Prognose über die Stromerzeugung einer Photovoltaikanlage mittels maschinellem Lernen

Bachelorarbeit

Vorgelegt von

Leopold Schmid

Matrikelnummer:79776

Fakultät für Elektronik und Informatik

Hochschule Aalen

Datum

Erklärung

Verzeichnis häufig verwendeter Symbole und Abkürzungen

Inhaltsverzeichnis

1 Einführung

In der größten bisher durchgeführten Studie zu Klimaangst von jungen Menschen behaupten 45% der Befragten, dass ihre Gefühle bezüglich des Klimawandels ihr tägliches Leben negativ beeinflussen. Ohne jede Zweifel stellt der Klimawandel uns und die uns folgenden Generationen vor eine nicht zu unterschätzende Herausforderung. Angefangen mit der Tatsache, dass die alleinige Diskussion darüber nicht selten zu einer Spaltung und Polarisierung der Gesellschaft führt. Folglich leitet die hitzige und emotionale Debatte dazu, dass die gegensätzlichen Lager sich immer weiter voneinander entfernen und somit jegliche Grundlage für eine zielführende Diskussion entreißen.

Auf der einen Seite werden die Befürchtungen der anderen für übertrieben, unwichtig und paranoid gehalten. Wissenschaftliche Unsicherheiten werden verwendet, um gesamte Ergebnisse von Studien als unzuverlässig einzustufen. Modelle werden in der Wissenschaft oft vereinfacht, denn die Isolation und Fokussierung auf bestimmte Aspekte kann helfen, um ein besseres Verständnis von Phänomenen in der Natur zu erlangen. Auf Grund von dieser Vereinfachung werden die Klimamodelle als zu ausdruckslos betitelt, um die Komplexität des Klimas von unserem Planeten widerzuspiegeln. Dementsprechend seien die Schlussfolgerungen aus den Studien unzutreffend. Investitionen mehrerer Milliardenbeträge seien nicht gerechtfertigt und politische Vorgaben schaden dem eigenem Land mehr, als dass sie dem Planeten helfen würden.

Am gegenüberliegenden Ufer wird gemahnt, die Ernsthaftigkeit der Situation nicht zu unterschätzen. Gewarnt wird, dass die Konsequenzen des Klimawandels irreversibel seien, weswegen Treibhausgase unverzüglich auf ein Minimum reduziert werden sollten. Um Folgen wie das Abschmelzen der Eisschilde und Gletscher, das Aussterben verschiedenster Tierarten und der Anstieg des Meeresspiegels zu verhindern, spielen erneuerbare Energien eine entscheidende Rolle. Im Gegensatz zu fossilen Energieträgern erzeugen die Erneuerbaren keine oder nur kaum Treibhausgasemissionen. Bis 2030 sollen die regenerativen Energien 80% des Strombedarfs Deutschlands decken. Da die Möglichkeiten der Energieerzeugung in Deutschland durch Wasserkraft und Biomasseverbrennung schon

heute nahezu ausgeschöpft sind, bilden Solar- und Windenergie einen wichtigen Grundpfeiler um die Ausbauziele zu erreichen.

Ebenso komplex wie die Klimamodelle, die die Verstrickungen verschiedenster Phänomene in unserer Natur berücksichtigen sollen, ist allerdings die Transformation unseres Stromnetzes. In einem Netz, welches dafür ausgelegt wurde, dass wenige größere Kraftwerke Strom einspeisen, werden künftig immer mehr kleinere, dezentrale Kraftwerke mitwirken. Als Folge kommen diverse Herausforderungen auf uns zu. Unter anderem verträgt nicht jedes elektronische Bauteil in einem Niederspannungsnetz Rücklaufstrom. (Quelle?) Allerdings könnte es genau zu einem solchen Rücklaufstrom kommen, wenn in dem Niederspannungsnetz mehr Strom erzeugt als verbraucht wird. Des Weiteren reagiert ein Stromnetz äußerst empfindlich auf Spannungsschwankungen. Auf Grund dessen müssen Netzbetreiber darauf achten, dass genauso viel Strom verbraucht wie erzeugt wird. Solar- und Windenergie zählen zu den fluktuierenden Energieerzeugern, ergo ist die Stromproduktion nur sehr begrenzt regulierbar und somit nicht an die aktuelle Marktnachfrage anpassbar.

Die Relevanz der Abschlussarbeit basiert auf der Annahme, dass eine genauere Prognose über die erzeugte Strommenge helfen würde das Stromnetz zu stabilisieren und Netzschwankungen, ergo das Risiko eines Brown- oder sogar Black-Outs zu minimieren. Anhand der neu gewonnenen Information könnten kurzfristige Stromimporte beziehungsweise -exporte besser reguliert werden. Ebenso wäre es denkbar, dass man für die Industrie und/ oder Endverbraucher mehr Anreize schafft ihren Stromverbrauch zu einem gewissen Grad an die Verfügbarkeit des Stroms anzupassen, wobei die Information über die Verfügbarkeit von Strom gleichfalls hilfreich wäre. Zuletzt könnten starke Abweichungen zwischen der Prognose und der tatsächlich erzeugten Strommenge ein Indiz darauf sein, dass die Solaranlage eine Wartung benötigt.

2 Stand der Technik

Im folgenden soll betrachtet werden, wie das Stromnetz von Deutschland im Moment aufgestellt ist und welche Maßnahmen ergriffen werden müssen, um den Bedarf an Strom zu decken und die Energiewende zu meistern. Zudem gilt es, den in Zukunft zu erwartenden Strombedarf zu analysieren.

2.1 Strombedarf in Deutschland

Im gesamten Jahr 2018 betrug die realisierte Stromerzeugung in Deutschland 543.053.395 MWh, wodurch gute 30 TWh Strom mehr erzeugt als benötigt wurden. 2022 hingegen sank der Bedarf und die Erzeugung auf ungefähr 490.000.000 MWh.

Tatsache ist jedoch, dass die Dekarbonisierung durch Elektrifizierung unserer Industrie, des Verkehrs als auch des privaten Sektors stattfinden soll, weshalb mit einer deutlichen Zunahme des Energiebedarfs zu rechnen ist. Allem voran kommen hierbei Elektroautos und Wärmepumpen ins Spiel, für dessen Betrieb elektrische Energie eine Grundvoraussetzung ist.

Die Wärmepumpen führen dazu, dass der Stromverbrauch im Winter deutlich steigen wird. Zwar kann diese ebenfalls im Sommer zum Kühlen der Räume verwendet werden, allerdings ist dabei der Stromverbrauch, zumindest bei einem Teil der Wärmepumpen, signifikant geringer. Nämlich reicht der Erdwärmepumpe eine kleine Umwälzpumpe aus, um die niedrigen Temperaturen aus den Tiefen zu holen und damit das Haus zu kühlen. Der Verdichter, der am meisten Strom verbraucht, wird nur von Luft-Wasser-Wärmepumpen zum Kühlen benötigt.

Bis 2030 sollen sich 15 Millionen Elektroautos auf den deutschen Straßen fortbewegen, damit wäre nahezu jedes vierte Auto im Mutterland der Automobilindustrie kein klassischer Verbrenner mehr, sondern von Energie aus der Wallbox abhängig.

Schließlich hat sich Deutschland es als Ziel gesetzt, bis 2030 65 Prozent weniger Treibhausgase zu emittieren als 1990. Hierzu soll der noch bisher hinkende Verkehrssektor durch die Elektrifizierung aufholen.

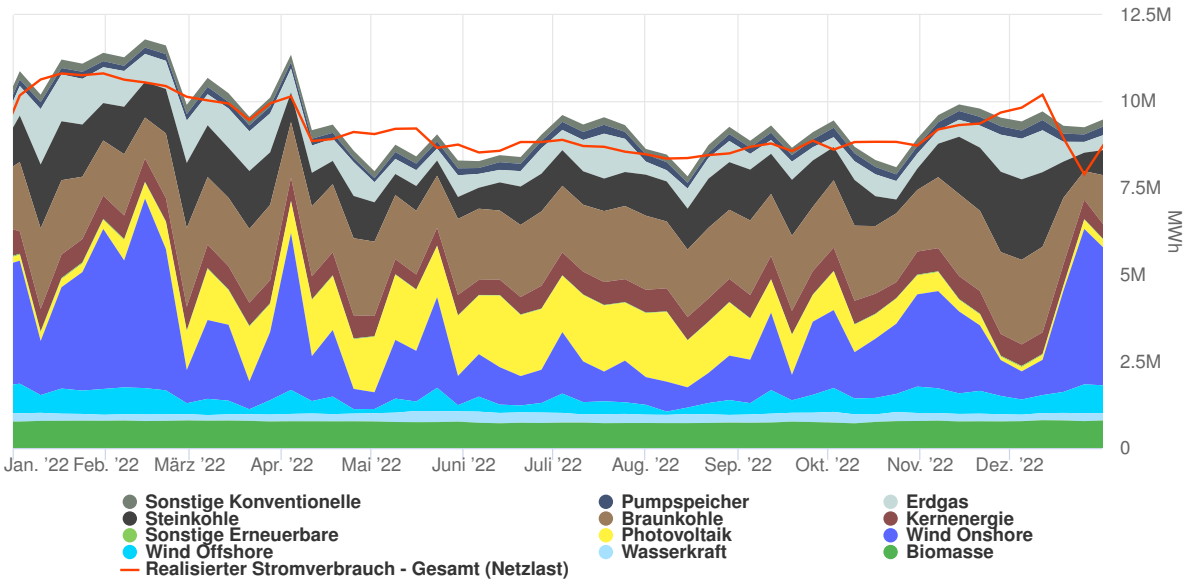

Abbildung 1: Stromerzeugung und -bedarf, Deutschland 2022

Um die Klimaneutralität bis 2045 zu erreichen, sollen parallel dazu ab 2024 500.000 neue Wärmepumpen pro Jahr installiert werden. Wärmepumpen werden mit elektrischer Energie betrieben und entziehen der Umweltwärme die Energie, um diese für die Wärmeversorgung von Gebäuden zu verwenden. So können wir uns von fossilen Energieträgern nach und nach freimachen.

Das Bunderministerium für Wirtschaft und Klimaschutz (BMWK) prognostiziert für 2030, dass der Bedarf an Strom auf 750 TWh ansteigen wird.

2.2 Strommix Deutschland

Laut dem Bundesministerium für Wirtschaft und Klimaschutz (Stand 1. April 2020) ist Deutschland im Besitz von Stromerzeugungsanlagen mit einer Netto-Nennleistung von effektiv 221,3 Gigawatt. Die Netto-Nennleistung bezeichnet die Energie, die eine Anlage unter Normalbedingungen ohne Beeinträchtigung der Lebensdauer vollbringen kann, bereits abgezogen ist der Eigenbedarf der Stromerzeugungsanlage. Im Falle einer Photovoltaikanlage ist Brutto-Nennleistung vor allem wegen des Wechselrichter höher.

Der Anteil der erneuerbaren Energien an Deutschlands Netto-Nennleistung beläuft sich dabei bereits auf mehr als die Hälfte, nämlich 121 Gigawatt. Füh-

rend unter den erneuerbaren Energien ist die On- und Offshore Windenergie, Energie aus Photovoltaikanlagen kommt mit 47,3 Gigawatt an zweiter Stelle. Mit 21,2 Prozent spielen sie dementsprechend eine entscheidende Rolle in der Energieversorgung unseres Industriestaates. Dabei ist jedoch zu beachten, dass die verschiedenen Stromerzeuger je nach Jahreszeit eine unterschiedlich wichtige Stellung einnehmen. Während Photovoltaik in der Winterzeit auf einen Bruchteil zurückgeht, erreicht die Windenergie im Winter seine Maximalwerte, die jedoch selbst in der wöchentlichen Betrachtung mit stärkeren Schwankungen einhergehen. In Abbildung 1 ist ersichtlich, dass temporär fast die Hälfte des Strommix durch auf dem Land befestigte Windkraftanlagen erzeugt wird. Ebenso ist erkenntlich, dass Deutschland auf den Import von Energie angewiesen ist, wenn eine Flaute aufzieht und die Erneuerbaren weniger Energie als üblich erzeugen. Ausschließlich Wasserkraft und Biogasanlagen sind in der Lage konstant Energie zu produzieren. Sowohl im Mai als auch Mitte Oktober 2022 konnten konventionelle Energieerzeuger die Diskrepanz nicht ausgleichen, weswegen die Vernetzung des Stromnetzes und dadurch möglichen Importe beziehungsweise Exporte unentbehrlich sind. Schließlich wären die Investitionskosten für Speicherkraftwerke oder überschüssige Kraftwerke, die selbst Dunkelflauten genügend Energie noch erzeugen können, weder ökologisch noch ökonomisch tragbar. Für eine gestärkte Versorgungssicherheit ist die Vernetzung elementar. Hilfreich ist, dass die unterschiedlichen geographischen Gegebenheiten der Länder verschiedene Arten der regenerativen Energieerzeugung ermöglichen. So verfügen die Alpenländer und Skandinavien eine erheblich bessere Grundvoraussetzung, um Wasserkraft Gebrauch zu machen. So ist das vernetze Stromnetz gegenüber Großwetterlagen, die zum Beispiel die Produktion von Solar- und Windkraftanlagen auf weiten Gebieten des Kontinents beeinflussen, weniger anfällig.

2.3 Aufbau des Stromnetzes

Auf Grund der geographischen Lage liegt Deutschland äußerst zentral in einem ineinander verstricktem europäischen Stromsystem. Somit nimmt es in Europa eine Schlüsselrolle ein und ist eine Drehscheibe für den Stromfluss innerhalb des Kontinents.

Um das Stromnetz genauer zu betrachten, muss zunächst zwischen Hochspannungsnetzen, Mittelspannungsnetzen und Niederspannungsnetzen differenziert werden. Mit 60 bis 220 Kilovolt Spannung ist der landesweite Transport des Strom signifikant effizienter, da durch die höhere Spannung die Stromstärke gesenkt werden. Schließlich nimmt der Verlust quadratisch mit der Stromstärke zu. Sozusagen sind Hochspannungsnetze die Autobahnen des Stromnetzes. (QUELLE)

Bisher wurden in der öffentliche Debatte die großen Stromtrassen thematisiert. Beschwerden, dass die Trassen zu nah am eigenen Grundstück wären, endeten im schleppenden Ausbau. Keineswegs zu unterschätzen sei jedoch die Wichtigkeit der lokalen Niederspannungsnetze. Die dezentrale Energieerzeugung führt dazu, dass die Energiewende genauso dort stattfinden muss. Leitungen, Trafos und Schaltkästen müssen weiter ausgebaut und erneuert werden. Nicht übertrieben wäre es zu sagen, dass die Energiewende im Verteilernetz stattfindet.

Heutzutage sind Verbraucher längst nicht mehr nur Verbraucher, stattdessen werden viele von ihnen zu sogenannten Prosumenten. Hinter dem Begriff verbergen sich die Wörter Konsument und Produzent. Bereits 2022 gehörten ungefähr 1,4 Millionen private Haushalte zu den eben diesen Prosumenten.

2.4 Schwankungen im Stromnetz

Dementsprechend ist es für uns unverzichtbar, die Wichtigkeit einer stabilen Stromerzeugung richtig einzuschätzen und anhand dessen Maßnahmen zu ergreifen, um Spannungsschwankungen im Netz so gut wie möglich zu reduzieren.

Der stetig schwankende Strombedarf und die ebenso fluktuierende Stromerzeugung führt dazu, dass von einzelnen Ländern in größeren Dimensionen Strom sowohl exportiert als auch importiert werden müssen. Dies hilft dabei, im gesamten Stromnetz die Netzspannung beständig und konstant zu halten. In 2019 konnte Deutschland 72,4 TWh Strom an verschiedene Nachbarländer exportieren und hat im selben Jahr etwa die Hälfte (39,8 TWh) davon importiert.

Auf Grund von Verbraucherverhalten sowie auch Teilen der Industrie ist der

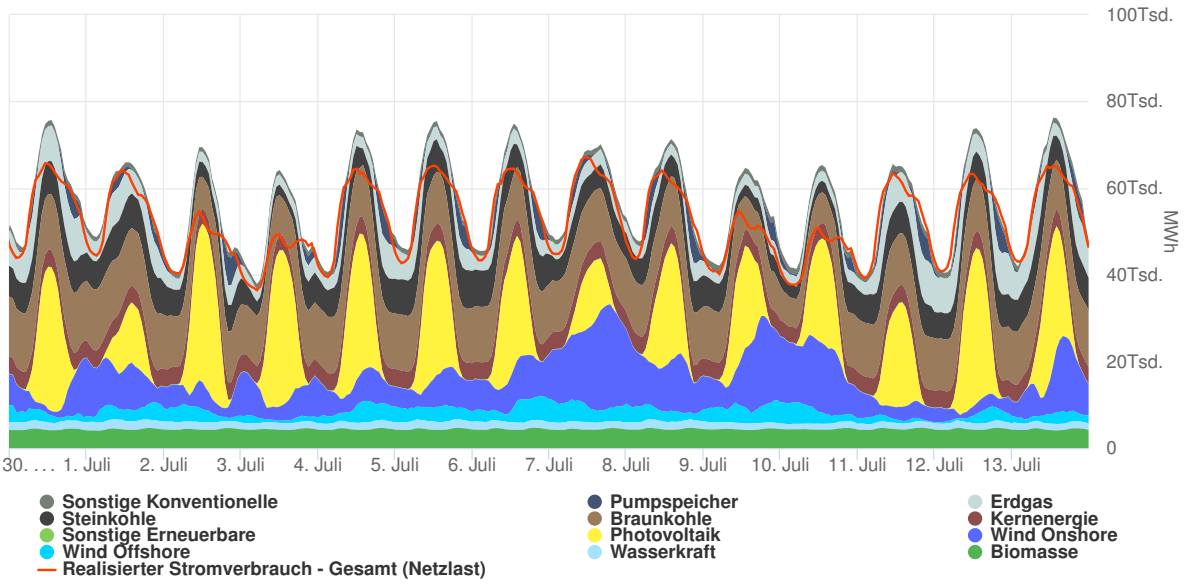

Abbildung 2: Realisierte Erzeugung Deutschland im Juli 2022

Strombedarf je nach Uhrzeit unterschiedlich. Äußerst gelegen ist die Sachlage, dass der Strombedarf tagsüber nahezu doppelt so hoch ist als während der Nacht. Von Frühling bis in die späten Sommermonate kann der zusätzliche Bedarf häufig durch Photovoltaik gedeckt werden und somit bildet Sonnenenergie eine gute Ergänzung für die anderen Energieerzeuger.

2.4.1 Redispatch

Zuvor wurde thematisiert, wie wichtig es ist, dass das Stromnetz auf internationaler Ebene miteinander verknüpft ist. Ebenso ist es notwendig, das Netz auf lokaler Ebene zu betrachten. Um die Netzstabilität zu gewährleisten, führen Übertragungsnetzbetreiber sogenanntes *Redispatching* durch. Dabei wird gezielt die Einspeisung durch Erzeugungsanlagen gedrosselt, sodass an Engpässen die Leitungen nicht überlastet werden. Da dies äußerst kurzfristig passiert und die zuvor gekaufte mit der tatsächlich erzeugten Strommenge übereinstimmen muss, beauftragen die Übertragungsnetzbetreiber andere Erzeugungsanlagen, die sozusagen hinter den Engpässen liegen, mehr Strom zu produzieren.

Eine Neuerung des Netzausbaubeschleunigungsgesetzes (NABEG) im Jahr 2019 wurde *Redispatch* weiterhin optimiert. Die Neuerung besagt, dass ab Ok-

tober 2021 die Netzbetreiber mehr Erzeugungs- und Speicheranlagen für das Unterbinden von Netzengpässen miteinbeziehen dürfen. Zuvor konnten sich Anlagen mit einer installierten Leistung unter 10 MW der Teilnahmepflicht am *Redispatch* entziehen. Der Grenzwert wurde nicht nur auf 100kW herabgesetzt, des Weiteren sind sämtliche steuerbaren Anlagen, unabhängig von deren installierten Leistung, ebenso an der Teilnahme verpflichtet.

3 Eigene Fragestellung und methodisches Vorgehen

Um die zukünftigen Herausforderungen der Energieversorgung zu bewältigen, befasst sich diese Arbeit damit den fluktuierenden Stromerzeuger, die Photovoltaikanlage, so gut wie möglich in unser bestehendes Stromnetz zu integrieren. Die zentrale Fragestellung lautet hierbei, wie man präzise den erzeugten Solarstrom prognostizieren kann. Um dieses Ziel zu erreichen, sollen zuerst entscheidende Parameter ermittelt werden, die die Stromerzeugung einer Solaranlage beeinflussen. Allerdings liegt der Fokus ebenfalls darauf, dass die Berechnung der Prognose für beliebige Photovoltaikanlagen anwendbar ist, indem nur wenige spezifische Informationen über die Anlage in die Berechnung einfließen. Selbstverständlich spielen verschiedenste Eigenschaften der Solaranlage eine entscheidende Rolle, jedoch werden diese im Datenmodell nicht berücksichtigt. Zunehmende Rechenkapazitäten und kostengünstigere Speichermöglichkeiten ermöglichen es, dass für jede einzelne Solaranlage ein eigenes Datenmodell angelegt wird. Somit können die Gegebenheiten der Photovoltaikanlage, wie Quantität und Qualität der Solarpanele, vernachlässigt werden. Durch die Trainingsdaten ist das Datenmodell selbstständig in der Lage die Leistungsfähigkeit der Solaranlage zu beurteilen. Ein weiterer Vorteil ist, dass dadurch schwer zu beschaffende Datensätze nicht weiter benötigt werden. Stattdessen wird das Modell fast ausschließlich von Wetterdaten trainiert, welche von verschiedensten Institutionen flächenmäßig und umfangreich aufgezeichnet werden. Die für das Trainieren des maschinellen Lernen Modells benötigten Wetterdaten werden von der Website *https://www.visualcrossing.com* extrahiert. Die Website bietet einen kostenfreien Zugang zu detaillierten historischen Wetterdaten, bis zu 50 Jahren vor heute.

HIER: QUELLE SOLARDATEN

Durch die Literaturrecherche sind bereits die relevanten Faktoren eingegrenzt wurden, allerdings ist eine exakte Analyse der Daten unumgänglich. Hierbei soll vor allem beachtet werden, ob die Aufnahme des Merkmals einen solch starken Einfluss auf die Stromerzeugung hat, sodass die zunehmende Komplexität des Modells gerechtfertigt ist. Herausfordernd ist dabei, ob die Informationen in man-

chen Merkmalen nicht bereits in anderen Merkmalen enthalten sind. Schließlich korrelieren einige Wetterdaten sehr stark miteinander. Ebenso gilt es zu überprüfen, ob sich die ausgewählten Merkmale auf unterschiedliche Solaranlagen auf die gleiche Art und Weise auswirken. Um dies zu überprüfen, sollen mehrere Solaranlagen an verschiedenen Standorten ausgesucht und getestet werden.

Sofern es gelingt, sämtliche sich variierende Parameter herauszufinden, die die Solarproduktion beeinträchtigen, ist maschinelles Lernen eine vielversprechende Technologie um die Stromerzeugung zu prognostizieren. Insbesondere ist zu erwarten, dass sich das Spektrum der Trainingsdaten und das Spektrum während der Produktion (Inbetriebnahme) sich nicht groß voneinander unterscheiden werden.

Eine Unvollkommenheit des Datenmodells ist jedoch die Alterung der Solaranlage, welche dazu führt, dass die Leistungsfähigkeit über die Lebenszeit sich reduziert. Wie stark die Alterung jedoch die Prognose beeinträchtigt, lässt sich über den Zeitraum dieser Arbeit von vier Monaten nur schwerlich beurteilen. Mögliche Konsequenzen wären ältere Daten kontinuierlich auszusortieren und mit das Modell mit den Neusten zu aktualisieren. Es ist davon auszugehen, dass der Alterungsprozess sich nur langsam in den Datensätzen bemerkbar machen wird, weshalb er die Ergebnisse über die kurze Zeitspanne kaum verfälschen dürfte.

4 Wichtige Faktoren bei der Stromerzeugung durch Solarenergie

Bevor der Stromertag einer Photovoltaikanlage mit Hilfe von künstlicher Intelligenz berechnet werden soll, ist es von immenser Bedeutung die maßgeblichen Faktoren herauszufinden. Wenn ein Modell eingelernt werden soll, können zu viele Faktoren zum *Fluch der Dimensionen* führen. Dieser besagt, dass zu viele Merkmale dazu verleiten, dass Muster und Strukturen sich schwerer erkennen lassen. Die Datenpunkte sind durch die Größe des mehrdimensionalen Raums weiter voneinander entfernt, wodurch die Interpretierbarkeit komplexer wird. Dementsprechend sind die Eingabedaten behutsam auszuwählen.

4.1 Wetter-unabhängige Faktoren

In Bezug auf die Rahmenparameter der Solaranlage ist die reine Größe selbstverständlich ein maßgeblicher Faktor. Hinzu kommt, dass sich dieses Merkmal für jede individuelle stark unterscheiden kann. Die Dimensionierung der Solaranlage hängt schließlich ebenfalls von verschiedenen Faktoren ab, darunter die verfügbare Fläche am Standort, dem Energiebedarf und dem Verwendungszweck.

4.1.1 Schatten

Durch die Aggregation von Metadaten der Solaranlagen könnte man die zuvor genannten Faktoren ebenfalls in das Datenmodell einfließen lassen. Die Komplexität des Modells würde erheblich zunehmen, jedoch wäre es somit auf verschiedene Solaranlagen verallgemeinerbar. Zwei Faktoren überdehnen jedoch nicht nur die Möglichkeiten der Anwendungen von maschinellem Lernen - oder würden die Vorhersagen des Modells zumindest signifikant beeinträchtigen. Des Weiteren stellen sie teilweise eine große Herausforderung beim Schritt der Datenerfassung dar. Erster Faktor ist die Umgebung der Solaranlage, die sich folglich nur auf jene Solaranlage auswirken. Umliegende höhere Gebäude, Hügel und Bäume können Schatten auf die Solarmodule werfen, wodurch die Sonnenstrahlung blockiert und die Stromproduktion reduziert wird. Gesondert anspruchsvoll wird es dadurch, dass der geworfene Schatten von der Elevation der Sonne abhängig

ist. Die Elevation der Sonne wiederum ist von der Jahreszeit abhängig. Die Wahrscheinlichkeit, dass die Solarpanele durch umliegende Strukturen verdeckt wird, ist somit im Winter höher.

4.1.2 Orientierung

Der zweite Faktor ist die Orientierung der Solarmodule, womit im Fachjargon der Architektur die Ausrichtung eines Baukörpers nach den Himmelsrichtung gemeint ist. Sofern die Module statisch befestigt sind, sollten die Module Richtung Süden ausgerichtet sein, um den höchsten Stromertrag zu erzielen. Allerdings lassen sich die örtlichen Gegebenheiten bei der Installation der Solarpanele nicht ignorieren. Vor allem Solaranlagen, die für den Eigenbedarf installiert wurden, sind häufig auf Dächern befestigt. Allein aus Sicherheitsgründen werden die Solarmodule in aller Regel flach auf den Dachziegeln des Schrägdachs montiert, da ansonsten starker Wind die Module aus ihrer Befestigung reißen könnte. Folglich gibt die Ausrichtung des Gebäudes in vielen Fällen die Ausrichtung der Solarmodule ohne großen Spielraum vor. Konsequenz dessen ist, dass die Leistungsfähigkeit einer Solaranlage so vielfältig sein kann, wie die Bedachung von Häusern individuell ist.

Die Ausrichtung der Solarmodule hat nicht nur auf die gesamte erzeugte Strommenge Auswirkungen, des Weiteren führt sie dazu, dass Solaranlagen mit vergleichbarer Gesamtleistung zu unterschiedlichen Uhrzeiten kontrastiert Strom produzieren. Für die Netzstabilität ist es jedoch zwingend notwendig, dass kontinuierlich für eine gleichbleibende Spannung gesorgt wird. Folglich interessieren wir uns nicht nur für die kumulativ erzeugte Strommenge einer Photovoltaikanlage, weitaus spannender sind die Echtzeit-Vorhersagen.

Sowohl die Ausrichtung der Solarmodule als auch die Umgebung der Anlage soll in das Modell einfließen, indem Uhr- und Jahreszeit zu der entsprechenden Strommenge festgehalten werden. Die Logik dahinter ist mit der Annahme verbunden, dass sich die beiden Faktoren durch Uhr- und Jahreszeit repräsentieren lassen, weil die Konstellationen wiederkehrend sind. Ein Haus oder ein Baum, dass zwischen 11:14 und 11:46 Uhr Schatten auf die Solaranlage wirft, wird mit

hoher Wahrscheinlichkeit die Stromproduktion der Solaranlage am nächsten Tag auf die gleiche Weise beeinträchtigen. Das Ziel ist der Forschung ist, dass das Datenmodell diesen Zusammenhang erkennt und eine niedrigere Stromproduktion prognostiziert als am Nachmittag, wenn sonst die gleichen Bedingungen vorliegen.

Die Jahreszeit bündelt mehrere Faktoren und nimmt somit Komplexität aus dem Modell, ohne dabei entscheidende Rahmenparameter zu vernachlässigen. Wie bereits erwähnt können sich die Umgebungsfaktoren über die Jahreszeiten hinweg verändern, zum anderen wandert die Sonne in einem anderen Winkel über die Solaranlage. Der Einstrahlungswinkel und die Umgebung sind entscheidend für die Stromproduktion und werden durch die Jahreszeit repräsentiert.

Die Kalenderwoche scheint ein guter Kompromiss zu sein, um die verschiedenen Faktoren abzubilden. Der *n-te* Tag des Jahres würde den Merkmalsraum des Modell deutlich vergrößern. Da die Dimension des Merkmals 1 bis 366 statt 1 bis 53 wäre, ist die Wahrscheinlichkeit, dass sich die einzelnen Datenpunkte in der Größe des Raums verlieren, deutlich höher. Die Elevation der Sonne unterliegt zwar einem ständigen Wandel, allerdings ist es zudem äußerlich fraglich, ob die minimalen Differenzen zwischen den einzelnen Tagen für die Prognose der Stromerzeugung einer Photovoltaikanlage überhaupt bemerkbar sind.

4.2 Nachgeführte Photovoltaikanlagen

Sogenannte nachgeführte Photovoltaikanlagen folgen selbstständig und automatisiert dem Sonnenstand, wodurch die Solarstromproduktion gegenüber stationären Anlagen verbessert wird. Optimalerweise trifft das Sonnenlicht fortwährend senkrecht auf die Solarmodule. Um dies zu bewerkstelligen, wird der Neigungswinkel und/ oder die Ausrichtung nach der Himmelsrichtung an den aktuellen Sonnenstand angepasst.

Die im letzten Unterkapitel genannten Faktoren werden aus dem Datenmodell mit der Begründung ausgeschlossen, weil für jede Solaranlage ein eigenes Modell angelegt wird und somit bei gleichen Wetterbedingungen dieselben Ergebnisse erzielt werden. Sowohl die Leistungsfähigkeit, der Wirkungsgrad des

Wechselrichters als auch die Umgebungsfaktoren sind nahezu feste Rahmenparameter. Bei nachgeführten Photovoltaikanlagen haben wir nun den Fall, dass sich ein wichtiger Faktor, die Position des Solarmoduls in Bezug auf die Sonne, stets verändert, ohne dass dies in den Eingabedaten des Modells bemerkbar ist. Um die Konsequenzen für unser Vorhersagemodell zu beurteilen, müssen wir zwischen zwei verschiedenen Methoden, um eine Photovoltaikanlage nachzuführen, unterscheiden. Zum einen gibt es die astronomische und die sensorische Steuerung von PV-Anlagen.

4.2.1 Astronomisch nachgeführte Photovoltaikanlagen

Bei der astronomischen Steuerung werden die Solarmodule kontinuierlich zur Sonne hin ausgerichtet, unabhängig von der Wolkendecke. In diesem Szenario ist die Ausrichtung der Solarmodule zwar dynamisch, allerdings wird mit Hilfe dieser Methode die Ausbeute der Stromproduktion stets auf die gleiche Art und Weise verbessert. Insofern macht es für das maschinelle Lernen keinen Unterschied, ob die Solaranlage stets zur Sonne gewandt oder statisch montiert ist. Den Sonderfall, dass die Getriebemotoren beschädigt sind und ausfallen, sei an dieser Stelle außen vorgelassen.

4.2.2 Sensorisch nachgeführte Photovoltaikanlagen

Spannender wird es bei der aufwendigeren Technologie, die einen lokal installierten Sensor die optimale Ausrichtung der Solarpanele ermitteln lässt. Der hellste Punkt am Himmel wird durch die Sensorsteuerung wahrgenommen und dementsprechend werden die Sonnenkollektoren ausgerichtet. Der hellste Punkt am Himmel kann sich jedoch sehr schnell ändern, da er von der aktuellen Wolkendecke abhängt. Vor allem bei einer durchwachsenen Wolkendecke lässt sich nur schwer ermitteln, ob die Solarpanele im Moment von einer Wolke bedeckt werden.

HIER: WOLKENKAMERA ERMÖGLICHT DIES, WEBSITE VERLINKEN

Folglich wird die Ausrichtung der Sonnenkollektoren durch den Sensor kontinuierlich an die aktuellen Gegebenheiten angepasst. Die exakte Wolkensituation

unterliegt einem stetiger Veränderung und wird keineswegs durch die Eingabedaten der Realität entsprechend repräsentiert, wodurch Ungenauigkeiten bei der Prognose entstehen können.

Die gleiche Problematik besteht auch bei herkömmlichen Solaranlagen, deren Module nicht sensorgesteuert ausgerichtet werden. Insbesondere werden die für dieses Projekt zur Verfügung stehenden Wetterdaten nicht in dem Intervall aktualisiert, indem sich die Wetterlage in der Wirklichkeit verändert. Da das Ziel dieser Arbeit ist, die gemittelte, erzeugte Strommenge stündlich vorauszusagen, soll uns die ständig schwankende Leistungsabgabe einer Solaranlage nicht fortführend stören.

Inwiefern die Prognose bei durch einen Sensor nachgeführten Photovoltaikanlage verschlechtert wird, gilt es zu untersuchen. Da vor allem die sensorgesteuerte Variante der nachgeführten Photovoltaikanlagen auch mehrere Nachteile, wie höhere Installations- und Wartungskosten mit sich bringt, wird der Marktanteil solcher Anlagen auf eher gering eingeschätzt. Folglich werden die Auswirkungen auf die Prognose nicht in dieser Arbeit untersucht.

4.3 Wetter-abhängige Faktoren

Verschiedene Wolkentypen wirken sich unterschiedlich auf die Sonneneinstrahlung aus. Zudem gehören Wolken zu den unbeständigen Faktoren, die sich sprunghaft auf die Stromerzeugung auswirken. Selbst wenn exakte, detailreiche Daten über die Bewölkung für die Forschung dieser Arbeit nicht vorliegen, soll der Effekt der unterschiedlichen Wolkentypen im folgenden kurz erläutert werden.

4.3.1 Wolkenarten

Stratuswolken sind flache, graue Wolken, die den Himmel oft bedecken. Sie bestehen aus Wassertröpfchen und liegen in niedriger Höhe. Stratuswolken blockieren die Sonneneinstrahlung und reduzieren die Helligkeit des Tageslichts erheblich. Sie haben eine kühlende Wirkung, da sie einen großen Teil der Sonnenenergie reflektieren.

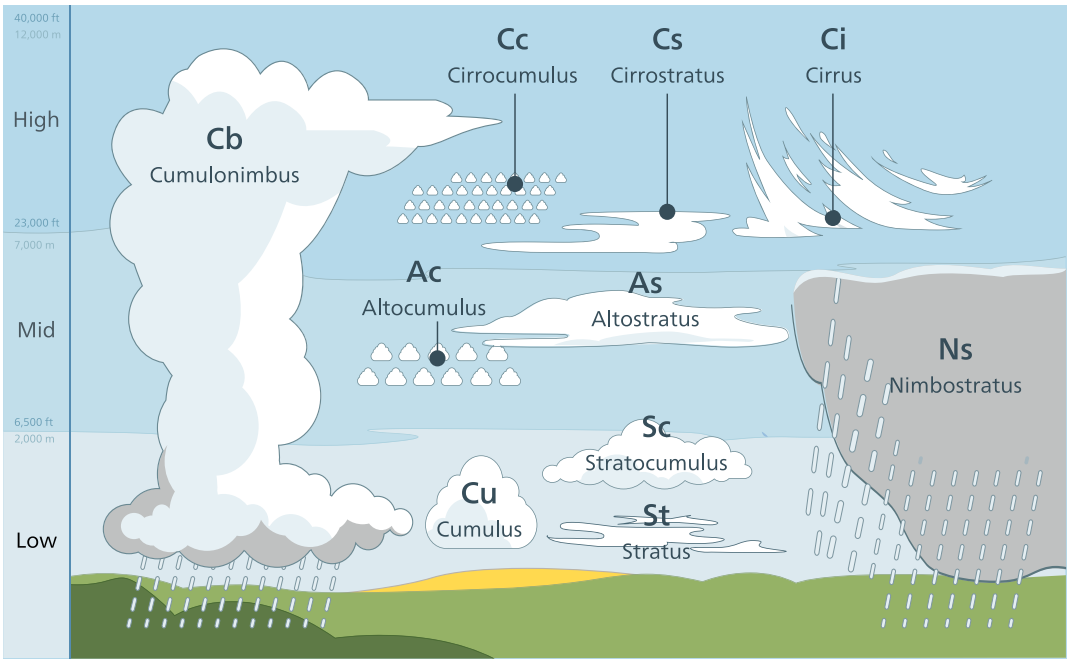

Abbildung 3: Wolkentypen

Cumuluswolken sind große, weiße, flauschige Wolken mit einer flachen Basis und einer kuppelförmigen Oberseite. Sie treten oft an sonnigen Tagen auf. Cumuluswolken können die Sonneneinstrahlung beeinflussen, indem sie sie teilweise reflektieren und teilweise absorbieren. Dadurch entstehen Schatten und Sonnenflecken auf der Erdoberfläche.

Cirruswolken sind dünne, faserige Wolken, die in großen Höhen schweben. Sie bestehen aus Eiskristallen und erscheinen oft als Federwolken oder Schleierwolken. Cirruswolken lassen viel Sonnenlicht durch und haben daher eine geringere Auswirkung auf die Sonneneinstrahlung. Sie können jedoch einen Schleier vor der Sonne bilden und das Licht diffus erscheinen lassen.

Nimbostratuswolken sind dichte, graue Wolken, die mit starkem Niederschlag verbunden sind. Sie erstrecken sich über große Gebiete und sind oft mit anhaltendem Regen oder Schneefall verbunden. Nimbostratuswolken blockieren die Sonneneinstrahlung weitgehend und führen zu trüben, düsteren Bedingungen.

4.3.2 Relative Luftfeuchtigkeit und Luftdruck

Im Allgemeinen lässt sich sagen, dass die Teilchen (Wassertröpfchen, Eiskristalle, Staub, Pollen, Meeressalze und Schadstoffe), aus denen die Wolken be-

stehen, die Sonnenstrahlen reflektieren oder absorbieren. Auf die gleiche Weise können Wasser- und Luftpartikel in der unteren Troposphäre die Strahlung beeinflussen. Folglich ist die gemessene Luftfeuchtigkeit und der Luftdruck ein weiteres Kennzeichen dafür, wie viele Teilchen sich in der Troposphäre befinden. Umso höher die beiden Werte sind, desto größer ist die Wahrscheinlichkeit, dass die Sonneneinstrahlung reflektiert oder absorbiert wird, wodurch wir die Erwartungen an den Stromertrag senken müssen.

4.3.3 Temperatur

Ein weit verbreiteter Glaube ist, dass die Stromproduktion durch Solaranlagen im Hochsommer am höchsten ist. In aller Regel ist dies jedoch nicht der Fall, da die Temperatur eine entscheidende Rolle spielt. Die hohen Temperaturen im Sommer beeinträchtigen den Wirkungsgrad der Solarzellen, welche die Schlüsselkomponente eines Photovoltaiksystems bilden. Die meisten Hersteller geben einen Temperaturkoeffizienten an, der spezifiziert inwiefern sich der Wirkungsgrad mit steigenden Temperaturen verändert. Gewöhnlich erhöht sich der Widerstand im Stromkreis der Solarzelle mit den höheren Temperaturen, wodurch die Gesamtleistung der Anlage zurückgeht. Folgende Temperaturkoeffizienten sind aus dem Datenblatt der Solarmodule *White* von dem renommierten, deutschen Solarmodulhersteller *Meyer Burger* entnommen:

Temperaturkoeffizient I_{SC}	α	$[\%/K]$	$+0,033$
Temperaturkoeffizient V_{OC}	β	$[\%/K]$	$-0,234$
Temperaturkoeffizient P_{MPP}	γ	$[\%/K]$	$-0,259$

Tabelle 1: Temperaturkoeffizienten Solarmodule Meyer Burger White

In Tabelle 1 ist ersichtlich, dass zwar der Kurzschlussstrom mit höherer Temperatur zunimmt, allerdings verringert sich die Leistung der Anlage. Dies ist am negativen Temperaturkoeffizienten P_{MPP} ersichtlich. Die Tatsache, dass auch die Wetter-abhängigen Faktoren sich unterschiedlich auf die Stromerzeugung einzelner Solaranlagen auswirken, bestärkt die Notwendigkeit für jede Solaranlage ein

eigenes Modell anzulegen.

Des Weiteren kommt hinzu, dass bei manchen Solaranlagen ein Kühlmanagementsystem verbaut ist, das die Effizienz der Solarmodule steigert. Ebenso ist die Art der Befestigung in Bezug auf die Temperaturentwicklung von Interesse. So herrscht ein stärkerer Luftzug, wenn unter den Solarmodulen ein Freiraum ist. Durch die spürbaren Auswirkungen der heißen Temperaturen, müssten bei einem allgemeinen Modell zusätzlich zu der sich variierender Temperatur auch die Leistung des aktiven beziehungsweise passiven Kühlsystems berücksichtigt werden, wodurch selbstverständlich die Komplexität weiter in die Höhe getrieben werden würde. Auch an dieser Stelle sei erwähnt, dass die Kühlleistung des Systems und die Art der Montage in aller Regel statisch ist, weswegen wir sie getrost nicht in den Eingabedaten unseres Datenmodells berücksichtigen müssen.

4.3.4 Wind

Wind kann als eine Art natürlicher Ventilator dienen. Die entstehende Luftbewegung um die Solarmodule trägt dazu dabei, dass stehende Hitze auf der Oberfläche der Panele abgeführt werden kann. Dies dient der Aufrechterhaltung niedriger Betriebstemperaturen und fördert somit die Effizienz der Anlage. Selbstverständlich hängt auch hier der Einfluss der Winde von den örtlichen Gegebenheiten ab. Luftschneisen können diesen Effekt verstärken, während umliegende Hindernisse den Luftstrom genauso blockieren können.

4.3.5 Sonneneinstrahlung

Die Sonne emittiert Energie in sämtliche Richtungen über den gesamten elektromagnetischen Strahlungsbereich. Ungefähr 20 Kilometer über der Erdoberfläche treffen im Mittel $S = 1367W/M$. Messungen im Weltraum haben gezeigt, dass dieser Wert nur geringfügig um vage 0,1% über die letzten Jahrzehnte geschwankt ist, weswegen er als die Solarkonstante betitelt wird.

In Abbildung 4 ist eine Karte von Deutschland zu sehen. Dargestellt wird die Monatssumme der Globalstrahlung. Deutlich zu erkennen ist, dass im Süden die Globalstrahlung im Mittel höher ist als im Norden. Folglich kommt die Sonnen-

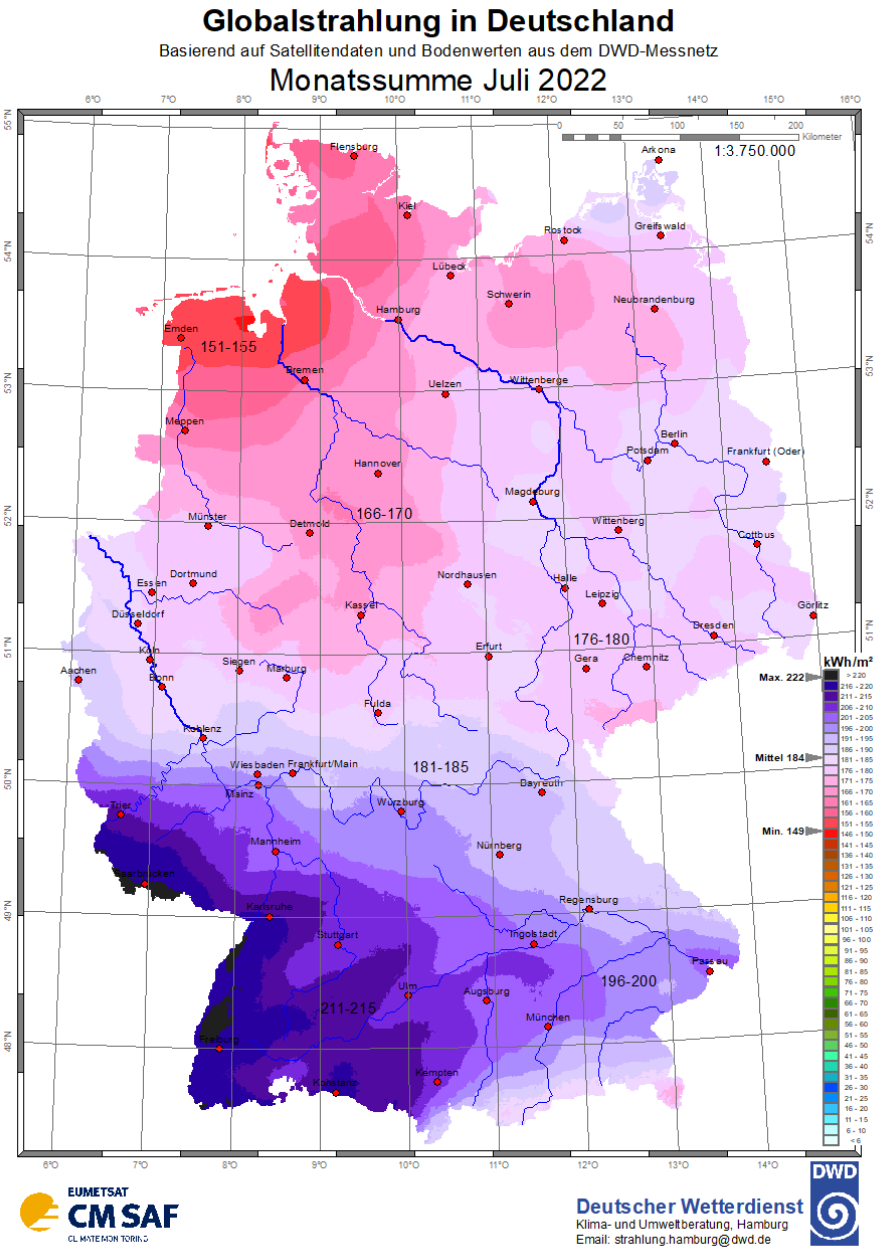

Abbildung 4: Sonneneinstrahlung Deutschland, Juli 2022

strahlung keineswegs gleichermaßen auf der Erdoberfläche an. Viele Faktoren tragen dazu bei, dass die Sonnenstrahlen unterschiedlich stark an verschiedenen Orten der Erdoberfläche auftreffen. Die Kugelform unseres Planeten ist der Grund, dass je höher der Breitengrad, desto flacher treffen die Sonnenstrahlen die Oberfläche der Erde, wodurch sich diese Strahlen auf eine größere Fläche verteilen. Auf die gleiche Weise wirkt sich die Topografie der Erde aus, wodurch es zu stärkeren, lokalen Unterschieden der Bestrahlung kommen kann.

Zum flacheren Einstrahlungswinkel in den höheren Breiten kommt hinzu, dass die Sonnenstrahlen eine größere Strecke zurücklegen müssen. Insbesondere auf dem letzten Abschnitt ihrer Reise, nämlich der Atmosphäre, kommen immer mehr Hindernisse wie Luftpartikel oder Wassertröpfchen hinzu, die die Strahlen reflektieren oder absorbieren können. Dass die Auswirkungen der Teilchen nicht zu unterschätzen sei, ist allein an der Temperaturabnahme bei zunehmender Höhenlage erkenntlich. Da die Energie der Sonnenstrahlen mit steigender Höhe von immer weniger Teilchen absorbiert werden kann, nimmt die Temperatur im Mittel $6,5K/1Kilometer$ ab, bis sie Zahlenwerte von unter $-50°C$ an der Tropopause erreicht.

Die auf der Erde gemessene Globalstrahlung setzt sich aus der Diffusstrahlung und Direktstrahlung zusammen. Unter Diffusstrahlung versteht man jene Strahlung, die auf ihrem Weg von der Sonne zur Erde an anderen Atomen reflektiert und dadurch gestreut wurde. Direktstrahlung hingegen wurde nicht abgelenkt und ist dadurch intensiver und gebündelt. Während die Globalstrahlung schon seit geraumer Zeit protokolliert wird, vermisst der *Deutsche Wetterdienst* erst seit wenigen Jahren auf die horizontale Ebene bezogene Diffusstrahlung. Aus der Differenz der beiden wird die Direktstrahlung berechnet.

5 Zusammenhänge der Merkmale

Die Qualität der Daten definiert maßgeblich die Leistungsfähigkeit des Prognose-Modells. Schließlich sind die Algorithmen im Bereich des maschinellen Lernens derart verallgemeinert, sodass sie die Bedeutung der Daten überhaupt nicht kennen müssen. Dies ist von enormen Vorteil, da ansonsten maschinelles Lernen kaum derart viele Anwendungsmöglichkeiten, wie Klassifizierung von Spam und Phishing-Mails, Erkennung von Objekten anhand von Bildern, Schätzung eines Immobilienwertes oder eben die Prognose der Stromerzeugung einer Photovoltaikanlage, bieten könnte.

5.1 Transformation der Daten zu Wissen

Im heutigen Zeitalter werden Unmengen von Daten gesammelt, die sowohl strukturiert als auch unstrukturiert sein können. Die Stärke der selbstlernenden Algorithmen ist es Erkenntnisse aus einer gewaltigen Datenmenge zu ziehen und anhand der Muster, die sich in den Daten befinden, bestimmte Vorhersagen zu treffen. In diesem Kapitel sollen die Zusammenhänge zwischen den einzelnen Merkmalen[1] und der Zielvariablen[2] im Detail betrachtet werden. Insbesondere sind wir daran interessiert, ob sich die zuvor beschriebenen Auswirkungen der Wetterfaktoren in den Datensätzen widerspiegeln.

Wie die Merkmale und die Zielvariable verknüpft sind, entscheidet über die Auswahl des passenden Lernalgorithmus. Im vorangehendem Kapitel haben wir bereits die Auswirkungen einzelner Faktoren erläutert. Auf Grund der Gegebenheit, dass die exakten Auswirkungen nicht umgehend erforscht sind, um sie beziffern zu können und sich zudem das gleiche Merkmal bei verschiedenen Solaranlagen unterschiedlich stark auswirken kann, ist ein Lernalgorithmus die richtige Wahl für diesen Anwendungsfall. Nicht zu vergessen, müssten bei einem konventionellem Algorithmus[3] Informationen über die Spezifikationen der Solaranlage

[1] Als Merkmale werden im Folgenden jegliche Daten bezeichnet, die die Grundlage für die Prognose bilden. Sie werden in den Lernalgorithmus eingespeist, um daraus eine Vorhersage zu erhalten.

[2] Als Zielvariable bezeichnen wir jenen Wert, für dessen wir eine Vorhersage treffen möchten.

[3] Für jeden Anwendungsfall werden bei einem konventionellem Algorithmus eine Reihe von Regeln festgelegt, die ein jeweiliger Programmierer umsetzt.

einfließen, wodurch die Programmierung des Algorithmus zunehmend komplexer werden würde.

Zuletzt gilt es zu überprüfen, wie stark die verschiedenen Merkmale tatsächlich mit der Zielvariablen korrelieren. Falls manche Merkmale zu irrelevant sind und dementsprechend nur geringe Auswirkungen auf die Stromerzeugung haben, kann die Eliminierung dieser Merkmale zu einer Verbesserung der Leistungsfähigkeit des Lernalgorithmus führen. Dies gilt insbesondere, wenn nur ein kleiner Datensatz für die Erstellung eines Modells zur Verfügung steht.

Um möglichst schnell für verschiedene Solaranlagen Prognosen erstellen zu können, könnte man zu Beginn absichtlich weniger relevante Merkmale ignorieren und sie erst zu einem späteren Zeitpunkt in das Modell einfließen lassen. Dies kann zu einer zu einer temporären Verbesserung der Ergebnisse führen.

5.2 Heatmap

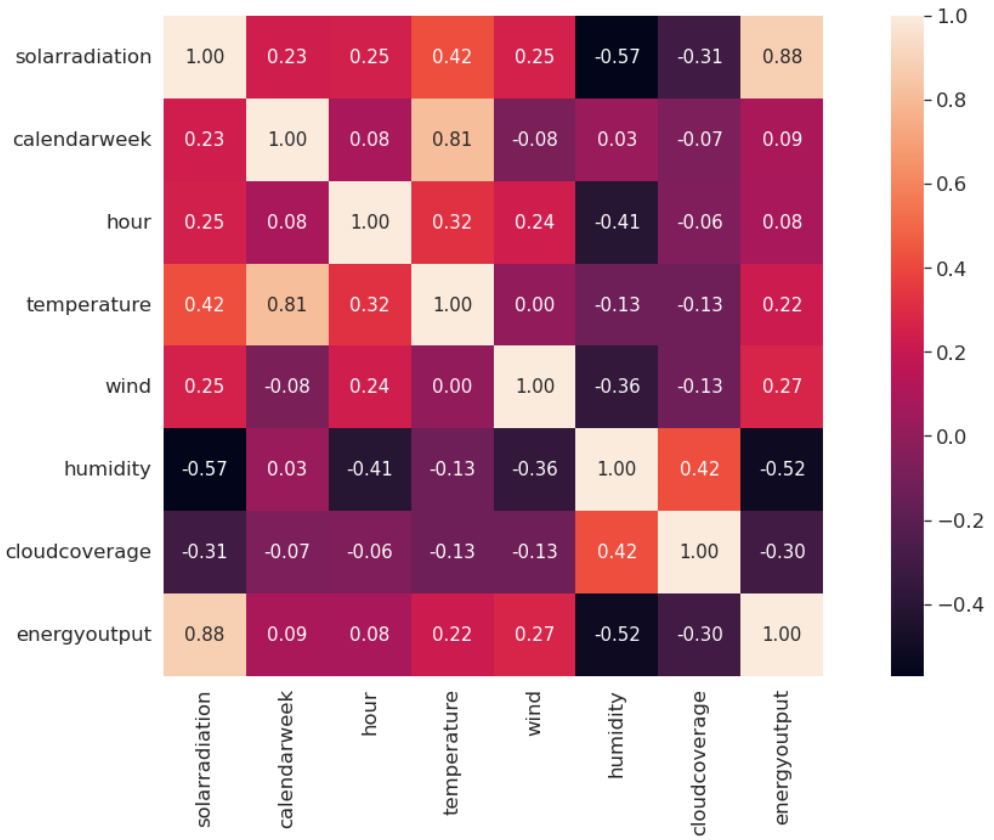

Abbildung 5: Heatmap zur Visualisierung der Korrelationen zwischen den Merkmalen

Eine Heatmap ist ein schnell und einfach zu erstellendes Diagramm, dass die Korrelationen zwischen den Merkmalen sowohl zueinander als auch zu der Zielvariablen aufzeigt. Die Zahlen auf den einzelnen Feldern, welche in Abbildung 5 zu sehen sind, sollen die Korrelation zwischen den jeweiligen zwei Variablen widerspiegeln. Die Zahlenspanne beinhaltet Werte zwischen -1 bis 1. Die entsprechenden Werte werden dabei wie folgt berechnet:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

Aus der Gleichung wird ersichtlich, dass das Ergebnis maximal $(= 1)$ wird, wenn die zwei Variablen x und y gleich stark steigen. Minimal hingegen wird das Ergebnis, wenn eine der Variablen steigt während die andere fällt. Der sogenannte *Pearson Korrelationskoeffizient* ***r*** soll somit abbilden, wie stark zwei Variablen miteinander korrelieren.

5.2.1 Korrelation und Kausalität

Anhand der Temperatur wird allerdings ersichtlich, dass die richtige Interpretation des Diagramms wichtig ist und teilweise die Aussagekraft nur sehr beschränkt ist. Wie wir wissen, leidet die Effizienz einer Solaranlage unter hohen Temperaturen. Dennoch vermittelt Abbildung 5 den Eindruck, dass die Stromproduktion mit steigender Temperatur zunimmt. Schließlich haben die Temperatur und die Solarengerie den Korrelationskoeffizienten 0.22. Jedoch darf man die Multikollinearität zwischen den anderen Faktoren nicht außer Acht lassen. So ergibt sich für die Temperatur und der Sonneneinstrahlung der Wert 0.42. Wenig überraschend steigt die Temperatur mit erhöhter Sonneneinstrahlung. Wie so häufig bedeutet Korrelation nicht gleich Kausalität.

Die tatsächliche Realität ist, dass Temperatur und der Stromertrag einer Photovoltaikanlage einen negativen Korrelationskoeffizienten haben. Dies soll später genauer untersucht werden.

5.2.2 Nicht-lineare Beziehungen

Ein Korrelationskoeffizient, der sich im Nullbereich befindet, muss nicht unbedingt bedeuten, dass die zwei Variablen nicht miteinander korrelieren. Der Koeffizient spiegelt nur wieder, ob es zwischen den Variablen eine lineare Beziehung vorhanden ist. Selbstverständlich ist die Tageszeit ein äußerst guter Indiz, um zu schätzen, wie viel Strom produziert wird. Nach dem die Sonne mittags ihren Zenit erreicht hat, fällt in der Regel die Stromproduktion wieder ab, währenddessen die Stunden weiter fortschreiten. Im Jahresdurchschnitt und etwas vereinfacht könnte die Beziehung zwischen Stromertrag und Uhrzeit durch eine umgedrehte Parabel beschrieben werden, wobei selbstverständlich diese unten abgeschnitten ist, da die Stromproduktion nicht negativ wird.

5.3 Streudiagramm

Da im Gegensatz zu Computern Menschen visuelle Grafiken an Stelle einer mit Zahlen gefüllten Matrix besser verstehen können, sind Streudiagramme eine gute Herangehensweise, um die Zusammenhänge innerhalb des Datensatzes besser zu verstehen.

Abbildung 6 zeigt die Daten einer Solaranlage in der Gemeinde Linthicum im Bundestaat Maryland[4] und die dazugehörigen Wetterdaten. In Abbildung 6 lässt sich die lineare Beziehung zwischen der Sonneneinstrahlung in der Nähe der Stadt Baltimore und dieser spezifischen Solaranlage beobachten. Allerdings sei an dieser Stelle erwähnt, dass nicht bei allen Solaranlagen das Streudiagramm ein solch deutliches Muster erblicken lässt. Dies kann man auf verschiedene Gründe zurückführen. Wechselrichter oder Solarmodule mit einer niedrigeren beziehungsweise höheren Effizienz könnten würden dazu führen, dass aus der gleichen Sonneneinstrahlung unterschiedlich viel in elektrische Energie umgewandelt werden kann. Ebenso gilt es zu bedenken, dass die Wetterdaten von einem separaten Dienst verwendet werden. Es ist nicht garantiert, das die Wetterstationen, die möglicherweise gar nicht exakt an diesem Standort stehen werden, für alle Orte auf der Welt die gleiche Qualität gewährleisten.

[4] Maryland ist ein Bundesstaat an der Ostküste der Vereingiten Staaten von Amerika.

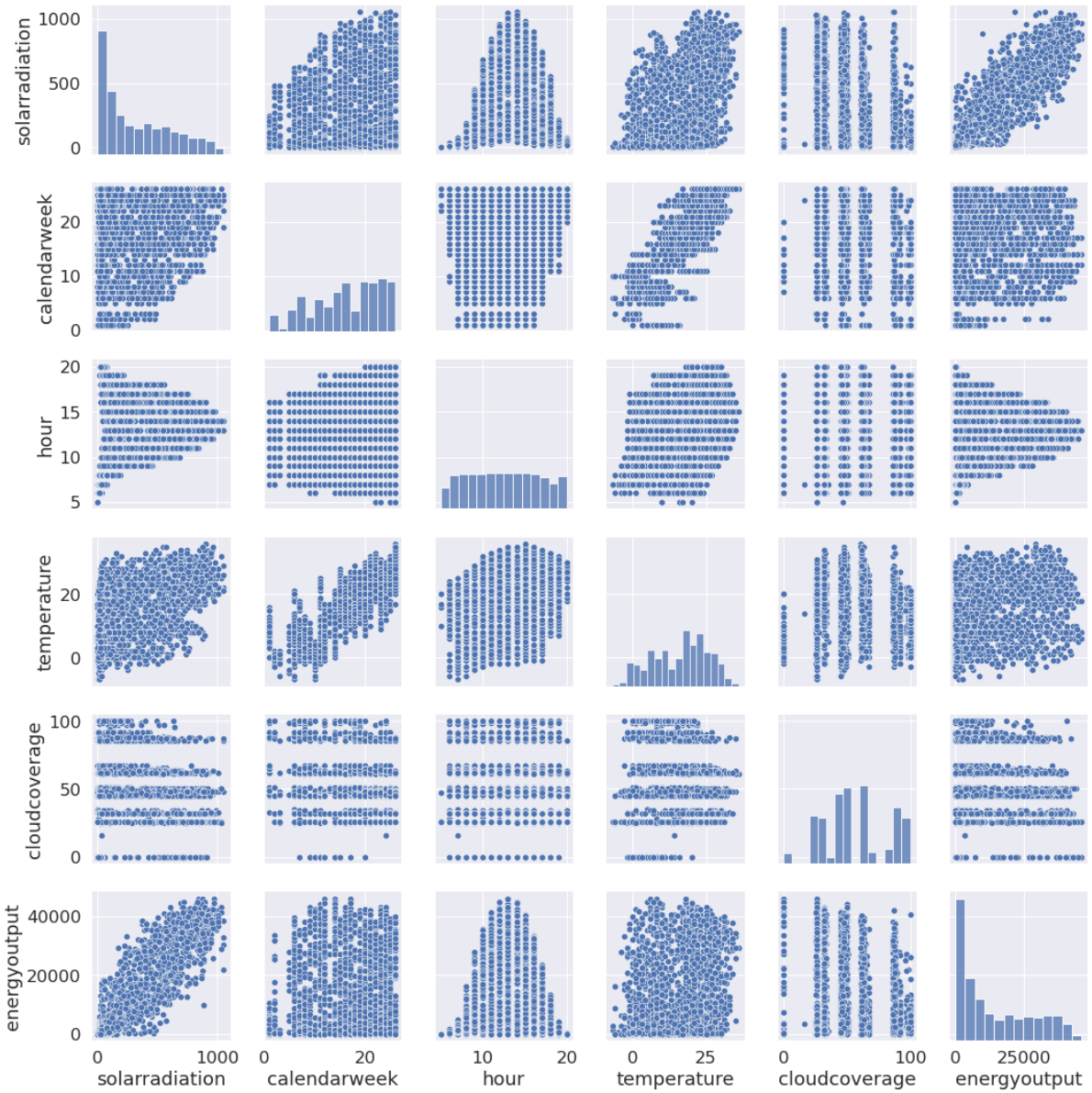

Abbildung 6: Streudiagramme zur Visualisierung der Korrelationen zwischen den Merkmalen

5.3.1 Qualität der Daten

In der Diagonalen der Abbildung 6 sind Balkendiagramme abgebildet, die die Verteilung der einzelnen Merkmale aufweisen. Besonders interessant ist hierbei das Balkendiagramm zum Bewölkungsgrad in Prozent. Die auffällige Verteilung spiegelt die in Unterunterabschnitt 4.3.1 bereits erwähnte Problematik die Bewölkung am Himmel anhand einer Prozentzahl darzustellen wider. Bei einem Index, der die Bewölkung einer Ortschaft wiedergeben soll, würden wir erwarten, dass die Verteilung keine größeren Ausreißer beinhaltet. Selbstverständlich ist die Bewölkung von Ort zu Ort unabhängig. So haben Städte in der Nähe von Flüssen in der Regel einen höheren Niederschlag, da sich über den Stadtflächen durch den Asphalt und die Betonflächen die Luft stärker erwärmt als über Grünflächen. Folglich verdunstet mehr Wasser, worauf die Wolken zu schwer werden können und letztendlich abregnen. Das bedeutet, dass es kein einheitliches Schaubild gibt, dass wir als Referenz verwenden könnten. Dass einige Werte jedoch eklatant auftreten, währenddessen deren benachbarte Werte überhaupt nicht in der Statistik vertreten sind, ist ein recht starkes Indiz dafür, dass es sich bei den Angaben zu der Bewölkung um Schätzwerte[5] handelt.

Da bei den anderen Werten keine Auffälligkeiten zu verzeichnen sind und deren Messung vergleichsweise einfach und standardisiert ist, können wir davon ausgehen, dass sie vertrauenswürdig sind. Für den Bewölkungsgrad könnte es eine mögliche Konsequenz sein ihn aus der Merkmalsliste des Lernalgorithmus zu streichen. Schließlich wird die Bewölkung, zwar nicht vollständig, aber zumindest teilweise durch die zuverlässigeren Messwerte über die Feuchtigkeit und des Einstrahlungswertes repräsentiert. In Abbildung 7 ist erkenntlich, dass ein höhere Feuchtigkeit mit mehr Bewölkung einhergeht, gleichzeitig kommen weniger Sonnenstrahlen auf der Erdoberfläche an.

[5] Die Dokumentation des Wetterdienstleisters beschreibt den Wert wie folgt: "Die Wolkenbedeckung ist der Anteil des Himmels, der von Wolken bedeckt ist, ausgedrückt in Prozent. Die Wolkenbedeckung gilt für alle Höhenlagen. Die Tageswerte umfassen den Mittelwert der stündlichen Werte der Wolkenbedeckung."
Wie dieser Wert zu Stande kommt oder ob verschiedene Wolkentypen unterschiedlich bewertet werden, bleibt dabei unklar.

5.3.2 Überflutung von Daten

Anhand der Kalenderwochen lässt sich ablesen, dass für die Erstellung der Streudiagramme verwendete Datensatz die Zeitspanne eines halben Jahres umfasst. Durch die große Menge an Datenpunkte werden Streudiagramme, die zwei schwach miteinander korrelierende Variablen beinhalten, häufig unübersichtlich. Die vielen Datenpunkte im Diagramm färben das gesamte Schaubild ein, sodass sich kein Muster mehr erkennen lässt. Wie bei der Heatmap kommt hinzu, dass die Sonneneinstrahlung andere Wetterfaktoren als auch selbst von diesen stark beeinflusst wird. Um Abhilfe zu schaffen, können wir die Daten filtern. Der Zusammenhang zwischen Sonneneinstrahlung und Stromproduktion ist gut erkenntlich. Ohne Zweifel besitzt dieses Merkmal die größte Hebelwirkung auf die Stromproduktion einer Solaranlage. Um besser die Auswirkungen der anderen Merkmale zu studieren, betrachten wir nun ausschließlich Datenreihen in denen der Einstrahlungswert zwischen 550 und 600 W/M^2 liegt. Zuvor umfasste der Datensatz Werte von 0 bis zu 1000 W/M^2.

Betrachten wir nun nochmals die Korrelation zwischen der Temperatur und der Stromproduktion, ist eine negative lineare Beziehung deutlich zu erkennen. Der negative Temperaturkoeffizient P_MPP zeigt sich in der Heatmap ebenfalls ersichtlich, dort erhalten wir diesmal den Wert -0.68. Nachdem wir den Datensatz nach Datenreihen mit ähnlicher Sonneneinstrahlung gefiltert haben, bilden Temperatur und Stromertrag den absolut höchsten Korrelationskoeffizienten innerhalb der Heatmap. Dementsprechend ist die Temperatur nach der Sonneneinstrahlung der zweitwichtigste Wetterfaktor. Wie bereits zuvor erwähnt können Windböen dabei helfen, die Wärme auf den Solarmodule abzutransportieren und somit die Effizienz der Module erhöhen. Inwiefern dies tatsächlich der Fall ist, soll auf die gleiche Art untersucht werden. Folglich schauen wir uns Datenreihen mit ähnlicher Sonneneinstrahlung ein geringer Temperaturschwankung an.

PLOT SIMILAR TEMPERATURE AND RADIATION

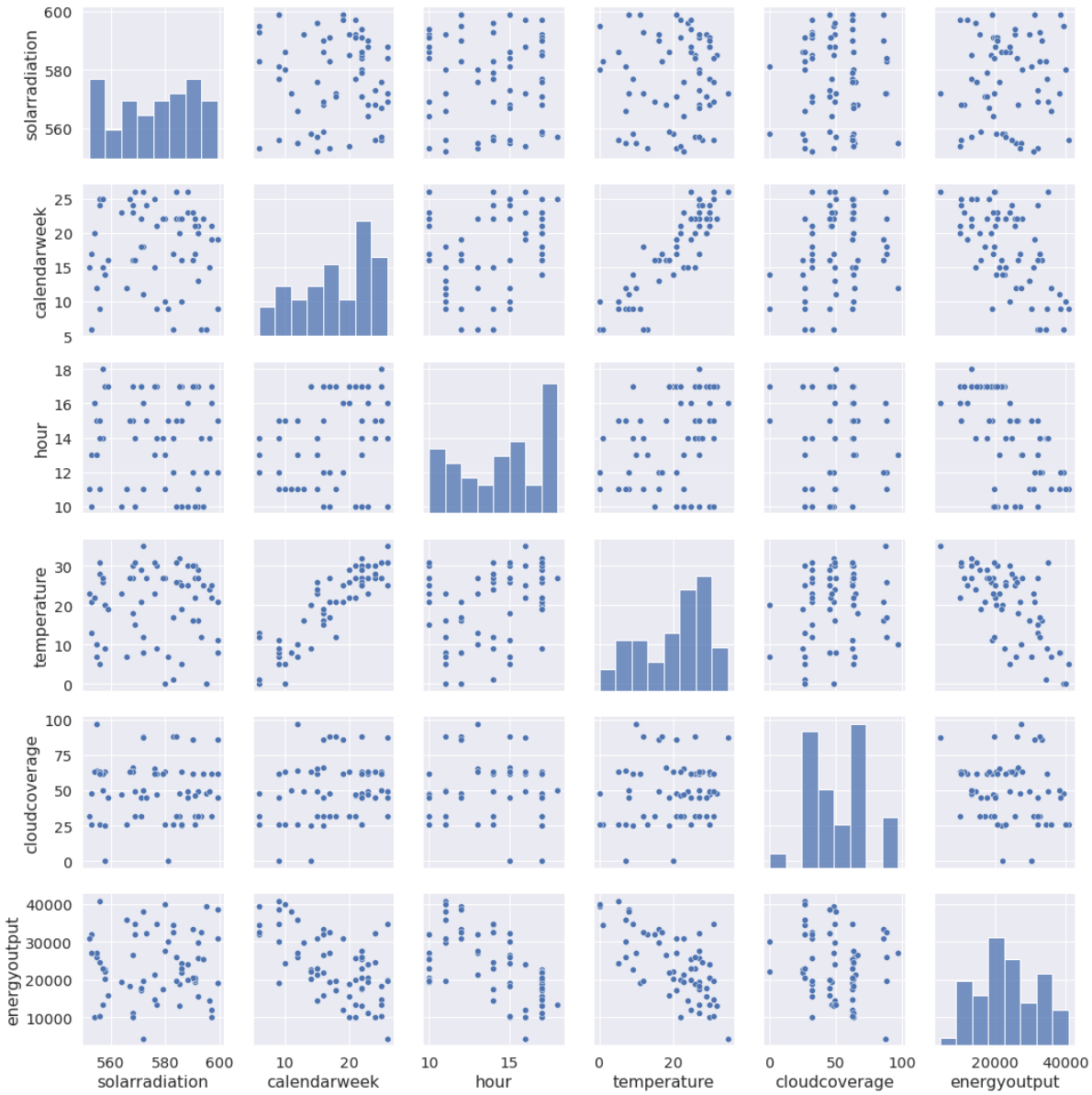

Abbildung 7: Streudiagramme zur Visualisierung der Korrelationen zwischen den Merkmalen bei einer Sonneneinstrahlung zwischen 550 und 600 W/M^2

6 Datenvorverarbeitung

Für die Forschung dieser Arbeit wurden ein paar Solaranlagen ausgewählt, um zu untersuchen, wie gut anhand von Wetterdaten sich die Solarstromproduktion prognostizieren lässt. Selbst bei den wenigen Solaranlagen, die für diese Forschung verwendet wurden, konnte festgestellt werden, dass die Zusammenhänge zwischen den einzelnen Wettermerkmalen und der erzeugten Strommenge sehr unterschiedlich sein kann.

Auf Grund der Gegebenheit, dass die Zusammenhänge zwischen den Merkmalen und der Zielvariablen je nach Solaranlage unterschiedlich sind, entscheiden wir uns für einen Lernalgorithmus, der möglichst wenig Datenvorverarbeitung benötigt. Dies soll zu dem Ziel dieser Arbeit beitragen, dass für beliebige Solaranlagen ein qualitativ hochwertiges Prognosemodell erstellt werden kann, dass in der Lage ist den erzeugten Solarstrom vorherzusagen.

Ziel der Datenvorverarbeitung ist es, dass die Daten für den jeweiligen Lernalgorithmus optimiert werden und somit die Leistungsfähigkeit des Modells gesteigert werden kann. Zudem soll somit vermieden werden, dass Unstimmigkeiten innerhalb des Datensatzes die Ergebnisse des Modells verzerren oder gänzlich verfälschen.

6.1 Ausreißer

Selbst wenn für ein Entscheidungsbaum die Datenvorverarbeitung nicht zwingend von Nöten ist, ist es dennoch zu empfehlen Unstimmigkeiten in den Daten zu beseitigen.

Unstimmigkeiten innerhalb des Datensatzes werden häufig als Ausreißer betitelt, da sich nicht zum Muster zwischen ihresgleichen und der Zielvariablen passen. Da solche Ausreißer unausweichlich und in sämtlichen Datensätzen auftauchen können, dürfen diese nicht ungeachtet bleiben. Ein möglicher Grund für etwaige Ausreißer könnte ein defektes Bauteil des Solarsystems sein. Ein nicht funktionierender würde dazu führen, dass trotz idealer Wetterbedingungen die erzeugte Wechselspannung ausbleibt. Umso länger das Solarsystem nicht re-

pariert wird, desto wahrscheinlicher ist es, dass der Fehler übersehen und die Kalenderwoche für die übermäßige Abweichung verantwortlich gemacht wird.

7 Lernalgorithmus

Die Konsequenz, dass die Zusammenhänge zwischen den Merkmalen und der Zielvariablen sehr verschieden sein können, ist es ein Modell auszuwählen, dass für seine Kontingenz gegenüber der Datenvorverarbeitung bekannt ist.

Ein Entscheidungsbaum *zu engl.: decision tree* hat eine Flussdiagramm ähnliche Baumstruktur, in der je nach ausgewählter Hyperparameter alle möglichen Ergebnisse, Eingabekosten und Nutzen dargestellt werden kann. Da das Modell sowohl für kategoriale als auch kontinuierliche Zielvariablen verwendet werden kann, eignet es sich für die Prognose verschiedener Solaranlagen.

7.1 Diskretisierung der Zielvariablen

Die Bedingungen werden innerhalb der Entscheidungsknoten aufgestellt. Der Entscheidungsbaum wird so lange durchgegangen, bis man an einem Endknoten angelangt ist. Die Endknoten eines Binärbaums werden als Blätter bezeichnet, sie bilden das Ergebnis für die jeweiligen Eingabedaten ab. Folglich müssen die Bedingungen des Binärbaum von oben nach unten durchgegangen werden, um zu einem Ergebnis zu gelangen. Ein Binärbaum hat die Eigenschaften, dass er maximal 2^n Blätter besitzen kann, wobei n für die Höhe des Baums steht. Indes es nur so viele Ergebnisse wie Blätter geben kann, muss die Wertemenge der Zielvariablen diskretisiert werden.

Die diskretisierten Werte werden so ausgewählt, sodass der mittlere quadratische Fehler am geringsten ist. Für die Berechnung des mittleren quadratischen Fehler eines Blattes werden ausschließlich die Trainingswerte verwendet, die eben diesem Blatt zuvor zugeordnet wurden. Auf Grund des geringen Rechenaufwands genießt diese Straffunktion in vielen Bereichen der Informatik größerer Beliebtheit. Schließlich genügt durch die Art wie ein Computer Zahlen speichert eine anspruchslose Operation, explizit gesagt eine Bitverschiebung nach links, um das Quadrat einer Zahl auszurechnen. Somit ist diese Methode äußerst effizient, um die Mitte zwischen 2 oder mehreren Punkten zu finden. Den mittleren quadratischen Fehler als Straffunktion zu verwenden birgt allerdings den Nach-

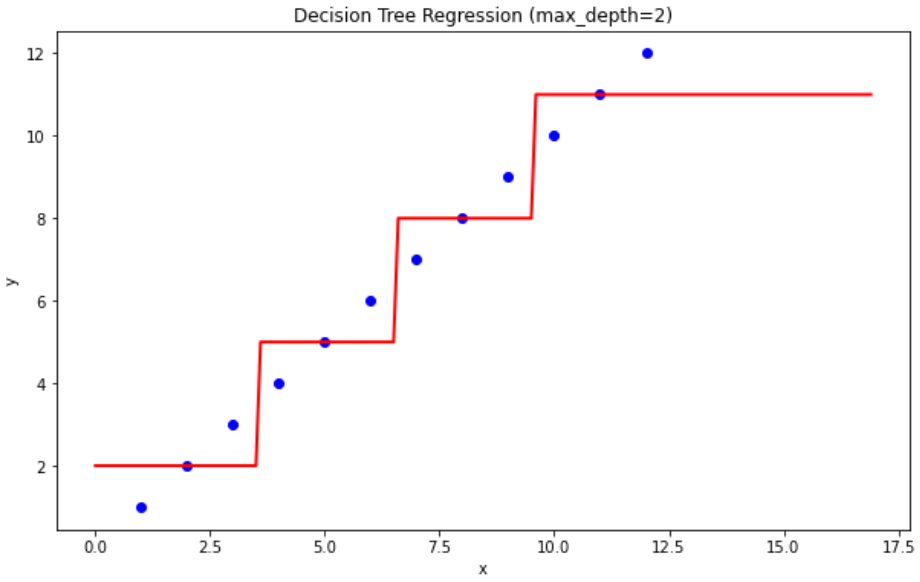

Abbildung 8: Einfacher Entscheidungsbaum mit einem Merkmal und der Tiefe 2

teil, dass Ausreißer die Vorhersage sehr stark verfälschen können, weshalb diese idealerweise aus dem für das Training zuständigen Datensatz entfernt werden.

7.2 Lineare Beziehungen

Wie so häufig kommen mit den Vorteilen eines Modells auch Nachteile einher. Zu einer der vielen Vorteile von Entscheidungsbäumen gehört die leicht verständliche Darstellung der Entscheidungsprozesse. Das Modell lässt sich als Flussdiagramm darstellen, wodurch sich die Vorhersagen des Modells gut nachvollziehen lassen. Dies ist bei der Fehlersuche und der Interpretierbarkeit von großem Nutzen.

Die Tatsache, dass nichtlineare Beziehungen gut erkannt werden können ist sowohl ein Nachteil als auch ein Vorteil. Viele der Wettermerkmale haben eine nicht-lineare Beziehung zu der erzeugten Strommenge einer Solaranlage. Vor allem sind die Beziehungen bei verschiedenen Solaranlagen unterschiedlich, wodurch die Vorverarbeitung der Daten sich als schwierig darstellt. Insofern ist es ein großer Vorteil, dass die Wetterdaten nicht modifiziert werden müssen. In Abbildung 8 ist jedoch ersichtlich, wie schwer sich ein Entscheidungsbaum mit linearen Beziehungen, wie beim Fall $x = y$, tun kann. Der Entscheidungsbaum kann den Zusammenhang $x = y$ nicht nur sehr limitiert wiedergeben, des Weiteren wird

bei gleicher Tiefe des Entscheidungsbaum der Fehler unabdingbar und maßlos größer, sobald Vorhersagen zu Daten getroffen werden soll, dessen Erwartungswert der Zielvariablen außerhalb des Wertebereichs des Trainingsdatensatz liegt. Das liegt daran, dass nach dem trainieren des Modells die Menge der möglichen Ausgangswerte bestimmt und statisch ist. Betrachte man nochmals Abbildung 8, so wird ersichtlich, dass für $x > 10$ stets $y = 11$ prognostiziert wird. Da $x = y$ ist, würde ein optimales Modell für $x = 111$ auch $y = 111$ vorhersagen. Wie bereits erwähnt würde unser Modell $y = 11$ prognostizieren, wodurch wir den quadratischen Fehler $100^2 = 10000$ erhalten.

7.3 Aufbau eines Entscheidungsbaum

Künstliche Intelligenz nimmt eine immer bedeutendere Rolle in unserer heutigen Gesellschaft ein. Bereiche unseres täglichen Lebens werden bereits durch sie erheblich erleichtert. Dennoch schreckt ein Teil der Gesellschaft vor den noch unerforschten Begleiterscheinungen der neuen Technologie zurück. Nicht selten sind die Ergebnisse dieser Algorithmen schwer zu verstehen, wodurch Ungewissheiten hervorgerufen werden. Ein Entscheidungsbaum ist ein Modell aus dem Bereich des maschinellem Lernens, welches wiederum ein Teilgebiet von künstlicher Intelligenz ist. Obwohl Entscheidungsbäume derzeit als äußerst mächtige Lernalgorithmen angesehen werden, sind ihre Prognosen äußerst leicht nachzuvollziehen. Im Grunde genommen ist ein Entscheidungsbaum eine Verkettung von *Wenn-Sonst-Anweisungen*, die schlussendlich zum Ergebnis führen.

Da der komplette Binärbaum mit der Tiefe 11 zu groß für eine Abbildung innerhalb dieser Arbeit ist, zeigt Abbildung 9 einen kleinen Ausschnitt des Entscheidungsbaum für die Solaranlage in Linthicum, Maryland, welcher der *CART-Algorithmus*[6] erstellt hat. In der ersten Zeile eines jeden Knotenpunktes steht die *Wenn-Sonst-Anweisung*, die den weiteren Verlauf bestimmt. Falls die Bedingung erfüllt wird, so wird die Bedingung des linken Kinderknoten als nächstes überprüft, oder die des rechten Kinderkonten, falls ihr nicht nachgekommen werden

[6]Die Funktion des *Classification and Regression (CART) Algorithmus* ist es, eine optimale zweiteilige Trennung zu finden, indem ein Merkmal und ein Grenzwert gesucht wird, wodurch der mittlere quadratische Fehler der zwei neuen Teilmengen minimal wird. Erstmals wurde dieser Algorithmus von Leo Breiman 1984 publiziert.

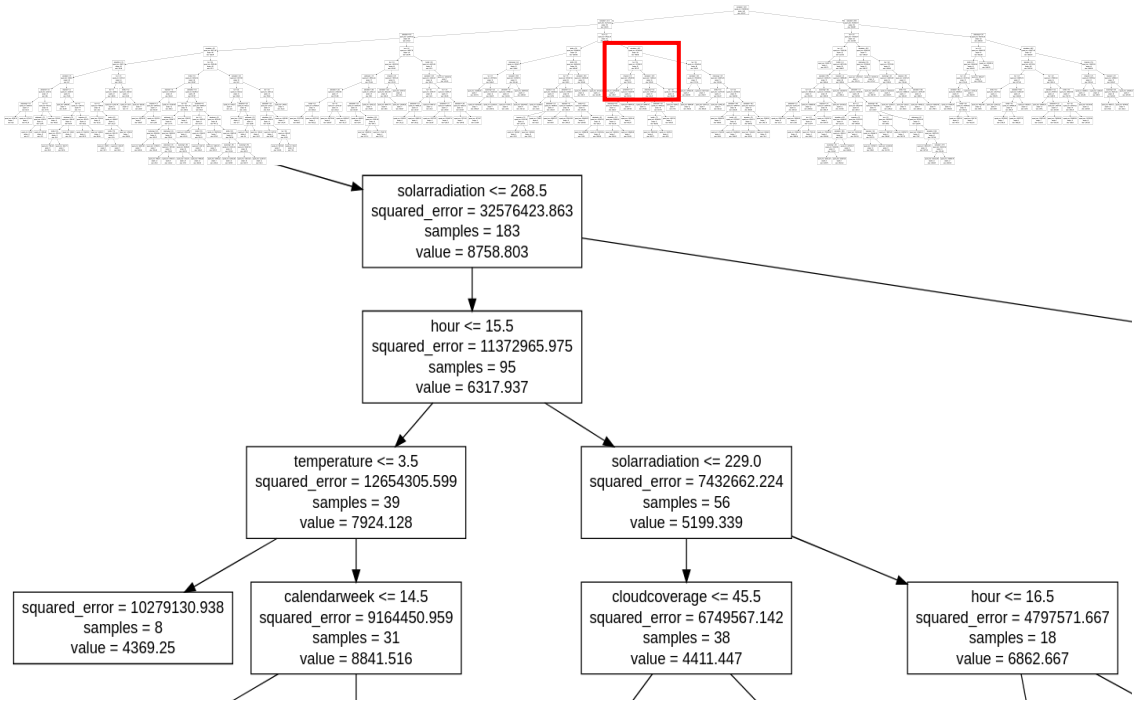

Abbildung 9: Ausschnitt des Entscheidungsbaum für die Prognose der Solaranlage in Linthicum, Maryland

konnte. Dieses Verfahren dauert so lange an, bis man beim Ende des Binärbaums, einem Blatt, angekommen ist.

Das Feld *squared_error* gibt den mittleren quadratischen Fehler der sich in diesem Feld befindenden Proben in Bezug auf den Wert des Feldes an. Wie viele Proben des Trainingsdatensatz diese Bedingungen und demzufolge die darüber liegenden Bedingungen ebenso erfüllen, lässt sich aus den Knotenpunkten ablesen. Der dort zu sehende Wert wäre die Prognose für die jeweiligen Datenreihen, wenn der Knotenpunkt keine Kinderknoten mehr hätte.

Die Zusammenhänge zwischen den Merkmalen und der Zielvariablen konnten weitestgehend gut nachgebildet werden. So wird die Prognose jedes mal herunter gesetzt, wenn die Sonneneinstrahlung weniger wird. Genauso fällt die Prognose ohne Ausnahme, wenn die Bewölkung oder die Luftfeuchtigkeit zunimmt.

Wie in Abbildung 9 zu erkennen ist, sind die Auswirkungen der Temperatur dem Modell nicht gänzlich vertraut. Fällt die Temperatur unter 3,5 °C, so wird die Prognose ebenfalls auf den Wert 4369 EINHEIT EINFÜGEN herabgesetzt. Bei vereinzelten Knotenpunkten des Entscheidungsbaum ist zu beobachten, dass ei-

ne niedrigere Temperatur zu einer geringeren Stromproduktion führt. Die Fehleinschätzung des Entscheidungsbaum ist vor allem dann zu beobachten, wenn der Grenzwert für die Temperatur besonders niedrig ist. Dies zeugt davon, dass die Multikollinearität zwischen Temperatur, Sonneneinstrahlung und Stromproduktion dem Entscheidungsbaum Schwierigkeiten bereitet. Da die Temperatur mit der Sonneneinstrahlung steigt, fehlinterpretiert der Algorithmus den Zusammenhang zwischen Temperatur und Stromproduktion. Da dies allerdings nur bei einem Teil der Knotenpunkte der Fall ist, soll die Temperatur nicht aus der Merkmalsmatrix ausgeschlossen werden. Unter dem Strich werden unter Berücksichtigung der Temperatur in Bezug auf die Prognose des Stromertrags dennoch bessere Ergebnisse erzielt. Dass die Temperatur ein wichtiger Parameter für die Prognose ist, liegt unter anderem daran, dass sie ebenfalls von verschiedenen Faktoren beeinträchtigt wird und somit, trotz der Korrelation zur Sonneneinstrahlung, nicht vernachlässigt werden darf. Zudem ist die Temperatur im Vergleich zur Sonneneinstrahlung viel träger.

Um ein besseres Verständnis für die Gewichtungen der einzelnen Merkmale zu bekommen, lassen wir uns ein Histogramm abbilden, dass die Anzahl der Verwendungen der einzelnen Merkmale bei den *Wenn-Sonst-Anweisungen* darstellt.

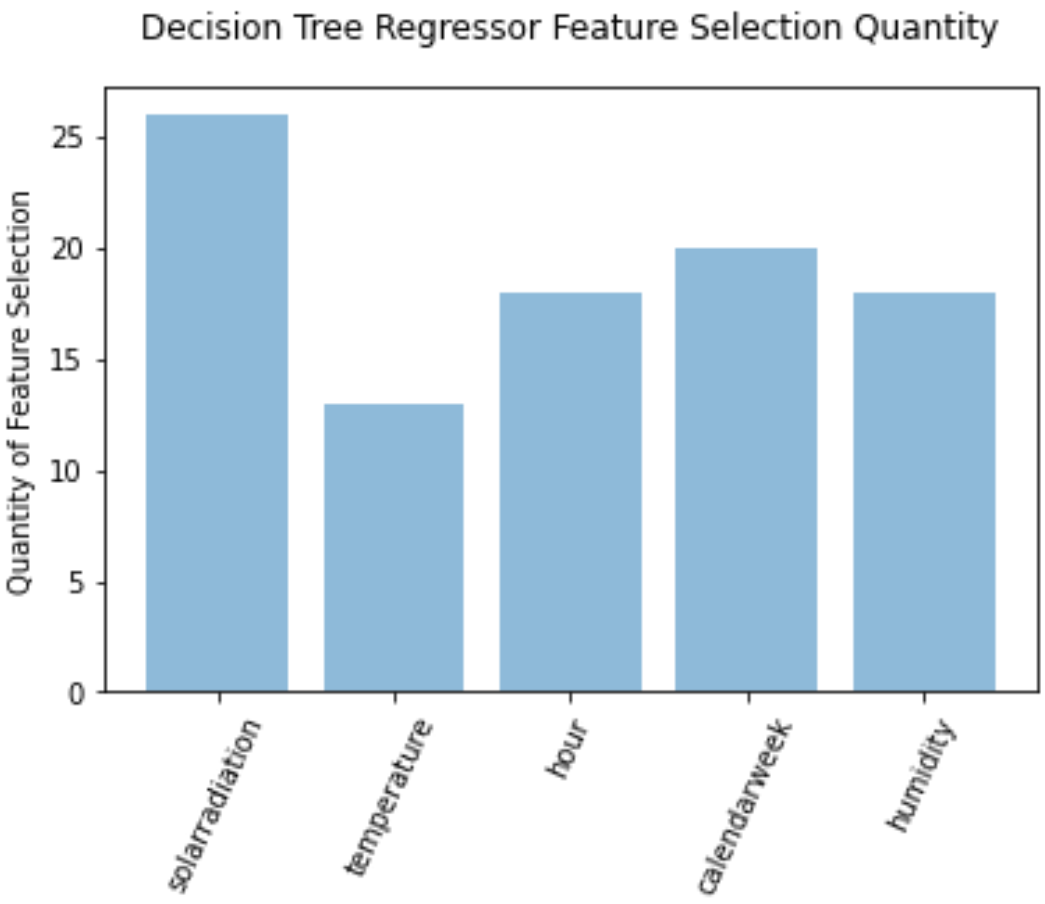

Abbildung 10: Verwendung der jeweiligen Merkmale im Entscheidungsbaum

Anhand der Häufigkeit, in der das Merkmal Sonneneinstrahlung auftritt, lässt sich die enorme Bedeutung für die Prognose der Stromerzeugung erahnen. Insbesondere die Tatsache, dass die Sonneneinstrahlung bei sämtlichen Modellen die Entscheidung für die obersten und somit grundlegendsten *Wenn-Sonst-Anweisungen* bildet, verleiht dem Gewicht der Sonneneinstrahlung weiteren Nachdruck.

7.4 Schlussfolgerungen für das Training des Modells

Aus dieser Eigenschaft von Entscheidungsbäumen müssen zwei Schlussfolgerungen für die Prognose über die Stromerzeugung von Photovoltaikanlagen gezogen werden.

7.4.1 Schwierigkeiten bei unbekannten Werten

Zum einem muss das Modell so lange trainiert werden, bis der Trainingsdatensatz alle möglichen Ausgangswerte der Zielvariablen beinhaltet. Da die Jahreszeit nicht nur eine erhebliche Rolle für die Vorhersage spielt, sondern den Maximalertrag eines Tages führend mitbestimmt, definiert sie folglich auch die Wertemenge der Zielvariablen. Somit sollte für jede Solaranlage mindestens ein Jahr lang Daten gesammelt werden. Gewiss kann das Modell bereits zuvor trainiert und für die Prognose verwendet werden, allerdings wird die Prognose weniger präzise, desto größer die Differenz zwischen den für das Training verwendeten Randwerten und den Test- beziehungsweise Produktionsdaten. Im Allgemeinen ist ein Entscheidungsbaum nicht gut darin Vorhersagen für Daten zu treffen, die zuvor nicht im Raum des Trainingsdatensatzes waren.

7.4.2 Limitierte Wertemenge für mögliche Prognosen

Die zweite Schlussfolgerung bezieht sich auf die die unterschiedliche Leistung von Solaranlagen. Selbstverständlich haben Photovoltaikanlagen mit qualitativ und quantitativen unterschiedlichen Solarpanelen verschieden hohe Maximalerträge. Folglich ist die Wertemenge der möglichen Ausgangswerte einer Photovoltaikanlage von einem Eigenheim mit weniger Solarmodulen kleiner als die einer häufig größeren, kommerziell genutzten Anlage. Um bei größeren Anlagen die gleiche Auflösung in Bezug auf die Vorhersage zu erhalten, muss die Tiefe des Entscheidungsbaums zunehmen. Da diese die Anzahl der Blätter und somit die Anzahl der möglichen Werte auf direkte Weise steuert. Eine Konsequenz dessen ist, dass die Komplexität zunimmt und das Modell in der Regel einen größeren Trainingsdatensatz benötigt.

8 Optimierung der Hyperparameter

Bei einem Entscheidungsbaum ist es auf Grund seiner Flexibilität gegenüber der verschiedenen Beziehungen zwischen den Merkmalen und der Zielvariablen von besonderer Wichtigkeit die Hyperparameter entsprechend anzupassen. Wenn ein Entscheidungsbaum in seiner Flexibilität nicht angepasst wird, neigt er dazu sich zu sehr an den Trainingsdatensatz anzupassen. Die Datenreihen des Trainingsdatensatz werden häufig perfekt abgebildet, sodass dieses Modell für den Trainingsdatensatz den Maximalwert bei der Evaluierung erzielt. In solchen Fällen schneiden zuvor unbekannte Daten, der sogenannte Testdatensatz, signifikant schlechter ab.

8.1 Tiefe des Entscheidungsbaum

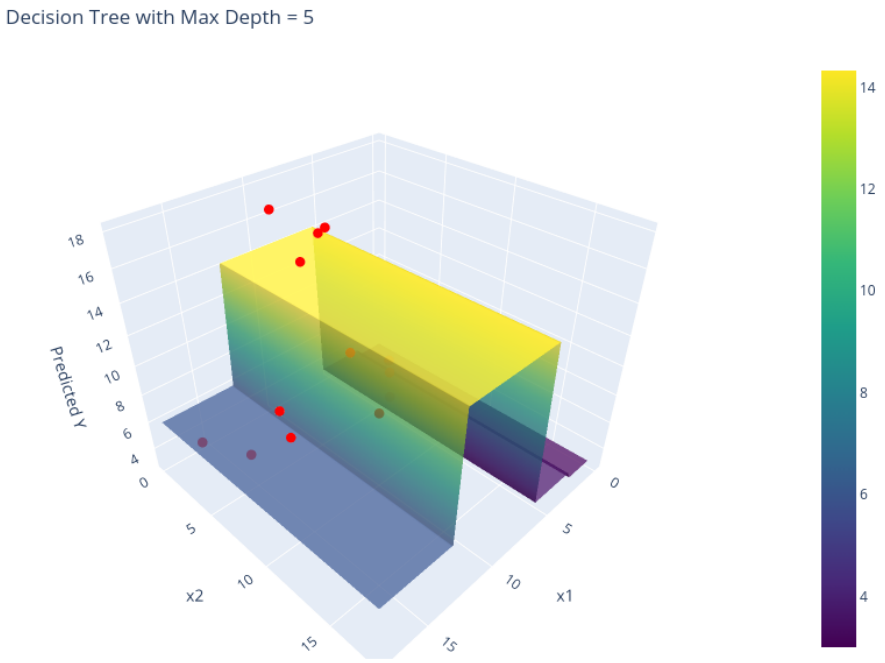

Abbildung 11: Einfacher Entscheidungsbaum mit 2 Merkmalen und der Tiefe 2

Der Mensch verfügt über eine begrenzte Sinneswahrnehmung, weswegen wir Schwierigkeiten dabei haben uns Dimensionen jenseits von drei bildlich vorzustellen. Sie sind für unser Gehirn ein unvertrautes Konzept, da sie über die alltäglichen Erfahrungen hinausgehen und es uns schlicht an Anschauungsobjekten

mangelt. Stattdessen fordert es abstraktes Denkvermögen und mathematische Konzepte. Um ein besseres Verständnis von Entscheidungsbäumen und die Auswirkungen der Hyperparameter zu bekommen, zeichnen wir zuerst das von einem Entscheidungsbaum erstellte Modell mit ausschließlich drei Dimensionen. Das in Abbildung 11 zu sehende Modell wurde mit der beschränkenden Eigenschaft, dass die Höhe des Entscheidungsbaum maximal zwei sein darf, erstellt. Folglich besitzt der Baum vier Blätter und somit vier mögliche Ausgangswerte für die Prognose.

Die roten Punkte stellen die Trainingsdaten dar, die für die Erstellung des Modells verwendet wurden. Trotz der geringen Tiefe des Baums, ist es dem Lernalgorithmus möglich die Zusammenhänge der Variablen gut widerzuspiegeln.

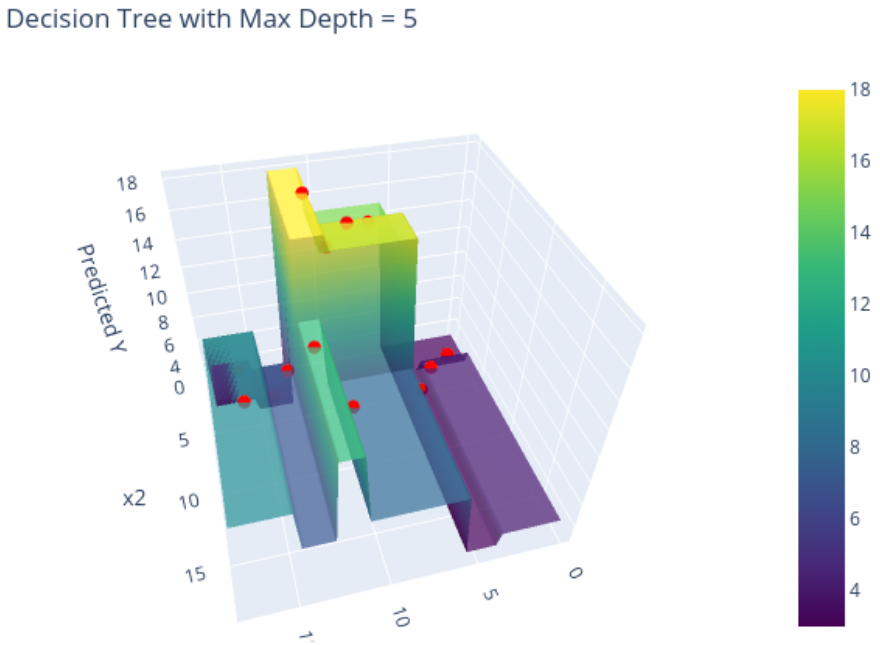

Abbildung 12: Einfacher Entscheidungsbaum mit 2 Merkmalen und der Tiefe 5

Abbildung 13 zeigt zum Vergleich ein Modell eines Entscheidungsbaums der Tiefe fünf. Anhand des Schaubilds lässt sich erkennen, dass sämtliche Trainingsdatenpunkte exakt auf den Ebenen des Modells liegen.

8.2 Weitere Hyperparameter

Sämtliche Hyperparameter haben den Sinn und Zweck, dass Modell bezüglich der Anpassung an die Trainingsdaten zu beschränken. Im Folgenden sollen ein paar der wichtigsten Hyperparameter detaillierter erläutert werden. **??** zeigt die Bestandteile eines Binärbaums.

8.2.1 Mindestanzahl an Proben pro Blatt

Dieser Parameter entscheidet, wie viele Proben mindestens in einem Blatt enthalten sein müssen, damit die Existenz dieses Blattes und die des Bruder-Blattes berechtigt ist. Wenn die Bedingung nicht erfüllt, teilt der sich darüber liegende Knotenpunkt nicht weiter auf und wird selbst zum Blatt. Da Auswirkung des Grenzwertes abhängig von der Größe des Datensatzes ist, ist es durchaus sinnvoll den Mindestwert ebenfalls abhängig von der Größe zu gestalten. Seit der Version 0.18 des Entscheidungsbaum aus der *scikit-learn* Bibliothek ist es möglich eine Fließkommazahl als Parameter anzugeben, woraufhin diese mit der Anzahl der Datenreihen multipliziert wird und somit den Grenzwert bildet.

8.2.2 Beschränkung der Merkmale

Bei jedem Knotenpunkt wird neu entschieden, welches der Merkmale sich am besten eignet, um den Datensatz weiter zu unterteilen. Da an jedem Knotenpunkt ein anderes Merkmal das Beste sein könnte, um die Daten aufzuteilen, muss die Berechnung immer wieder erneut durchgeführt werden. Um die Rechenzeit zu verkürzen, lässt sich die Anzahl der zu berücksichtigten Merkmale beschränken. Wenn die Anzahl der zu berücksichtigen Merkmale limitiert werden soll, wird die maximale Anzahl an Merkmalen zufällig aus der Gesamtliste der Merkmalen ausgewählt.

An dieser Stelle sei angemerkt, dass ein sogenannter *gieriger Algorithmus*[7]

[7] *Gierige Algorithmen* entscheiden schrittweise über nächst besten Folgeschritt. Folglich können wir uns bei dem Ergebnis, welches durch den Algorithmus gefunden wurde, nicht sicher sein, dass es sich um das optimale Ergebnis handelt. Der Vorteil von dieser Art von Algorithmen vergleichsweise schnell und leicht zu implementieren sind. Zudem sind die Ergebnisse in der Regel gut genug, um den Anforderungen gerecht zu werden.

den Entscheidungsbaum erstellt. Für den Entscheidungsbaum bedeutet dies, dass bei jedem Knotenpunkt das momentan beste Merkmal aus der eventuell temporär begrenzten Merkmalsliste ausgewählt wird. Durch die Begrenzung der Merkmale kann es vorkommen, dass ein Merkmal nicht berücksichtigt wird, dass zwar im aktuellen Schritt für die Teilung des Datensatzes das beste Ergebnis erzielt hätte, jedoch nicht für das optimale Endergebnis.

Ob dieses Szenario tatsächlich eintreten könnte, lässt sich nur schwerlich voraussagen. Dementsprechend soll für die Suche nach den besten Merkmalen für die optimale Teilung, die bei jedem Knotenpunkt im Entscheidungsbaum stattfindet, die zusätzliche Rechenzeit in Kauf genommen werden. Stattdessen beschränken wir die Rechenoperationen an anderer Stelle, indem wir die automatisierte Rastersuche auf die aus Erfahrung gut geeigneten Hyperparametern beschränken.

8.2.3 Maximalanzahl an Blättern

Die Maximalanzahl an Blättern ähnelt dem Parameter, der die maximale Höhe des Baums bestimmt. Schließlich lässt sich die Anzahl der Blätter bei einem vollständigem Binärbaum durch die Formel $n = 2^h$ beschreiben lässt. Folglich können beide Parameter sich gegenseitig begrenzen.

Durch die Tatsache, dass es sich bei dem erstellten Entscheidungsbaum nicht zwingend um einen vollständigen Binärbaum handeln muss, können die zwei sehr ähnlichen Parameter den Entscheidungsbaum auf verschiedene Weise beeinträchtigen. Zusätzlich dazu wurden Zahlen ausgewählt, die möglichst weit von der nächsten Zweierpotenz entfernt sind. So wird im Rahmen der Rastersuche nach den optimalen Parametern sichergestellt, dass voneinander differenzierte Entscheidungsbäume erstellt und überprüft werden.

8.2.4 Minimaler gewichteter Anteil der Proben

8.2.5 Aufteilung der Proben

8.3 Hyperparameter auswählen

Da im Vorhinein schwer zu beurteilen ist, welche Hyperparameter am besten zum jeweiligen Datensatz passen, ist eine Rastersuche eine Methode um diese bestmöglichen Parameter zu finden. Eine Rastersuche ist ein Vorgehen, bei der mit roher Gewalt *(zu engl.: brute-force)* versucht wird, die beste Kombination von Hyperparametern zu finden, indem alle Variationen durchprobiert werden.

Da die *Brute-Force-Methode* mit erheblich mehr Rechenaufwand verbunden ist - schließlich muss für jeden Kombination das Modell trainiert und evaluiert werden -, ist eine verkürzte Rastersuche eine beliebte Alternative. Dabei werden nicht alle Variationen durchprobiert, sondern eine zufällige Kombination verschiedener Hyperparameter getestet. Sofern man in Kauf nehmen kann, nicht das bestmögliche Ergebnis zu erzielen, kann durch die randomisierte Auswahl der Kombinationen an Hyperparametern erheblich Rechenzeit eingespart werden.

Um reproduzierbare Ergebnisse zu erhalten, verwenden wir die *Brute-Force-Methode*, um die optimalen Parameter für das jeweilige Modell zu finden. Bei verschiedenen Solaranlagen wurden zwar unterschiedliche Kombinationen an Hyperparametern als optimal bewertet, allerdings waren diese selten weit voneinander entfernt. Folglich können wir das Raster auf ein paar wenige Kombinationen reduzieren und verringern damit gleichzeitig den Rechenaufwand.

OPTIMIERUNG DES RASTERS UNTERSUCHEN UND DIE GEFUNDEN HYPERPARAMETER HIER ABBILDEN

9 Evaluierung des Lernalgorithmus

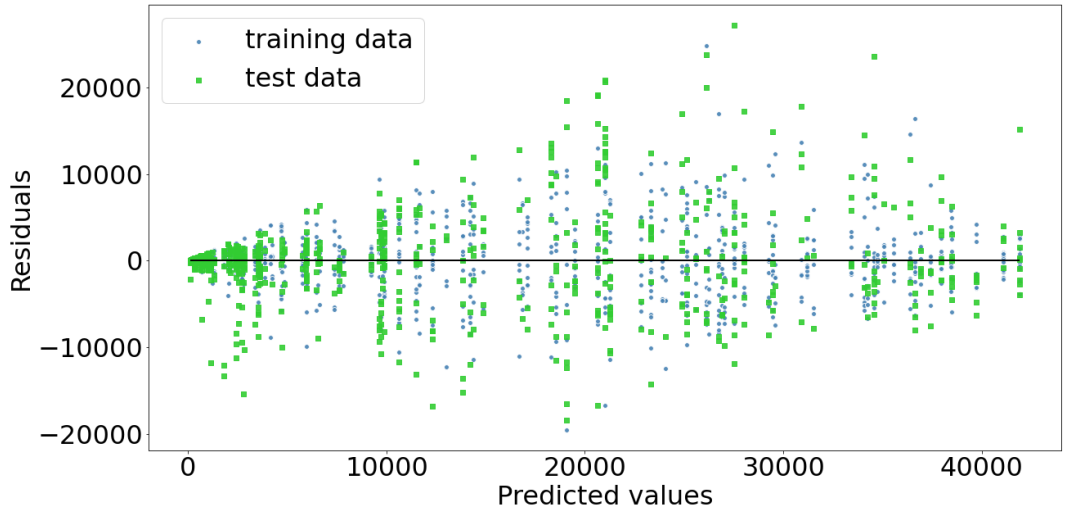

Abbildung 13: Einfacher Entscheidungsbaum mit 2 Merkmalen und der Tiefe 5

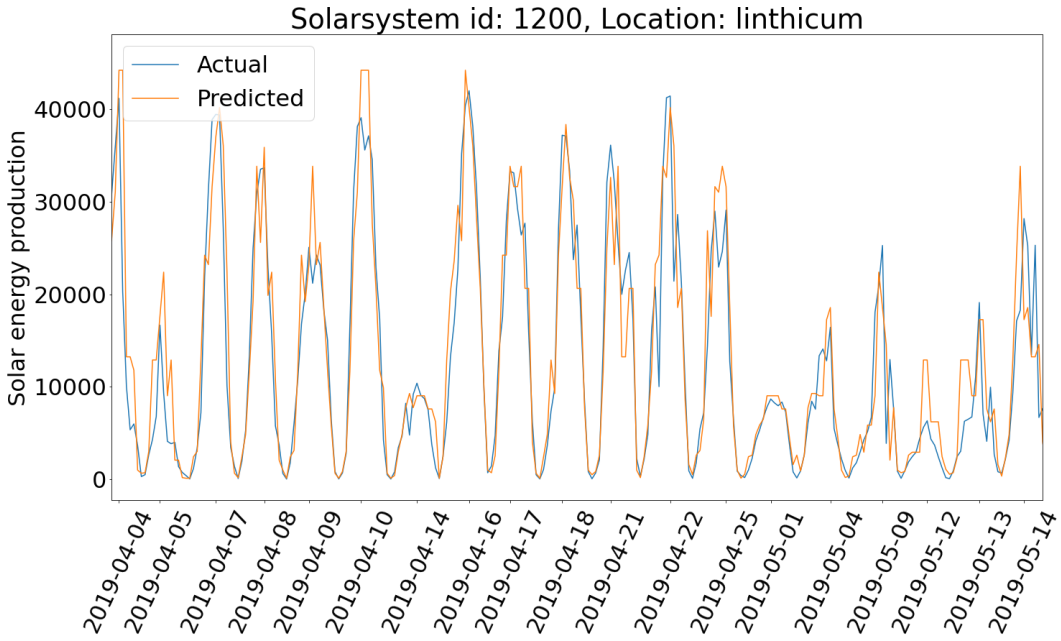

Abbildung 14: Einfacher Entscheidungsbaum mit 2 Merkmalen und der Tiefe 5

9.1 Evaluierung der einzelnen Merkmale

9.1.1 Bewölkungsgrad

Die im Unterunterabschnitt 5.3.1 Qualität der Daten zuvor geäußerten Bedenken, dass die Angaben des Wetterdienst bezüglich des Bewölkungsgrad zu fragwürdig

sind, sollen anhand der Evaluierung des Lernalgorithmus untersucht werden. Dazu trainieren wir das Modell einmal mit und einmal ohne den Bewölkungsgrad als Teil der Merkmalsmatrix. Anschließend vergleichen wir den Determinierungskoeffizienten. Dabei gilt allerdings zu beachten, dass die Aufteilung des gesamten Datensatzes zu einem Test- und einem Trainingsdatensatz randomisiert stattfindet. Dadurch sind die Ergebnisse nicht reproduzierbar und die Evaluierung der unterschiedlichen Modelle nur bedingt vergleichbar. Um dies Genüge zu leisten, teilen wir die Datensätze in etwa zwei gleichgroße Teilmengen auf. So sinkt die Wahrscheinlichkeit, dass Ausreißer im Testdatensatz den mittleren quadratischen Fehler übermäßig verfälschen. Ausreißer im Trainingsdatensatz werden durch Hyperparameter wie die Mindestanzahl an Proben pro Blatt behandelt und das Modell sodurch geglättet.

Zudem führen wir die Evaluierung mehrmals durch und erhalten dadurch mehrere Vergleichswerte. Lassen wir den Bewölkungsgrad aus der Merkmalsmatrix außen vor, so erhalten wir für den Determinierungskoeffizienten des Entscheidungsbaum Werte zwischen 0.82 und 0.89. Im Durchschnitt schneidet das Modell um 0.04 besser ab, als wenn der Bewölkungsgrad für die Prognose berücksichtigt wird.

10 Zusammenfassung und Fazit

11 Ausblick