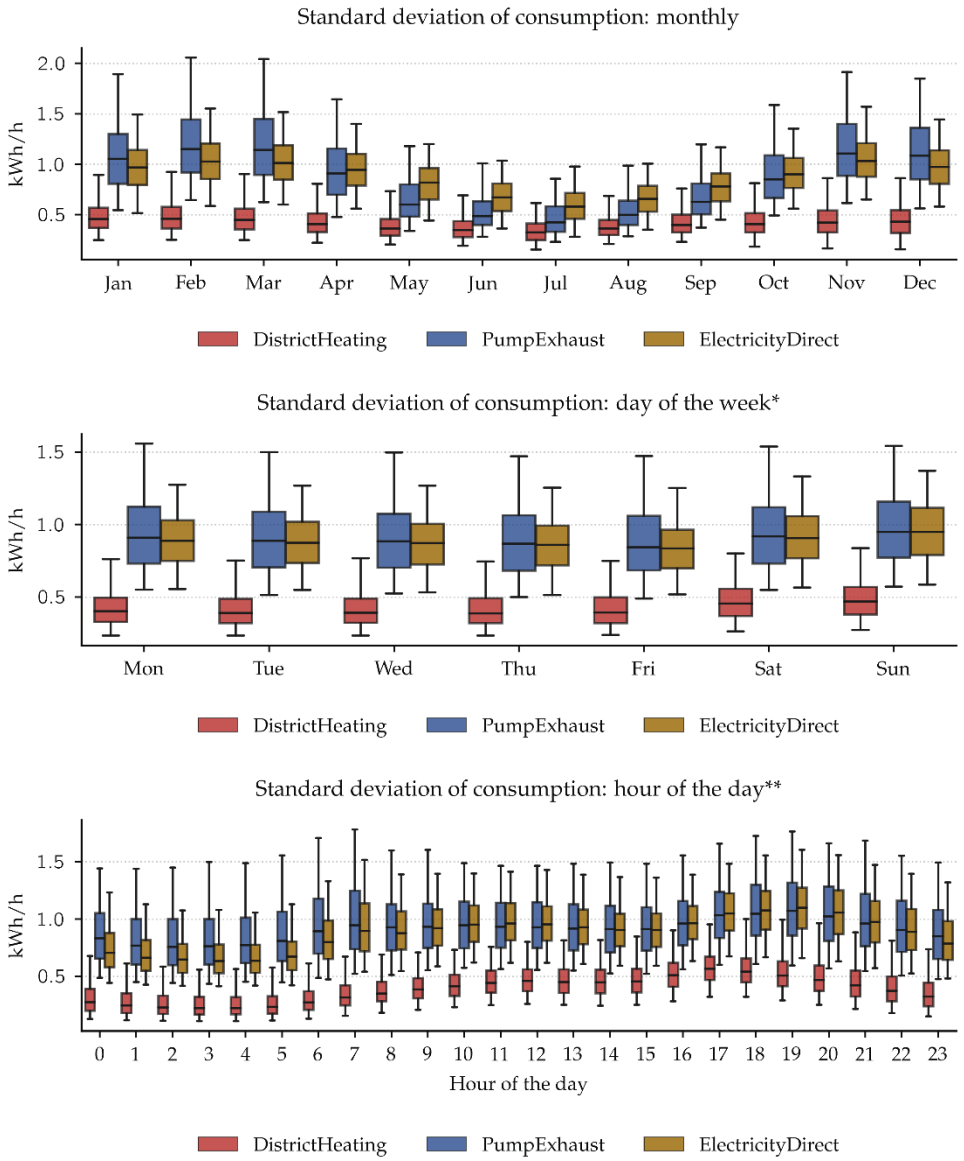


Figure 4.2 Monthly variation of the standard deviation of the consumption for one family households in Gothenburg the year 2018. Upper: monthly variation, Middle*: day of the week variation after *removing* the monthly trend, Lower: hour of the day variation after removing monthly and day of the week trends. The colors represent the different heating sources used in the buildings, where buildings with only one heating system are included. The boxplot shows the 25th, 50th, and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data sources: [8], [10]**

4.1.4 Peak consumption

The peak load P_{peak} is when the electricity consumption demand is high, often for a short period. It is here represented as the 95th percentile of the electricity consumption samples for a given time window,

$$P_{\text{peak}} = p_{.95}(P_t), \quad \forall\, t \in \mathrm{T}$$

where P_t is the measured consumption at time t, $p_{.95}(P_t)$ the 95th percentile of the consumption, and T the observations for the given time window.

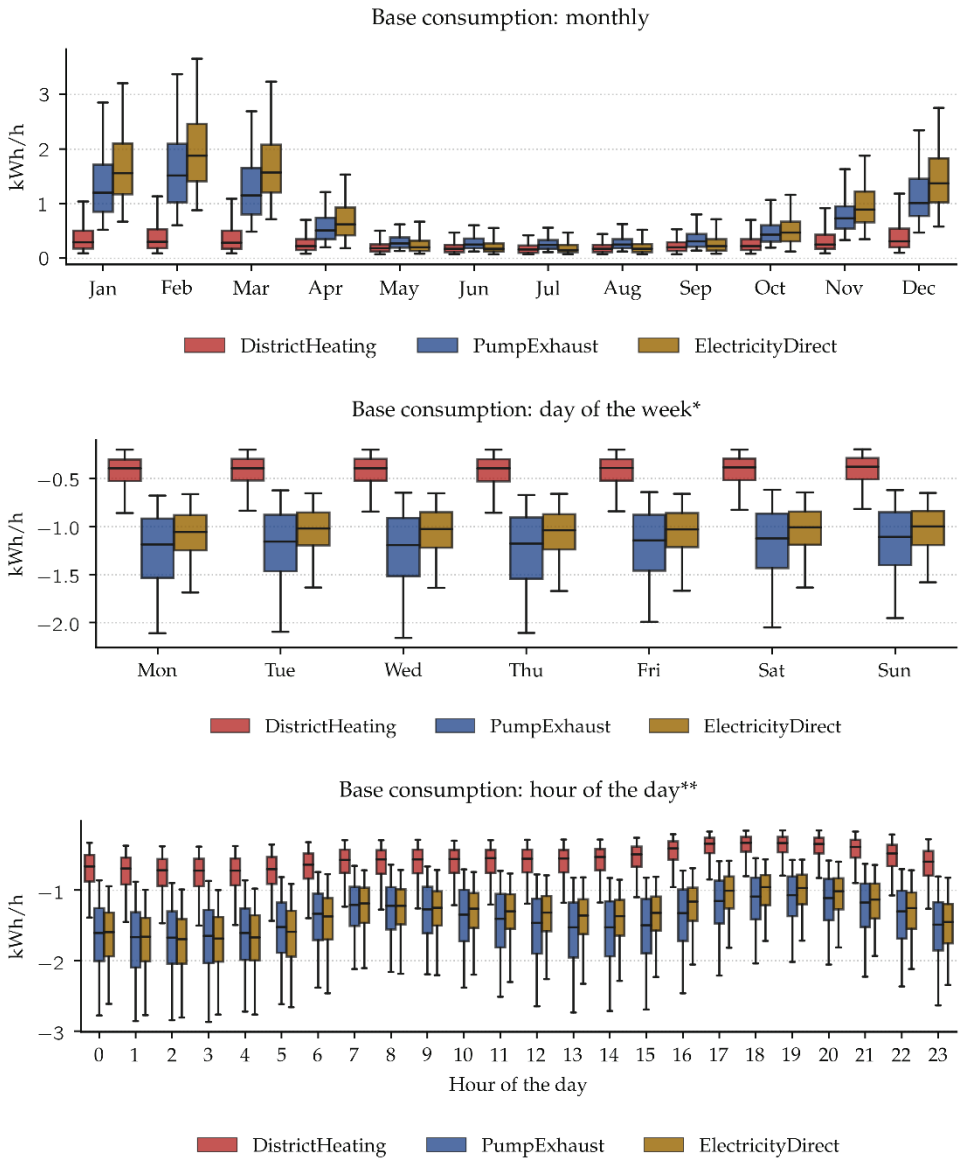


Figure 4.3 Monthly variation of the base consumption for one family households in Gothenburg the year 2018. Upper: monthly variation, Middle*: day of the week variation after *removing* the monthly trend, Lower: hour of the day variation after removing monthly and day of the week trends The colors represent the different heating sources used in the buildings, where buildings with only one heating system are included. The boxplot shows the 25th, 50^{th,} and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data sources: [8], [10]**

Figure 4.4, presents the corresponding results for the peak electricity consumption. In the case of peak load, it is mainly the hourly variations of the day, especially in the morning, that has a different characteristic compared to the previous feature components, where the exhaust air heat pump shows on average a higher night/morning peak compared to direct electricity.

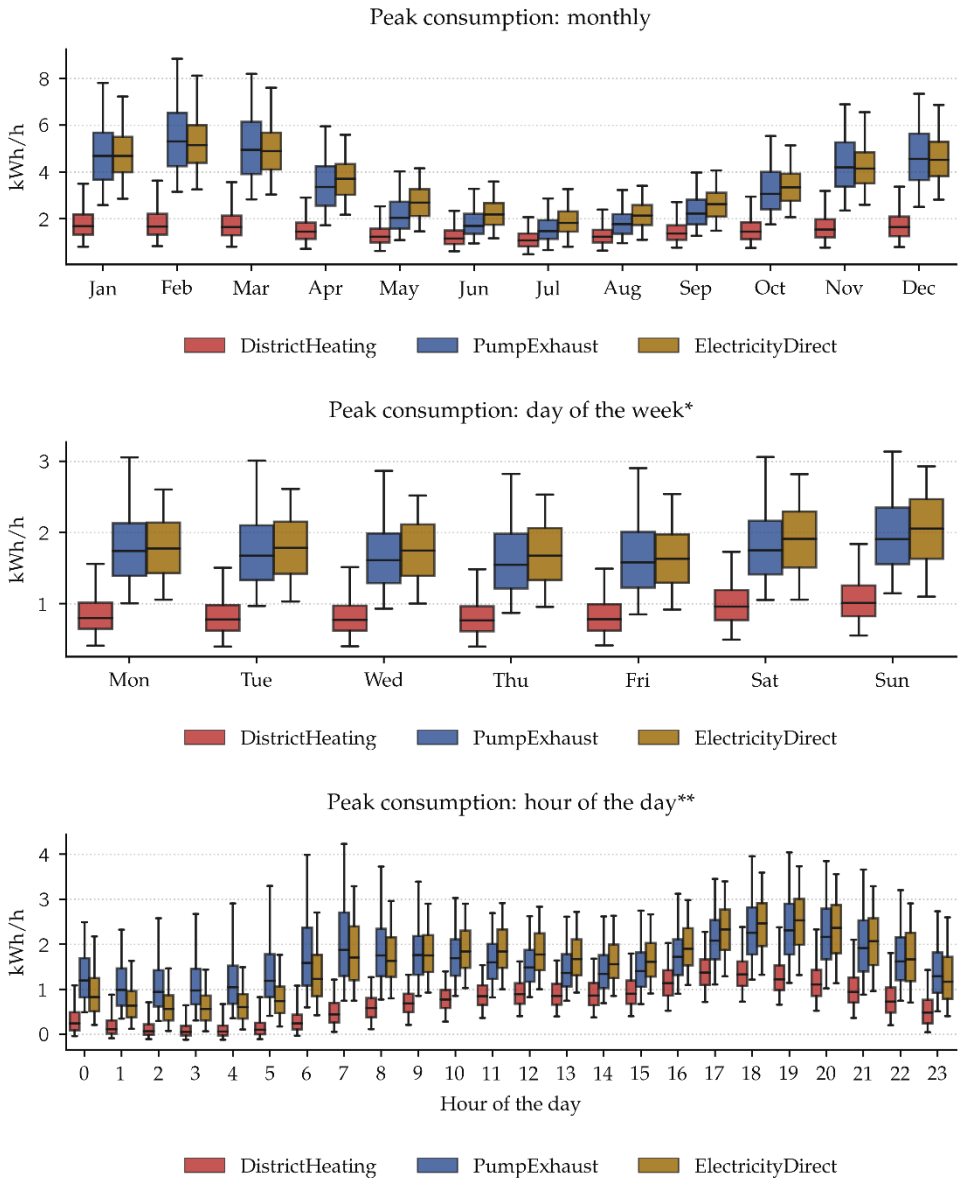


Figure 4.4 Monthly variation of the peak consumption for one family households in Gothenburg the year 2018. Upper: monthly variation, Middle*: day of the week variation after *removing* the monthly trend, Lower: hour of the day variation after removing monthly and day of the week trends The colors represent the different heating sources used in the buildings, where buildings with only one heating system are included. The boxplot shows the 25th, 50^{th,} and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data sources: [8], [10]**

4.2 BUILDING CHARACTERISTICS

4.2.1 Buildings heated area

The heating demand in a building is affected by multiple factors, see Table 2.1, where the heated area is an important factor. This can be viewed in Figure 4.5, where the heating demand shows a positive linear trend with the heated area, i.e. the heating demand increases with the heated area. Note that the heating demand in the figure is based on measurements from the energy declaration. Furthermore,

the smart meters only give the value of total electricity consumption, where electricity consumption for heat demand is unknown. A straightforward way to compensate for the increased demand is to divide the electricity consumption measurements by the heated area. The idea is to make loads of the same class more similar to each other, reducing the spread due to the sizing of the building. However, as the different heating types use a different amount of electricity, there is also a risk to make loads of the same class more unlike each other, e.g. for district heating which is not using electricity as a heating source.

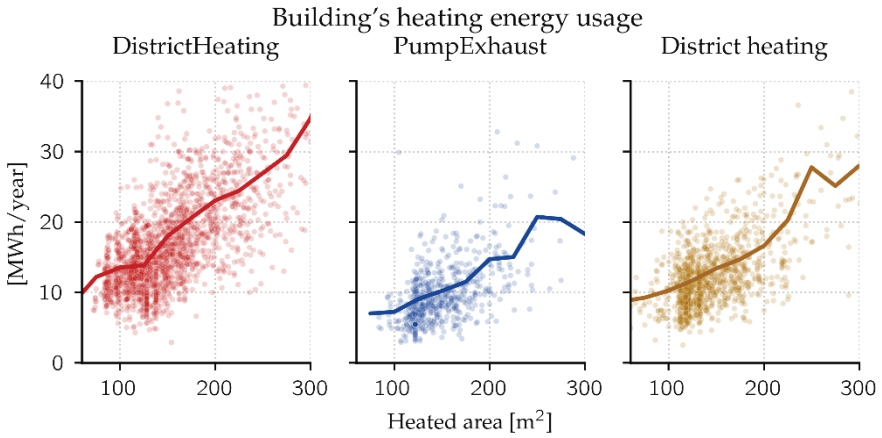


Figure 4.5 The building's heated area against measured annual energy consumption for *heating*, where the lines show the trend. The data is not scaled to a normal year. Data source: [9]

4.2.2 Building year (why not to include directly)

By using the building year, see Figure 2.7 and Figure 4.6, as the only feature for the classifier, a 93.1% test accuracy can be achieved for identifying direct electricity and exhaust air heat pump. These are the two heating types that are more difficult to separate as can be seen in the following sections. However, as one of the purposes of the classification result is to identify if a consumer has changed heating type, there is a risk that the classifier will misclassify consumers that have a building that was built in the '70s, and where they have changed their heating system. Instead, the year of the building could be used indirectly. For example, if buildings of a certain age and heating system are more probable of upgrading their heating system. This is not included in the report.

4.3 MODEL SETTINGS

The classification model settings used for the classification of the different heating types can be seen in Table 4.1. A 5x2-fold cross-validation with 5 respectively 2 folds is used for training, hyperparameter tuning, and evaluation of the model, see Section 3.3. The 5-folds in the outer loop splits the data into train and test sets (80%/20%). The test set is a holdout set and is only used to test the general performance of the classifier. For comparison between classification models, the average performance of the 5-folds is used. The 2-folds in the inner loop split the training set into a training subset and a validation set (50%/50%), which is used for tuning of the hyperparameters C and γ in the support vector machine (SVM). The

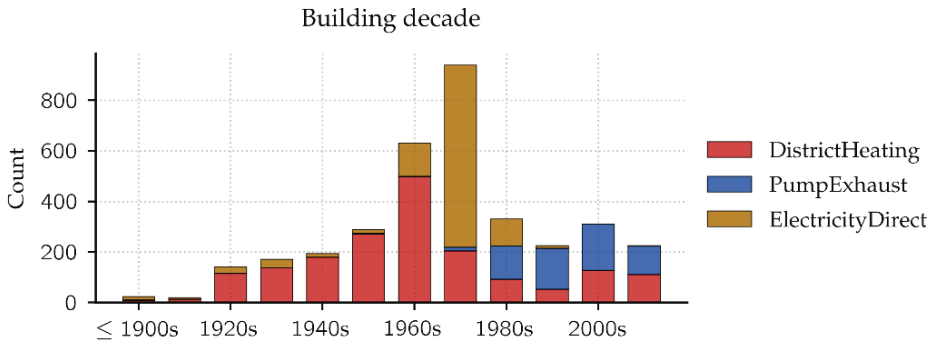


Figure 4.6 Installed heating source sorted after building decade for one family households in Gothenburg. Only buildings with one heating type, including district heating, exhaust air heat pump, and direct electric heating, are included. Data source: [10]

optimal values for the hyperparameters are then used to train the model with the complete training set, before evaluating general performance on the test set. The split into the folds are random, however, the *same* random split should (and is) used when comparing the performance of different models. For further information, see Chapter 3.

Table 4.1 Classification model settings used for the classification of heating types in this work. See Chapter 3 for a more detailed description

Model parameter	Setting
Classifier:	Support vector machine
Multiclass:	One-vs-rest
Class weight	Balanced
Kernel	Radial basis function (RBF)
Feature scaling	Standardization
Train/test split:	Stratified 5-fold cross-validation
Hyperparameter tuning:	Method: Grid-search Split: stratified 2-fold cross-validation $C \in [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$ $\gamma \in [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$
Performance score	Accuracy
Library	LIBSVM [18]

4.4 CONSUMER CLASSIFICATION USING SMART METER DATA AND THE BUILDINGS HEATED AREA

In this section, the effect of different time resolution on the classification result is evaluated, with and without scaling to the heated area. The analyses are divided into different time resolutions of the year, the time resolution of a week, and the time resolution of a day, where each is individually assessed. By increasing the level of detail, the number of features increases, and thereby also the computational cost, the interpretability, and the risk of the curse of dimensionality. On the other hand, by increasing the level of detail, the temperature variation can indirectly be captured. The features from the smart meter data include the mean, peak, base, and standard deviation of the load consumption.

4.4.1 Features reflecting seasonality

In Table 4.2, the result of different time resolutions of the year is presented. The different time resolution of the year include: the entire year, seasons {Winter, Spring, Summer, Autumn}, months {January,…,December}, weeks {week 1, …, week 52}, days {day 1, …, day 365} and hours {hour 1,…, hour 8760}. Note that the smart meter measurements are hourly values, hence the mean, base, peak and standard deviation of the consumption within the hour are not accessible.

Table 4.2 Classification accuracy (average ± std) of the 5x2 fold cross-validation (CV) for different time resolution within a year, with and without scaling to the heated area. The features include the mean, base, peak, and standard deviation of the electricity consumption. Two classes including district heating, electricity-based heating source, where electricity-based heating sources include exhaust air heat pump, direct electricity. Three classes include district heating, exhaust air heat pump, direct electricity. The smart meter data is from the year 2018. Data sources: [8], [10]

Time resolution*	Number of features	CV accuracy: two classes		CV accuracy: three classes	
		No scale	Heated area	No scale	Heated area
Dummy classifier**	-	.517 ± .001	.517 ± .001	.517 ± .001	.517 ± .001
Annual	4 (1y·4)	.943 ± .010	.944 ± .014	.843 ± .017	.839 ± .015
Seasonal	16(4s·4)	.957 ± .008	.954 ± .008	.887 ± .019	.883 ± .011
Monthly	48(12m·4)	.961 ± .011	.963 ± .008	.897 ± .012	.901 ± .009
Weekly	208(52w·4)	.962 ± .006	.966 ± .006	.917 ± .012	.918 ± .013
Daily	1460 (365d·4)	.963 ± .006	.967 ± .008	.915 ± .008	.919 ± .009
Hourly***	8760 (8760h·1)	.889 ± .013	.872 ± .017	.739 ± .012	.722 ± .012

* See Appendix B for an example of feature components with a seasonal time resolution

** Predicts the most frequent class in the training set

*** $C = 100$, $\gamma = 0.001$ was manually selected (based on the experience of the other feature components) due to extensive computational costs

By scaling with the heated area, there is a marginal or no increase in classification accuracy for the different feature components. By simply scaling to the heated area, it would not help to discriminate the loads. Hence, this approach will not be further evaluated in this report. A more sophisticated approach may be needed to account for the heated area. On the other hand, for samples with similar feature values and similar heated areas, no additional information is gained by taking the heated area into account.

Nonetheless, by increasing the level of detail and number of features, the performance generally increases. For two classes, the performance stagnates at a seasonal level. The results also indicate that even though we are increasing the number of features, with an increased possibility of the curse-of-dimensionality, the SVM with an RBF kernel shows to generalizes well on the test set. Though, by using all the individual hourly measurements of the year directly, the performance is drastically reduced. The features are not able to reflect the general characteristics of the heating types.

4.4.2 Features reflecting individual days of a week

At the monthly resolution, it is possible to increase the time resolution by separating data according to weekday/weekend or even Monday to Sunday. In Table 4.3, the result of different time resolutions of the week is presented. It shows that including the time resolution of a week, it does not give a big improvement in the classification accuracy, neither for two-classes nor three classes classification. This is understandable, as it was shown in Figure 4.1-Figure 4.4 that the features did not discriminate the different classes.

Table 4.3 Classification accuracy (average ± std) of 5x2 fold cross-validation (CV) for weekly time resolution with monthly variation as a reference. The features include the mean, base, peak, and standard deviation of the electricity consumption. Two classes including district heating, electricity-based heating source, where electricity-based heating sources include exhaust air heat pump, direct electricity. Three classes include district heating, exhaust air heat pump, direct electricity. The data is from the year 2018. Data sources: [8], [10].

Feature component*	Number of features	CV accuracy Two classes	CV accuracy Three classes
Monthly**	48(12m·4)	.961 ± .011	.901 ± .009
Monthly with Weekday/Weekend	96(12m·2d·4)	.965 ± .009	.905 ± .014
Monthly with Monday-Sunday	336(12m·7d·4)	.962 ± .008	.912 ± .013

* See Appendix B for an example of feature components with a seasonal time resolution

** Results same as in Table 4.2

4.4.3 Features reflecting individual hours of a day

The classification result for the time resolution of the day is presented in Table 4.4. The time resolution of a day includes hour 0 to hour 23 for an hourly measurement level of detail. The same reference case as in 0 is used, i.e. the monthly variation of a year. Comparing with and without hourly variation, there is an improved classification rate for both two and three-classes classification, which shows the highest accuracy noted. However, the number of features is increased by 24 times, which reduces the interpretability and the computational cost. As previously, the results show that the SVM with RBF kernel generalizes well, even by increasing the number of features by 24.

Table 4.4 Classification accuracy (average ± std) of 5x2 fold cross-validation (CV) for daily time resolution with monthly variation as a reference. The features include the mean, base, peak, and standard deviation of the consumption. Two classes including district heating, electricity-based heating source, where electricity-based heating sources include exhaust air heat pump, direct electricity. Three classes include district heating, exhaust air heat pump, direct electricity. The data is from the year 2018. Source: [8]

Feature component*	Number of features	CV accuracy Two classes	CV accuracy Three classes
Monthly**	48(12m·4)	.961 ± .011	.901 ± .009
Monthly with hourly variations	1052(12m*24h·4)	.971 ± .004	.924 ± .014

* See Appendix B for an example of feature components with a seasonal time resolution

** Results same as in Table 4.2

4.5 CONSIDERING OUTDOOR AIR TEMPERATURE VARIATIONS

By only considering the smart meter data, the outdoor air temperature is indirectly captured by the consumer characteristics. The outdoor air temperature time series can also be used directly to capture the outdoor air temperature dependency, which will be analyzed in this section. The daily average outdoor air temperature in Gothenburg for the year 2018 can be viewed in Figure 4.7.

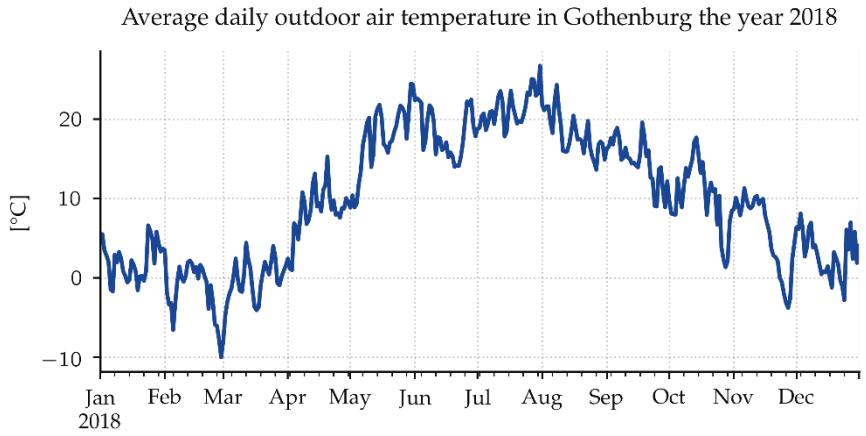


Figure 4.7 Average daily outdoor air temperature in Gothenburg the year 2018. Data source: [9]

4.5.1 Analysis of data

The outdoor air temperature dependency is analyzed by considering the slope of the linear regression between electricity consumption and the outdoor air temperature. That is, how the consumption is changing to the outdoor air temperature. The linear regression line is defined as [19]

$$P(T) = P_0 + p_T T$$

where P is the power, T the outdoor air temperature, P_0 the interception and p_T the slope.

An example is shown in Figure 4.8, where the daily average outdoor temperature is plotted against the daily average electricity consumption for two consumers, one with district heating (left) and one with an exhaust air heat pump (right). In [19], the daily average temperatures above 6°C were excluded for the regression as the regression slope goes to zero at high temperatures. In this report, for the annual data, daily average temperatures above 10°C were excluded to disregard the non-linearity between warm and cold outdoor air temperature. The choice of this threshold could have an impact on the classification; however, this was not verified. However, when splitting the data into seasons and months, days with an average temperature above 10°C is also included.

For extreme cold outdoor air temperature, the consumption can behave non-linear due to an installed max capacity of the heating system, isolation properties of the

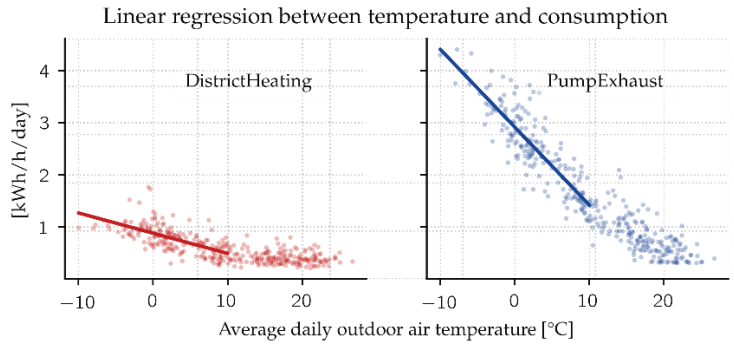


Figure 4.8 Linear regression fit between average daily outdoor temperature (<10°C) and average daily electricity consumption of one consumer with district heating (left), and one consumer with exhaust air heat pump (right). Data sources: [8], [9].

building, reduced efficiency of heat pumps, etc. This can be especially important when the analyzed heating systems include heat pumps that take heat from the outdoor air, where a reduced outdoor air temperature can reduce the coefficient of performance (COP) of the heat pump substantially [19]. However, in this analysis, outdoor air heat pumps are not included, and the outdoor air temperature for the analyzed year and geographical location is not considered extreme, see Figure 4.7.

The linear regression is performed for different time scales: year, seasonal and monthly, with and without hourly variations within a day. This to reduce the effect of other factors that can have a seasonal correlation with electricity consumption, e.g. daytime. The boxplot in Figure 4.9 shows how the slope coefficient of the linear regression is distributed for the different months of the year. During the summer period {June, July, August}, the slope of the regression is similar between the three heating types, which is expected as the space heating is normally not active during this period. During the winter period {January, February, December}, the features show the maximum separability between the classes.

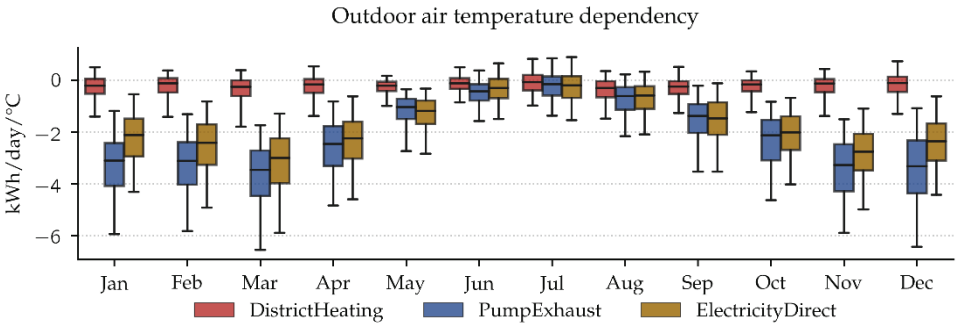


Figure 4.9 The slope coefficient of the linear regression between daily electricity consumption and average daily outdoor temperature. The colors represent the different heating sources used in the buildings, where buildings with only one heating system are included. The boxplot shows the 25th, 50^{th,} and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data sources: [8], [9]

For future work, it is also possible to do a more in-depth analysis of the temperature effect. For example, how the loads behave when it is very cold, how many days in a row there is with cold temperatures, time shift of temperature change and consumption change, etc.

4.5.2 Consumer classification: *temperature*

The classification result by considering the linear trend between electricity consumption and outdoor air temperature is presented in Table 4.5. By only considering the linear regression slope for the annual data (excluding days where the temperature is above 10°C), the 2-class classification shows similar performance as by only considering smart meter data with multiple features. As only one feature, this is easily interpretable and can be used to see if a consumer has changed their heating system from district heating to an electricity-based heating source, or vice-versa. However, with three classes, the performance is poor. This big difference in accuracy between 2 and 3-class classification indicates that the single linear regression slope does not discriminate between exhaust air heat pumps and direct electricity sufficiently well. Increasing the level of detail on the other hand gives a substantial improvement on the accuracy for the 3 classes. Nonetheless, the accuracy is less than by only considering smart meter data.

Table 4.5 Classification accuracy (average ± std) of 5x2 fold cross-validation (CV) by considering the linear regression slope between smart meter data and outdoor air temperature. Without considering the hourly variation, the daily average electricity consumption and outdoor air temperature are used for the linear regression. Two classes including district heating, electricity-based heating source, where electricity-based heating sources include exhaust air heat pump, direct electricity. Three classes include district heating, exhaust air heat pump, direct electricity. The data is from the year 2018. Data sources: [8], [9]

Feature component*	# of features	CV accuracy 2-class	CV accuracy 3-class
Annual**	1	.958 ± .008	.772 ± .009
Seasonal	4	.959 ± .015	.818 ± .011
Monthly	12	.963 ± .008	.863 ± .006
Monthly with hourly variations	288 (24h·12m)	.967 ± .009	.894 ± .003

* See Appendix B for an example of feature components with a seasonal time resolution

** Average daily outdoor air temperature ≤ 10°C

4.6 MODEL AND ERROR ANALYSIS

An error analysis is performed to better understand the misclassifications. Only the classification when all three heating types are used is shown, as this also indicates the result of the two-class classification {District heating, Electric-based heating source}. The feature component with the best performance is selected for this analysis, i.e. smart meter data with hourly time resolution for each month of the year, see Table 6.4. Figure 4.10 shows the confusion matrix with the given feature component. The confusion matrix shows the sum of the individual confusion matrices in the 5-fold cross-validation test set. Pump exhaust has the lowest accuracy for the three classes, where 81.2% is classified correctly, 3.9% misclassified as district heating, and 14.8% misclassified as direct electricity.

Figure 4.11 shows the classifier's probability estimation according to [18] of a sample belonging to a specific heating type, given the input features space. For those heating types that were correctly classified, the model shows a higher

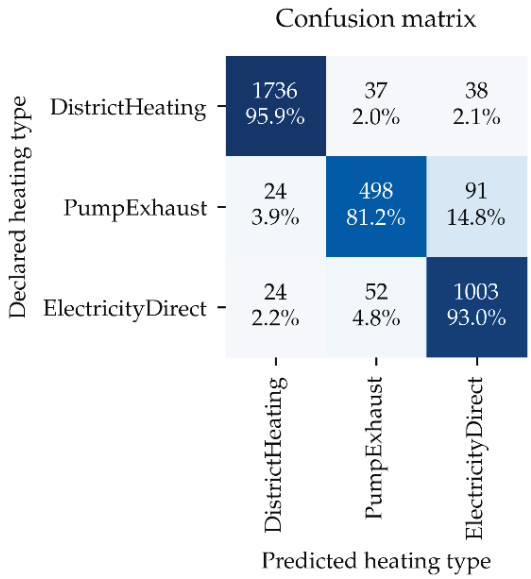


Figure 4.10 The sum of the confusion matrices of the 5-fold cross-validation. The percentages show the shares of the predicted heating type, given the declared heating type.

probability of the sample belonging to the correct class, as compared to samples that were misclassified. For 50% of the district heating class that were classified correctly, the model shows a close to a 100% probability for those samples. This can be compared for the 50-70% probability for incorrect classified samples with district heating as the main heating source. If a sample is incorrect classified, but the classifier shows a high probability for the misclassified class, this could be due to for example that the features are not representative, the classification model is not able to capture the complexity of the feature components, the consumer has changed their class and heating system, etc.

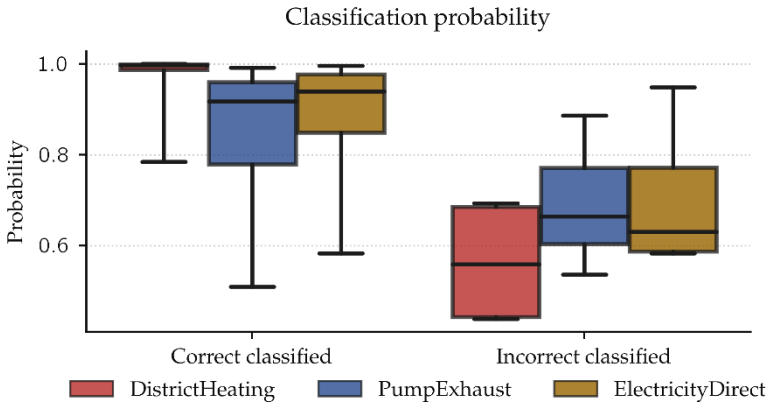


Figure 4.11 The class probability of the test set, with an hour of the day variation for each month of the year for one family household in Gothenburg the year 2018. The boxplot shows the 25th, 50^{th,} and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data sources: [8], [10]

The following error analyses will be broken down into two parts: geographical analysis and energy declaration analysis.

Geographical area

In Figure 4.12, a geographical pie chart of the classification error is presented. Each pie is the representation of correct and incorrect classified samples for a given zip-code. The area of the pie is proportional to the zip code with the largest number of samples (consumers). The six zip codes marked with red stands for 26.7% of the classification error, but only for 9.4% of the overall samples. We will analyze the first and third areas further.

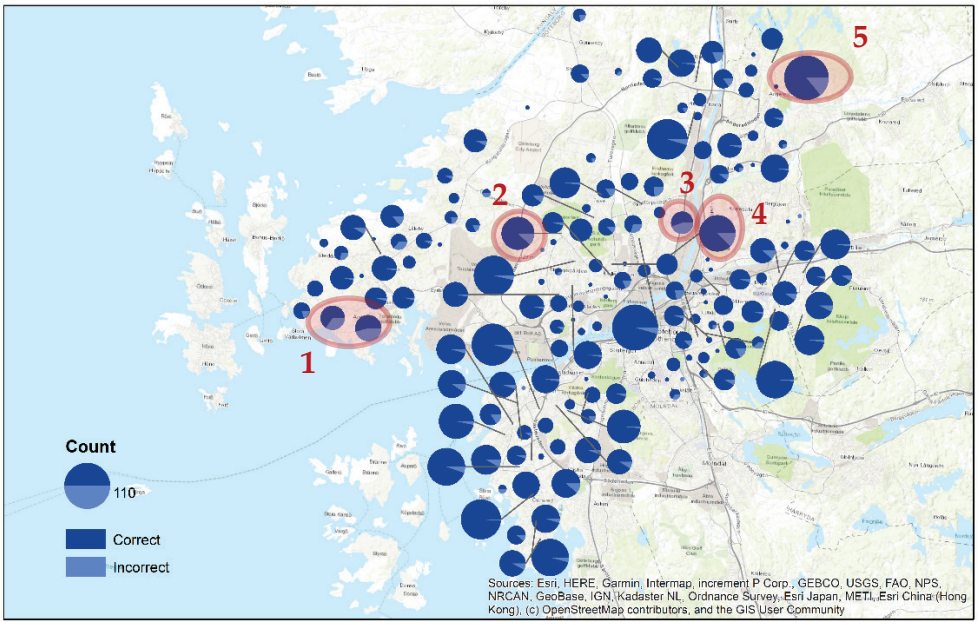


Figure 4.12 Geographical pie chart of correct and incorrect classified samples, where each pie represents one zip-code. The area of the pies is sized proportionally to zip-code with the largest number of samples, i.e. 110 consumers. The red circles represent the zip-codes that includes samples that sum up together as 26.7% of the classification error.

In area 1, 81.2% of the buildings in the data set is declared to have an exhaust air heat pump, where the rest is declared to have direct electricity. 42.9% (24/56) of the buildings with exhaust air heat pumps are incorrectly classified. That is given the declared heating type. This can be compared with the overall test accuracy for the exhaust air heat pump that is 81,2%, see Table 4.4. No factors that explain this have been found and it has to be analyzed further.

In area 3, 9/10 of the consumers with district heating are incorrectly classified. Analyzing this further, it was shown in the energy declaration that there is a shift in heating systems in the year 2015. For energy declarations before 2015, district heating was the most common heating source in this area. For energy declarations after 2015, district heating was not declared as a heating source, where instead exhaust air heat pumps are the most common heating system. A deeper analysis showed that in this specific area, there is a community association with detached and row houses. It was confirmed with Bällskärs Norra Samhällighetsförening's board that in 2015, 136 out of 139 buildings in the community association shifted their district heating to a waterborne exhaust air heat pump system. Hence, there has been a shift in the heating system since the energy declaration was performed for these buildings. The given example shows that the energy declaration can be outdated. This can impose errors in the training process, and especially in the

evaluation of the model. On the other hand, it also shows that the classification model can identify changes in the heating system. To mitigate this error, one possibility is to only look at energy declarations from the last year/s or to use older energy declarations with smart meter data from the same year as the declaration.

Energy declaration analyses

There are various parameters to analyze further from the energy declaration. Some of the key features that could describe the misclassifications are analyzed in this section.

As scaling to the heated area did not improve the accuracy, it was left out of the classification model. As previously discussed, that heated area could still be of great importance for the classification, but simply scaling all power measurements of the year is not enough. In Figure 4.13 (left), the boxplot of the heated area of the building for the different heating types and correct/incorrect classification is presented. For buildings with a larger heated area, the model shows an increased probability of misclassify district heating loads. For the other two classes, the difference in the heated area is insignificant between correct and incorrect samples.

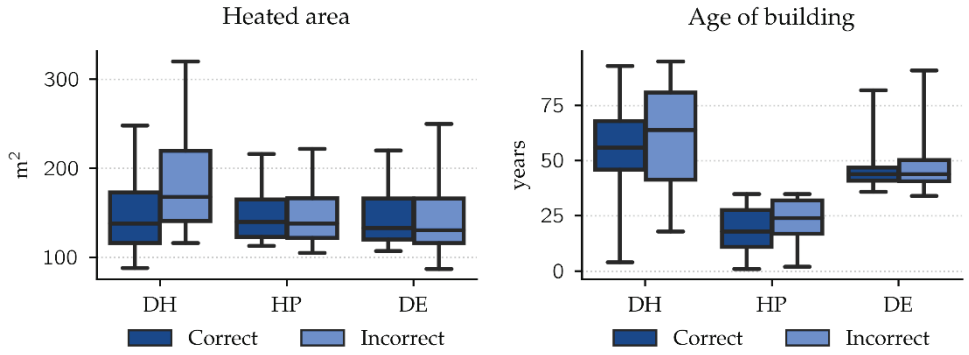


Figure 4.13 Left: Heated area of the building. Right: the age of the building. The boxplot shows the 25th, 50th, and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data source: [10]

As discussed in 4.2.2, the age of the building could indicate if there is an increased likelihood of changing the heating system. The results, however, do not give a strong indication of that, see Figure 4.13 (right).

Figure 4.14 presents the building's energy performance. It is based on the *energy consumption* of the *building*, including the building's electricity and heat consumption. The heat consumption is geographically adjusted, and the different energy carriers are weighted differently in the building primary energy values [13]. It is here used to compare loads of the same heating type where the same set of weights are used. However, no major difference within a class can be seen in the building's energy performance.

Lastly, the number of buildings with installed mechanical ventilation with a heat exchanger (FTX) and proposed action to change the heating system is presented in Figure 4.15. For consumers with only district heating, reduced energy consumption does not affect electricity consumption. However, mechanical ventilation increases

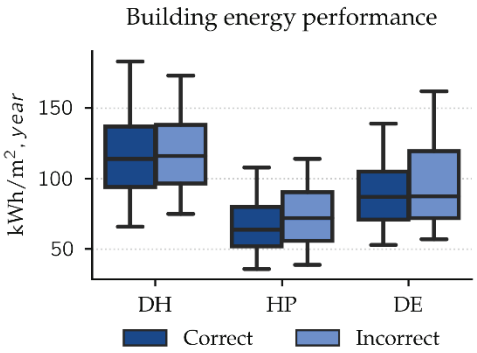


Figure 4.14 The building energy performance. The boxplot shows the 25th, 50th, and 75th percentile of the dataset, the whiskers show the 5th and 95th percentile. Outliers are excluded from the graph. Data source: [10]

electricity demand. Otherwise, it is a low proportion of the buildings with an FTX-system, and it appears to not have a significant impact on the overall accuracy of the model. Furthermore, Figure 4.15 (right) shows that a relatively large share of the buildings with direct electricity is proposed by the independent energy expert to change their heating system [10]. This is only a proposed action, but it can indicate that some of the samples with direct electricity could have changed their heating system, thus it will affect the training and testing of the model.

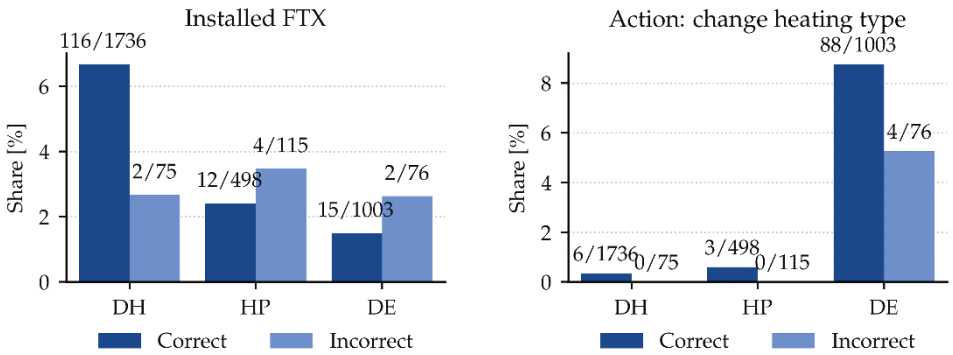


Figure 4.15 Left: number of customers with mechanical ventilation with a heat exchanger (FTX). Right: proposed action to change the heating system. Data source: [10]

5 Conclusions and future work

5.1 CONCLUSIONS

In this report, the classifier to distinguish three different electricity consumer types has been successfully developed by using the support vector machine algorithm. This report focuses on buildings with only one heating system, which is either district heating, exhaust air heat pump, or direct electricity. The feature components used in the model development include the mean, standard deviation, base, and peak electricity consumption. The work has also analyzed the impact of different time resolutions of the feature components on the classification accuracy.

The result shows that increasing the data time resolution increases the generalization performance of the classifier. The exception is when using all the hourly measurements as features, which will reduce the performance substantially due to model overfitting. Specifically, hourly variations for each month of the year gives the best generalization performance, where the average performance of the 5-fold cross-validation was 97.1% for classifying consumer with district heating from consumers with electricity-based heating sources; whereas the average accuracy is reduced to 92.4% if the classifier is to tell further if the consumer uses exhaust air heat pump or direct electricity. However, the interpretability is reduced, and it is difficult to point out what features affect the classifier. Linear regression slope between temperature and power as a single feature showed to have a good performance with an accuracy of 95.8% when classifying consumer with district heating from consumers with electricity-based heating sources. It can be used to see if a consumer has changed their heating system from district heating to an electricity-based heating source, or vice-versa. However, for classifying the three analyzed heating types, this feature alone showed a poor performance of 77.2%

The misclassification was analyzed for different characteristics of the buildings. From the result, it is difficult to draw any major conclusion. Heated area showed some difference for district heating, where the misclassified samples showed a larger heated area compared to the correct classified loads. Scaling to the heated area, however, did not show to increase the performance.

An example also shows that the energy declarations can be outdated and that the model was able to indicate the change in the heating system, even though wrongly labeled samples were included in the training of the classifier.

5.2 FUTURE WORK

In general, it is difficult to find features to improve the separation between the exhaust air heat pump and direct electricity as the *major* difference between these two heating types is the power-to-heat ratio. The model complexity will increase if more consumer classes, and/or a consumer with multiple heating types are included. For future work of feature/classification development:

- further analyses of the time series, utilizing expert knowledge
- combining feature components

- data transformation to find new patterns
- tree-structured classification model
- deep learning to automatically find the features that separate the heating types

It was also noted that the heating type has changed for some customers/building since the energy declaration was performed. The classifier can for example be trained with only the last year's energy declaration. To capture the change over time, the analysis can also be formed for each year where the energy declaration can be seen as the ground truth for the year it was conducted. For the following year, the classifier is retrained to see if it is probable that the load has changed its heating type.

In this report, only a limited selection of 1-2 family households was used for the analysis. For future work, this is to be extended with:

- all 1-2 family households
- all heating types, including buildings with multiple heating sources
- other types of consumers/buildings, e.g. apartments, offices, schools, etc.

The complexity of this depends on how detailed the classes need to be, e.g. non-electric, heat pump, full-electric heating source, or a more detailed one with district heating, oil, gas, woodchips, firewood, ground source heat pump, exhaust air heat pump, air-to-air heat pump, air-to-water heat pump, direct electricity, electricity-to-water, electricity-to-air.

6 References

[1] C. Beckel, L. Sadamori and S. Santini, "Automatic socio-economic classification of households using electricity consumption data," *e-Energy 2013 - Proceedings of the 4th ACM International Conference on Future Energy Systems,* pp. 75-86, 2013.

[2] K. Hopf, M. Sodenkamp, I. Kozlovkiy and T. Staake, "Feature extraction and filtering for household classification based on smart electricity meter data," *Computer Science - Research and Development,* vol. 31, p. 141–148, 2016.

[3] J. Helbrink, M. Lindén, M. Nilsson, D. Pogosjan and J. Ridenour, "Kategorisering av Elkunder Utifrån Förbrukningsprofil," Energiforsk, Stockholm, 2017.

[4] D. Vercamer, B. Steurtewagen, D. Van den Poel and F. Vermeulen, "Predicting Consumer Load Profiles Using Commercial and Open Data," *IEEE Transactions on Power Systems,* vol. 31, no. 5, pp. 3693-3701, 2016.

[5] C. Sandels and J. Widén, "End-User Scenarios and Their Impact on Distribution System Operators - A techno-economic analysis," Energiforsk, Stockholm, 2018.

[6] Z. Jiang, R. Lin and F. Yang., "A Hybrid Machine Learning Model for Electricity Consumer Categorization Using Smart Meter Data," *Energies,* vol. 11, no. 9, p. 2235, 2018.

[7] Svenska Elverksföreningen, "Belastningsberäkning med typkurvor," Svenska Elverksföreningen, Stockholm, 1991.

[8] Göteborg Energi Nät AB, GENAB. [Online]. Available: https://www.goteborgenergi.se. [Accessed 17 04 2020].

[9] Sveriges Meteorologiska och Hydrologiska Institut, "Lufttemperatur timvärde," 2019. [Online]. Available: https://www.smhi.se/klimatdata/meteorologi/temperatur. [Accessed 15 03 2019].

[10] Boverket, 01 01 2020. [Online]. Available: https://www.boverket.se. [Accessed 03 04 2020].

[11] SFS 2006:985, "Lag om energideklaration för byggnader," [Online]. Available: https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/lag-2006985-om-energideklaration-for_sfs-2006-985.

[12] A. Mikola and T.-A. Kõiv, "The Efficiency Analysis of the Exhaust Air Heat Pump System," *Engineering,* vol. 6, pp. 1037-1045, 2014.

[13] Boverket, *Boverkets byggregler, BBR 29 - BFS 2020:4,* Yvonne Svensson, 2020.

[14] M. M. Tom, Machine Learning, New York: McGraw-Hill, 1997.

[15] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, New York: Springer, 2017.

[16] M. Fernandez-Delgado, E. Cernadas, S. Barro and D. Amorim, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?," *Journal of Machine Learning Research,* vol. 15, pp. 3133-3181, 2014.

[17] C. Corinna and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, p. 273–297, 1995.

[18] C. Chih-Chung and L. Chih-Jen, "LIBSVM: A Library for Support Vector Machines," 29 11 2019. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf. [Accessed 24 11 2020].

[19] C. Dahlström, E. Eriksson, P. Fritz and P. Lydén, "Framtagande av effektprofiler samt uppbyggnad av databas över elanvändning vid kall väderlek. Elforsk rapport 11:12," Elforsk, 2011.

[20] B. Fulcher and N. Jones, *Highly Comparative Feature-Based Time-Series Classification,* vol. 26, 2014.

[21] J. Reunanen, "Overfitting in Making Comparisons between Variable Selection Methods," *J. Mach. Learn. Res.,* vol. 3, p. 1371–1382, 2003.

[22] M. Kuhn and K. Johnson, Applied predictive modeling, New York: Springer, 2016.

[25] Bällskärs Norra Samfällighetsförening (BNS) , *Personal communication with the BNS's board,* 2020.

Appendix A: Grid search hyperparameters

Figure A.1 shows an example of a grid-search approach to find the optimal hyperparameters for the support vector machine. The example shows two features (synthetic), where the blue dots represent class A and the red dots class B. The colored area shows the decision boundary for the classifier.

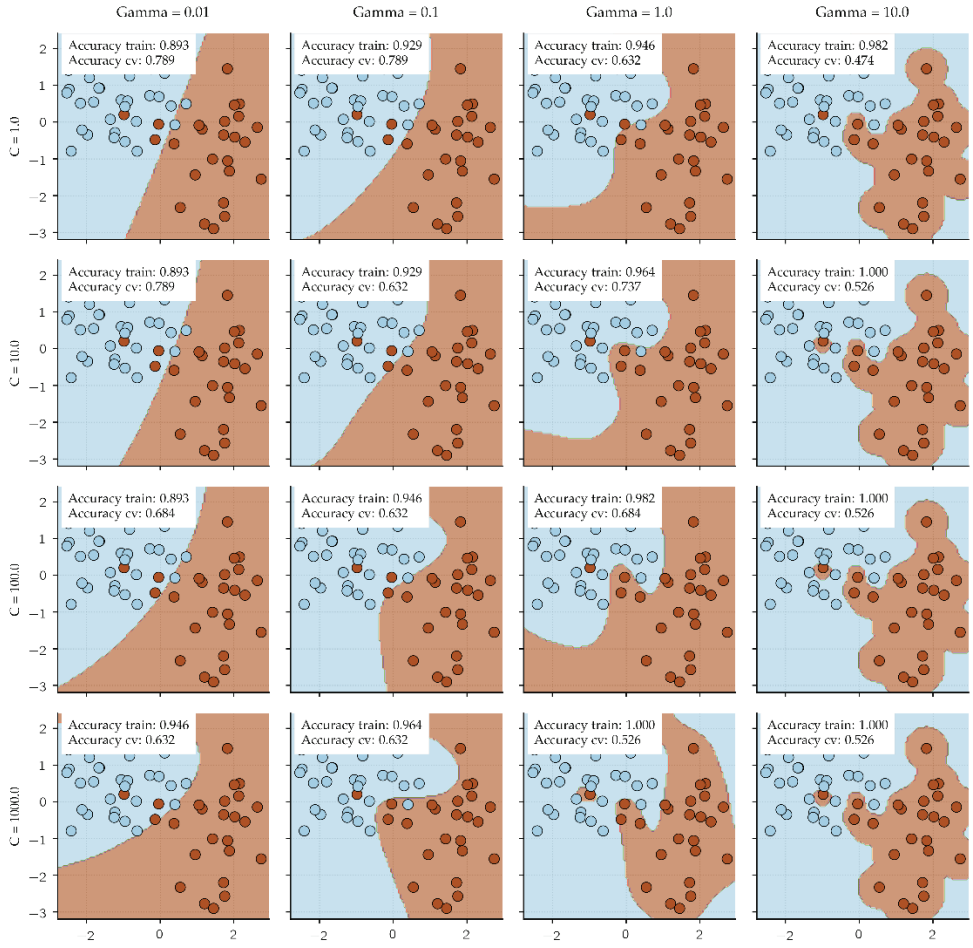


Figure A.1 An example of support vector machine grid search for hyperparameters C and γ. The example shows two features, where the blue dot represents class A and the red class B. The colored area shows the decision boundary for the classifier.

Appendix B: Feature components

The feature $\boldsymbol{x^{(i)}}$ is the feature vector of d-dimensions of the i^{th} consumer, including D feature components, and is given as

$$\boldsymbol{x}^{(i)} = \left[\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \dots, \boldsymbol{x}_D^{(i)}\right]^T = \left[x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}\right]^T$$

where $\boldsymbol{x}_D^{(i)}$ is the D^{th} feature component and $x_d^{(i)}$ is the d^{th} feature. Table B.1 shows the feature component and corresponding feature vector for a seasonal timescale, including the mean $\bar{P}$, base P^{base}, peak P^{peak}, standard deviation of the electricity consumption s, and the slope of the linear regression between outdoor air temperature and electricity consumption p. Similar applies to other timescales. The feature matrix for N samples becomes

$$\boldsymbol{x} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{bmatrix} \in \mathbb{R}^{\boldsymbol{N \times d}}$$

Table B.1 Time-dependent features with monthly variations

Description	Feature component and corresponding vector
Seasonal average electricity consumption	$\boldsymbol{x}_1^{(i)} = \left[\bar{P}_{\text{winter}}, \bar{P}_{\text{spring}}, \bar{P}_{\text{summer}}, \bar{P}_{\text{autumn}}\right]$
Seasonal standard deviation of electricity consumption	$\boldsymbol{x}_2^{(i)} = \left[s_{\text{winter}}, s_{\text{spring}}, s_{\text{summer}}, s_{\text{autumn}}\right]$
Seasonal base load electricity consumption	$\boldsymbol{x}_3^{(i)} = \left[P_{\text{winter}}^{\text{base}}, P_{\text{spring}}^{\text{base}}, P_{\text{summer}}^{\text{base}}, P_{\text{autumn}}^{\text{base}}\right]$
Seasonal peak electricity consumption	$\boldsymbol{x}_4^{(i)} = \left[P_{\text{winter}}^{\text{peak}}, P_{\text{spring}}^{\text{peak}}, P_{\text{summer}}^{\text{peak}}, P_{\text{autumn}}^{\text{peak}}\right]$
Seasonal linear regression between outdoor air temperature and electricity consumption	$\boldsymbol{x}_5^{(i)} = \left[p_{\text{winter}}, p_{\text{spring}}, p_{\text{summer}}, p_{\text{autumn}}\right]$

ELECTRICITY CONSUMER CLASSIFICATION USING SUPERVISED MACHINE LEARNING

Här har en klassificeringsmodell för att skilja mellan tre olika elkonsumenttyper utvecklats med hjälp av data för att klassificera enfamiljshushåll.

Ett datadrivet tillvägagångssätt har använts för att klassificera enfamiljshushålls huvudsakliga uppvärmningstyp, där uppvärmningstypen samlades in från byggnadens energideklaration och egenskaperna hos elkonsumenterna extraherades från smart elmätare.

Resultatet visar att en ökad datatidsupplösning ökar klassificeringens prestanda, där prestandan är baserad på konsumenter med ett uppvärmningssystem som är okänt för modellen.

En analys av felklassificeringarna visar också att energideklarationerna kan vara föråldrade och att modellen kan indikera förändringar i uppvärmningsmetoden, även om felmärkta exempel ingår vid träningen av klassificeringsmodellen.

Energiforsk is the Swedish Energy Research Centre – an industrially owned body dedicated to meeting the common energy challenges faced by industries, authorities and society. Our vision is to be hub of Swedish energy research and our mission is to make the world of energy smarter!

Energiforsk