Exploring Weather Data Patterns for Energy Consumption Predictions



Bachelor’s thesis

Degree Programme in Computer Applications

Spring 2024

Ilyass Souhail



Degree Programme in Computer Applications Abstract

Author Ilyass souhail Year 2024

Subject Exploring weather data patterns for energy consumption predictions

Supervisors Esa Huiskonen

ABSTRACT

The thesis investigates the relationship between weather conditions and energy use, analysing historical data and exploring the potential of machine learning techniques to improve energy consumption prediction. The research aims to identify patterns between weather and energy consumption, which could be used to predict future energy demand based on weather forecasts. This could be a game-changer for utilities, businesses, and individuals, enabling them to optimise energy usage and reduce costs. The research methodology involved a comprehensive study of the Finland case, using Power BI, Excel, and Python to analyse energy consumption and weather data.

The thesis found a strong correlation between specific weather conditions and energy demand, suggesting that incorporating weather forecasts into energy prediction models can significantly improve accuracy. The thesis recommends expanding energy forecasting models to include weather data, refining models to incorporate factors like building type and occupancy patterns, and testing models in real-world scenarios. These practical applications could revolutionise energy consumption, increase efficiency, cost savings, and reduce environmental impact.

Keywords Supervised learning, Mean Squared Error, Energy efficiency, Sustainability.

Pages 44 pages and appendices 1 page

Glossary

F Fahrenheit

C Celsius

K Kelvin

ICZ Intertropical Convergence Zone

HAWTs Horizontal-axis wind turbine

VAWTs Vertical-axis wind turbine

UV Ultra-Violet

GHPs Geothermal heat pumps

EGS Enhanced geothermal systems

BECCS Bioenergy with carbon capture and storage

IPCC Intergovernmental Panel on Climate Change

CCUS Carbon Capture, Utilization, and Storage

VPPs Virtual Power Plants

ETL Extract Transform Load

BI Buseniss Intellegence

AI Artificiel Intellegence

KNN K-Nearest Neighbors

DB DataBase

SARSA State Action Reward State Action

SVM Support Vector Machines

SVR Support Vector Regression

SVC Support Vector Classification

RBF Radial Basis Function

MSE Mean Squared Error

OECD The Organization for Economic Cooperation and Development

FMI the Finnish Meteorological Institute

Contents

[1 Introduction 1](#_Toc166760033)

[2 Weather and energy 2](#_Toc166760034)

[2.1 Weather Aspects 2](#_Toc166760035)

[2.1.1 Temperature 2](#_Toc166760036)

[2.1.2 Precipitation 3](#_Toc166760037)

[2.1.3 Winds 3](#_Toc166760038)

[2.1.4 Cloud Covers 4](#_Toc166760039)

[2.1.5 Weather in Finland 5](#_Toc166760040)

[2.2 Energy Aspects 5](#_Toc166760041)

[2.2.1 Kinetic Energy 5](#_Toc166760042)

[2.2.2 Radiant Energy 6](#_Toc166760043)

[2.2.3 Biomass 6](#_Toc166760044)

[2.2.4 Geothermal Energy 7](#_Toc166760045)

[2.2.5 Electrical Energy 7](#_Toc166760046)

[2.2.6 Energy Consumption in Finland 8](#_Toc166760047)

[2.3 Weather energy interaction 8](#_Toc166760048)

[2.3.1 The Relation between Weather and Energy 9](#_Toc166760049)

[2.3.2 Tools and Technology 9](#_Toc166760050)

[2.3.3 Challenges 10](#_Toc166760051)

[2.3.4 Solutions 11](#_Toc166760052)

[3 Data definitions 12](#_Toc166760053)

[3.1 Data and information 12](#_Toc166760054)

[3.2 Data types 12](#_Toc166760055)

[3.3 Dataset 13](#_Toc166760056)

[3.4 Data analysis 13](#_Toc166760057)

[3.4.1 Data Analysis Process 14](#_Toc166760058)

[3.4.2 Data Analysis Types 14](#_Toc166760059)

[3.5 Data science 15](#_Toc166760060)

[3.5.1 Data science process 15](#_Toc166760061)

[3.5.2 Data Science Techniques 15](#_Toc166760062)

[3.6 ETL Process 16](#_Toc166760063)

[3.6.1 Extraction-Transformation-Load 16](#_Toc166760064)

[3.6.2 Benefits and Challenges 16](#_Toc166760065)

[4 Machine Learning types and techniques 18](#_Toc166760066)

[4.1 Machine learning 18](#_Toc166760067)

[4.1.1 Supervised Machine Learning 18](#_Toc166760068)

[4.1.2 Unsupervised Machine Learning 19](#_Toc166760069)

[4.1.3 Semi-Supervised Learning 19](#_Toc166760070)

[4.1.4 Reinforcement Machine Learning 19](#_Toc166760071)

[4.2 Classifications in Machine Learning 20](#_Toc166760072)

[4.2.1 Linear regression 20](#_Toc166760073)

[4.2.2 Support Vector Machines 21](#_Toc166760074)

[4.2.3 Random Forest 21](#_Toc166760075)

[4.2.4 Operations needed in modeling 22](#_Toc166760076)

[5 Methods and tools 23](#_Toc166760077)

[5.1 Microsoft Excel 23](#_Toc166760078)

[5.2 Power BI 23](#_Toc166760079)

[5.3 Visual Studio code 24](#_Toc166760080)

[5.4 Python and Jupyter Notebook 24](#_Toc166760081)

[5.5 Python libraries 24](#_Toc166760082)

[5.6 Data Retrieval 25](#_Toc166760083)

[6 Data Processing 26](#_Toc166760084)

[6.1 Weather data 26](#_Toc166760085)

[6.2 Energy consumption data 27](#_Toc166760086)

[6.3 The weather and Energy consumption relation 31](#_Toc166760087)

[7 Machine learning by Python 32](#_Toc166760088)

[7.1 Preparation of Environment 32](#_Toc166760089)

[7.2 Studies of data from 2020 to 2023 33](#_Toc166760090)

[7.2.1 Linear Regression 33](#_Toc166760091)

[7.2.2 Support vector Regression with RBF kernel 34](#_Toc166760092)

[7.2.3 Random Forest Model 36](#_Toc166760093)

[7.3 Testing Results 37](#_Toc166760094)

[8 Results 39](#_Toc166760095)

[9 Summary 40](#_Toc166760096)

[10 References 41](#_Toc166760097)

[Figure 1: Cloud Types (L.S. Gardiner, 2024) 4](#_Toc166345039)

[Figure 2: Benefits and Challenges (Yung, 2023) 17](#_Toc166345040)

[Figure 3: Linear relationship i the feature space (Linear Regression (Python Implementation), 2017) 20](#_Toc166345041)

[Figure 4: 3 rows of Temperature Table 26](#_Toc166345042)

[Figure 5: Temperature data by months 27](#_Toc166345043)

[Figure 6: 1st quarter of 2022 Energy consumption 28](#_Toc166345044)

[Figure 7: Electricity consumption in households from 2022 28](#_Toc166345045)

[Figure 8: Electricity price by consumption in January 2022 29](#_Toc166345046)

[Figure 9: Energy consumption data by months 29](#_Toc166345047)

[Figure 10: Energy consumption data in households in 2022 29](#_Toc166345048)

[Figure 11: Electricity prices in 2022 30](#_Toc166345049)

[Figure 12: Wood Fuel Consumption and Temperature 31](#_Toc166345050)

[Figure 13: the call of Excel files 32](#_Toc166345051)

[Figure 14: The merge operation 32](#_Toc166345052)

[Figure 15: Linear Regression call 33](#_Toc166345053)

[Figure 16: Interception, coefficient and mean squared error calculation 33](#_Toc166345054)

[Figure 17: Linear regression plot for 2020-2022 34](#_Toc166345055)

[Figure 18: Energy prediction in -5 and 15 degrees using RL 34](#_Toc166345056)

[Figure 19: the call of SVR with RBF kernel 35](#_Toc166345057)

[Figure 20:SVR plot for 2020-2022 35](#_Toc166345058)

[Figure 21: Energy prediction in -5 and 15 degrees using SVR 36](#_Toc166345059)

[Figure 22: the call of random forest model 36](#_Toc166345060)

[Figure 23: Random forest plot 37](#_Toc166345061)

[Figure 24: Energy prediction in precipitation of 60.9 mm 37](#_Toc166345062)

[Figure 25: prediction of energy in -5.5 degrees 37](#_Toc166345063)

[Figure 26: Prediction of energy in 37.9 mm 38](#_Toc166345064)

# Introduction

This thesis investigates how weather affects energy use. Understanding these patterns can help predict energy consumption more accurately. By analysing different weather data, the aim is to create better models for forecasting energy needs. This research can assist in smarter energy management and contribute to environmental sustainability efforts.

Weather impacts energy consumption in residential, commercial, and industrial sectors. Temperature, humidity, and wind speed affect heating, cooling, and energy demands. Extreme weather can cause supply shortages or higher costs. Understanding these dynamics is crucial for resource management, distribution optimisation, promoting sustainability, and transitioning to renewable energy sources.

While weather conditions significantly influence energy usage, existing forecasting models may lack precision or fail to capture the complex interplay between meteorological factors and energy demand. This gap hinders effective energy management strategies and can lead to resource allocation and distribution inefficiencies.

Comprehending the complex correlation between meteorological phenomena and energy usage is imperative for efficient resource allocation and ecological endeavours. The following research issues are intended to be addressed by this study:

* Is there a significant relationship between Weather and Energy consumption?
* What are the available tools and ways of profiting from the atmosphere that enhance energy production?
* **How can machine learning techniques enhance energy consumption prediction using weather data?**
* What impact does a better understanding of weather-energy dynamics have on optimising energy distribution and promoting environmental sustainability?

# Weather and energy

This chapter examines the correlation between energy and weather and the weather-energy relationship, particularly emphasising Finland as a Nordic cold country.

## Weather Aspects

Weather is a collection of daily occurrences in the atmosphere. It varies from region to region and can change over minutes, hours, days, and weeks. Most weather occurs in the troposphere, the region of the atmosphere closest to the Earth's surface.( Center for Science Education, 2024)

### Temperature

As an indicator of heat energy transfer, temperature is unrelated to the amount of matter present. Kelvin (K), Fahrenheit (°F), and Celsius (°C) are three broadly employed temperature scales. Due to the enormous scale and thermal capacity of the world's oceans, the average annual surface temperature of the Earth has risen by approximately 2 degrees Fahrenheit (1 degree Celsius) since the pre-industrial era. This additional heat causes seasonal and regional temperature extremes, alters the habitat ranges of plants and animals, and reduces snow cover and sea ice. It also intensifies heavy rainfall. Although the Arctic is warming faster than most other regions, terrestrial areas have warmed more rapidly than maritime areas. (The Editors of Encyclopædia Britannica, 2024g)

Although it may appear peculiar, the notion of a global "average" temperature serves a practical purpose in identifying and monitoring temporal variations in the Earth's energy budget. To calculate the global average temperature, scientists use temperature data collected at various locations worldwide. In the 2023 Global Climate Report, each month was ranked among the seven warmest for that month, with the months comprising the latter half of the year being their warmest on record. (JV, 2022)

Agriculture may profit from global warming in the near and long term, as longer warm seasons and an earlier start to spring may help farmers in the short run by facilitating crop growth. However, there are drawbacks, such as weeds, alien plant species, insect pests, reduced water availability, and negative effects of drought and heat stress's consequences. Over time, benefits to the shipping business could come from extended Northwest Passage times due to melting Arctic sea ice. (Herring, 2020)

### Precipitation

Precipitation comprises a variety of solid and liquid water particles that descend from clouds and settle on the ground. It includes hail, snow, snow granules, and ice crystals. There are two stages involved in the transformation of a cloud from consisting solely of cloud droplets to incorporating precipitation particles: first, incipient precipitation elements are generated in the vapour state; second, these elements proliferate via aggregation and collision with cloud droplets; and finally, they combine with precipitation particles. Precipitation via ice crystals is contingent on the ability of cloud particles to spontaneously solidify at temperatures below -40°C or -40°F. Small solution droplets are also significant as precursor particles to precipitation, given that soluble chemical substances are present in the atmosphere in many small particles. The predominant mechanisms by which precipitation particles grow are collision and coalescence. (The Editors of Encyclopædia Britannica, 2024f)

The Department of Physical Geography and Ecosystem Science has identified a critical breakpoint for annual precipitation in the Amazon forest, which is drier than the current climate. If annual precipitation exceeds 2,000mm, growth and other functions remain unaffected. The future of the Amazon depends on more than just annual precipitation. A warmer climate increases water demand for plants, and deforestation also affects precipitation in the central Amazon region. Accurate projections and conclusions on causes and solutions are crucial for understanding ecosystem changes and avoiding inaccurate outcomes. (Noomi, 2024)

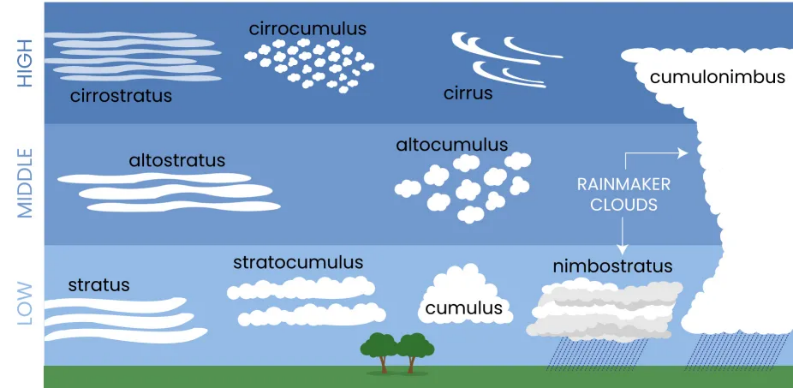
### Winds

Winds are the differential motion of air to the surface of the Earth; they substantially influence the regulation and regulation of climate and weather. As a result of horizontal and vertical pressure differences in the atmosphere, wind systems rotate clockwise around highs in the Southern Hemisphere and counterclockwise around lows in the Northern Hemisphere. Standing waves are the most sizable wave patterns in the middle and upper troposphere, where pressure systems are arranged in high-pressure ridges and low-pressure troughs. At high latitudes, winds tend toward the east near the surface. However, in low, tropical, and equatorial latitudes, trade winds that are northeasterly to the north of the intertropical convergence zone (ICZ) and southeasterly to the south of the ICZ travel in the direction of the ICZ. On either side of the ICZ, winds generate cumulonimbus clouds and heavy precipitation vertically, delineating a constricted region called the doldrums. Local winds, including foehn winds, katabatic winds, sea and land breezes, and mountain and valley breezes, indicate particular geographic regions and are shaped by topographical characteristics. As a result of the Sun's thermal energy causing the air to be overturned, wind gusts and intensity are typically greatest during the day. In contrast, during the night, gusts subside and winds are generally weaker. (The Editors of Encyclopædia Britannica, 2024b)

### Cloud Covers

Cloud cover influences the sky conditions, precipitation forecasts, and temperature regulation of a region, all of which are critical for comprehending and forecasting the weather. In the atmosphere, clouds manifest in diverse geometries, including cumulus, cirrus, and stratus (Figure 1). High-level cirrus or low-level stratus clouds indicate precipitation from incoming low-pressure systems, whereas low-level cumulus clouds indicate favourable weather and sunny skies. Most clouds assume these forms and are categorised as nimbo-form if they produce precipitation. (L.S. Gardiner, 2024)

Figure 1: Cloud Types (L.S. Gardiner, 2024)



Clouds regulate the Earth's surface temperature by functioning as an atmospheric covering. Daytime solar energy is reflected into space, averting excessive global warming. Additionally, clouds warm the lower regions of the atmosphere by absorbing heat emitted by the surface and radiating it back to Earth. The variety of clouds present in the atmosphere determines whether clouds heat or chill the Earth's surface. (Brainly, 2022)

Radiative cooling involves the discharge of solar heat absorbed by the ground during the day, which can be hindered by cloud cover. Dense cloud cover entraps and reflects a portion of the heat, thereby maintaining the surface temperature above that of a nocturnal period devoid of cloud cover. (National Geographic Society, 2023)

### Weather in Finland

The climate of Finland exhibits notable variations based on the time of year and geographical location, owing to its extensive coastline and dynamic climate patterns influenced by ongoing climate change. July is the warmest month, and weather forecasting is difficult because of its fluctuations. (Visit Finland, 2024)

The summer season in Finland, which usually spans from late May to mid-September, is characterised by moderate temperatures and 24-hour daylight. In Lapland, the summer season begins a month earlier and ends earlier compared to the south coast. The onset of autumn in Finland often occurs in late August, characterised by the emergence of vibrant foliage and increased precipitation, while snowfall is observed in Lapland during September and in other regions during October or November. Finland experiences a winter season that spans 200 days in Lapland and 3-4 months in southern Finland. This season is marked by cold temperatures, snowfall, and the stunning Northern Lights, creating a magical winter landscape. The spring season in Finland, commencing in April rejuvenates both nature and individuals with a surge in sunlight. The main attraction is the 1st of May, when Finnish people congregate for outdoor meals in parks, irrespective of the prevailing weather conditions. (Visit Finland, 2024)

## Energy Aspects

Energy is the capacity to perform labour in various forms—including potential, kinetic, thermal, electrical, chemical, and nuclear—in physics. Potential energy can be transformed into kinetic energy, causing the box and slope to become heated. Energy is incapable of creation or destruction; rather, it is limited to form transformations per the first law of thermodynamics or the principle of energy conservation. ( *Britannica*, 2024)

### ****Kinetic Energy****

Kinetic energy is the form of energy that the motion of an object or particle generates. The translation of a quantity into vibration, rotation about an axis, or any combination of these motions is contingent upon its mass. The translation kinetic energy of an object is calculated by dividing the product of its mass and velocity by half. The concept of relativistic kinetic energy is defined as the product of the square of the speed of light and the increase in the mass of a particle. The joule is the unit of energy, whereas the erg is utilised in the centimetre-gram-second system. The rotational kinetic energy of a body in motion is calculated by dividing the product of its moment of inertia and the square of its angular velocity by half. (The Editors of Encyclopædia Britannica, 2024a)

Wind turbines convert wind energy into electricity, with two main types: horizontal-axis (HAWTs) and vertical-axis (VAWTs). HAWTs have multiple blades, while VAWTs harness wind from any direction. Some are less economically competitive. Wind turbine siting is a concern due to environmental impacts, public perception, and return on investment. (The Editors of Encyclopædia Britannica, 2024e)

### Radiant Energy

Radiant energy transmitted via electromagnetic waves, including thermal radiation, X-rays, light, and gamma rays, can be characterised as continuous electromagnetic waves or discrete photons. Under the principle of energy conservation, any radiant energy that a system emits or absorbs must be accounted for in the overall energy. (The Editors of Encyclopædia Britannica, 2024d)

Sunlight, visible solar radiation, is divided into visible, ultraviolet, and infrared components. The visible portion makes up nearly half of Earth's total radiation. UV light is crucial for vitamin D production. The atmosphere absorbs and weakens the radiation, with short wavelengths scattering more easily. Ozone, water vapour, and carbon dioxide are effective absorbers. The Eppley pyrheliometer measures the duration and intensity of sunlight by measuring the electromotive force produced by the sun's rays. (The Editors of Encyclopædia Britannica, 2024d)

### Biomass

Biomass refers to the total quantity of living organisms in a unit area or habitat, often measured in units of energy or weight. It includes species biomass, which refers to the total amount of organic material produced by living organisms, and community biomass, which refers to plant materials and animal waste used as fuel. (The Editors of Encyclopædia Britannica, 2024c)

Wood has historically been utilised as an energy source in various forms, such as divided logs that have been air-seasoned, shavings, briquettes, and pellets. As an expanding source of energy, waste timber necessitates sophisticated incineration technologies. For sustainable forestry, wood is considered a climate-neutral energy source; however, its combustion is intricate and produces air pollutants. Fuel assortments, combustion reactions, and factors influencing wood oxidation are discussed in this chapter. The article also addresses the advancement of combustion facilities and operational strategies for sustainably utilising biomass fuels. (Marutzky, 2023)

### Geothermal Energy

Geothermal energy is an inherent resource utilised for cooking, bathing, space heating, and electrical power generation, among other applications. It captures and harnesses heat energy from the Earth's subsurface. Volcanic eruptions, geysers, fumaroles, hot springs, and mud pots are all examples of surface phenomena caused by interior Earth heat. Surface energy recovery and utilisation is estimated to be 4.5 × 106 exajoules or approximately 1.4 × 106 terawatt-years. The quantity of practical energy derived from geothermal sources depends on extraction methods and depth. The most prevalent applications do not require specialised apparatus and involve directly using heated water from the earth. Utilising the comparatively stable moderate temperatures within the first 300 meters of the surface, geothermal heat pumps (GHPs) provide heating and cooling for structures in winter and summer, respectively. GHPs are energy-efficient, consuming 25–50% less electricity and emitting up to 72% fewer greenhouse gases than electric resistance heating systems and conventional heating and conditioning systems, respectively. (The Editors of Encyclopædia Britannica, 2016)

Geothermal power facilities generate electricity by utilising heat beneath the Earth's surface. Flash-steam plants utilise naturally occurring sources of underground heated water and steam, whereas dry-steam power plants utilise underground sources of steam. During the Arctic winter, flash-steam plants are prevalent in the Philippines and Iceland, where they heat frigid sidewalks and parking lots. Binary cycle power plants effectively manage water usage by heating subterranean water to a range of 107°-182°C while producing steam that drives a generator. Enhanced geothermal systems (EGS) supply fluid and permeability to regions with heated but desiccated subsurface rock via injection, drilling, and fracturing. As a result of the potential for seismic activity induced by the injection process, certain Basel, Switzerland projects may be cancelled. (National Geographic Society, 2024)

### Electrical Energy

Electrical energy is a convenient and versatile form used in various applications such as running computers, operating appliances, home heating, and transportation. It accounts for 18% of end-use energy worldwide and is primarily used by industry, households, and businesses. The energy is derived from the movement and configuration of electric charge, which can accumulate on a capacitor and be physically carried in electric and magnetic fields. Electrical conduction allows for easy transportation of electricity through conductor wires, forming the electrical grid. It is not a primary energy source but an energy currency, with primary energy sources like wind or natural gas converted into electricity for practical use and transport. The flexibility of electricity encourages the production of as much energy as possible, and its use rate is growing faster than global energy use. (the Organization for Economic Cooperation and Development, 2014)

Emissions from electricity generation vary based on factors like generation amount, technologies, and air pollution control devices. Estimating household emissions and using tools like Power Profiler can help reduce personal impact. Electrical safety is crucial when using electrical energy, including electric shock and burns, arcing injuries, fires from faulty equipment, and explosions caused by unsuitable electrical apparatus. (US EPA, 2015)

### Energy Consumption in Finland

Compared to the previous year, Finland's total energy consumption decreased by 6% in 2020, with incineration accounting for a substantial portion of the decline. Carbon dioxide emissions from fuel use decreased by 10%, while renewable energy sources such as hydro, wind, and solar power increased. The consumption of renewable energy exceeded that of peat and fossil fuels combined. Wood fuels decreased by 6% due to the contraction of the forest industry and mild weather. The generation of hydropower increased by 28%, that of wind energy by 30%, and that of methane decreased by 3%. (Sandberg, 2022)

Peat consumption decreased by a quarter compared to 2019 when fossil fuels and peat decreased by 10%. A substantial 22% decline in coal consumption was attributed to prohibiting hard coal usage in energy generation. Additionally, the coronavirus decreased fossil fuel consumption by 6%. In 2020, Finland's total electricity production decreased by 6% to 81 TWh, primarily due to warmer weather and decreased industrial consumption. The industry accounted for 45% of the total final consumption. (Sandberg, 2022)

In 2020, a variety of energy products worth 6.7 billion euros were imported into Finland, with the majority of these items coming from Russia. Energy products exported to OECD countries comprised 74% of the total value of exports. (Sandberg, 2022)

## Weather energy interaction

The weather influences numerous spheres, including energy production and consumption, agriculture, and the planet's geology; therefore, it is of considerable interest to those who profit substantially from the weather.

### The Relation between Weather and Energy

Weather plays a crucial role in daily activities and energy consumption. Temperature variations, seasonal changes, and energy-dependent sources like solar panels and wind power can all impact energy consumption. Strategies include regulating indoor temperatures, using energy-efficient insulation, and utilising renewable energy sources. Extreme weather events can damage infrastructure, while heat waves strain power plants. Weather forecasting and energy management systems can help predict energy demand. Sustainable solutions include integrating renewable energy, enhancing energy efficiency, and promoting responsible consumption. A more sustainable future can be created by understanding the relationship between weather and energy consumption. (Bureau of Meteorology, 2023)

Zhang's study investigates the impact of temperature fluctuations on electricity consumption in rural China. It suggests that a one-degree increase in summer temperatures can increase electricity consumption by 0.015% per capita, while a one-degree decrease in winter temperatures can increase it by 0.002%. The northern region consumes more electricity during colder conditions. (Zhang et al., 2019)

### Tools and Technology

Nuclear fusion releases energy from hydrogen atoms in the sun's core. Renewable energy sources like solar panels, wind turbines, and hydropower generate over 3.5% of the world's electricity. Technological advancements are improving these sources, with solar panels using hexagonal lenses, thin silicon layers, and perovskite crystals. Wind turbines are made cheaper, more efficient, and more powerful using 3D printing, curved tips, and smart blades. Hydropower technology is also improving, with low-head hydropower generating electricity from gentle slopes. Analytical tools are improving hydro plant efficiency, especially in light of climate change and variable water flows. These advancements are crucial to meeting the United Nations' Sustainable Development Goal. (Heggie, 2023)

Electrification is using low-carbon electricity to meet conventional fossil fuel requirements, such as heat pumps for building heating and electric vehicles to replace conventional fuels. Its projected contribution of 35% towards the necessary emissions reductions for a net zero energy system by 2050 represents a potential reduction of 19% in emissions. Renewable energy sources, including solar and wind, have the potential to supply a substantial amount of low-carbon electricity. Over the last decade, the installed capacities of these sources have increased in every country. Decarbonising the current electricity supply will not suffice, however, because the worldwide electricity demand is projected to double by 2050. (Esin, 2023)

There are primarily three approaches to CO2 removal from the atmosphere: technological solutions, enhancing natural processes, and solutions founded on nature. Enhanced natural processes involve land management and the inclusion of biochar, whereas nature-based solutions involve afforestation and reforestation. Technological solutions include Direct air and bioenergy with carbon capture and storage (BECCS). According to the IPCC, agriculture, forestry, and land-use modifications could eliminate between 1 billion and 11 billion tonnes of CO2 annually by 2050. BECCS, an established technique for carbon removal from biomass-burning processes, encounters obstacles to the infrastructure and sustainable biomass availability. Alternative pathways are excluded from the IPCC because they have not yet reached a mature stage. (Budinis, 2020)

### Challenges

Safe, inexpensive, and low-carbon alternatives to large-scale energy sources are currently absent globally. The world will persist in grappling with its two existing energy challenges unless these alternatives are used widely. The most discussed energy issue is the connection between energy access and greenhouse gas emissions. However, an equally significant global energy issue exists: hundreds of millions of individuals are completely deprived of adequate energy, which has dire consequences for their well-being and the environment. (Roser, 2024)

Energy consumption is a significant issue, contributing to pollution and global warming. Inefficient energy distribution increases power consumption, causing pollution and ejecting over 32.5 billion tons of CO2 into the atmosphere annually. This damage affects ecosystems, agriculture, and human health. Major energy sources include transportation, residential and commercial buildings, and industrial manufacturing. The U.S. Energy Information Administration highlights transportation, residential and commercial buildings, and industrial manufacturing as major sources of energy consumption. It is crucial to reduce energy consumption and resources to mitigate the effects of power production and distribution on the climate. (Wang, 2020)

Emissions rebounded in 2021, reaching 2019 levels. To meet 1.5° pathway requirements, mature economies need to accelerate emissions decline and emerging economies transition to lower-carbon growth paths. Renewables account for over 30% of global investments by 2035, twice as high as conventional power generation. Sustainable fuels, including biofuels and power-to-gas, are expected to contribute between 6% and 37% of transportation energy demand by 2050. CCUS, a niche decarbonisation technology, is projected to scale significantly by 2050. Sustainable fuels are expected to triple in demand over 20 years, with $40-$50bn investment by 2025. Hydrogen demand growth is projected to accelerate after 2035, with road transport and new industrial uses accounting for over 50% of growth. Renewables dominate the power generation mix by 2050, with solar and onshore wind contributing the most. (Ellis, 2022)

The grid stability is impacted by the intermittent and variable nature of renewable energy sources, necessitating the development of energy storage solutions. Initial capital investments continue to be substantial despite cost reductions, and the levelized cost of electricity must be comparable to that of fossil fuels. Technology advancements and economies of scale are indispensable for achieving cost reductions. Transitioning to renewable energy necessitates the modernisation of transmission networks and pre-existing power grids, among other infrastructure developments. To overcome NIMBY sentiment, which is comprised of concerns regarding environmental, pollution, and visual impacts, public acceptance and policy support are crucial. (Kolkowska, 2023)

### Solutions

Energy storage technologies play a pivotal role in mitigating the challenges of intermittent renewable energy generation. Grid-scale systems and batteries facilitate the effective utilisation of surplus energy. Hydrogen shows great potential for a wide range of energy applications and long-term storage. Digital transformation and intelligent infrastructures enhance demand-side management, capacity balancing, and energy waste reduction. Governments, businesses, and academic institutions must work together to foster innovation. Grid flexibility and virtual power plants (VPPs) optimise energy generation and consumption through the integration of multiple renewable energy sources, energy storage systems, and demand response mechanisms. Blockchain technology presents viable resolutions for the obstacles encountered in the energy sector. These include streamlined grid administration, transaction traceability and transparency, and peer-to-peer energy trading. Predictive analytics and sophisticated monitoring systems can optimise performance, identify defects, and facilitate maintenance. By analysing enormous quantities of data, artificial intelligence and machine learning algorithms can optimise the generation and consumption of renewable energy. This improves grid management and efficiency by enabling precise weather forecasting, energy distribution, and demand forecasting. (Kolkowska, 2023)

# Data definitions

This chapter provides definitions of data and numerous data types, examines the meaning, process, and categories of data analysis and data science, respectively, and describes the ETL procedure and extract transformation and load.

## Data and information

Data is a distinct category of structured information that is further subdivided into programs and data. Data science is a multifaceted field that utilises mathematical prowess, programming expertise, programming processes, systems, algorithms, and processes to extract knowledge and insights that can be implemented. This information is subsequently implemented in an extensive range of contexts and applications. Proficiency in the foundational principles of data and data science is imperative for success across many disciplines. (Simplilearn, 2020)

Information is structured data that is significant to consumers and utilized in decision-making. To be of substantial value, it must satisfy the criteria of timeliness, accuracy, and comprehensiveness. Completeness ensures the information is comprehensive and pertinent, whereas accuracy guarantees its availability when required. (Simplilearn, 2020)

## Data types

Quantitative data is another name for quantitative data; although it can be classified into distinct groups, it is not amenable to measurement or quantification. Nominal, ordinal, and continuous data are the three primary categories of categorical knowledge. Nominal data consists of categories or identifiers lacking intrinsic significance, ranking, or numerical order. Frequently assigned arbitrarily, it lacks an intrinsic hierarchy or classification within its categories. (Jagdeesh, 2023)

Ordinal data is characterized by values arranged in meaningful order or classification yet lacking a consistent and evenly spaced numerical distinction among them. There are non-numerical labels, ordered categories, and no fixed intervals. Ordinal data renders arithmetic operations such as addition and subtraction meaningless, as the intervals between categories lack quantifiability. (Jagdeesh, 2023)

Quantitative data consists of measurable quantities. Discrete data comprises unique, discrete categories or values, typically enumerated and whole integers. Important attributes encompass measurable values, discrete categories, inter-value gaps, and visual depictions via bar charts or histograms. (Jagdeesh, 2023)

Continuous data, commonly represented numerically, can assume an infinite number of values within a specified range. It is also referred to as continuous variables or quantitative data. The essential attributes encompass boundless values, exactitude, the absence of voids or discontinuities, and visual depiction through scatter plots or line charts. (Jagdeesh, 2023)

Time-series data are obtained by collecting or recording information at regular intervals, illustrating the temporal variation of a specific variable or set of variables. They find application in a multitude of disciplines, including finance, meteorology, epidemiology, and manufacturing, where they facilitate the comprehension of process dynamics over time, trend monitoring, and prediction. (Jagdeesh, 2023)

## Dataset

A dataset is a compilation of data organised into rows and columns, utilised for insights, decision-making, and algorithm training in industries such as machine learning, business, and government. Data cleansing and preprocessing are necessary to assure the quality and suitability of the information for analysis or modelling, contingent upon its size and complexity. (GeeksforGeeks, 2023b)

Datasets are numerical, categorical, web, time series, image, ordered, partitioned, file-based, bivariate, and multivariate datasets. A dataset's features are critical for deploying models and predicting new data points. Key features include numerical and categorical data points, metadata, data size, formatting, target variables, and data entries. Metadata provides a general description of the dataset, while size refers to the number of entries and features. Formatting options include JSON, CSV, XML, DataFrame, and Excel Files. (GeeksforGeeks, 2023b)

## Data analysis

A methodical procedure, data analysis consists of scrutinising, purifying, converting, and modelling data to extract actionable insights, formulate conclusions, and facilitate decision-making. It enables organisations to enhance operational efficiency, forecast trends, and make well-informed decisions. In addition to defining objectives, data acquisition, data cleansing, data analysis, data interpretation and visualisation, and data narrative, the procedure consists of multiple stages. (Crabtree, 2023)

### Data Analysis Process

Initiating the data analysis procedure with formulating objectives and inquiries establishes the trajectory for the entire undertaking. This entails comprehending the issue, discerning the requisite data, and establishing metrics for evaluating results. The information is then gathered quantitatively or qualitatively via databases, interviews, or surveys. Data cleansing is a crucial step in verifying data for errors and inconsistencies to ensure its quality and dependability. Following this, statistical or mathematical methods are applied to the data to identify patterns, relationships, or trends. Specialised applications such as SPSS and SAS, in addition to Python, R, and Excel, are utilised for this objective. (Crabtree, 2023)

The findings are then illustrated with data visualisation and interpretation, which simplify the comprehension of complex data. The last stage involves data storytelling, which delivers discoveries in a captivating and comprehensible narrative structure. This is vital to conveying findings to non-technical recipients and formulating decisions based on data. (Crabtree, 2023)

### Data Analysis Types

Descriptive analysis is a method that summarises and interprets raw data, focusing on historical events to identify patterns and trends over time. For instance, a business might use descriptive analysis to understand monthly sales for the past year. (Crabtree, 2023)

Diagnostic analysis is a more detailed method that aims to understand the cause of an outcome by comparing different data sets, such as a company's sales drop in a specific month, rather than using descriptive analysis. (Crabtree, 2023)

Predictive analysis is a method that uses statistical models and forecasting techniques to predict future events. It is often used in risk assessment, marketing, and sales forecasting. It involves using past data to predict future outcomes, such as a company forecasting the next quarter's sales based on historical data. (Crabtree, 2023)

Prescriptive analysis is a sophisticated data analysis method that uses advanced tools like machine learning and artificial intelligence to predict future outcomes and suggest actions to benefit from these predictions, such as suggesting the best marketing strategies to boost future sales. (Crabtree, 2023)

## Data science

To derive meaningful insights from data, data science is an interdisciplinary discipline that integrates statistics, science, computation, and machine learning. It emphasises transforming data into knowledge to facilitate informed decision-making. Data drive our world, and its increased adoption and utilisation result from the necessity for businesses, governments, and individuals to comprehend and utilise it effectively. (Abdullahi, 2023)

### Data science process

Data science involves a unique process that starts with identifying the primary purpose of the collected and analysed data. This knowledge is crucial for accurate analysis and questioning, enabling data scientists to gather accurate and qualitative insights from valid sources. (Abdullahi, 2023)

Data collection involves cleaning, correcting errors, removing duplicates, and identifying inconsistencies to prepare it for analysis. Data scientists then interpret and report results using graphical, visual, or storytelling patterns to aid decision-making and understanding of the data. (Abdullahi, 2023)

### Data Science Techniques

Data scientists analyse and enhance information using various methods. Machine learning entails developing models and algorithms that learn from experience and analyse data according to predefined metrics. Data analysis, summarisation, and interpretation through the utilisation of classification and regression analysis are all functions of statistics. (Abdullahi, 2023)

Data mining reveals latent patterns and correlations to discern trends and enhance the accuracy of predictions. Deep learning, which is a subfield of machine learning, trains models to recognise patterns and produce outcomes using a variety of learning techniques. Data visualisation effectively communicates the outcome comprehensibly, enabling observers to discern recurring patterns and trends. (Abdullahi, 2023)

## ETL Process

ETL emerged with the proliferation of centralised data repositories in the 1970s. In the late 1980s and early 1990s, software developed specifically for loading data into data warehouses emerged. Early ETL tools were rudimentary yet effective, capable of processing limited quantities of data. With the proliferation of data volumes, ETL software tools underwent sophisticated advancements. By the conclusion of the 20th century, data storage and transformation were predominantly conducted in on-premises data warehouses.(Matillion, 2023)

### Extraction-Transformation-Load

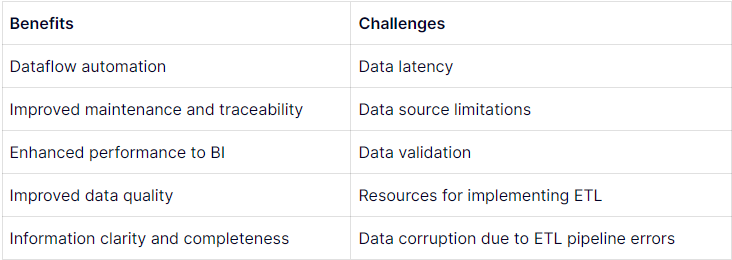
Data analysis tools are employed to extract data to fulfil business intelligence requirements. Periodically, data engineers extract information in accordance with business requirements and available sources. Validation standards are implemented regularly to filter unprocessed data. Online extraction, notifying systems, incremental updates, and full extraction are all methods. The transient data storage in a "Warehouse Staging Area" is permitted for audit reports. The extracted data is subsequently transformed from the source files into warehouses and data marts. (Tatvasoft, 2015)

Transforming extracted data into a structured format compatible with a predefined Data Warehouse format is crucial for business processes. This involves cleaning, deriving, sorting, joining, and validating. Cleaning involves mapping values by code, generating new values, sorting to ensure fast retrieval, joining integrated data elements, and generating surrogate key values for uniformity. Validation rules preserve old values and replace original values to handle inconsistencies during data refreshment. (Tatvasoft, 2015)

Data warehouses load structured and transformed data into tables, maintaining consistency and referential integrity. Record-by-record or bulk loading is preferred for better performance. Table partitioning, depending on time intervals, can improve query performance. (Tatvasoft, 2015)

### Benefits and Challenges

A reliable, automated ETL solution is crucial for organisations to understand their data and make informed decisions. Automation saves time and avoids errors. Proper ETL processes can lead to positive outcomes, such as customer retention or revenue growth, as seen with Tradezella and Mailtrap. Figure 2 shows the benefits of ETL and its Challenges. (Yung, 2023)

Figure 2: Benefits and Challenges (Yung, 2023)

# Machine Learning types and techniques

This chapter defines machine learning and describes its four distinct types: supervised, unsupervised, semi-supervised, and reinforcement. In addition to discussing the operations required for modelling, machine learning encompasses the classifications utilised in the implementation, such as random forest, SVM, and linear regression.

## Machine learning

A subfield of artificial intelligence, machine learning empowers machines to simulate human behaviour by solving complex problems like humans do. It entails providing computers with experience-based programming instruction, commencing with text, images, or statistics data. The greater the quantity of data, the more effective the program. The data is provided by programmers, who then select a machine-learning model and allow it to train itself to recognise patterns or generate predictions. As time passes, the human programmer can modify the model by adjusting its parameters to improve its accuracy. (Brown, 2024)

### Supervised Machine Learning

A supervised learning approach entails training a model using a labelled dataset comprising input and output parameters. The algorithms use training and validation datasets to discover how to map points between inputs and accurate outputs. Constructing an image classifier to distinguish between canines and felines serves as an illustration. When presented with new images, the computer predicts whether the subject is a dog or a cat, having learned to differentiate between these images. (GeeksforGeeks, 2023a)

supervised learning encompasses two primary categories: classification and regression. The classification field investigates the determination of category target factors, such as the risk of cardiac disease or spam. Classification methods include Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), and Naive Bayes. On the contrary, regression analysis involves attempting to estimate continuous objective variables such as product sales or property prices. (GeeksforGeeks, 2023a)

### Unsupervised Machine Learning

Unsupervised learning is a machine learning methodology that uncovers patterns and relationships from unlabeled data. The objective is to detect latent patterns, clusters, or similarities in the data that can be utilised for data exploration, visualisation, and dimensionality reduction. Clustering can, for instance, categorise customers with comparable purchasing patterns, thereby exposing prospective clients who do not possess predetermined labelling. (GeeksforGeeks, 2023a)

Assigning and clustering are the two primary classifications of unsupervised learning. Clustering entails classifying data points according to their similarity, utilising algorithms such as the K-Means Clustering algorithm, Mean-shift algorithm, DBSCAN Algorithm, Principal Component Analysis, and Independent Component Analysis. Association rule learning identifies rules that specify the probability that one item's presence implies another's presence, thereby revealing relationships between items in a dataset. (GeeksforGeeks, 2023a)

### Semi-Supervised Learning

Utilising both labelled and unlabelled data, semi-supervised learning is an algorithm for machine learning that is practical when acquiring labelled data would be prohibitively expensive, time-consuming, or resource-intensive. It is especially beneficial when dealing with partially labelled data that lacks the remainder. In particular, unsupervised methods can generate label predictions fed to supervised methods on image datasets in which not every image is labelled. Graphic-based, label propagation, co-training, self-training, and generative adversarial networks (GANs) are all techniques utilised in semi-supervised learning. Label propagation propagates labels according to similarities, co-training trains models on distinct unlabeled data subsets, self-training trains models on labelled data and predicts labels for unlabeled data, and graph-based learning employs graphs to represent relationships between data points. (GeeksforGeeks, 2023a)

### Reinforcement Machine Learning

A learning approach known as reinforcement machine learning employs delay, trial, and error to enhance performance via reward feedback. Prominent algorithms for reinforcement learning consist of Deep Q-learning, SARSA, and Q-learning. SARSA modifies the Q-function based on the action performed, whereas Deep Q-learning integrates Q-learning and deep learning by implementing a neural network. (GeeksforGeeks, 2023a)

Positive reinforcement, which rewards agents for desired actions, and negative reinforcement, which eliminates undesirable stimuli to encourage or discourage the behaviour, are the two primary categories of reinforcement learning. Instances include supplying a point in a game or rewarding a dog for resting. (GeeksforGeeks, 2023a)

## Classifications in Machine Learning

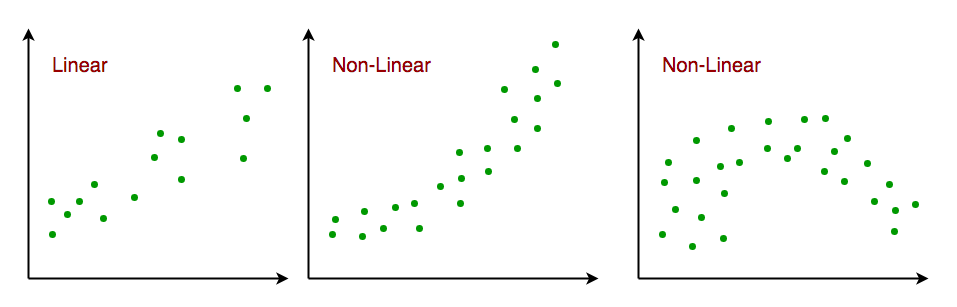
Classification involves supervised machine learning and aims to forecast the class of data points based on their association with targets, labels, or categories. Artificial neural networks, K-nearest neighbours, decision trees, and naive Bayes are prevalent classification algorithms. As an illustration, email service provider spam detection is accomplished through binary classification using known and non-spam emails as training data. Classification application extends to various domains, including credit approval, medical diagnosis, and target marketing. (Sidath, 2022)

### Linear regression

Linear regression is a supervised machine learning algorithm that determines the relationship between a dependent variable and one or more independent features by applying observed data to a linear equation. It is crucial due to its straightforwardness, interpretability, and integration into sophisticated models such as support vector machines and regularisation. (GeeksforGeeks, 2017)

The scatter graphs in Figure 3 verify the linear relationship between response and feature variables, which is the assumption underlying linear regression models. Variables depicted in the first figure are presumably non-linear, whereas those in the second and third figures are linearly related. (GeeksforGeeks, 2017)

Figure 3: Linear relationship i the feature space (Linear Regression (Python Implementation), 2017)

 It is also a cornerstone in assumption testing, allowing researchers to validate key assumptions about the data. Linear regression is used when there is only one independent feature, Univariate Linear Regression, or multiple dependent variables. (GeeksforGeeks, 2017)

### Support Vector Machines

Supervised learning algorithms, such as support vector machines (SVM), are employed to address classification and regression challenges, particularly on smaller datasets. This is especially true for support vector classification (SVC) and support vector regression (SVR), which require more processing time. Support Vector Machines (SVM) operate by identifying a hyperplane that optimally partitions the features into distinct domains. Due to their prevalence, Gaussian distributions will be the primary focus. In SVM models, the Gaussian RBF is a prevalent kernel technique whose value is proportional to the distance from the point or origin. (Yadav, 2018)

SVMs frequently employ the RBF kernel to represent intricate relationships between input and output variables. The similarity between the two data elements is assessed using the Euclidean distance. The selection of gamma and C values substantially influences the model's efficacy. The gamma parameter determines the impact of a solitary training instance; a smaller value corresponds to reduced accuracy, while a greater value signifies increased accuracy. Values in the middle serve as effective decision boundaries. (Kumar, 2020)

The model's accuracy diminishes as gamma increases, underscoring the criticality of choosing suitable values. The C parameter functions as a regularisation element, establishing the model's threshold for misclassification to minimise generalisation error. A greater C value leads to a reduced tolerance, resulting in a classifier with a maximum margin. A reduced value of C promotes an increased margin, which leads to a more straightforward decision function but compromises training accuracy. As C increases, model accuracy improves, as depicted in the decision boundary diagram; conversely, a reduced value of C permits more misclassification, which results in decreased accuracy. (Kumar, 2020)

### Random Forest

Leo Breiman and Adele Cutler created the random forest algorithm for machine learning. It generates a solitary outcome by merging several decision trees. Its adaptability and user-friendliness in the context of classification and regression challenges have contributed to its widespread adoption. The random forest algorithm generates an uncorrelated forest of decision trees by combining bagging and feature randomness, an extension of the bagging method. Instead of decision trees, random forests select a subset of potential feature divisions, thereby ensuring that decision trees lack correlation. (IBM, 2024)

For classification or regression problems, random forest algorithms are implemented with three hyperparameters: number of trees, number of features sampled, and node size. Each decision tree is composed of an out-of-bag sample and a bootstrap sample. Randomness is incorporated via feature aggregation, correlation reduction, and adding diversity. Predictions differ based on the nature of the problem; regression necessitates the utilisation of averaged decision trees, whereas classification relies on majority ballots. (IBM, 2024)

### Operations needed in modeling

The mean squared error (MSE) measures the proximity of a regression line to a set of points by squaring the distances from the points to the regression line, removing negative signs and giving more weight to larger differences. A lower MSE indicates a better forecast. Figure 3 contains the MSE formulas with the explanation of notations. (Panik, 2020).

Commonly associated with linear models, a category of algorithms utilised for regression and classification, are intercepts and coefficients, which are fundamental to machine learning. Coefficients are employed by linear models to allocate numerical values to each input variable, thereby signifying the significance or weight of each feature in the prediction process. These coefficients determine the direction or slope of the relationship between each feature and the objective variable. The intercept, alternatively referred to as the bias or constant, is a constant term that denotes the value of the objective variable in linear model equations when every feature is set to zero. Offering an offset or baseline prediction enables the model to detect patterns that may not be directly associated with the input features. (UMME HABIBA, 2023)

# Methods and tools

This chapter lists the environment's required tools, including Microsoft Excel, Power BI, Jupiter Notebook, Visual Studio Code, and Python libraries. It also specifies the data sources and the retesting process.

## Microsoft Excel

The Microsoft Excel spreadsheet application enables users to perform calculations, formatting, and data organisation within a spreadsheet. It is compatible with other Office suites and is a member of the Office product family. Excel's arrangement of numerous cells into rows and columns facilitates the observation of updated or added information by data analysts and other users. It is accessible on the following platforms: Windows, macOS, Android, and iOS. (Gillis, 2020)

Excel is an extensively utilised application within business environments, serving various purposes such as performance reporting, human resource management, and operations management. A vast collection of cells is employed to structure and manipulate data, which can be modified using formulas, graphing tools, and pivot tables. Visual Basic for Applications is an additional macro programming language in Microsoft Excel. It serves various purposes such as data collection, verification, and analysis. (Gillis, 2020)

## Power BI

Power BI is an adaptable and robust application that allows users to exploit their data's capabilities. In the current competitive environment, organisations can achieve business success, obtain valuable insights, and make data-driven decisions by leveraging the functionalities of Power BI. (Smith, 2024)

Power BI is a powerful data analysis tool that connects users to various data sources, enables data transformation, and offers a range of visualisation options. It also integrates AI capabilities to enhance data analysis and facilitates team collaboration. Benefits of Power BI include increased efficiency, better decision-making, enhanced visibility, and cost savings. Automating repetitive data analysis tasks helps organisations streamline decision-making processes, drive business growth, and track progress towards goals. Its cloud-based service offers flexible pricing options and scalable infrastructure, making it a cost-effective solution for organisations of all sizes. (Smith, 2024)

## Visual Studio code

Visual Studio Code is a robust, lightweight source code editor compatible with Raspberry Pi OS, Windows, macOS, and Linux. In addition to TypeScript, Node.js, and JavaScript, it also provides extensions for runtimes, environments, programming languages, and clouds. Git support is included, in addition to IntelliSense code completion and graphical debugging. (Heller, 2022)

Visual Studio Code boasts a substantial user base of millions, including data scientists and developers affiliated with Google and Facebook. It inspects repositories, and it provides a lightweight editor, syntax check, code completion, refactoring, and diagnostics. In addition to supporting Docker, Kubernetes, and major clouds, it integrates with Git to facilitate collaboration. (Heller, 2022)

## Python and Jupyter Notebook

Jupyter Notebook is a scientific computing and data science-oriented open-source web application that enables users to generate and distribute documents comprising narrative text, live code, equations, and visualisations. Python IDLE is an integrated development environment (IDE) that offers a rudimentary interface for creating and executing Python code, specifically engineered for scientific computation. Both scientific computing and data science require the use of both instruments. (Singh, 2023)

Offering an interactive user interface, live code execution, integration with multiple programming languages, simple collaboration, data visualisation, and markdown support, Jupyter Notebook is a popular tool among data scientists and researchers. However, it is resource-intensive for large datasets, has limited diagnostic tools, presents version control challenges, and executes code from untrusted sources, which presents security concerns. Additionally, it encounters security concerns when executing code originating from untrusted sources. (Singh, 2023)

## Python libraries

Multiple libraries are utilised for machine-learning objectives in this thesis. The open-source Python library Numpy enables the execution of scientific and mathematical computations. It provides math.fsum and math.frexp functions, which facilitate complex computations involving multidimensional arrays and matrices. Pandas, an additional indispensable library, is dedicated to manipulating and analysing data. It supports data extraction tasks by offering a variety of data structures, functions, and components that assist in data set preparation and training. (Shinde, 2023)

Matplotlib is an exceptionally potent data visualisation application utilised in the thesis. It empowers programmers to analyse data patterns efficiently. Functionality is further enhanced by features such as Pyplot, which provides options for line designs and font management, which are especially advantageous when generating two-dimensional graphs. On the contrary, Scikit-Learn distinguishes itself as a sciPy-based ML library that is quick and intuitive. It features extensive data classification, segmentation, and regression algorithms. In addition, it provides a supportive developer community that can offer assistance in resolving any issues that may arise. (Shinde, 2023)

## Data Retrieval

The objective of this thesis is to examine patterns in weather data. Therefore, the aim is to identify a dataset that provides comprehensive information on weather conditions throughout the year. Regarding energy consumption, the focus is specifically on 2020 to 2023. The statistics on energy use were sourced from the website Statistics Finland, renowned for its reliable database encompassing several aspects of Finland's affairs. Due to the impracticality of collecting meteorological data for the entire nation, the focus is on gathering data for a specific city, such as one in the Uusimaa region (Helsinki, Vantaa, Espoo). The meteorological data was sourced from the Finnish Meteorological Institute (FMI).

A repository file was created to store the datasets that would be utilised, including precipitation, weather data from the previous three years in Helsinki, and energy consumption. The plan is to convert the data points into Excel format for ETL (Extract, Transform, Load) purposes and import them into Power BI for visualisation, analysis, and modelling. The upgraded dataset file may then be saved to Google Drive, which will be utilised in Visual Studio code to implement machine learning using Python, utilising the upgraded table and data.

The 2022 data focuses on snow depth, cloud cover, wind speed, humidity, temperature, precipitation, and air pressure. It also covers energy usage, including heating and electrical use, gasoline consumption, price variations, and various energy sources like nuclear, fossil fuels, renewables, and peat.

# Data Processing

This chapter comprises the analysis portion of the thesis, in which tables are cleansed, modelled, analysed, and presented as graphs; these graphs will aid in comprehending and obtaining the necessary statistics, whether they pertain to energy consumption data, meteorological data, or the correlation between the two.

## Weather data

To avoid any missing data, the weather data for 2022 was divided among multiple sources, as attempting to gather all the details from a single source could result in some incomplete results. Consequently, the approach involved creating separate tables for each element and combining them by utilising the date as the primary key.

Initially, a particular data point is chosen, such as the temperature. The table includes columns for the year, month, day, hour, and average air temperature (measured in degrees Celsius). The subsequent step involves gathering data and selecting an Excel file for importing, specifically one containing precipitation statistics. Subsequently, the Navigator presents us with the choice to load or convert the data. Converting the data is chosen due to the imperative nature of ensuring that both datasets remain in harmony. Queries can now be combined in the Power Query Editor. Figure 10 has selected the specific year, month, and day for the current table and implemented a filter on the database to exclusively incorporate the pertinent weather component.

To correlate snow depth with precipitation, the cell format should also be a numerical value with one decimal place. In addition, it is necessary to replace any negative precipitation and snow depth values with 0. Furthermore, the snow depth values should be converted from centimetres to millimetres and added to the corresponding column. Finally, the result in Figure 4 is the temperature table with features in the column those features are Month, Day, Wind speed mean, Air pressure mean, Air temperature mean, snow depth in mm and cm, precipitation amount and cloud cover in each day of the year 2022.

Figure 4: 3 rows of Temperature table



After removing the data from the Excel spreadsheet, the tables can be duly organised, enabling the utilisation of Power BI to facilitate the modelling, analysis, and presentation of the data. The Excel worksheet was formerly utilised to perform the ETL process manually.

Figure 5 shows in terms of precipitation, February receives the most quantity with a total of 97.5 mm, which is also the same for snow depth at 704.1 cm. Additionally, the lowest temperature occurs in January at -22.8°C, while in July it is 6.9°C. These statistics are expected, given that they fall in the heart of the winter season.

Regarding cloud cover, March experiences the highest number of clear sky days (21 days), followed by June, July, and August, with 20 to 18 days, indicating the summer season. In contrast, January, December, and November are predominantly cloudy.

Figure 5: Temperature data by months

A screenshot of a graph



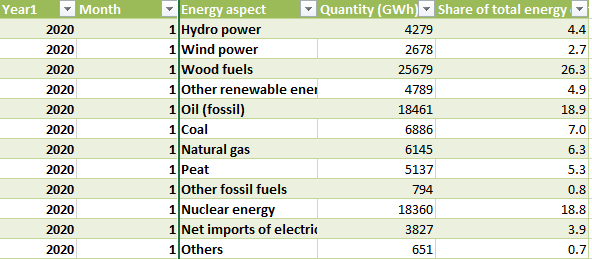
The mean wind velocity is 3.09 m/s, with a peak value of 9.4 m/s documented in March. In January, the average wind speed reaches a maximum of 4.09 m/s.

Clouds and air pressure are interconnected. Cloudy days exhibited the lowest mean air pressure, measuring at 1008.12 hPa, and the highest average wind speed, reaching 3.6 m/s.

## Energy consumption data

The first table shown in Figure 6 contains the total energy consumption by energy source by quarter and energy source and information for the years 2020 to 2023. There are three datasets in total. The data is manageable for the upcoming phases and free of null values.

Figure 6: 1st quarter of 2022 Energy consumption

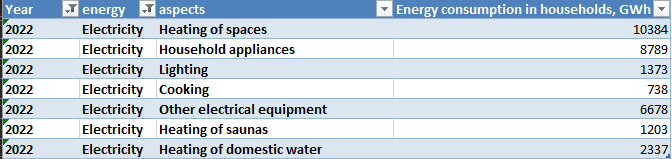


The second table, shown in Figure 7, contains data on home energy usage from 2019 to 2022, categorised by year, energy source, energy end use, and relevant statistics.

For instance, the use of wood does not impact cooking and lighting, resulting in various qualities having null values. Consequently, modifying the energy consumption component based on the energy source is more logical. Upon implementing a year-based filter, it is evident that the displayed rows are limited to those that contain the year 2019 in the column denoting the current year. This is because the tables are not organised in chronological order.

The Energy column comprises the following elements: Wood, Electricity, Peat, Heavy fuel oil, Ambient energy, Natural gas, and District heat. To address this issue, the approach is adding more rows, which also applies to other tables.

Figure 7: Electricity consumption in households from 2022



The 3rd table in Figure 8 contains 2 types of customers, households and enterprise and corporate clients, and each has its Limitation of the use annually.

Figure 8: Electricity price by consumption in January 2022

A screenshot of a computer

Description automatically generated

The extraction of those datasets into Power BI will facilitate data analysis, especially when focusing on the year 2022.

A line and clustered column chart in Figure 9 is generated based on the given data to facilitate the calculation of various values for each energy use and enable a comparison. Wood fuels account for the largest total energy consumption, amounting to 101,960 TWh. The chart exhibits a gradual decline, reaching a minimum average of 6925 GWh by the midpoint of the third quarter, and subsequently rises to a value of 7926.17 in the last quarter of the year. The fluctuation in average quantities throughout the year is also apparent in the first quarter, where the average reached 8081.17 GWh.

According to the statistics, most energy is consumed during the first and final quarters of the year, which correspond to the autumn and winter seasons.

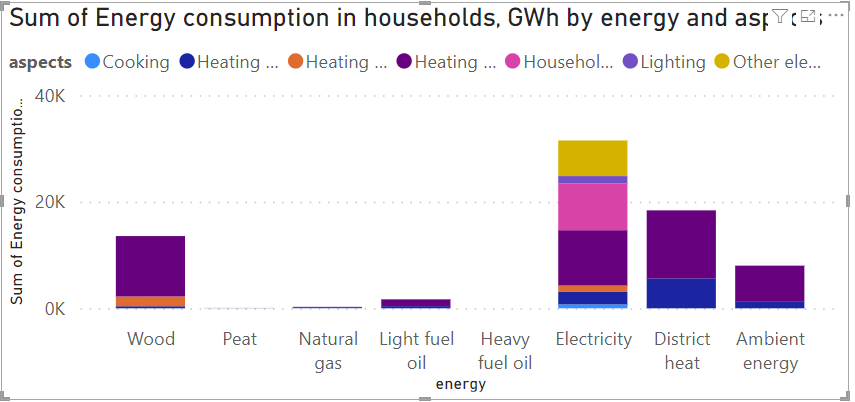
Figure 9: Energy consumption data by months

A graph of different colored lines

Description automatically generated

Figure 10 illustrates the prevalence of electricity for heating areas and powering home appliances in a stacked column graphic. District heat accounts for a larger proportion of heat spaces than electricity. This indicates that the main purpose of energy consumption in a household is heating.

Figure 10: Energy consumption data in households in 2022



The line and clustered column chart in Figure 11 indicate that prices remain relatively consistent from January to April, fluctuating between 14.5 and 13 €/KWh. Subsequently, there is an upward trend in pricing, reaching an average of 21.68 €/KWh in December.

Figure 11: Electricity prices in 2022

A graph with colorful lines and text

Description automatically generated

Furthermore, families with an annual usage of less than 1GW incur the highest price, reaching 35.75 €/KWh in November. Enterprise and corporate clients are charged the highest rate, which is 21.35 €/KWh if their yearly usage is less than 20 MWh.

## The weather and Energy consumption relation

An earlier investigation of meteorological and energy use data has revealed that the time of year is characterised by interconnected elements such as price fluctuations, variations in consumption, and changes in temperature, precipitation, or snow depth. The weather, namely the transition from winter to summer, with its shift from cold to hot and from dark to light, is the main catalyst for energy use. It provides information on the amount of thermal energy utilised in residential buildings, the quantity of wood consumed, and the expenses incurred for fuel during the winter months of December, January, and February. This graph in Figure 12 illustrates a prominent utilisation of wood fuels in low-temperature applications.

Figure 12: Wood Fuel Consumption and Temperature

A graph showing the average of the energy level



There is a direct relationship between the amount of wood fuel used and the average minimum ground temperature. The places with the lowest yearly temperature have the highest concentration of wood fuel use.

The objective is to predict energy consumption by analysing meteorological data patterns, namely those that are relevant, and taking into account the disparity between energy production and consumption. Although it is challenging to forecast the weather accurately in specific regions, it is possible to make general predictions on a broader scale. For example, precisely predicting low temperatures, cloud cover, and precipitation totals in the beginning and latter quarters of the year is challenging. However, energy consumption and production can be accurately forecasted by utilising these trends and employing predictive analysis techniques. By constructing a machine learning model using the data from 2022, it is possible to forecast the energy consumption in 2023.

# Machine learning by Python

This chapter details the Python code implementation for predictive analytics and machine learning from 2020 to 2023. The results are calculated using random forest for precipitation and linear regression and SVM with RBF kernels for the temperature case and precipitation case, respectively.

## Preparation of Environment

In this step, fundamental libraries, including pandas and numpy, are verified in the Jupyter Notebook. The Sklearn package encompasses machine learning methods like Support Vector Machines (SVM) and regression models. The library dedicated to visualising plots is named Matplotlib. Accessing the necessary Excel file in Figure 13 initially retrieves Finland's 2022 meteorological data and energy usage figures.

Figure 13: the call of Excel files

code

Given that the tables are configured in the notebook, combining the two tables can help with modelling and manifesting data. In Figure 14, the main keys "Year" and "Month" are the emphasis of the jointure.

Figure 14: The merge operation



Finland's weather and energy usage tables are now included in the table with the same Year (2022) and Month columns. Some weather values are repeated because each month has a different energy consumption, but the weather values fluctuate daily. An example of this can be seen in the January 1, 2022 data: the energy component exhibits variation in each row, whereas the meteorological values remain constant.

The answer is to eliminate the days column by calculating the average weather data for each month to prevent this issue. Solving this is to convert them into fractions (cloudy=1, mostly cloudy=0.8, partially cloudy=0.6, mostly clear=0.4, clear=0.2), as they are expressions but still include the cloud cover elements.

So, using a pivot table in Excel will ease the operations of the average and sum of values. The same thing applies to the Energy consumption table, which focuses on the sum of the quantity of the whole energy aspect.

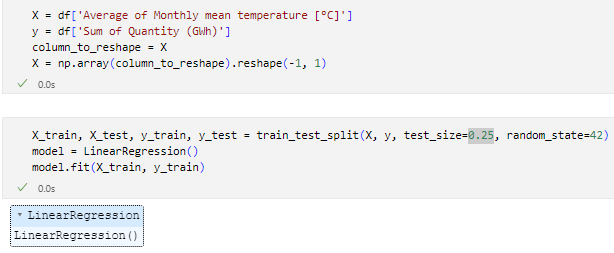
## Studies of data from 2020 to 2023

There is an observation that the data is weak and has few values after taking the year 2022 into consideration to alter tables and utilise machine learning models for that year. For this reason, it is increasingly preferred to include the years 2020, 2021, and 2023 in research.

### Linear Regression

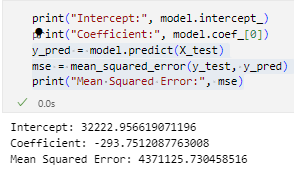
Let's stick to the same characteristics, including total energy consumption [GWh] and the average monthly mean temperature [°C]. The procedure is the same as in earlier chapters, with the columns being called an axis and reshaped as seen in Figure 15.

Figure 15: Linear Regression call



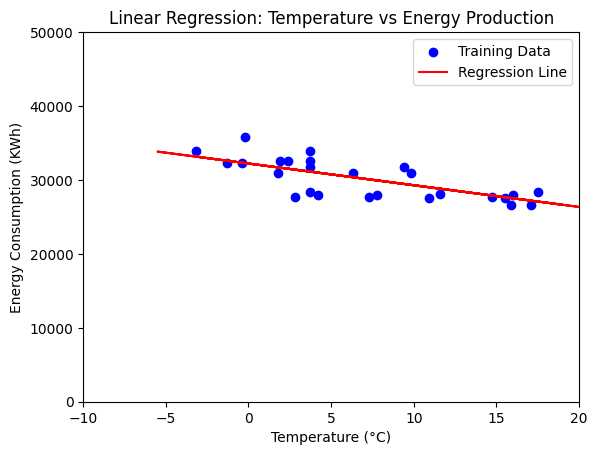
In this instance, a fit linear regression model in Figure 16 is also called for, along with selecting 25% of the data for testing. The next step is to determine the coefficient and the interception.

Figure 16: Interception, coefficient and mean squared error calculation



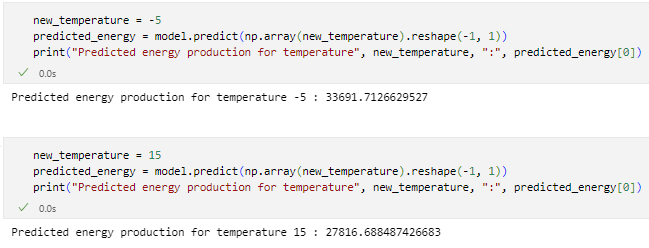
This plot in Figure 17 contains the red scatter for the regression, the line that represents the predicted results of energy consumption, and the blue for the trained data, which is the actual data by temperature between -10 and 20 degrees and energy consumption between 10 GWh and 50 GWh.

Figure 17: Linear regression plot for 2020-2022



For instance, estimate how much energy will be used at 15 and -5 degrees (Figure 18). The prediction makes sense if energy consumption is higher at lower temperatures.

Figure 18: Energy prediction in -5 and 15 degrees using RL

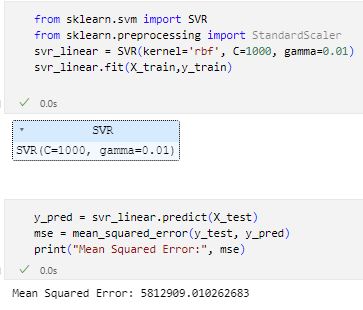


The code indicates that 33691.71 kWh of energy will be used at -5 degrees. While 27816.68 kWh at 15 degrees. Although the forecast is undoubtedly inaccurate, the previously stated condition is met since energy usage is lower at 15 degrees than at -5 degrees.

### Support vector Regression with RBF kernel

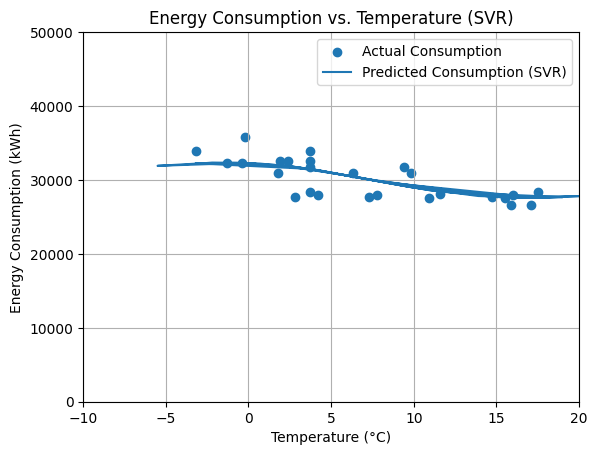
The most difficult part of utilising SVM with an RBF kernel is figuring out the right C and gamma values to obtain the lowest Mean Squared Error. The gamma in Figure 19 is 0.01 and the C is 1000.

Figure 19: the call of SVR with RBF kernel



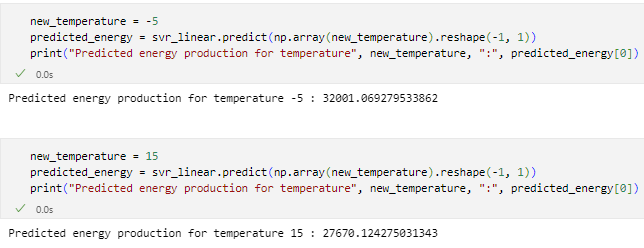
The SVR model is called after the training data are fitted, and the linear regression's MSE is used to calculate the reasonably high Mean Squared Error. The plot in Figure 20 shows the blue scatter for the anticipated consumption and the blue spots for the actual consumption.

Figure 20:SVR plot for 2020-2022



It is better to compute the estimated amount of energy consumption at -5 and 15 degrees (Figure 21) to compare the forecast findings. The outcome differs from linear regression, but it's still very close, and the condition is still met. As before, the forecast may be off, but it makes sense.

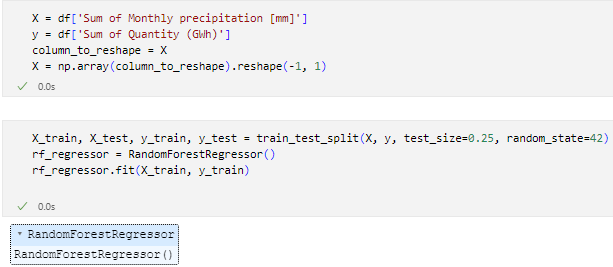
Figure 21: Energy prediction in -5 and 15 degrees using SVR



### Random Forest Model

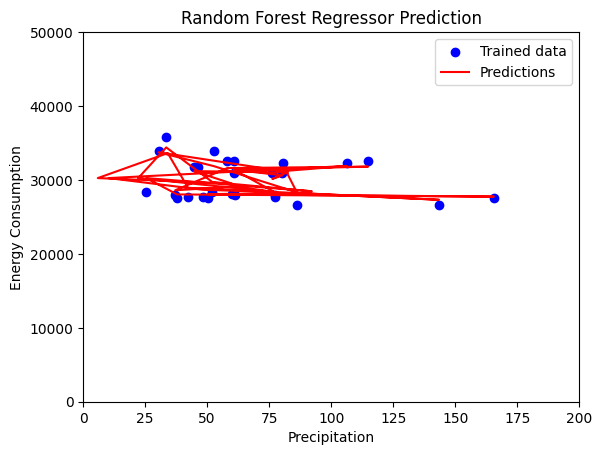
Using the Sum of Monthly precipitation [mm], since the values are very random, the suitable model to use is random forest. The procedure is the same as in earlier chapters, with the columns being called an axis and reshaped as seen in the code in Figure 22.

Figure 22: the call of random forest model



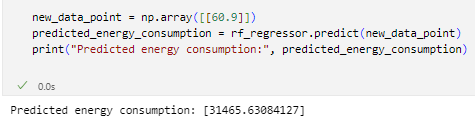
In this instance, a fit Random Forest model is also called for, along with selecting 25% of the data for testing. This plot in Figure 23 contains the red scatter for the regression and the blue for the trained data.

Figure 23: Random forest plot



Given the stochastic nature of precipitation values, 60.9 mm can be used as a predictive value for the energy consumption quantity, as shown in the code in Figure 24.

Figure 24: Energy prediction in precipitation of 60.9 mm

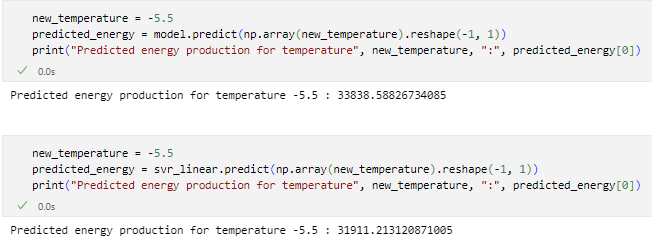


## Testing Results

The testing phase is required to determine the level of prediction accuracy. Based on the graphs presented in the previous chapters, it can be seen that both the linear regression and SVM with RBF kernel models have a margin of error because some training data is far from the prediction line scatter. Additionally, the mean squared error is high, particularly in the random forest model, indicating that they can be approximated while the results are not exact.

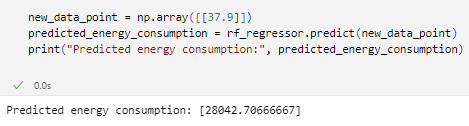
Consider the code in Figure 25. At a temperature of -5.5 degrees, linear regression predicts a quantity of consumption of 33838.58826734085 GWh; however, the actual data shows a different amount of energy consumption at -5.5 degrees: 35779 GWh. The estimated quantity of consumption is 31911.213120871005 GWh when using the SVM with RBF kernel for the same degree of -5.5; this number is significantly different from the prior model. This indicates that the linear regression model is a good choice for predicting energy usage based on temperature.

Figure 25: prediction of energy in -5.5 degrees



This code in Figure 26 indicates that the random forest model was chosen for the precipitation in the previous chapter because the correlation between precipitation and energy consumption is wide and the number chosen for the total precipitation is 37.9. The real amount is 27622.666667 GWh, whereas the desired result is 28042.70666667 GWh.

Figure 26: Prediction of energy in 37.9 mm



# Results

The evolution of the IT sector is crucial to daily life, as it affects all aspects that can be modified, including healthcare, sports, and entertainment; technology is no exception. As mentioned in this thesis, data engineering development has facilitated data presentation and its subsequent modelling, analysis, visualisation, and prediction.

The tools employed in this thesis included Excel, which facilitated the creation and cleansing of data tables, Power BI for data analysis and visualisation, and a Jupyter Notebook integrated with Visual Studio Code and the Python libraries sckilearn, pandas, Numpy, and Matplotlib to generate the graphs. While certain tables may not be utilised in Python's machine learning framework, they are still beneficial for data analysis as they aid in determining which attributes exhibit the strongest correlations.

As a result of the utilisation of wood fuels for heating purposes during the colder months of Finland and the subsequent increase in electricity prices due to the lengthier nights, it was determined, after cleaning and modelling multiple datasets, that energy consumption increases during the autumn and winter.

The environmental data for implementing a linear regression model and SVM with RBF kernel in addition to random forest, consists of energy consumption in Finland and temperature for 2020 and 2023. MSE is a significant indicator of the accuracy of the results; however, the deviation of the predicted values from the actual values is not substantial. This suggests that the prediction remains feasible and applicable for individuals with sufficient datasets who are interested.

However, the correlation between weather patterns and energy usage is also contingent upon the government's ability to profit from favourable weather conditions to generate energy and effectively distribute it to conserve it for colder seasons, Particularly in light of the existence of certain technologies and instruments, also applied machine learning for the prediction to get better insight.

# Summary

Climate variables, including temperature, wind speed, and sunlight, substantially influence energy usage across multiple industries. Heating and ventilation requirements are influenced by temperature, whereas wind speed affects buildings' efficacy and the performance of renewable energy sources. The influence of sunlight on both natural and artificial illumination impacts energy usage. Additionally, seasonality impacts energy consumption patterns. Comprehending this correlation is of the utmost importance in energy planning, demand forecasting, and the formulation of approaches to enhance energy efficiency and weather-related resilience. Hydropower, solar panels, wind turbines, and nuclear fusion collectively produce more than 3.5% of the world's electricity. Technological advances contribute to achieving the Sustainable Development Goals of the United Nations. The demand for electricity on a global scale is projected to double by 2050. The three primary approaches to CO2 removal from the atmosphere are technological solutions, augmentation of natural processes, and nature-based approaches. By harnessing meteorological data, machine learning methodologies have the potential to enhance energy consumption forecasting through the identification of critical energy consumption drivers, facilitation of proactive energy management, and demand response measures. Gaining insight into the interplay between weather and energy can positively affect resource allocation, environmental sustainability, and energy distribution. This knowledge can facilitate the development of policies, demand response initiatives, energy efficiency improvements, climate resilience planning, emissions reduction, and the integration of renewable energy sources.

This is an excellent opportunity to employ what I have learned and mastered regarding data analysis and machine learning. Specifically, I have gained insight into the challenges of obtaining appropriate datasets and establishing correlations among disparate tables. Despite what I have learned, there are still many things I am unaware of, which motivates me to continue my research and complete additional projects incorporating diverse data and content.

# References

Abdullahi, A. (2023, December 15). *What is Data Science? Benefits, Techniques and Use Cases*. TechRepublic. https://www.techrepublic.com/article/what-is-data-science/

Brainly. (2022). *III. How Clouds Affect the Weather ​—Brainly.ph*. https://brainly.ph/question/29549768

Britannica. (2024, April 22). *Energy | Definition, Types, Examples, & Facts | Britannica*. https://www.britannica.com/science/energy

Brown, S. (2024, May 6). *Machine learning, explained | MIT Sloan*. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Budinis, S. (2020, January 31). *Going carbon negative: What are the technology options? – Analysis*. IEA. https://www.iea.org/commentaries/going-carbon-negative-what-are-the-technology-options

Bureau of Meteorology. (2023, December 7). *The Relationship Between Weather and Energy Consumption: Unveiling the*. Bureau of Meteorology. https://bureauofmeteorology.org/blogs/news/the-relationship-between-weather-and-energy-consumption-unveiling-the-connection

Center for Science Education. (2024). *What Is Weather? | Center for Science Education*. https://scied.ucar.edu/learning-zone/how-weather-works/weather

Crabtree, M. (2023, July). *What is Data Analysis? An Expert Guide With Examples*. https://www.datacamp.com/blog/what-is-data-analysis-expert-guide

Ellis, D. (2022, May 4). *Top 10 issues facing the energy industry*. https://energydigital.com/top10/top-10-issues-facing-the-energy-industry

Esin, S. (2023, February 13). *What technology do we need to cut carbon emissions?* Grantham Research Institute on Climate Change and the Environment. https://www.lse.ac.uk/granthaminstitute/explainers/what-technology-do-we-need-to-cut-carbon-emissions/

GeeksforGeeks. (2017, March 19). *Linear Regression (Python Implementation)*. GeeksforGeeks. https://www.geeksforgeeks.org/linear-regression-python-implementation/

GeeksforGeeks. (2023a, August 21). *Types of Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/types-of-machine-learning/

GeeksforGeeks. (2023b, September 8). *What is a Dataset: Types, Features, and Examples*. GeeksforGeeks. https://www.geeksforgeeks.org/what-is-dataset/

Gillis, A. S. (2020). *What is Excel? Everything You Need to Know - Definition by WhatIs.com*. Enterprise Desktop. https://www.techtarget.com/searchenterprisedesktop/definition/Excel

Heggie, J. (2023, November 6). *Technology: Powering the Future of Energy*. Science. https://www.nationalgeographic.com/science/article/paid-content-technology-powering-the-future-of-energy

Heller, M. (2022, July 8). *What is Visual Studio Code? Microsoft’s extensible code editor*. InfoWorld. https://www.infoworld.com/article/3666488/what-is-visual-studio-code-microsofts-extensible-code-editor.html

Herring, D. (2020, October 29). *Are there positive benefits from global warming? | NOAA Climate.gov*. http://www.climate.gov/news-features/climate-qa/are-there-positive-benefits-global-warming

IBM. (2024, April 2). *What Is Random Forest? | IBM*. https://www.ibm.com/topics/random-forest

Jagdeesh. (2023, September 15). Types of Data in Statistics—A Comprehensive Guide. *Machine Learning Plus*. https://www.machinelearningplus.com/statistics/types-of-data-in-statistics/

JV. (2022). *Fossil CO2 emissions at record high in 2023*. Global Carbon Budget. https://globalcarbonbudget.org

Kolkowska, N. (2023, May 31). Challenges in Renewable Energy. *Sustainable Review*. https://sustainablereview.com/challenges-in-renewable-energy/

Kumar, A. (2020). *SVM RBF Kernel Parameters With Code Examples—DZone*. Dzone.Com. https://dzone.com/articles/using-jsonb-in-postgresql-how-to-effectively-store-1

Marutzky, R. (2023). Utilization of Wood for Energy. In P. Niemz, A. Teischinger, & D. Sandberg (Eds.), *Springer Handbook of Wood Science and Technology* (pp. 1695–1712). Springer International Publishing. https://doi.org/10.1007/978-3-030-81315-4\_31

Matillion. (2023). *What is ETL? The Ultimate Guide*. Matillion. https://www.matillion.com/blog/what-is-etl-the-ultimate-guide

National Geographic Society. (2023, October 19). *Cloud Cover*. https://education.nationalgeographic.org/resource/cloud-cover

National Geographic Society. (2024). *Geothermal Energy*. https://education.nationalgeographic.org/resource/geothermal-energy

Noomi, E. (2024, March 26). *The importance of precipitation for ecosystems on Earth | Lund University*. https://www.lunduniversity.lu.se/article/importance-precipitation-ecosystems-earth

Panik, M. (2020). *Mean Squared Error: Definition and Example*. Statistics How To. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/

Roser, M. (2024). The world’s energy problem. *Our World in Data*. https://ourworldindata.org/worlds-energy-problem

Sandberg, A. (2022, April 5). *Statistics Finland—Energy supply and consumption*. Statistics Finland. https://stat.fi/til/ehk/2020/04/ehk\_2020\_04\_2021-04-16\_tie\_001\_en.html

Shinde, S. (2023, January 11). Top 10 Machine Learning Python Libraries: Uses and Examples. *Emeritus Online Courses*. https://emeritus.org/blog/machine-learning-python-libraries/

Sidath, A. (2022, November 15). *Classification in Machine Learning: An Introduction | Built In*. https://builtin.com/machine-learning/classification-machine-learning

Simplilearn. (2020, March 3). *What Is Data: Types of Data, and How to Analyze Data [Updated]*. Simplilearn.Com. https://www.simplilearn.com/what-is-data-article

Singh, V. (2023, September 7). *Difference Between Jupyter Notebook and Python IDLE - Shiksha Online*. https://www.shiksha.com/online-courses/articles/jupyter-notebook-vs-python-idle/

Smith, J. (2024). *Unlocking Data Insights with Power BI. Data Analytics Today,*.

Tatvasoft. (2015, July 14). *ETL Process (Extract Transform Load)*. TatvaSoft Blog. https://www.tatvasoft.com/blog/etl-process-extract-transform-load/

The Editors of Encyclopædia Britannica. (2016). *Geothermal energy—Renewable, Heat, Power | Britannica*. https://www.britannica.com/science/geothermal-energy/History

The Editors of Encyclopædia Britannica. (2024a, March 7). *Kinetic energy | Definition, Formula, Units, Examples, & Facts | Britannica*. https://www.britannica.com/science/kinetic-energy

The Editors of Encyclopædia Britannica. (2024b, March 12). *Wind | Meteorology, Renewable Energy & Climate Change | Britannica*. https://www.britannica.com/science/wind

The Editors of Encyclopædia Britannica. (2024c, March 15). *Biomass | Definition, Types, & Facts | Britannica*. https://www.britannica.com/science/biomass

The Editors of Encyclopædia Britannica. (2024d, March 17). *Sunlight | Definition, Wavelengths, & Facts | Britannica*. https://www.britannica.com/science/sunlight-solar-radiation

The Editors of Encyclopædia Britannica. (2024e, March 19). *Wind turbine | Renewable Energy, Efficiency & Design | Britannica*. https://www.britannica.com/technology/wind-turbine

The Editors of Encyclopædia Britannica. (2024f, March 21). *Precipitation | Rain, Snow, Sleet & Hail | Britannica*. https://www.britannica.com/science/precipitation

The Editors of Encyclopædia Britannica. (2024g, March 29). *Temperature | Definition, Scales, Units, & Facts | Britannica*. https://www.britannica.com/science/temperature

the Organization for Economic Cooperation and Development. (2014, October 30). *Electrical energy—Energy Education*. https://energyeducation.ca/encyclopedia/Electrical\_energy

UMME HABIBA, T. (2023, September 23). What is the coefficient and intercept in Machine Learning and why do we use this concept in ML ? *Medium*. https://medium.com/@umme7072/what-is-the-coefficient-and-intercept-in-machine-learning-and-why-do-we-use-this-concept-in-ml-c1e5a4c7c36d

US EPA, O. (2015, August 10). *Learn about Energy and its Impact on the Environment* [Overviews and Factsheets]. https://www.epa.gov/energy/learn-about-energy-and-its-impact-environment

Visit Finland. (2024). *Weather and climate in Finland*. https://www.visitfinland.com/en/practical-tips/climate-and-weather-in-finland/

Wang, E. (2020, November 24). Problem Statement. *Grand Challenges Energy Usage Research Project*. https://medium.com/grand-challenges-energy-usage-research-project/problem-statement-5bd0e6805592

Yung, Z. (2023, August 4). *ETL Explained: Extract - Transform - Load*. Coupler.Io Blog. https://blog.coupler.io/what-is-etl/

Zhang, C., Liao, H., & Mi, Z. (2019). Climate impacts: Temperature and electricity consumption. *Natural Hazards*, *99*(3), 1259–1275. https://doi.org/10.1007/s11069-019-03653-w

**Appendix: Material Management Plan**

I made a file to stock the datasets I will use, such as the Weather in Helsinki for the past three years, precipitation, and energy consumption. I plan to manifest those data and use them in Excel for ETL and pass them to Power BI for modelling, analysis, and visualisation. Afterwards, the upgraded table and data will be used in Visual Studio code for applying machine learning by Python.

**Development project:**

I may save the folder named “Thesis Documents” that contains datasets and it’s upgraded in Google Drive exactly in this following link: <https://drive.google.com/drive/folders/1ziF2TmgjmrkxqXZohIC64PCdsm44nQfR?usp=sharing>

The thesis documents contain a folder of numbered figures used in the thesis, a Jupyter notebook opened in Visual Studio Code, where I wrote multiple codes to make the predictions using models, and a Power BI reports file.

**Research work:**

All the research used in the thesis has been done using books or open-source documentation. The Excel files were adjusted and added sheets to help with Power BI and machine learning in Python. The first 2 figures were from open sources, but all the rest are either in Excel, Power BI, or VS code.