Fachhochschule Nordwestschweiz
Hochschule für Technik

Künstliche Intelligenz & Skitouren

Automatische Bestimmung des Schwierigkeitsgrades von Skitouren

Diplomarbeit von
Fabian Brunner 2015ibb
Michel Bittel 2015ibb

FHNW
Hochschule für Technik
Studiengang Informatik
Betreuender Dozent:
Csillaghy Andre

Windisch, 28 September 2019

Abbildung 1 Berglandschaft [1]

Kontaktdaten

Fabian Brunner
Bellevuestrasse 26
2540 Grenchen
fabian.brunner@windowslive.com
+41 79 266 20 63

Michel Bittel
Meisenweg 2
3422 Kirchberg
michel.bittel@bluewin.ch
+41 79 653 17 13

Betreuer:

Csillaghy Andre
Bahnhofstrasse 6
5210 Windisch
andre.csillaghy@fhnw.ch
+41 56 202 76 85

Kunde:

Günter Schmudlach
Im Sydefädeli 19
8037 Zürich
schmudlach@gmx.ch
+41 77 471 90 39

Experte:

Oliver Walzer
Hardturmstrasse 161
8005 Zürich
oliver.walzer@ssuf.ch
+41 78 770 03 64

Danksagung
An dieser Stelle sprechen wir unseren Dank an allen Beteiligten aus, welche uns bei der Fertigstellung der Arbeit unterstützt haben.

Namentlich danken wir Herr Csillaghy für die tatkräftige Betreuung der Arbeit und den vielen aufschlussreichen Einwänden. Des Weiteren möchten wir uns bei Herr Schmudlach für die Gelegenheit bedanken, dass wir dieses Projekt durchführen konnten. Zusätzlich haben uns die Erklärungen bezogen auf die Skitouren immer neue Einblicke geliefert.

Ein besonderer Dank gilt Juan Camilo Osorio Iregui und Lorenzo Gatti, welche uns in unzähligen Stunden mit Erklärungen und Rat zu Seite standen.

Ebenfalls möchten wir uns bei Anita Gertiser und Thomas Fluri für das Korrekturlesen unserer Bachelorarbeit bedanken.

Abstract
Seit Jahren werden bestehende Skitouren von Experten und Expertinnen in literarischen Werken beschrieben und bewertet. Diese Bewertungen basieren auf Erfahrungswerten und einer subjektiven Wahrnehmung diverser Kriterien. Im Gegensatz dazu soll ein Modell entwickelt werden, das basierend auf GPS-Daten eine Klassifikation einer Route vornimmt. Diese Arbeit befasst sich mit der Analyse der Daten und einer entsprechenden Entwicklung eines Modells, welches die Routen nach ihrem Schwierigkeitsgrad in mehrere Klassen einteilt. Machine Learning ist in der heutigen Zeit omnipräsent und wird in allen möglichen Bereichen angewandt. Die Datenerhebung basiert auf den GPS-Daten, welche manuell erhoben wurden. Für die Klassifikation empfehlen sich die Modell Logistic Regression, naives Bayes und Random Forest. In einer ersten Phase wurde eine Genauigkeit von 27.2% erreicht mit einer durchschnittlichen Abweichung von 1.5. Mit einer Reduktion der Klassen wurden die Resultate weiter verbessert bis zu einer Genauigkeit von 74.3%, bei einer Abweichung von 0.99.

Glossar

Begriff	Begriffe
Alpha	Schwellwert für die statistische Relevanz
BIAS	Systematische fehlerhafte Neigung
Correlation Matrix	Vergleich der Einflüsse einzelner Feature auf die Vorhersage
Data Augmentation	Die künstliche Anpassung vorliegender Daten
Decision Tree	Ein Entscheidungsbaum über dessen Knoten ein Problem in Teilprobleme abgearbeitet wird
Feature	Eine messbare Eigenschaft der observierten Daten
Imbalance	Nicht alle Klassen haben die gleiche Anzahl an Daten
Klasse	Ein Wert der Vorhersage des Modells, welches mehrere Daten zusammenfasst. Beispiel mehrere Routen haben den Schwierigkeitsgrad 5. Der Schwierigkeitsgrad ist hierbei eine Klasse
Level	Siehe Klasse
Lineare Regression	Machine Learning Modell mit linearem Hintergrund
Modell	Algorithmus für Machine Learning
p-value	Wahrscheinlichkeit des Eintritts der Antithese
Random Forest	Machine Learning Modell mit Decision Tree als Grundlage
Sampling	Die Trennung der Daten in einzelne Sets
Schwelle	Grenzwert, welcher nicht über- oder unterschritten werden darf.
Set	Auswahl von Datensätzen
Supervised Learning	Ansatz bei welchem dem Modell die Rahmenbedingungen und das Resultat mitgegeben wird
TSFresh	Tool zur Feature Erweiterung durch mathematische Merkmale der Daten
Unsupervised Learnings	Ansatz bei welchem das Modell die Zusammenhänge und Resultate selber ermitteln soll

Tabelle 1 Glossar

Inhaltsverzeichnis

1 Einleitung

Im folgenden Kapitel werden die Ausgangssituation und die definierten Ziele festgehalten und Problemstellungen festgehalten. Auch soll ein Verständnis für die Thematik vermittelt werden.

1.1 Ausgangslage

Die Web-Plattform «www.skitourenguru.ch» hat das Ziel, Wintersportlern bei der Auswahl von geeigneten Skirouten behilflich zu sein. Dabei unterstützt sie Wintersportler bisher in folgenden Bereichen:

1. Eine Hilfestellung für Skitouren Einsteiger
2. Eine Berechnung des aktuellen Lawinenrisikos
3. Eine Übersicht über die aktuellen Schneemengen in Alpenraum Schweiz
4. Eine Sammlung von automatisch generierten Skitouren
5. Eine Reihe von Empfehlungen für Skirouten, welche ein niedriges Lawinenrisiko aufweisen.
6. Einen von Experten manuell festgelegten Schwierigkeitsgrad pro Route, welcher sich nach dem Schweizer Alpenclub (SAC) richtet.

Ein grosser Teil dieser Aufgaben findet automatisiert statt. Dies heisst das ein grosser Teil der Risikokarten und Routen von Computern generiert wird. [1]

Trotz anfänglicher Skepsis von Skitouren Experten, konnte sich die Plattform mit dieser automatisierten Methode etablieren und wird heute vom Schweizer Alpenclub für die Planung von Skitouren sogar empfohlen. Durch diese Etablierung ist der Wunsch nach einem neuen Produkt entstanden. [1]

In einem weiteren Schritt soll nun die bislang manuelle Einschätzung des Schwierigkeitsgrads einzelner Routen, ebenfalls automatisiert werden. Dies soll eine bisherige Inkonsistenz bei der Bewertung vermindern, da eine manuelle Einschätzung durch Einzelpersonen immer auch eine subjektive Komponente mit sich bringt. Aufgrund dessen wurde die bisherigen erfasst GPS-Daten von der Web-Plattform ans uns weitergegen. Diese Sammlung beinhaltet mehr als 1000 Skitourenrouten. Diese automatische Bestimmung des Schwierigkeitsgrades auf Grunde der übermittelten Daten ist der Kern der vorliegenden Thesis.[2]

1.2 Kontext

Doch was genau ist eine Skitour?

Skitouren (auch als Skibergsteigen bekannt) beschreibt das Besteigen eines Berges mittels Tourenski. Diese Skier sind kürzer und breiter als gewöhnlich um die Abfahrt im Tiefschnee zu erleichtern.

Beim Besteigen des Bergs werden Felle oder andere Stoffe um die Skier gewickelt, um einen festen Halt im Schnee zu gewährleisten. Ausserdem ist die Ferse nicht in der Bindung fest eingeschlossen, sondern kann sich heben, um den Aufstieg zu erleichtern. Dies ermöglicht das erklimmen von steilen Berghängen auf eine sehr Energieeffiziente Art.
Oben angekommen werden dies Felle entfernt und die Fahrt zurück ins Tal beginnt.[3],[4]

Skitouren an sich sind jedoch nichts neues. Man geht davon aus, dass die Menschen bereits vor mehr als 7000 Jahren Ski-ähnliche Gerätschaften genutzt haben, um schneller durch den Schnee zu kommen. Grosses Interesse hatte auch schon sehr früh das Militär, da in Bergregionen eine schnelle Aufklärungstruppe viele Vorteile brachte, besonders als die Technik noch nicht so weit

fortgeschritten war wie sie es heute ist. Daraus entstand in den 1920er Jahren die ersten zivilen Skitourenrennen und schliesslich eine winterliche Freizeitbeschäftigung.[3]

Bei Skitouren ist je nach Route einiges an Vorbereitung gefragt. Dies beginnt bei der Ausrüstung: Neben offensichtlichen Dingen wie die Skiausrüstung selbst, Verpflegung und Werkzeugen, sind Lawinensuchgeräte, Lawinensonden, sowie Lawinenschaufeln bei den meisten Touren Standard. Zudem sind Informationen zu aktuellen Lawinengefahren, Schwierigkeit und Begehbarkeit äusserst wichtig. Und daher sind Angebote wie die erwähnte Seite «www.skitourenguru.ch», eine grosse Hilfe und müssen entsprechend verlässlich und präzise sein.[3]

1.3 Ziele

Das Hauptziel der Arbeit ist, wie bereits erwähnt, mit den vorhandenen Daten ein Modell zu erstellen, welches Skirouten gemäss der Skala des SAC klassifizieren kann.

Daraus ergeben sich folgende Unterziele:

1. Es muss eine geeignete Methode oder ein Modell gefunden und erstellt werden, um aus den vorliegenden Routeninformationen, eine Einschätzung des Schwierigkeitsgrades machen zu können.
2. Die Kriterien, welche das SAC für die einzelnen Schwierigkeitsgrade vorgibt, sollen nach Möglichkeit eingehalten werden.
3. Die Einstufung soll eine Genauigkeit erreichen, mit der ein verlässlicher Service geboten werden kann.

Basierend auf diesen Zielen, ergeben sich folgende Fragestellungen, die innerhalb dieser Arbeit gelöst und geklärt werden sollen:

1. Mit welchem Modell wird eine verlässliche Genauigkeit erreicht?
   (Kundenziel)
2. Sind noch andere Daten vorhanden, direkt oder indirekt, welche bei der Bewertung von Skitouren hilfreich sein könnten? (Akademisch)
3. Sind zwischen den verschiedenen Attributen innerhalb der Daten bisher noch unbekannte Beziehungen vorhanden?
   (Akademisch)

Eine ergänzende Fragestellung, welche als Input durch den Kunden gestellt wurde:

4. Haben einzelne Personen eine Auswirkung auf die Bewertung des SAC?
   Werden als Beispiel Routen im Tessin eher einfacher bewertet als die im Graubünden?
   (Kundenziel)

1.4 Herausforderungen

Die Hauptherausforderung ist die Genauigkeit der Vorhersagen, da wie in den Zielen definiert, die Einschätzung verlässlich sein soll. Das Augenmerk liegt hierbei bei den Daten. Dabei sind folgende zwei Hauptproblematiken innerhalb der Daten vorzufinden:

1. Beschaffung der Daten
   Da die Schwierigkeit der Routen bisher von einzelnen Personen geschätzt und keine Route doppelt eingestuft wurde, schwanken die Einschätzungen der Schwierigkeitsgrade durch die Experten stark. Um das vereinfacht darzustellen: Was für den einen eine einfachere Route ist, mag für den nächsten eine mittlere Route sein.
   Diese Schwankungen erschweren das Herausfiltern der Eigenschaften einer Route für einen gegebenen Schwierigkeitsgrad. Oder anders formuliert: Was muss eine Route aufweisen, um als einfach eingestuft zu werden?

2. Menge an Daten
   Die vorliegende Datenmenge ist mit 1203 Routen eher bescheiden. Zwar ist die Zahl für die kleine Region der Schweizer Alpen beeindruckend, für eine automatisierte Bestimmung jedoch eher gering.

1.5 Resultate

Aufgrund der Arbeit konnte in folgenden Bereichen neue Erkenntnisse gewonnen werden, welche im Laufe der Arbeit erläutert und nachgewiesen werden.

- Verbesserungsvorschläge für die Datenerfassung
- Qualität der Bewertungsskala des Schweizer Alpen Clubs
- Beziehung der Daten zueinander
- Modell mit einer 95% Genauigkeit
- Entwicklungsmöglichkeiten des Modells

1.6 Berichtsstruktur

Die Struktur des Berichtes beginnt mit einer Beschreibung der Ausgangslage. Die Ausgangslage setzt sich mit dem Anwendungsbereich des Schwierigkeitsgrades von Skitouren auseinander. Des Weiteren wird auf den aktuellen Stand der Forschung von Machine Learning Bezug genommen. Dieser Abschnitt leitet nachher zur Datenanalyse über. Im Abschnitt der Datenanalyse werden die erhaltenen Daten analysiert und mit verschiedenen Grafiken dargestellt und veranschaulicht.

Anschliessend befasst sich der Bericht mit der Datenaufbereitung, bei welcher die erhaltenen Daten weiterverarbeitet werden. Dies beinhaltet auch die Anpassung existierender Daten. Darauffolgend befasst sich der Bericht mit der Modell Auswahl. An dieser Stelle werden verschiedene Modelle ausprobiert und deren Eigenschaften und Optimierungen beschrieben.

Am Ende des Berichtes befindet sich die Auswertung, in welchem das gewählte Modell genau beschrieben wird. Zusätzlich wird ein Ausblick auf die Entwicklungsmöglichkeiten des Modells gegeben.

2 Forschungsstand

Dieser Abschnitt beinhaltet einen Überblick der Theorie und der eingesetzten Technologien und Algorithmen. Zusätzlich wird das bisherige Bewertungsverfahren der Skitouren dokumentiert. Des Weiteren werden ein paar der wichtigsten Mathematischen Funktionen erläutert.

2.1 Bewertung von Skitouren

Zum aktuellen Zeitpunkt werden Skitouren in der Literatur von Experten bewertet. Diese Bewertung erfolgt basierend auf der Skala, welche im Anhang zu finden ist (A). Die jetzige Bewertungsskala überlässt dem Experten einen relativ grossen Interpretationsspielraum bei der Bewertung. Die Hauptkriterien bei der Bewertung der verschiedenen Routen sind die Faktoren, Steigung, Ausgesetztheit und die Anzahl und die Länge von Engpässen. Von diesen Kriterien ist nur die Steigung mit einem Messbaren vordefinierten Wert beschrieben während der anderen Faktoren mit Worten umschrieben werden. Dies führt zu eben beschriebenen Interpretationsspieleraum. [5], [1]

Die Datenerfassung der Routen wurde in der Vergangenheit mit einem Foto der Seitenansicht des Berges und einem Stift realisiert. Dies wird in nachfolgenden Abbildung 2 veranschaulicht.

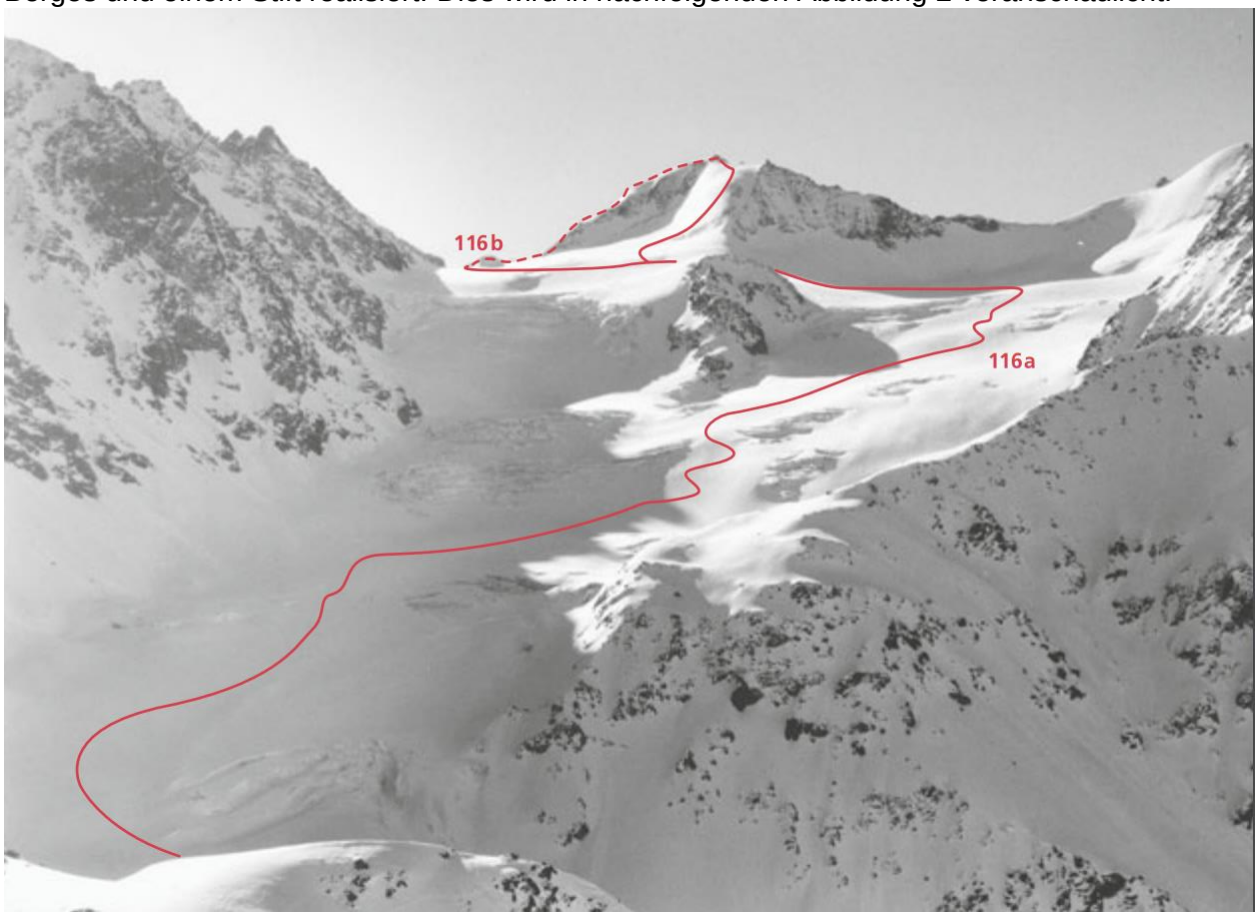


Abbildung 2 Route von Combin de Boveire von «die klassischen Skitouren» [6, S. 107]

Die Route auf Abbildung 2 wurden von einem Experten von Hand auf das Foto bezeichnet. Es soll einem Anfänger einen Leitfaden einer Route liefern, welche ein Experte mit mehr Erfahrung wählen wurde. Dabei kann es jedoch vorkommen, dass Abschnitte überhaupt nicht ersichtlich sind, wenn diese von der Seitenansicht nicht einsehbar ist. Diese Methode ist sehr ungenau und erschwert es einem Anfänger die entsprechende Route basierend auf diesem Bild korrekt zu wählen. Häufiger musste man sich mündlich bei erfahrenen Kollegen erkundigen, diese konnten den Weg präziser erläutern.

Das heisst, die jetzige Routenansicht ist nur eine ungenaue Wegbeschreibung, an welche die eigene Route angepasst wird. Es gab aber in der Vergangenheit Anstrengungen diese Routen mittels GPS zu vermessen. Diese Daten waren aber bisher nicht öffentlich zugänglich.

Bei einer Skitour kann es ausserdem vorkommen, dass teilweise Abschnitte zu Fuss oder mit Kletterausrüstung zurückgelegt werden. Für diese Abschnitte besteht jeweils eine eigene Skala. Um Fusspassagen und Kletterpassagen zusammenfassen zu können wird der Begriff alpintechnische Schwierigkeit eingeführt. Im Rahmen unserer Arbeiten werden wir nicht genauer darauf eingehen.

2.2 Machine Learning

Machine Learning beschreibt die Methode einem Computer das Lernen zu ermöglichen, ohne dies explizit zu programmieren. Beispielweise bei der Bewertung von Skitouren könnten wir ein Programm schreiben, welches definiert sobald der Durchschnitt der Steigung einer Route den Wert 30% überschreitet, so ist diese Route als schwer zu betrachten. Im Gegensatz dazu steht Machine Learning wo als Beispiel eine Route mit verschiedenen Eigenschaften mitgegeben und ein Schwierigkeitsgrad zurückgegeben wird. Hier werden die Schwellenwerte wie im Beispiel die 30% nicht selbst definiert. Die Entscheidung des Machine sind auf komplexe statische Operationen gestützt und unterscheiden sich je nach Algorithmus stark voneinander. Des Weiteren wird dabei das Modell «trainiert», damit es lernt eine möglichst genaue Schätzung abzugeben, respektive, welche Feature welchen Einfluss dazu haben, um den Schwierigkeitsgrad zu bestimmen. Unter Training ist zu verstehen, dass eine gewisse Anzahl Beispielsdaten eingelesen werden, auf deren Basis eine empirische Risikominimalisierung ausgewählt wird, um die Bewertung durchzuführen. Das Risiko wird durch einen Fehlentscheid bei der Bewertung verkörpert. Der Begriff des Trainings wird in weiteren Verlauf der Arbeit immer wieder auftauchen.[7]

Das Modell erhält also keine Information darüber “wie” das Problem zu lösen ist. Lediglich einen Input (die Eigenschaften einer Skiroute) und in manchen Fällen einen erwarten Output (der Schwierigkeitsgrad einer Skiroute), an welchem dieser sich richten kann. Je komplexer die Probleme werden, desto schwieriger wird es für uns Menschen diese Werte ohne Erfahrungswerte zu definieren. [7]

Da abgesehen von der Steigung keine andere Routeneigenschaft nummerisch klar auf einem Schwierigkeitsgrad abgebildet wurde, ist ein herkömmlicher Weg, über eine vordefinierte Skala nicht möglich. Beispielsweise wird die Absturzgefahr als «Engpässe kurz, aber steil» beschrieben. Was dies nun in Zahlen zu bedeuten hat, muss ein Modell selbst lernen. Daher wird im weiteren Verlauf den Ansatz des Machine Learnings verwendet.[7]

2.2.1 *Supervised Learning*

Die verschiedenen Methoden im Machine Learning lassen sich in zwei Klassen einteilen. Das «Supervised Learning» oder auch überwachtes lernen ist die häufigste angewandte Methode des Lernens. Hierbei beinhalten die Daten neben den Input Variablen auch «Features» genannt, auch die Output Variablen. Die Skiroute A besitzt folgende Input Variablen (Steigung, planare Krümmung, Ausgesetztheit und Korridorbreite), beinhaltet aber auch die aktuelle Bewertung der Route mit dem Schwierigkeitsgrad «leicht». [8]

Wird das Modell nun trainiert, kann sich dieses an dem erwarteten Resultat orientieren. Es wird also eine Funktion gesucht, welche die Eingangswerte (X) möglichst genau auf die erwarteten Ausgangswerte (Y) abbildet. Dabei wird die Funktion Schritt für Schritt verbessert und nähert sich somit (im optimalen Fall) immer näher an den erwarteten Wert an.[7]

Bei dieser Annäherung wird häufig eine Grenze oder Schwelle definiert. Über eine «Cost Function» wird die Genauigkeiten der Hypothese durch die durchschnittliche Differenz ermittelt. Sobald diese Differenz unter eine vorher festgelegt Grenze fällt, oder sich verschlechtert, wird das

Lernen abgebrochen. Dazu wird auch die Grösse der einzelnen Schritte über Alpha, die sogenannte Learning Rate, festgelegt. Diese Grösse gibt an, wie gross die Schritte eines Modells sein sollen, auf dem Weg zum Minimum der Cost Function. Verschlechtern kann sich die Differenz, wenn über den optimalen Punkt “hinaus gelernt” wird. Dies kann zum Beispiel passieren, wenn Alpha zu gross gewählt wird.[7]

2.2.2 *Unsupervised Learning*

Beim «Unsupervised Learning», auch unüberwachtes Lernen genannt, wird dem Modell das erwartete Resultat nicht mitgeteilt. Das Modell muss selbst Informationen und Zusammenhänge erkennen.[7]

Ein Kleinkind, welches das Wort “Hund” noch nicht kennt, kann dennoch Hunde als solche erkennen. Es erkennt zum Beispiel Eigenschaften, welche die meisten Hunde bisher aufwiesen, wie z.B. zwei Ohne, Augen, die typische Nasenspitze, vier Beine, Zähne, Geräusche wie Bellen usw.) Das Kind weiss also nicht wie das Tier genannt wird, kann jedoch trotzdem einen Hund erkennen, wenn es ihn sieht. Dies wird auch Clustering genannt was mit einer Klassifikation gleichzusetzten ist.[7]

So ähnlich funktioniert Unsupervised Learning. Dem Modell wird nicht gesagt, was erwartet wird oder was die Daten bedeuten. Es soll selbst Zusammenhänge erkennen und daraus Schlüsse ziehen. Meisteins wird diese Art bei der Datenanalyse eingesetzt, um versteckte Zusammenhänge zwischen den Daten aufzuzeigen.[7]

Der grössten Herausforderung hierbei ist die Validierung des Resultates. Bei der Supervised Methode kann die Vorhersage mit den eigentlichen Daten vergleichen. Diese Möglichkeit gibt es beim Unsupervised Learning nicht. Dies macht es sehr schwierig den Output auf seine Korrektheit zu überprüfen. [8]

3 Datenübersicht

Im folgenden Abschnitt werden die Daten erklärt. Dabei werden Aussagen zum Messverfahren sowie die verschiedenen Eigenschaften der Daten definiert. Des Weiteren befasst sich das Kapitel mit der Struktur, der Art und der Menge der Datensätze.

3.1 Rohdaten

Die zur Verfügung gestellten Daten bestehen aus einem QGIS Projekt. QGIS ist eine Open-Source Software, mit welcher geografische Informationen verarbeitet und dargestellt werden können. Die vorliegenden Routen sind darin wie folgt abgebildet:

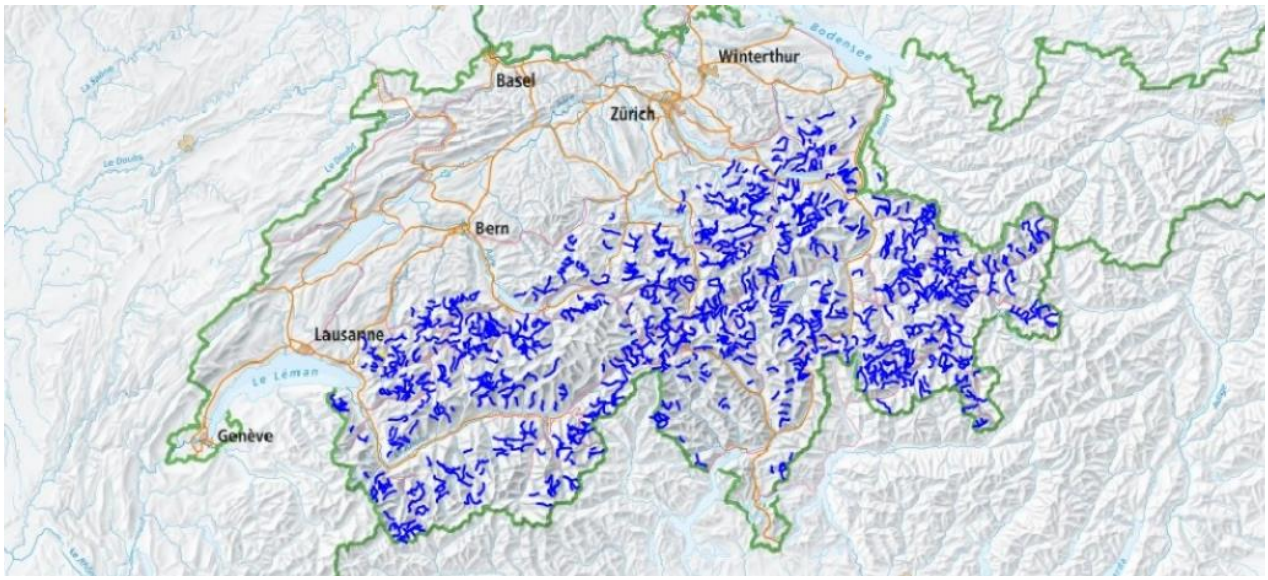


Abbildung 3 Übersicht der Routen

Die dazugehörigen Daten liegen als SQLite lokal auf der jeweiligen Arbeitsstation und können entsprechend exportiert werden. In dem Projekt enthalten ist eine Hintergrundkarte der Schweiz, die einzelnen Routen und dazu die verschiedenen Properties jedes einzelnen Punktes dieser Routen.

Struktur der Daten im QGIS:

- Base Maps
  Das sind die Hintergrundkarten. In diesem Fall eine Grundlegende Karte der Schweiz. Diese liegt als einfache Grafik vor und dient der Orientierung und ist für diese Arbeit jedoch nicht weiter von Relevanz.

- Routes
  Linien Features, welche Start- und Endpunkt, sowie Region und den festgelegten Schwierigkeitsgrad beinhalten. Diese Daten werden mit einem 10 Meter Abstand gesampelt und daraus entstehen die einzelnen Punkte, welche zusammen mit den Route Properties die Rohdaten für diese Arbeit ergeben.

- Route_Properties
  Enthalten die einzelnen Eigenschaften, die eine Route aufweisen kann. Darin enthalten ist pro Eigenschaft eine Tabelle, mit der ID der Route und der entsprechenden Eigenschaft für den jeweiligen Punkte.

Geliefert wurden 1203 Routen, welche sich aus insgesamt 632’520 Punkten zusammensetzen.

3.2 Zusammenhang der Daten

Jede Route ist durch eine ID gekennzeichnet. Jeder gesampelte Punkt aus einer Route erhält diese ID ebenfalls mitgeliefert. Eine Route hat dementsprechend 1 bis n Punkte.

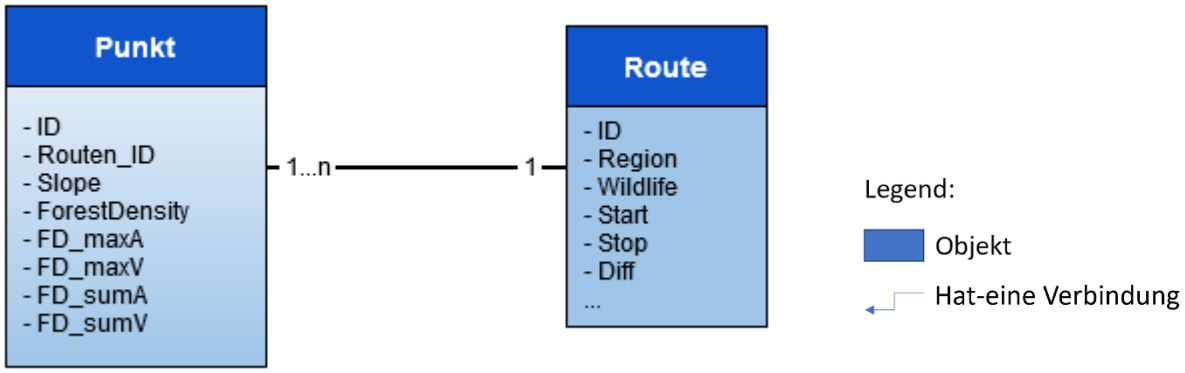


Abbildung 4 Datenstruktur

Darüber hinaus gibt es für jede Eigenschaft, welche ein Punkt aufweisen kann, eine eigene Tabelle. Es gibt also eine Tabelle, welche die Steigungswerte für jeden einzelnen Punkt beinhaltet. Eindeutig gekennzeichnet sind die Punkte durch die Kombination aus ihren Koordinaten und der zugehörigen Route.

Um die Daten nun weiter zu verwenden, werden alle Tabellen mit Eigenschaften zu einer grossen Gesamttabelle zusammengeführt. Über die Routen ID können bei Bedarf weitere Informationen wie der Schwierigkeitsgrad der Route oder Start und Stopp Punkt aus der Routen Tabelle ermittelt werden.

Diese Datenlage ist nachfolgend nochmal visuell dargestellt: Auf der rechten Seite sind die einzelnen Tabellen mit den einzelnen Punkten, welche dann über eine Join Operation zu einer, auf der linken Seite dargestellten, Gesamttabelle zusammengefasst werden.

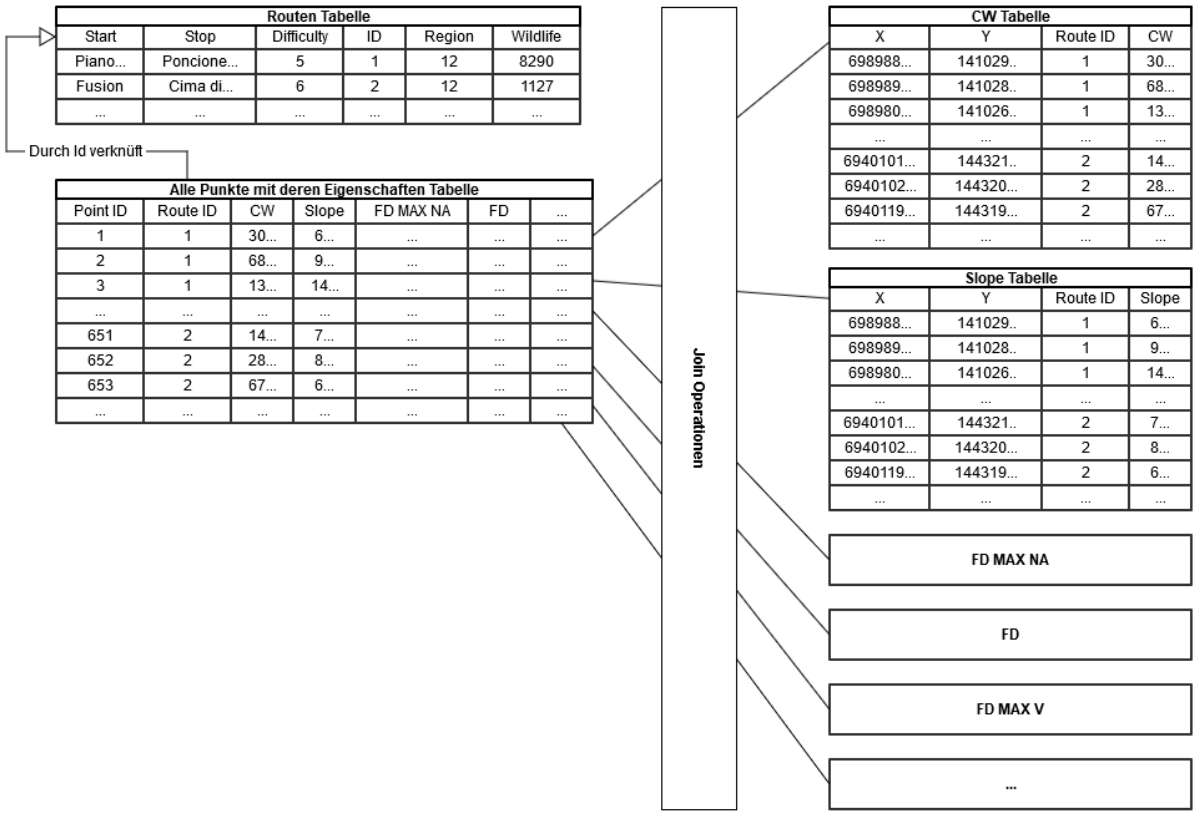


Abbildung 5 Zusammenführung der Daten

3.3 Routeninformationen

Die Routentabelle liefert weitere Informationen, welche abgesehen vom Schwierigkeitsgrad jedoch keinen direkten Mehrwert darstellen. Start- und Endpunkt sind beispielsweise für eine Suche auf der Karte interessant, für die Einschätzung des Schwierigkeitsgrades aber nicht relevant.
Die Region wurde ebenfalls nicht verwendet. Zwar wäre zur Analyse der subjektiven Vergabe der Schwierigkeitsgrade durch die einzelnen Experten eine Unterteilung in die einzelnen Regionen interessant, jedoch ist die Datenmenge als Ganzes bereits sehr klein und würde durch eine weitere Unterteilung nicht mehr ausreichen für die Erstellung eines Modells. Daher wird diese Information erst gegen Ende zur Klärung der Fragestellung umgesetzt, und nicht um Daten zu korrigieren.

Feldinhalt	**Datentyp**	**Skalierung**
Schwierigkeitsgrad	ganzzahlige Zahl	0-18
Start- und Endpunkt	Text	-
RegionsID	ganzzahlige Zahl	Autogen
Wildlife	ganzzahlige Zahl	nicht interpretierbar bis jetzt

Tabelle 2 Informationen der Routentabelle

3.4 Schwierigkeitsgrad

Die Skala für den Schwierigkeitsgrad entspricht der offiziellen Skala des SAC für Skitouren. Diese ist dem Anhang zu entnehmen (A1 **Error! Reference source not found.**).

Grad	Steigung	Ausgesetztheit	Geländeform	Engpässe
L L+	bis 30°	keine Gefahr	weich, hügelig [...]	keine
WS- WS WS+	ab 30°	kürzere Rutschwege	[...] offene Hänge, begehbare Hindernisse	kurze, wenig steil
ZS- ZS ZS+	ab 35°	längere Rutschwege mit Bremsmöglichkeiten	kürzere Steilstufen, mässig steiles Gelände	kurz, aber steil
S- S S+	ab 40°	lange Rutschwege (Lebensgefahr)	Steilhänge, viele Hindernisse	lang und steil
SS- SS SS+	ab 45°	Rutschwege in Seilstufen abbrechend (Lebensgefahr)	anhaltend steiles Gelände, viele Hindernisse	sehr lang und steil. Quersprünge nötig
AS- AS AS+	ab 50°	äusserst ausgesetzt	äusserst steile Flanken. Keine Erholungs-möglichkeiten	lang und sehr steil. Nur Quersprünge und Abrutschen möglich
EX	ab 55°	extrem ausgesetzt	extreme Steilwände	evtl. Abseilen über Felsstufen nötig

Tabelle 3 Skala zum Schwierigkeitsgrad des SAC

Dabei wird der Schwierigkeitsgrad, welcher bei SAC als eine Kombination aus Buchstaben und den Zeichen + und - dargestellt wird, in eine Zahl umgewandelt. Dies da es für Laien einfacher ist eine Zahl von 1-18 zu interpretieren, als Buchstaben mit + und -.

L	L+	WS-	WS	WS+	ZS-	ZS	ZS+	S-	S	S+	SS-	SS	SS+	AS-	AS	AS+	EX
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Tabelle 4 Zuordnung des Schwierigkeitsgrades Test zu Nummer

Wert	Datentyp	Wertebereich	Messbereich
diff	Faktor	1,2,3,4,5,6,7,8,9,10,11,12,13,15,16,17,18	Route

Tabelle 5 Wertebereich der diff Eigenschaft

Wie bereits in der Problemstellung erwähnt, wurden die Schwierigkeitsgrade der Routen durch verschiedene Experten geschätzt. Jedoch wurde jede Route bloss einmal geschätzt und dies immer nur von einer Person oder Gruppe. Nun ist es so, dass verschiedenen Experten Routen leicht unterschiedlich einschätzen. Was für den Ersten eine 3 ist, mag für den nächsten eine 4 sein.

Dies führt dazu, dass es nicht möglich ist, einen Durchschnitt für die jeweiligen Routen zu nehmen. Eine regionale Analyse der Routen wird ebenfalls durch die geringe Menge der Daten erschwert. Mit deutlich mehr Daten könnte eine regionale Analyse durchgeführt werden, um Tendenzen

festzustellen und so die Daten zu korrigieren, falls z.B. das Tessin im Schnitt immer einen Schwierigkeitsgrad unter dem Durchschnitt liegt.

Aufgrund der geringen Datenmenge ist keine aussagekräftige regionale Analyse, ohne ein vorheriges Modell zu haben, jedoch nicht möglich.

3.5 Eigenschaften der Punkte

Die einzelnen Punkte weisen Eigenschaften auf, die die Route beschreiben. Diese Eigenschaften werden im nachfolgenden erläutert.

3.5.1 *Steilheit (Slope)*

Slope beschreibt die Steigung. Diese wird immer in Richtung des Gradienten berechnet und in Grad angegeben. Ab einer Steigung von 18° beginnt der Skitourer den Berghang in einem Zick-Zack zu besteigen, um die Steigung zu bewältigen. Dies ist Aufgrund der physikalischen Limitierung der Steigfelle notwendig.

Das Skitourenguru Tool verlängert hierbei ab einer Steigung von 18° die Strecke, da davon ausgegangen wird, dass hier im Zick-Zack gestiegen wird, wie in Abbildung 6 veranschaulicht wird.

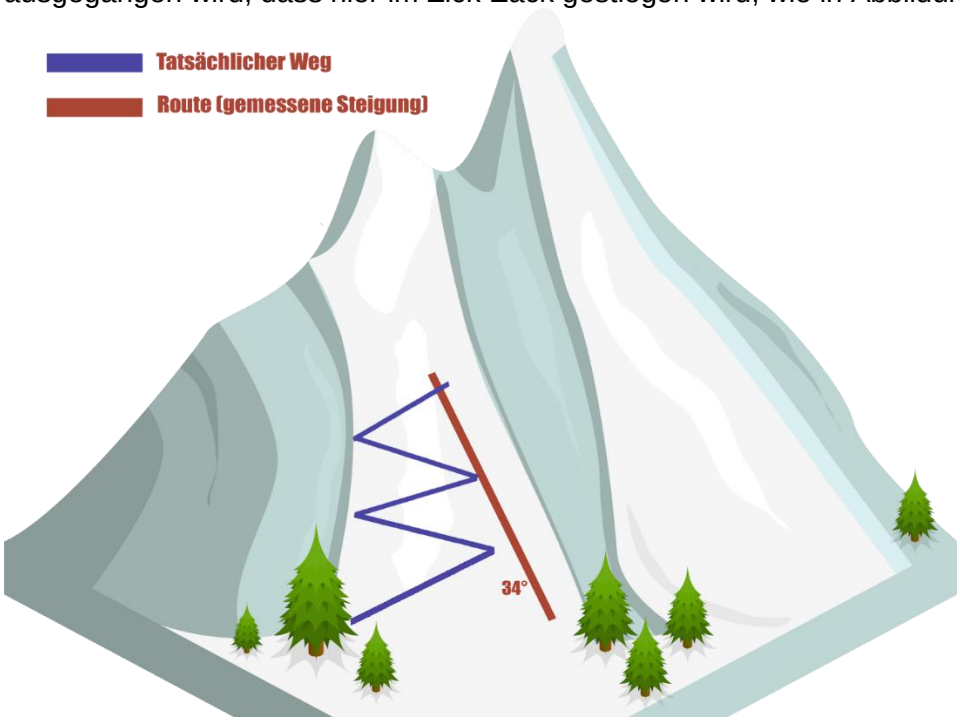


Abbildung 6 Neigungsbeispiel

Dieser Aspekt wurde nicht direkt verarbeitet, da die Länge der Route keinen Einfluss auf den Schwierigkeitsgrad aufweist. Daher lassen wir das Modell selbstständig Steigungen im Bezug zur Schwierigkeit einschätzen.

Wert	Datentyp	Wertebereich	Messbereich
Slope	Fliesskommazahl	0 bis ∞	Punktuelle

Tabelle 6 Dateneigenschaften der Slope Werte

3.5.2 *Planare Krümmung*

Die Planare Krümmung verläuft senkrecht zum Steigungswert. Sie beschreibt die Abweichung der Kurve von der Geraden. Der Vektor, welcher sich senkrecht zur Steigung befindet, kann sich im zweidimensionalen Raum in zwei Richtungen beugen. Daher kann diese Eigenschaft positive wie auch negative Werte annehmen.

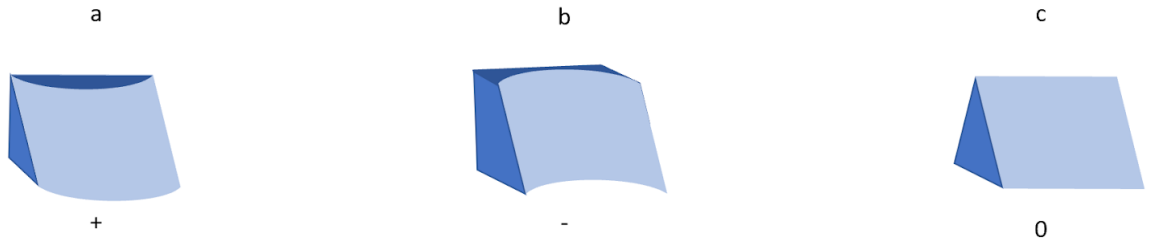


Abbildung 7 Beispiel der planaren Krümmung

Wert	Datentyp	Wertebereich	Messbereich
Planc	Fliesskommazahl	-∞ bis ∞	Punktuelle

Tabelle 7 Dateneigenschaften der Planc Werte

- Sehr stark konvex (sehr spitzer Grat): [-unendlich.-100]
- Stark konvex (spitzer Grat): [-100..-50]
- Konvex (runder Grat = Rücken): [-50..-20]
- Uniformer Hang: [..-20, +20]
- Konkav (runde Rinne) [+20..+50]
- Stark konkav (spitze Rinne): [+50..+100]
- Sehr stark konkav (sehr spitze Rinne) [+100..+unendlich]

3.5.3 *Ausgesetztheit*

Die Ausgesetztheit beschreibt was passieren könnte, sollte der Skitourer abstürzen. Dafür wird einerseits die Beschleunigung, die auf einen fallenden Körper wirken, sowie die Geschwindigkeit gemessen. Wobei bei der Beschleunigung, nur die nach unten wirkender Kraft verwendet wird, da diese schlussendlich die gefährliche ist. Hierbei werden verschiedene Messwerte verwendet:

Summe der Beschleunigung [m/s2]	FS_SUM_NA	Die Summe aus allen gemessenen Beschleunigungen, im Abstand von 10 Metern während des Falles.
Maximale Beschleunigung [m/s2]	FD_MAX_NA	Die höchste Beschleunigung, welche erreicht wurde.
Summe der Geschwindigkeit [m/s]	FS_SUM_V	Sie Summe aus allen gemessenen Geschwindigkeiten, im Abstand von 10 Metern während des Falls.
Maximale Geschwindigkeit [m/s]	FD_MAX_V	Die höchste Geschwindigkeit, welche erreicht wurde.

Tabelle 8 Bedeutung der Verschiedenen Werte zur Ausgesetztheit

Wert	Datentyp	Wertebereich	Messbereich
FS_SUM_NA	Fliesskommazahl	0 bis ∞	Punktuelle
FD_MAX_NA	Fliesskommazahl	0 bis ∞	Punktuelle
FS_SUM_V	Fliesskommazahl	0 bis ∞	Punktuelle
FD_MAX_V	Fliesskommazahl	0 bis ∞	Punktuelle

Tabelle 9 Dateneigenschaften der Ausgesetztheit

3.5.4 *Walddichte (Forest Density)*

Dieser Wert beschreibt die Walddichte des anzutreffenden Punktes. Bäume können die Sicht auf den weiteren Routenverlauf einschränken, die Beschaffenheit des Schnees kann in Waldnähe verschieden sein und sie werden zu Hindernissen bei der Abfahrt.

Der Wertebereich ist wie folgt definiert 0-100. Diese Zahl gibt an wie viel Prozent des Punktes mit Bäumen bedeckt sind.

Wert	Datentyp	Wertebereich	Messbereich
density	ganzzahlige Zahl	0-100	20mx20m

Tabelle 10 Dateneigenschaften der Walddichte

3.5.5 *Korridorbreite*

Die Korridorbreite ist im Grunde die Wegbreite in Metern.

Ist links und rechts viel Platz, um einem möglichen Hindernis oder einer potenziellen Absturzstelle auszuweichen, so ist die Route einfacher einzuschätzen, als wenn steile Engpässe bewältigt werden müssen, die dann oft mit einer Absturzgefahr verbunden sind.

Wert	Datentyp	Wertebereich	Messbereich
CW	Fliesskommazahl	0 bis ∞	Punktuelle

Tabelle 11 Dateneigenschaften der Korridorbreite

4 Statistische Auswertungen

Die vorhandenen Daten werden in den nachfolgenden Kapiteln auf Unterschiede, Gemeinsamkeiten, Trends und allgemeine Charakteristiken untersucht.

4.1 Verteilung der Daten pro Klasse

Werden die Daten für jeden Schwierigkeitsgrad / Klasse gruppiert, so ergibt sich nachfolgendes Bild:

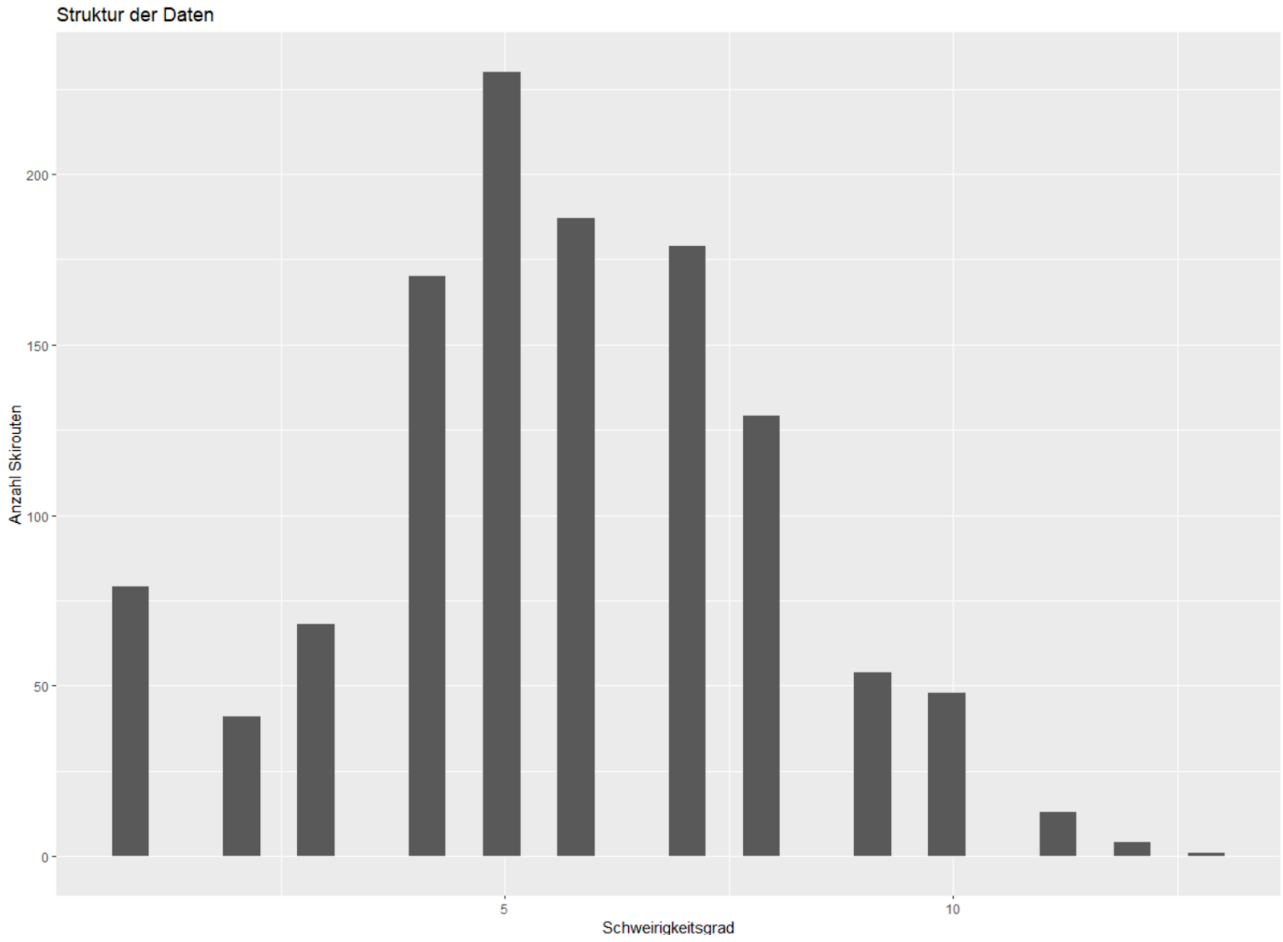


Abbildung 8 Histogramm über die Verteilung der Schwierigkeitsgrade

Dies zeigt sehr deutlich eines der Hauptprobleme. Die Datenmenge an sich ist bereits gering, doch zeichnet sich hier eine weitere Herausforderung ab. Die Daten sind sehr unterschiedlich verteilt. Die Schwierigkeitsgrade 4 - 8 sind ausreichend vertreten, doch für die anderen sind wenig bis zu bloss einem einzigen Eintrag auf der Stufe 13 vorhanden.

4.2 Slope

Die Steigung ist einer der Schlüsselwerte und daher auch interessant für eine Analyse bezüglich eines Zusammenhangs zwischen den Werten der Steigung und dem Schwierigkeitsgrad.

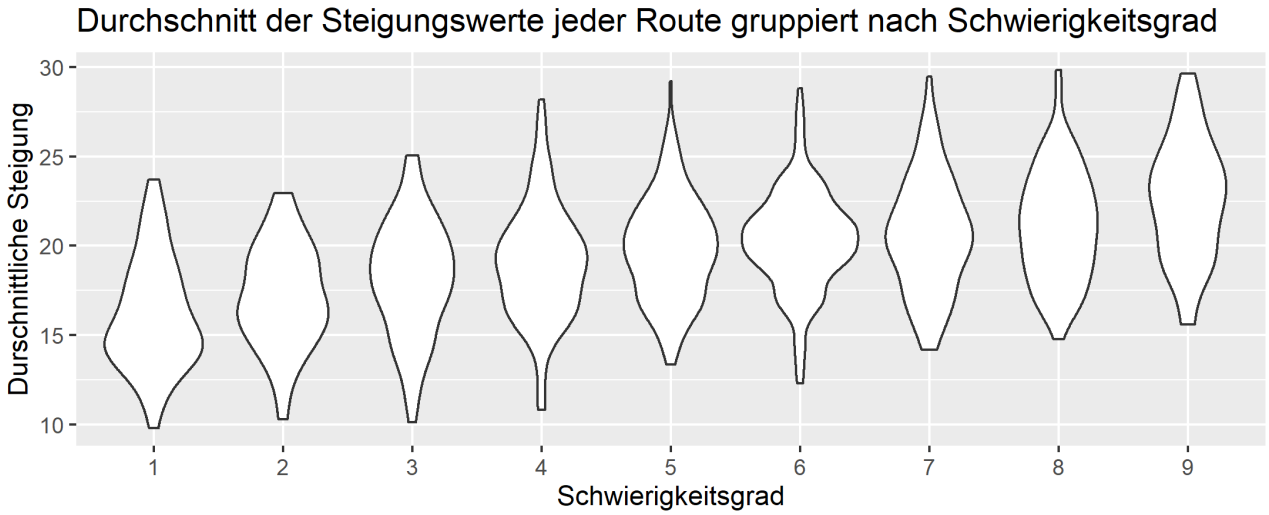


Abbildung 9 Diagramm der durchschnittlichen Steigung pro Route gruppiert nach Schwierigkeitsgrad

Von den Schwierigkeiten 1-10 ist eine klare Verschiebung der Slope Werte nach oben ersichtlich. Für die darüberliegenden Schwierigkeitsgraden ist die Menge der Daten nicht ausreichend, um ein Muster zu erkennen.

Die Verschiebung ist jedoch äusserst gering, sodass weitere unterstützende Features nötig sind oder die Daten weiter gefiltert werden müssen.

4.3 Corridor Width

Die Wegbreite ist ebenfalls eines der Schlüsselattribute. Auch hier sieht die Übersicht ähnlich aus wie davor:

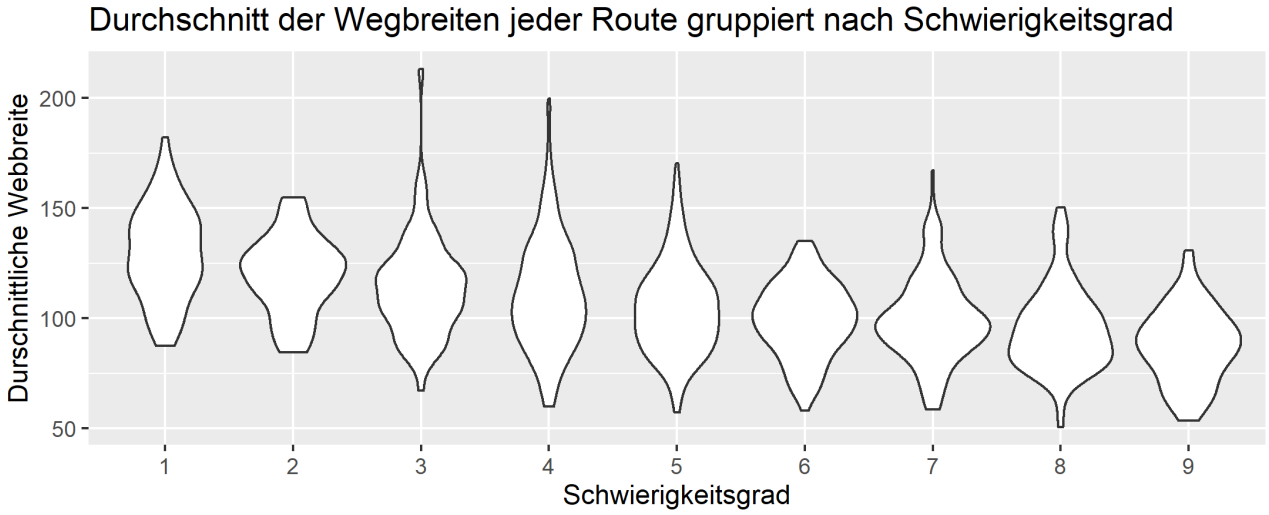


Abbildung 10 Diagramm der durchschnittlichen Wegbreite pro Route gruppiert nach Schwierigkeitsgrad

Es ist eine Verschiebung ersichtlich. Je schmaler die Route, desto schwieriger. Doch auch hier ist die Abstufung äusserst gering und die Streuung enorm.

4.4 Planc

Die Ausnahme der Regel bildet der Planc Wert. Während die anderen Features eine Tendenz nach unten oder oben aufweisen mit steigendem Schwierigkeitsgrad, ist eine solche Tendenz beim Planc Wert nicht ersichtlich.

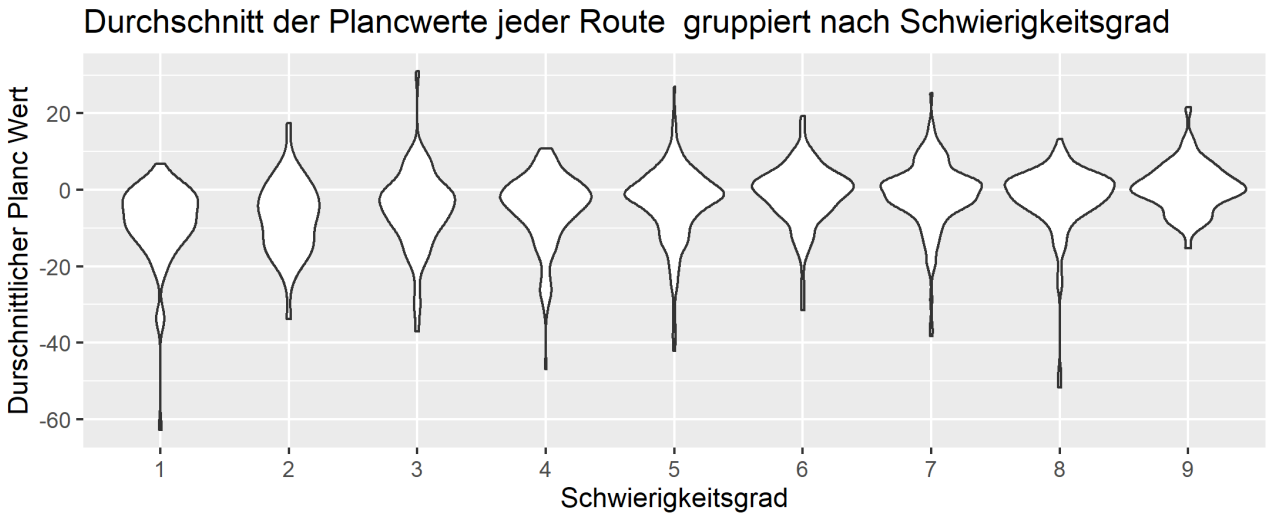


Abbildung 11 Diagramm der durchschnittlichen Planc Werte pro Route gruppiert nach Schwierigkeitsgrad

Der Schwierigkeitsgrad 3 sieht von der Verteilung ähnlich aus, wie die 9. Und auch wenn die 10 deutlich höher ist als die davor, ist keine klare Tendenz erkennbar.

4.5 Länge

Die Information zur Länge einer Route ist nicht direkt als Information gespeichert, ergibt sich jedoch aus der Anzahl der Punkte * 10 [Meter], da diese immer einen 10 Meter Abstand aufweisen.

Ziel der Untersuchung war die Erkundung eines möglichen Zusammenhangs zwischen der Schwierigkeit und der Länge einer Route. Dazu wurde die Längen aller Routen berechnet und entsprechend ihres zugewiesenen Schwierigkeitsgrades geplottet.

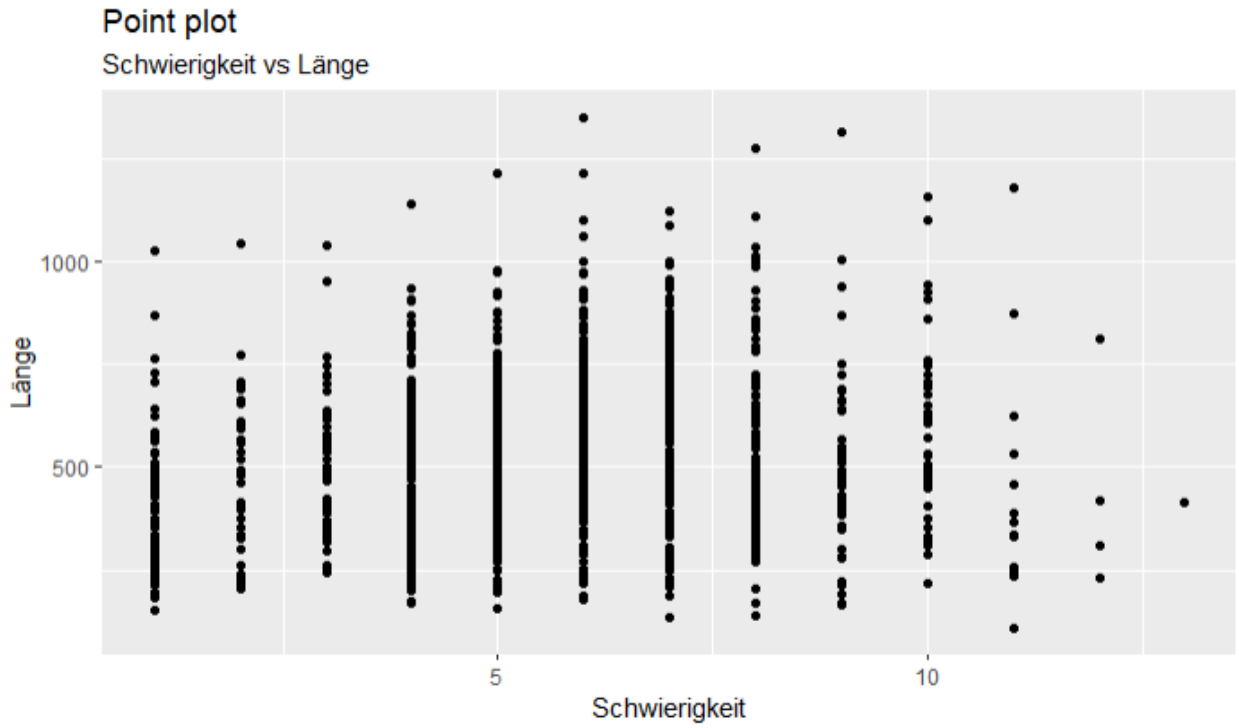


Abbildung 12 Diagramm Schwierigkeit im Verhältnis zur Länge

Dieser Plot zeigt deutlich, dass kein erkennbarer Zusammenhang zwischen der Länge einer Route und deren Schwierigkeitsgrad existiert. So sind die sehr schweren Routen im Schnitt sogar weniger lang, als diejenigen mit einem mittleren Schwierigkeitsgrad.

Dies deckt sich mit der Aussage von Experten, nach welchen die Länge kein Kriterium ist, nach welchem die Routen bewertet werden.

Daher wird die Länge für die Modell vernachlässigt.

4.6 Lineare Charakteristika

Zur weiteren Analyse wurden Durchschnittswerte der einzelnen Features genommen und nach Schwierigkeitsgrad gruppiert. Ziel dabei ist die Erkennung von Eigenschaften der einzelnen Schwierigkeitsgrade. Rein intuitiv würde man damit rechnen, dass eine Route schwieriger eingestuft wird, je steiler, gefährlicher und unwegsamer sie ist. Dies deckt sich mit dem Bewertungsschema des SAC.

Dabei wurden die Schwierigkeitsgrade 11, 12 und 13 nicht berücksichtigt, da die geringe Menge an Daten hierbei kein aussagekräftiges Resultat zulassen.

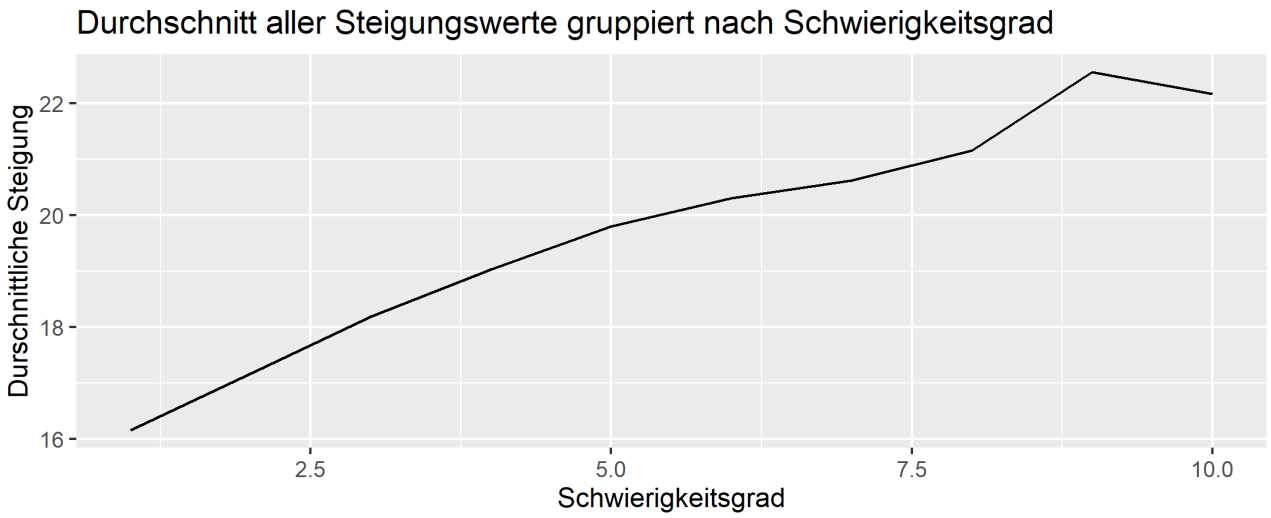


Abbildung 13 Diagramm Durschnitt aller Steigungswerte gruppiert nach Schwierigkeitsgrad

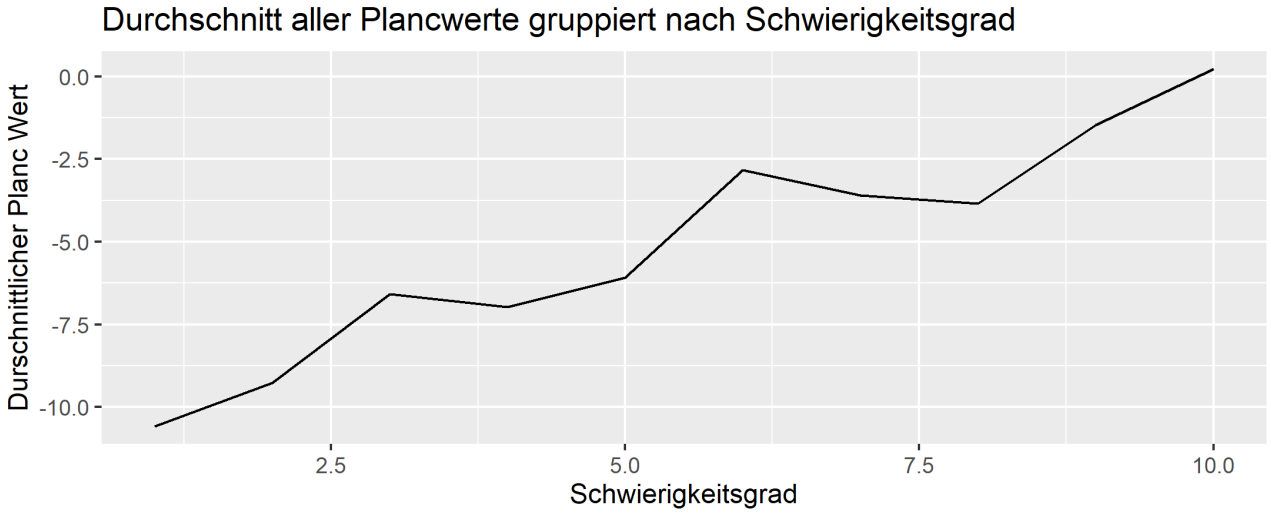


Abbildung 14 Diagramm Durschnitt aller Plancwerte gruppiert nach Schwierigkeitsgrad

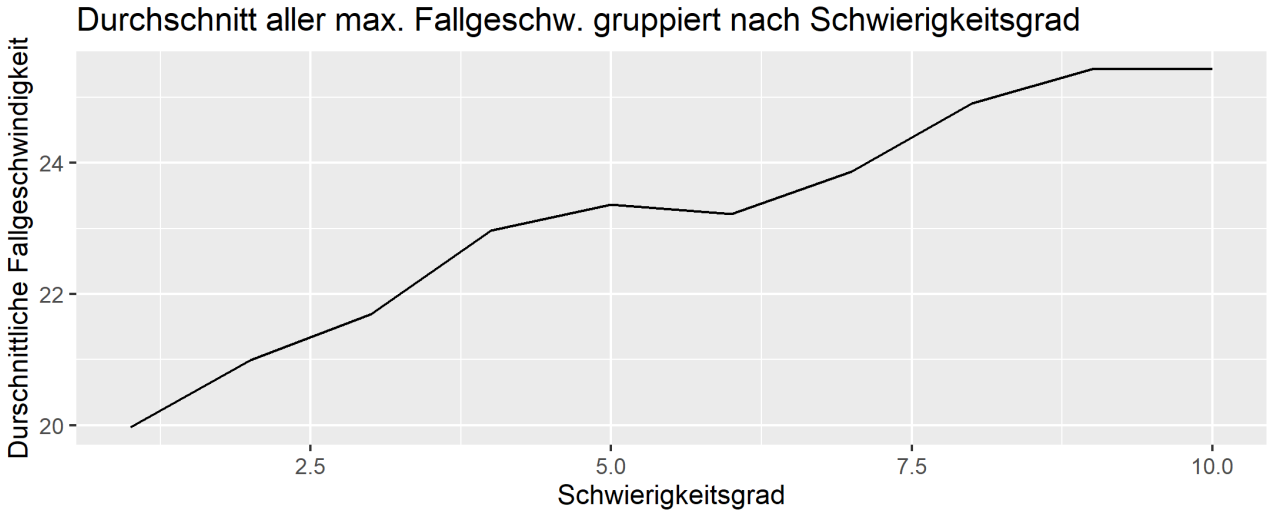


Abbildung 15 Diagramm Durschnitt aller max. Fallgeschwindigkeiten gruppiert nach Schwierigkeitsgrad

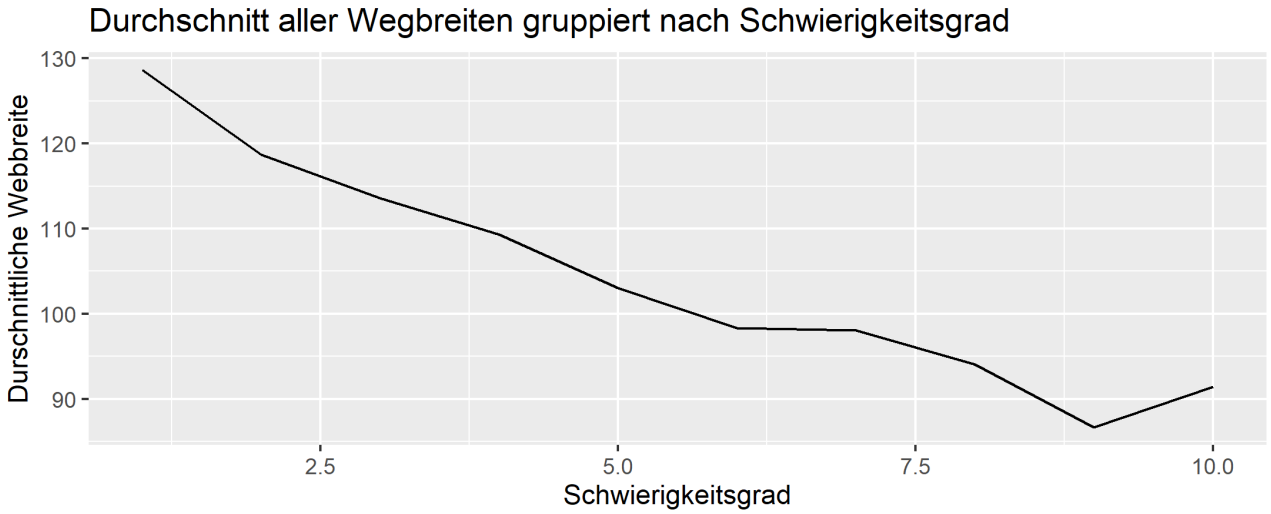


Abbildung 16 Diagramm Durschnitt aller Wegbreiten gruppiert nach Schwierigkeitsgrad

Wie in Abbildung 13, Abbildung 14, Abbildung 15 und Abbildung 16 erkennbar ist, gibt es einen linearen Zusammenhang zwischen den Features und dem Schwierigkeitsgrad. Dieser unterliegt zwar leichten Schwankungen, ist jedoch trotzdem signifikant.

5 Datenaufbereitung

Das Kapitel der Datenaufbereitung befasst sich mit dem Prozess die Daten in ein gewünschtes Format zu verpacken. Die Beinhaltet die Bereinigung von Nullwerten, sowie die Anwendung mathematischer Funktionen auf die Punktreihen. Darüber hinaus wird der Filter der Fusspassagen spezifiziert. Abschliessend wird die Data Augmentation Methoden angeschaut und die Aufteilung der Daten in verschiedene Sets, auch Sampling genannt.

5.1 Rohdaten

In einem ersten Schritt wurden die Rohdaten angepasst. Diese beinhalten Extremwerte, welche als Null Werte zu interpretieren sind.

-9.99999 Werte sind als «Not a Number» zu interpretieren.

Diese Werte treten bei den vier Geschwindigkeitsangaben auf und wurden daher an diesen Stellen durch NA ersetzt, wodurch die Diagramme nicht länger verzerrt werden. Für die Algorithmen im späteren Verlauf wurden selbst die NA Werte entfernt.

```
fd_max_a <- na_if(fd_max_a, -9.999000e+03)

fd_max_v <- na_if(fd_max_v, -9.999000e+03)

fd_sum_a <- na_if(fd_sum_a, -9.999000e+03)

fd_sum_v <- na_if(fd_sum_v, -9.999000e+03)
```

5.2 Feature Erweiterung mit TS Fresh

Die Daten aller Punkte einer Route entsprechen den Eigenschaften einer Timeseries. Die X-Achse ist hierbei jedoch nicht mit der Zeit, sondern dem Raum verbunden. Diese Time Series erlaubt die Extraktion von weiteren Eigenschaften.

Über das TS Fresh Paket (Time Series Feature Extraction) können zahlreiche statistische Merkmale der Daten extrahiert werden. Dabei sind generelle Werte enthalten wie z.B. der Durchschnitt der Steigungswerte, aber auch erweiterte Werte wie z.B. die Varianz, Interquartile, Range und Standardabweichung.

Damit wird aus vielen einzelnen Punkten jeweils ein neues Feature erzeugt, welches den Charakter der Route repräsentiert. Dies hilft im späteren Verlauf beim Trainieren der Modelle. Dafür werden die einzelnen Punkte dann nicht länger verwendet.

Das Resultat ist also ein «flaches» Feature. Auf diese Weise wurden 658 neue Features hinzugefügt.

5.3 Correlation Matrix

Um den Zusammenhang der verschiedenen Features darzustellen wird die mathematische Funktion der Korrelation verwendet. Diese ist in R wie folgt definiert:

$$r = Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Ziel dieser Funktion ist es eine lineare Abhängigkeit von zwei Variablen zu ermitteln. Dies äussert sich mit einem Outputwert, welcher sich von 1 bis -1 bewegen kann. Dabei steht die 1 für eine starke Abhängigkeit des Features zur Vorhersage. Die -1 hingegen symbolisiert praktisch eine negative Abhängigkeit. Der Null Wert ist als Abwesenheit von Abhängigkeit der Werte zu

verstehen. Ziel der Methode war es die Qualität der generierten Features zu validieren und allenfalls noch weitere Analysen vorzunehmen. Dazu werden von allem Feature der Korrelation Koeffizient zum Schwierigkeitsgrad berechnet. Je näher der Wert an der 1 respektive der -1 ist desto stärker ist der Einfluss auf den Schwierigkeitsgrad. Trotz dieser mathematischen Abhängigkeit müssen die Daten anschliessend noch von Hand validiert werden. Da Anhand des Wertes keine Aussage über die Verteilung der Daten gemacht werden kann. Das eine manuelle Analyse notwendig ist, lässt sich mit Abbildung 17 erläutern.

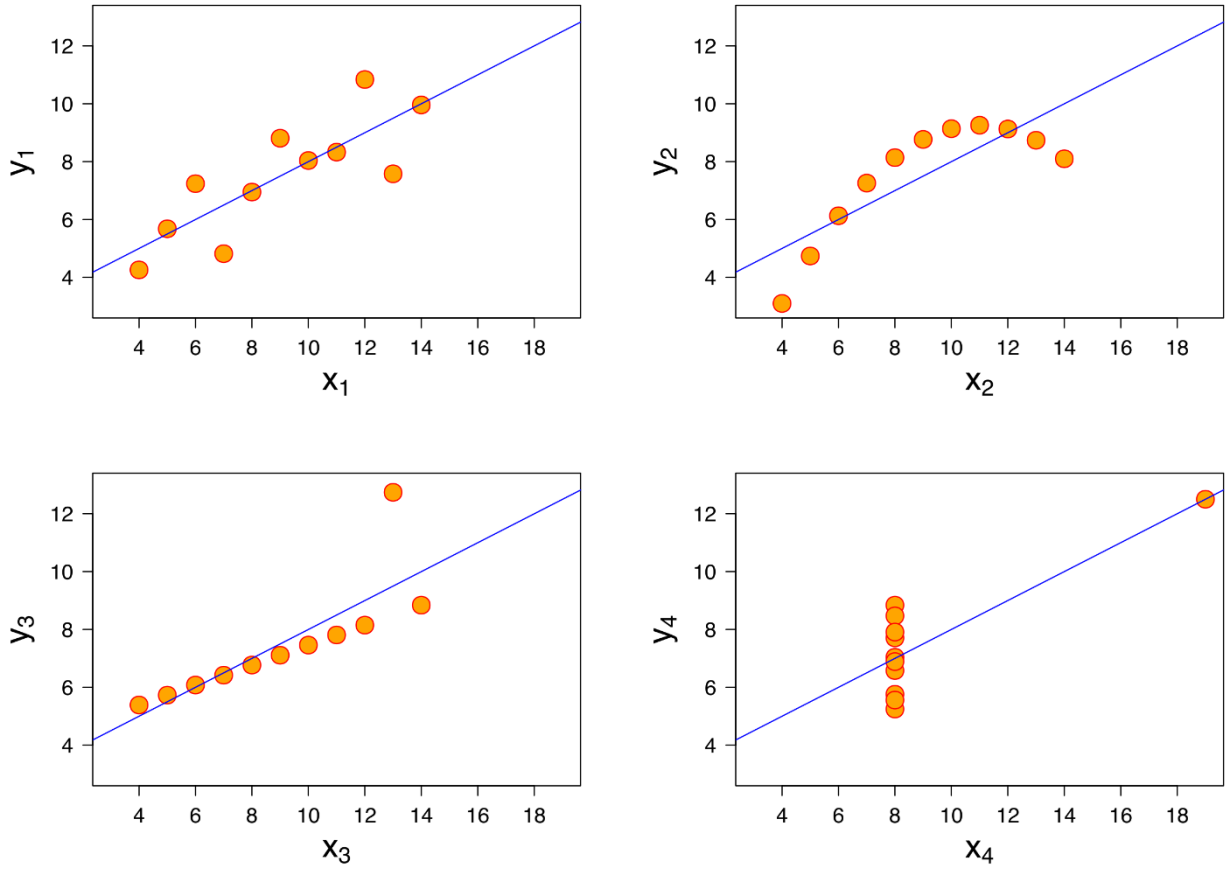


Abbildung 17 Korrelation Koeffizienten Beispiel [9]

In Abbildung 17 sind jeweils vier Berechnung des Korrelationskoeffizient visuell dargestellt. Dabei symbolisiert die blaue Linie den Koeffizienten. Alle vier Grafen haben dabei als Resultat denselben Korrelationskoeffizienten von 0.816. Diese Berechnung berücksichtigt die Verteilung der Daten nur teilweise. Wie an der Abbildung gut zu erkennen ist, können die Daten unterschiedlich verteilt sein. Diese Verteilungen lassen aber unterschiedliche Schlussfolgerungen über die Daten zu. Daher ist eine manuelle Sichtung der wichtigsten Features essenziell. Die Berechnungsmethode geht immer von einer Normalverteilung aus. Daraus lässt sich schliessen das eine Verifizierung durch die visuelle Sichtung der Features mit den grössten Korrelation Koeffizienten notwendig ist. [8]

Mit den erweiterten Features können 320356 verschiedene Kombinationen von Features Paare gebildet werden. Anschliessend wurden die Kombinationen auf den Schwierigkeitsgrad reduziert. Somit wurden die Beobachtungen auf 1131 beschränkt. Die Werte bewegten sich zwischen 0.62 und -0.50 um die Relevanz der Werte besser aussortieren zu können wurde ein weiterer Wert bei der Beobachtung verwendet die statistische Signifikanz.

Die statistische Signifikanz gibt Auskunft drüber, ob eine Aufgestellte These oder Risiko eintrifft. In diesem Fall wäre die These: «Die Daten haben eine Beziehung zu einander» die Antithese dazu wäre «Die Daten haben keine Beziehung zu einander». Normalerweise gilt ein Resultat als statistisch relevant sobald der p-value unter dem festgelegten Alpha ist. Dies ist ein Schwellwert, welcher in den meisten Studien mit 0.05 initialisiert wird. Alpha ist als ein Schwellwert zu interpretieren. Sobald dieser überschritten wird gilt das Resultat als nicht mehr relevant. Der «p-value» gibt also Auskunft über die Wahrscheinlichkeit, dass der berechnete Korrelation Koeffizient auf einen Datensatz nicht zutrifft. Sind Beispielweise die Slopewerte einer Route hoch, so steigt auch der Schwierigkeitsgrad entsprechend. Dies lässt sich mit logischen Argumenten erklären, da die körperlichen Anforderungen an den Sportler steigen und ist somit nachvollziehbar. Anbei die Tabelle mit den Korrelation Koeffizienten.[10]

Feature X	**Feature Y**	**Korrelation Koeffizient**	**p-value**
diff	diff	1.0000000	*NA*
diff	f.quantile_90_slope	0.6282782	0.000000e+00
diff	f.quantile_95_slope	0.6237922	0.000000e+00
diff	f.quantile_95_fd_maxv	0.5648108	0.000000e+00
diff	f.quantile_90_fd_maxv	0.5567591	0.000000e+00
diff	f.quantile_80_slope	0.5527161	0.000000e+00
diff	f.quantile_75_slope	0.5234741	0.000000e+00
diff	f.max_slope	0.4963837	0.000000e+00
diff	f.sd_fd_sumv	0.4958817	0.000000e+00
diff	f.quantile_70_slope	0.4933942	0.000000e+00
diff	f.sum_fd_sumv	0.4887035	0.000000e+00
diff	f.quantile_95_fd_sumv	0.4878558	0.000000e+00
diff	f.quantile_80_fd_maxv	0.4730303	0.000000e+00
diff	f.sd_slope	0.4699055	0.000000e+00
diff	f.quantile_90_fd_sumv	0.4647083	0.000000e+00
diff	f.mean_fd_sumv	0.4643773	0.000000e+00
diff	f.sum_fd_sumna	0.4627264	0.000000e+00
diff	f.max_fd_maxv	0.4600795	0.000000e+00
diff	f.sd_fd_maxv	0.4588615	0.000000e+00
diff	f.mean_fd_sumna	0.4568775	0.000000e+00
diff	f.mean_slope	0.4528767	0.000000e+00

diff	f.range_slope	0.4528619	0.000000e+00
diff	f.max_fd_sumv	0.4484423	0.000000e+00
diff	f.range_fd_sumv	0.4479695	0.000000e+00
diff	f.variance_slope	0.4477761	0.000000e+00
diff	f.quantile_60_slope	0.4396838	0.000000e+00
diff	f.quantile_95_fd_maxna	0.4379071	0.000000e+00
diff	f.variance_fd_maxv	0.4369403	0.000000e+00
diff	f.mean_fd_maxna	0.4359561	0.000000e+00
diff	f.range_fd_maxv	0.4323014	0.000000e+00
diff	f.quantile_75_fd_maxv	0.4273503	0.000000e+00
diff	f.coeff.variation_CW	0.4269503	0.000000e+00
diff	f.quantile_90_fd_maxna	0.4260273	0.000000e+00

Tabelle 12 Übersicht zu den Features mit den besten Korrelationskoeffizienten

In nachfolgender Abbildung 18 ist eine klar lineare Abhängigkeit zwischen dem Feature «f.quantile_95_slope und dem Schwierigkeitsgrad ersichtlich.

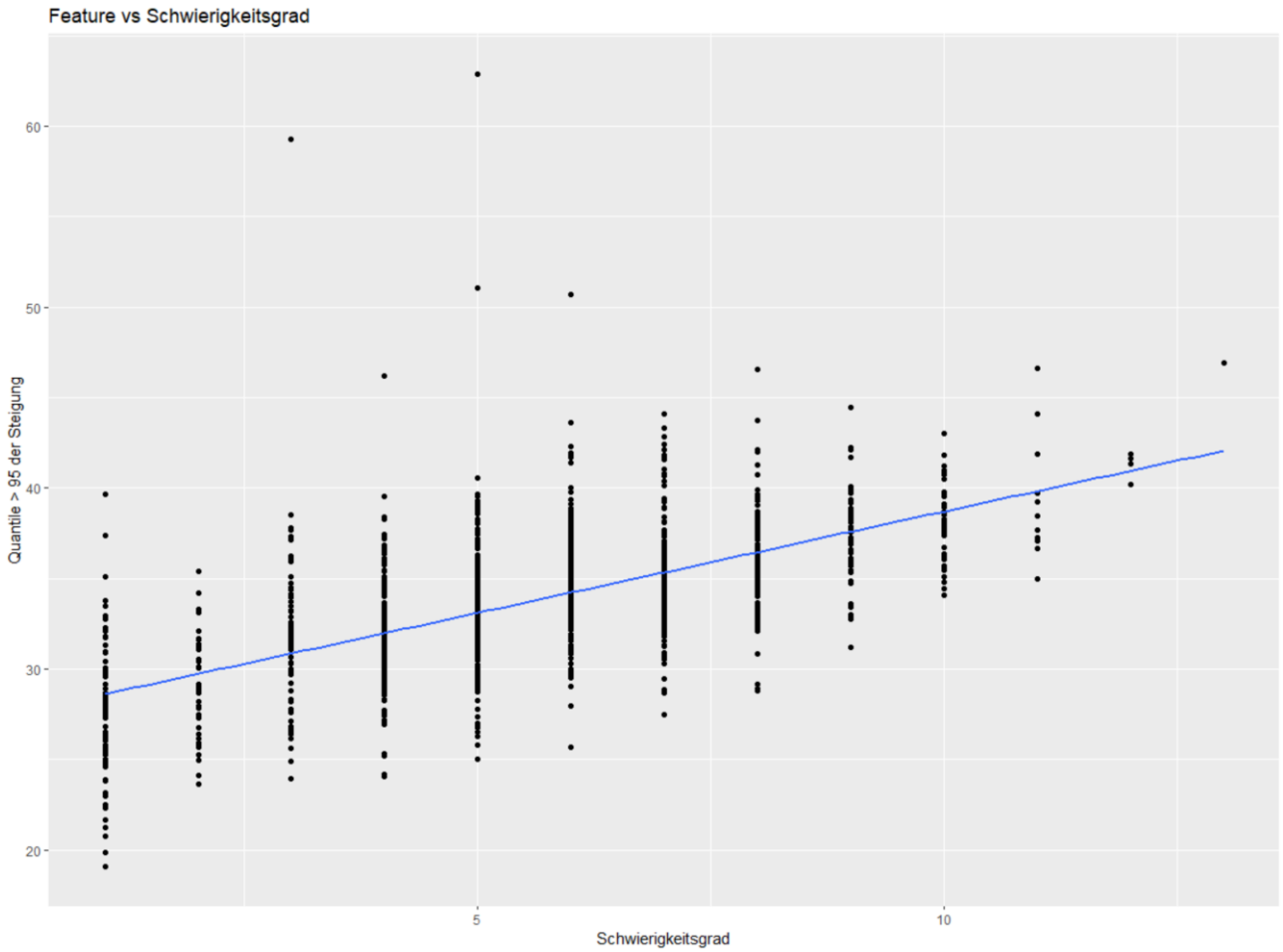


Abbildung 18 Diagramm zu der Korrelation

Die Abbildung ist wie folgt zu interpretieren, jeder Punkt repräsentiert eine Route und deren Quantile, welche die Daten in zwei Hälften teilt, wobei eine dieser Hälften mehr als 95% der Daten beinhaltet. Es ist also eine klare Steigung zu erkennen. Die blaue Linie symbolisiert den Korrelation Koeffizienten.

5.4 Data Augumentation

Aus Kapitel 4.1 ist zu entnehmen, dass sich die Daten sich in einem «imbalanced» Zustand befinden. Dies löst einen Nachteil bei der Beurteilung der Routen aus, da dem Modell eine Tendenz zu den mehrheitlich vertretenen Schwierigkeitsgraden eintrainiert wird. Data Augumentation beinhaltet die Methoden mit welcher solche Datenset durch «künstliche Generierung von Daten» ausbalanciert werden können. [11]

Hierbei gibt es zwei Möglichkeiten man könnte den dominanten Schwierigkeitsgrade verkleinern in dem erfasste Routen beim Training weglassen werden. Diese Methode wird auch «under sampling» genannt. Ein grosser Nachteil dieser Methode ist das die Informationen, welche in den weggestrichenen Datensätzen enthalten sind nicht ins Modell einfliessen. Angesicht der kleinen Daten Menge von 1203 Routen müsste man gemäss Abbildung 8 ungefähr 300 Routen wegstreichen. Dies entspricht einem Viertel der Daten. Es würden also wichtige Erkenntnisse und Informationen verloren gehen.[12]

Beipsiel Undersampling:

	Vorher	**Nacher**
Positive Fälle	**90 (90%)**	**20 (66%)**
Negative Fälle	**10 (10%)**	**10 (33%)**
Gesamt Anzahl Daten	**100**	**30**

Tabelle 13 Beispiel Undersampling

Eine andere Methode ist, bei welcher kein Datenverlust auftritt ist ,das «over sappling» hierfür werden künstliche neue Daten in diesem Fall Routen generiert. Die einfachste Form davon ist die Anpassung von bestehenden Datensätzen mit einer Zufallszahl Operation. Als Beispiel könnte eine Addition oder Subtraktion einer Zufallszahl für den Wert des kopierten Datensatzes genommen werden. [12]

Beispiel Oversampling:

	Vorher	**Nacher**
Positive Fälle	**90 (90%)**	**90 (50%)**
Negative Fälle	**10 (10%)**	**90 (50%)**
Gesamt Anzahl Daten	**100**	**180**

Tabelle 14 Beispiel Oversampling

Bei «SMOTE» handelt es sich um einen hybride Variante dieser obenerwähnten Methoden. Es wurden also Klassen teilweise reduziert und künstlich erstellt. Dabei werden die küntlichen Daten nicht rein Zufällig erstellt. Mit Hilfe der k gwählten nächsten Nachbarpunkte wird eine Verbindung erstellt. Die nachher generierten Punkte werden sich auf dieser Linie befinden. Dies hat den Vorteile das die Daten sich in der bisherigen Abweichungen bewegung und so für eine Verdicktung des Bildes sorgen. In Abbildung 19 wurde dieser Prozess visualisiert um den Prozess zu veranschaulichen.[11]

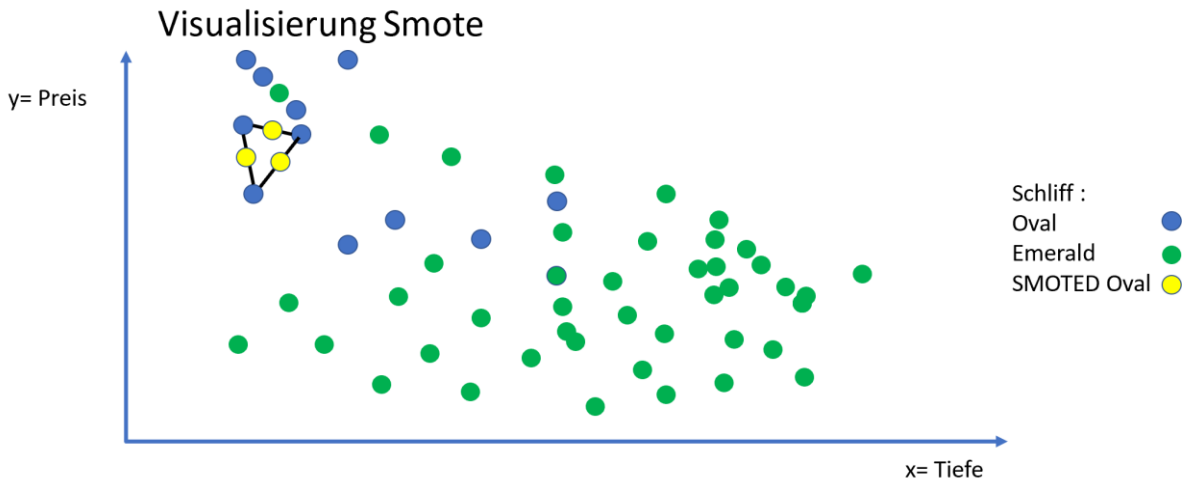


Abbildung 19 Visualisierung Smote

Für die Grafik wurden Daten von verschiedenen Diamenten verwendet. Es ist klar ersichtlich das die Oval-Schliffart unterrepräsentiert ist. Nun werden mit Smote weitere gelbe Datenpunkte generiert. Diese Punkte werden aber die bisher abgesteckt Fläche der Punkte nicht verlassen. Sie werden also die vorhandenen Daten künstlich verdichten, um den Schwellenwert für eine beispielsweise lineare Vorhersage nach linksoben zu korrigieren.

In diesem Projekt wurden die unteren Schwierigkeitsgrade durch neu generierte Daten ergänzt und beispielsweise die Daten des Schwierigkeitsgrad 5 reduziert. Dies resultiert in einer neuen Datenverteilung, welche deutlich ausgeglichener ist, wie in Abbildung 20 ersichtlich.

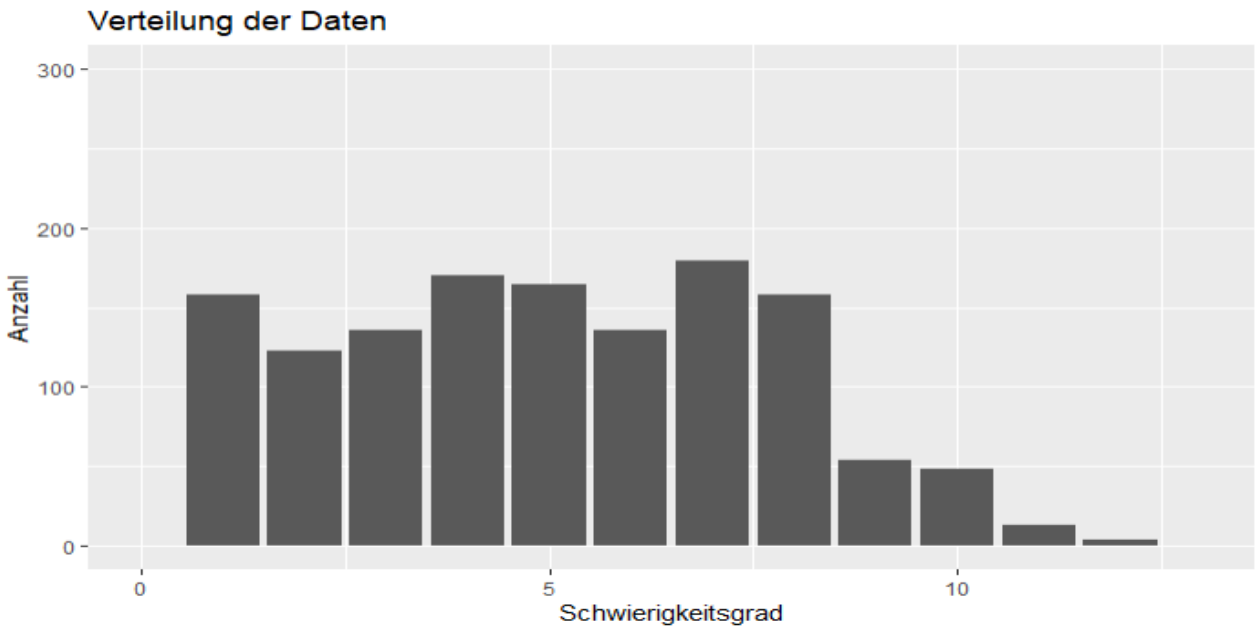


Abbildung 20 Diagramm zur Verteilung der Daten nach dem Schritt der Data Augmentation

Diese Erweiterung kann jedoch nicht beliebig oft vorgenommen werden, da durch zu viele künstliche Daten die Genauigkeit des Modells verschlechtert wird. Das betrifft hierbei die Schwierigkeitsgrade 9, 10, 11, 12 und 13.

5.5 Fusspassagen trennen

Es kann vorkommen, dass gewisse Passagen der Route zu Fuss begangen werden. Dies kann nach Aussage von Experten vorkommen, wenn der Slope (Steigungswert) mehr als 40 Grad **oder** der Planc Wert weniger als -50 annimmt.

Dabei gibt es zwei verschiedenen Arten von Fusspassagen.

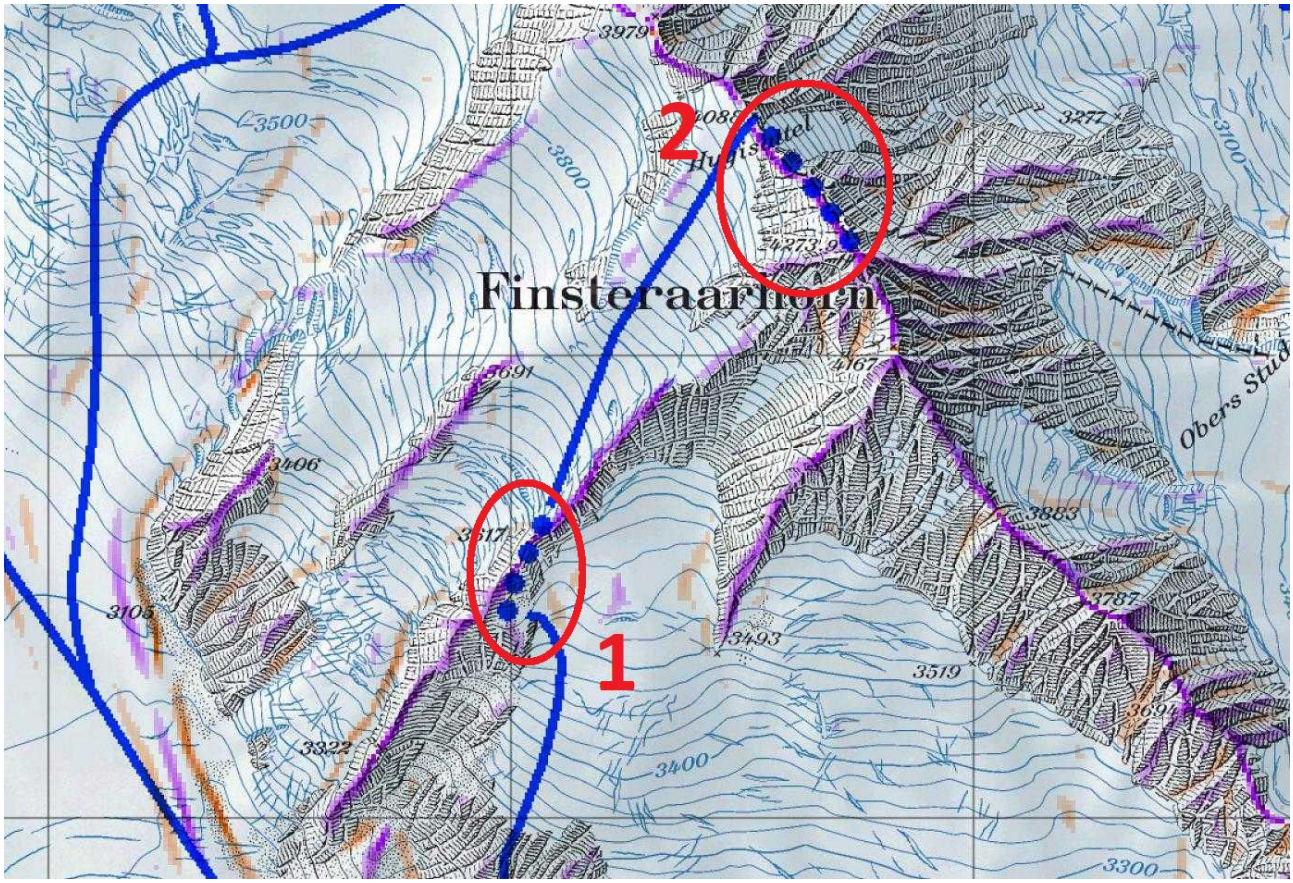


Abbildung 21 Beispiel einer Fusspassage

1. Obligatorische Fusspassagen

Hierbei handelt es sich um Abschnitte, welche inmitten der Route liegen, und passiert werden müssen, um das Ziel zu erreichen.

Diese werden für die Einstufung der Schwierigkeit beim SAC berücksichtigt.

2. Optionale Fusspassagen

Diese Abschnitte sind nicht zwingend nötig, da sie nur noch auf den letzten Abschnitt des Gipfels führen. Man könnte die Wanderung auch am letzten Punkt vor der Fusspassage beenden und bereits von da mit den Ski hinunterfahren.

5.5.1 *Vorgehen*

Diese Abschnitte werden beim SAC ignoriert für die Bewertung der Route und müssen daher auch bei unseren Daten entfernt werden, da sonst die extreme Steigung und Planc Werte das Ergebnis beeinflussen können.

Entfernen und nicht «abflachen», aus dem Grund, da die Länge der Route keinen Einfluss auf den Schwierigkeitsgrad hat. (Siehe Statistische Auswertung - Länge)

Leider gibt es keine genaue Definition, ab wann ein Abschnitt eine Fusspassage ist und die Punkte sind auch nicht als solche gekennzeichnet. Daher mussten Fusspassagen manuell auf der Karte untersucht werden, um Schwellwerte zu finden.

Slope > 40° ***oder*** *Planc < -50*

Ausserdem müssen die Abschnitte geglättet werden, um ganze Passagen und nicht bloss einzelne Punkte zu finden. Dies wird durch ein Sampling von je 50 Metern erreicht. Innerhalb dieser 50 Meter wird der Schnitt der Steigung und der Plancwerte angeschaut. Sind diese über den Schwellenwerte, werden die 50 Meter davor angeschaut usw. bis beide Werte unterhalb der festgelegten Grenze fallen. Damit wird der Punkt ermittelt, ab welchem die Punkte als Fusspassage deklariert werden und damit entfernt werden.

Nun wird beim Gipfel geschaut, ob diese Werte überschritten werden und falls ja, wird nach unten alles abgetrennt, bis die Werte wieder unter den Schwellenwert fallen.

Ein gutes Beispiel hierfür ist die Route mit der ID 451 vom Gemmipass zum Daubenhorn.

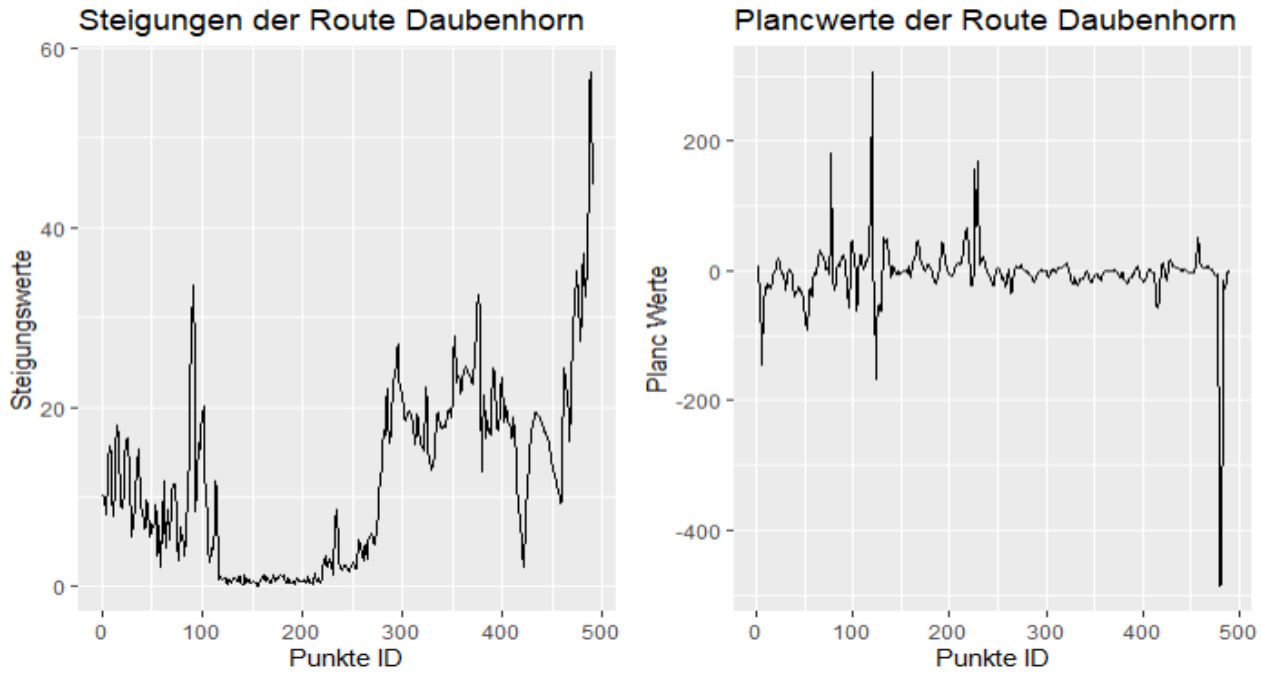


Abbildung 22 Fusspassagen in den Daten

Diese Route erhielt vom SAC einen Schwierigkeitsgrad von 3, welcher Steigungen bis zu +/-35 Grad erlaubt. Durch die Fusspassage am Ende erreicht die Route jedoch Steigungen bis zu 55 Grad und einen Planc Wert von unter -450.

Dazu nachfolgend die letzten 15 Punkte der Route und mit rot die entsprechenden Grenzwerte eingezeichnet:

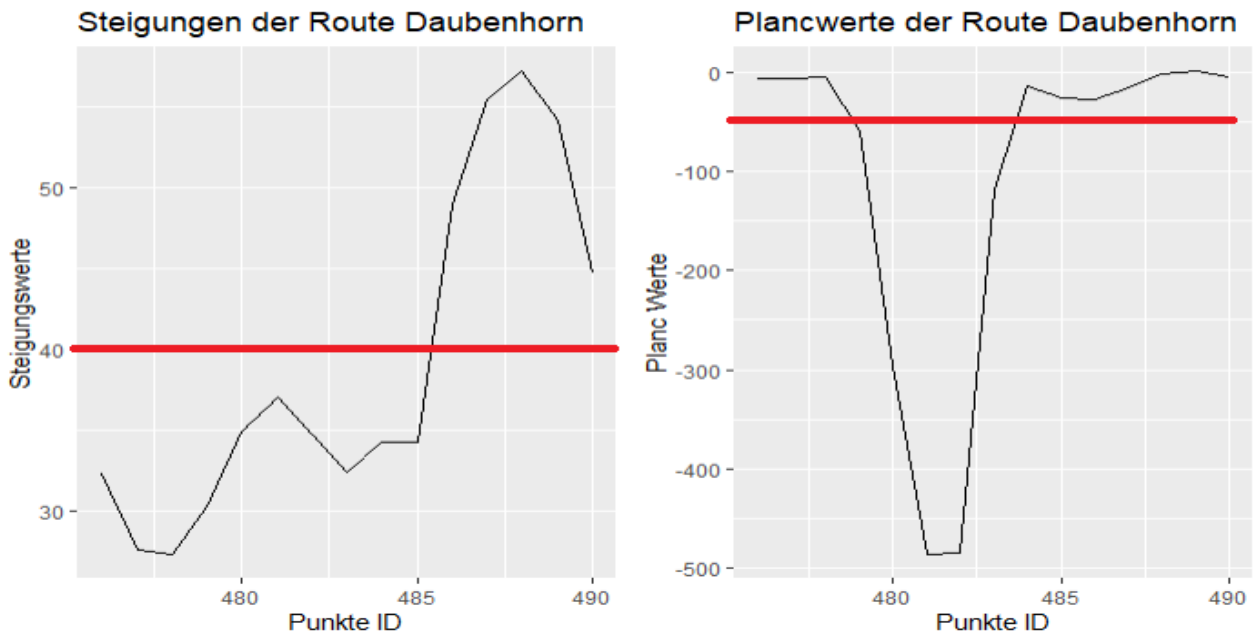


Abbildung 23 Abschnitt der Route mit ID 451

Daraus ergibt sich:

- von Punkt ~485 - 490 ist eine steigungsbedingte Fusspassage.
- von Punkt ~478 - 483 ist eine Planc bedingte Fusspassage.

Der Algorithmus macht daraus eine Fusspassage von 475 - 490. Diese wird entfernt und die Route erhält ein entsprechend realistisches Bild für eine Route mit dem Schwierigkeitsgrad 3:

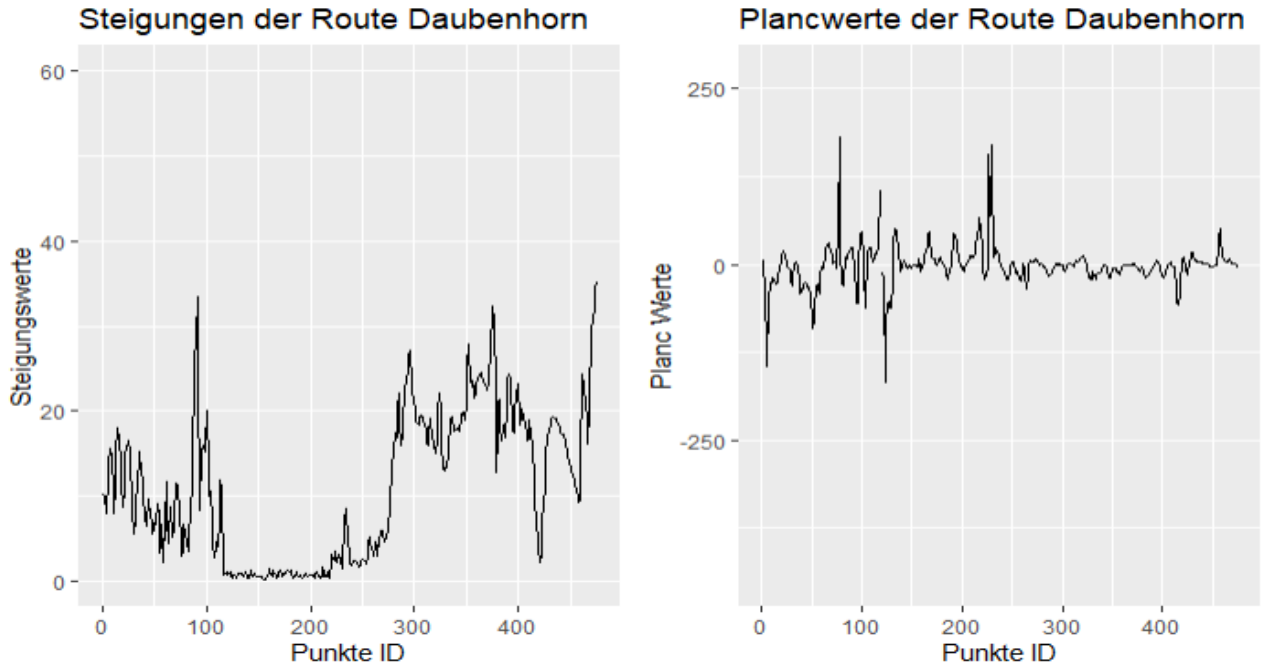


Abbildung 24 Output nach der Bearbeitung

Da besonders die Features, welche den oberen Quantilen der Steigung repräsentieren eine hohe Korrelation aufweisen und damit grossen Einfluss auf die Vorhersage haben, profitiert ein Modell von dieser Bereinigung.

In Abbildung 25 ist dieses Feature aufgeführt, vor der Entfernung der Fusspassagen.

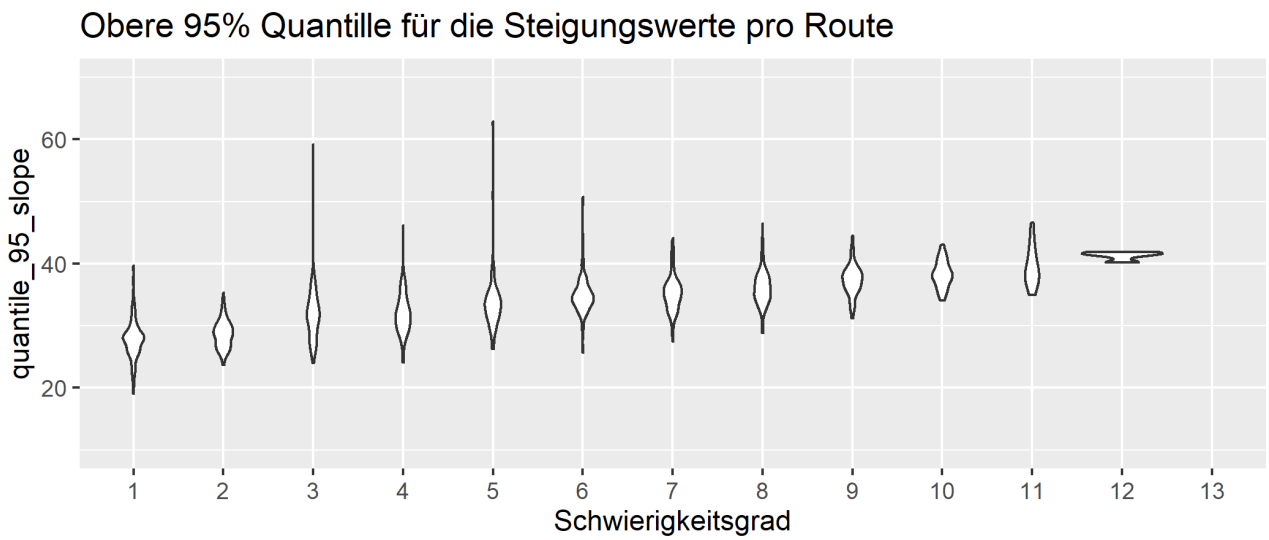


Abbildung 25 Diagramm der 95% Quantile vor der Entfernung der Fusspassagen

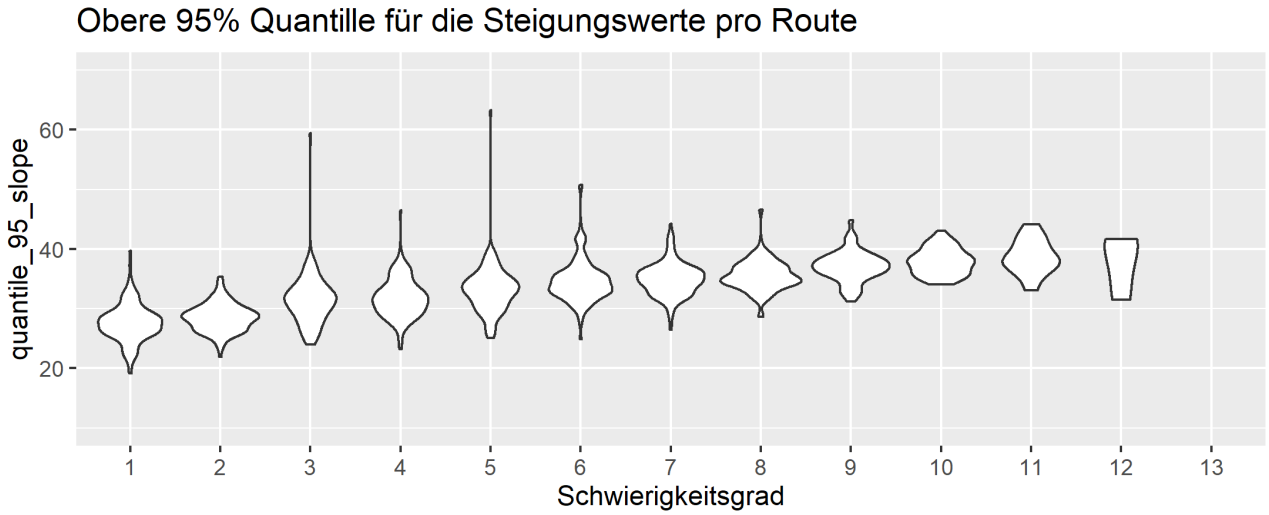


Abbildung 26 Diagramm der 95% Quantile nach der Entfernung der Fusspassagen

Abbildung 26 zeigt das Feature nach der Entfernung der Fusspassagen. Durch diese Massnahme wurde das Feature dichter, wovon zukünftige Modelle profitieren. Es sind immer noch Ausreisser vorhanden, dies daher, da in Mitten der Route immer noch Fusspassagen vorkommen können, die jedoch nach SAC in die Bewertung der Schwierigkeit mit einfliessen und damit auch nicht entfernt werden.

5.5.2 *Resultatsübersicht*

Die Anzahl der behandelten Routen ähnelt der Gesamtübersicht der Anzahl Routen. Es wurden 963 der insgesamt 1203 Routen angepasst.

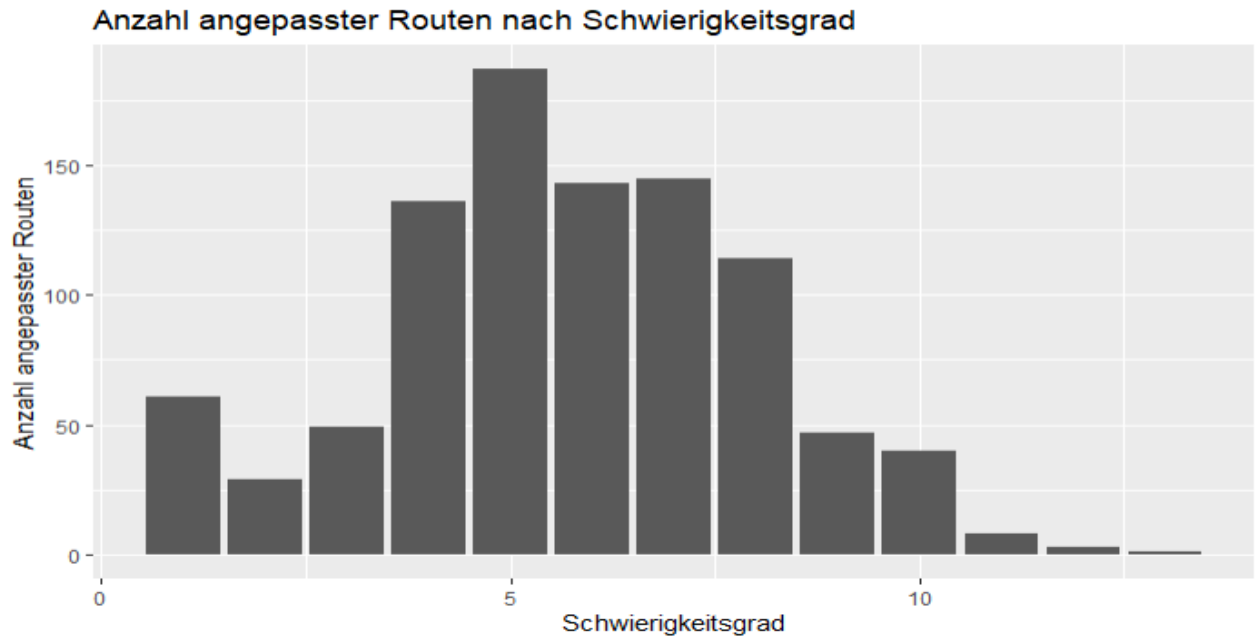


Abbildung 27 Diagramm angepasste Routen

Davon wurde jedoch nur bei einem geringeren Teil mehr, als bloss die letzten 5 Meter angepasst.

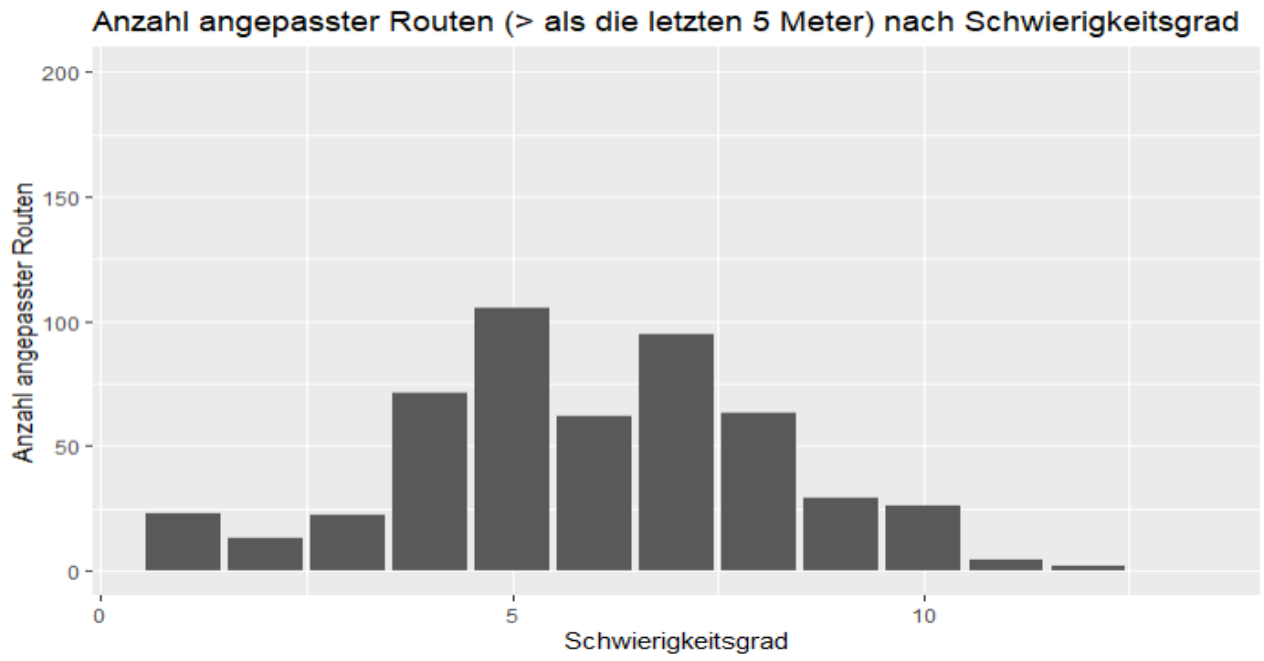


Abbildung 28 Diagramm angepasste Routen >5m

Insgesamt wurden bei 515 Routen mehr als die letzten 5 Meter der Route angepasst. Diese Werte dienen als wichtige Grundlage für spätere Analysen zum Einfluss der Fusspassagen auf die Genauigkeit der Vorhersagen.

5.6 Klassenreduktion

Da eines der Hauptprobleme die geringe Datenmenge ist, was besonders bei den unteren und ganz besonders bei den oberen Klassen durch die ungleichmässige Verteilung noch gravierender wird, wurden die Klassen reduziert.

Dabei wurden zwei Grundideen umgesetzt.

5.6.1 *Abschneiden von Randklassen*

Hierbei werden alle Klassen, die unter oder über einer festgelegten Schwelle liegen, schlichtweg entfernt. Damit bleiben bloss noch die Klassen übrig, welche über eine ausreichende Datenmenge verfügen. Wobei ausreichend hier relativ ist.

5.6.2 *Zusammenfassen von Randklassen*

Hierbei werden die Klassen unter oder über dem festgelegten Schwellenwert gruppiert und zu einer Klasse zusammengefasst. Dadurch sind zwei Klassen mehr möglich als durch das abtrennen, jedoch sind damit auch grössere Unterschiede in den einzelnen Routen der Randklassen vorhanden.

Übersicht

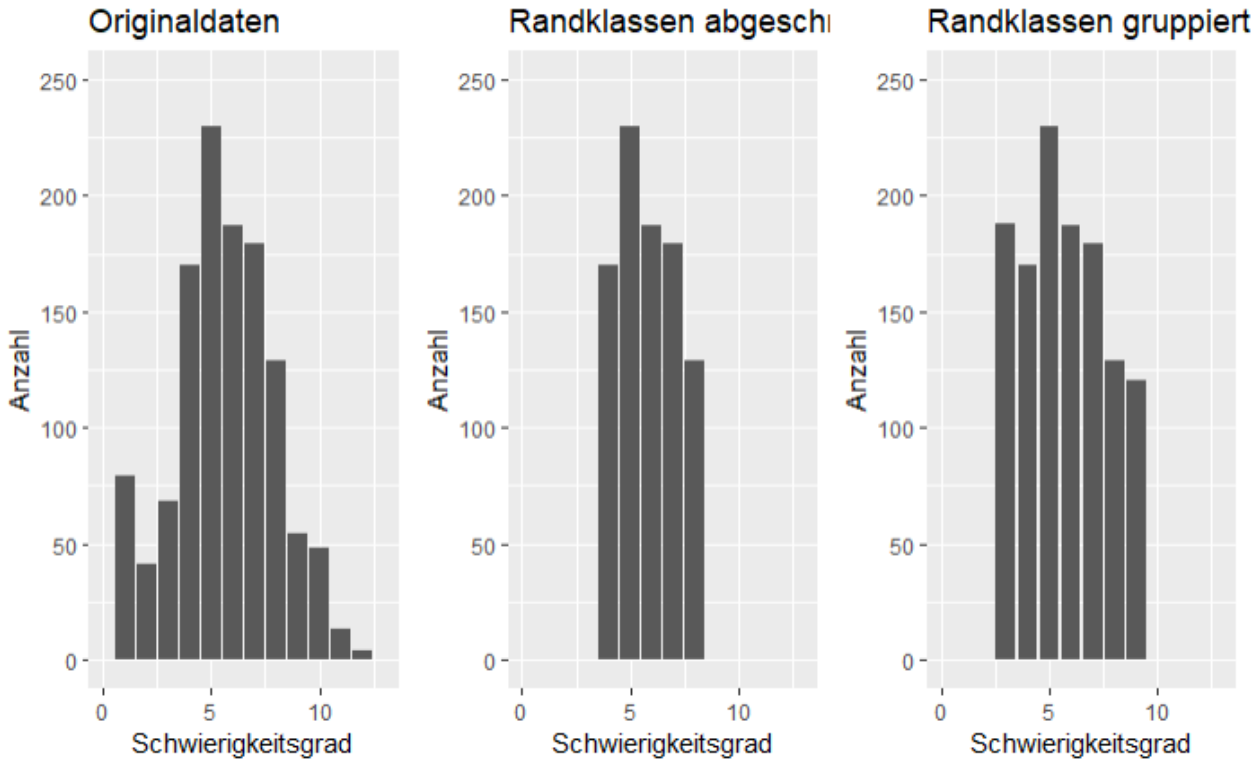


Abbildung 29 Diagramm über die neue Definition der Klassen

5.7 Sampling

Das Sampling der Daten sollte optimalerweise eine gleichmässige Verteilung über alle Klassen aufweisen. Trainings- und Testdaten sollten also proportional die gleiche Menge an Daten pro Klasse haben.

Wird, wie in der Praxis üblich, die 70/30 oder 60/20/20 Unterteilung vorgenommen, ohne nach Klassen zu gruppieren, so kann es vorkommen, dass eine Klasse komplett in den Trainingsdaten vorkommt, nicht jedoch in den Testdaten, und umgekehrt.

Daher wurde ein Sampling pro Gruppe umgesetzt. Daraus ergibt sich dann folgendes Bild:

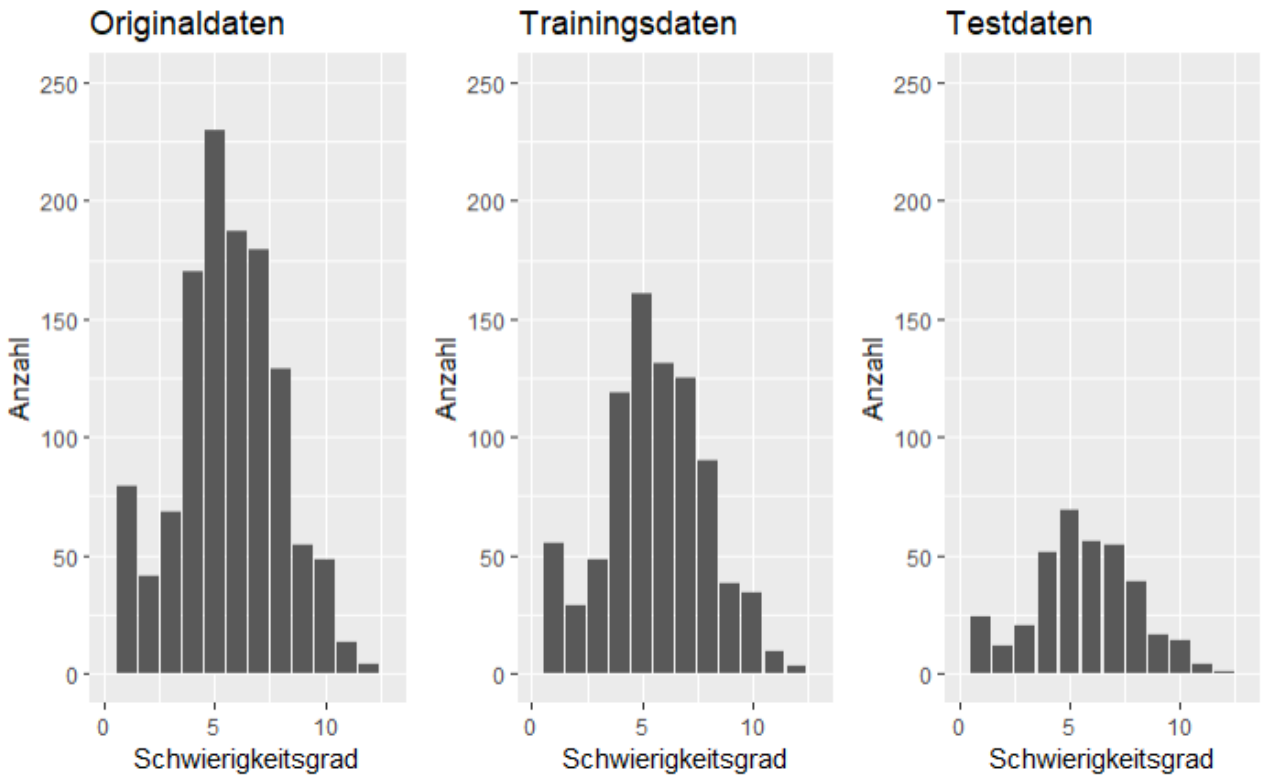


Abbildung 30 Verteilung der Datensets

Damit ist das Verhältnis der Trainings und Testklassen gleichmässig verteilt und bietet so eine bessere Grundlage für das Training und Testen der Modelle.

Ausnahme ist hierbei der Schwierigkeitsgrad 13, da es für diesen bloss einen Eintrag gibt. Mehr dazu im vorhergehenden Kapitel (Siehe: Klassenreduktion).

5.8 Normalisierung

Verschiedenen Daten haben in der Regel unterschiedliche Wertebereiche. So kommt die Durchschnittliche Steigung auf Werte von ~9.8 bis zu ~31. Planc hingegen beginnt im negativen Bereich und erreicht auch deutlich höhere Werte.

Für gewisse Modelle sind standardisierte Daten erforderlich. Dies bedeutet, dass egal wie der ursprüngliche Wertebereich aussieht, sie werden so umgewandelt, dass anschließend alle Daten zwischen 0 und 1 liegen.

Nachfolgende Grafiken verdeutlichen dies anhand der durchschnittlichen Steigung aller Routen.

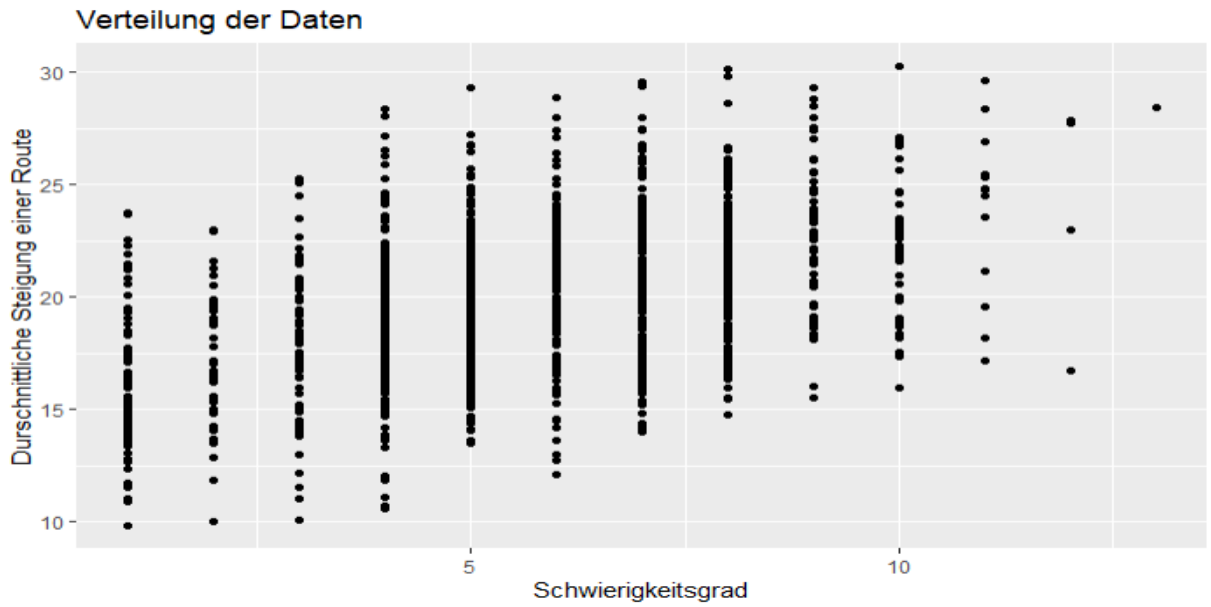


Abbildung 31 Diagramm Verteilung der Daten in Bezug auf die durchschnittliche Steigung

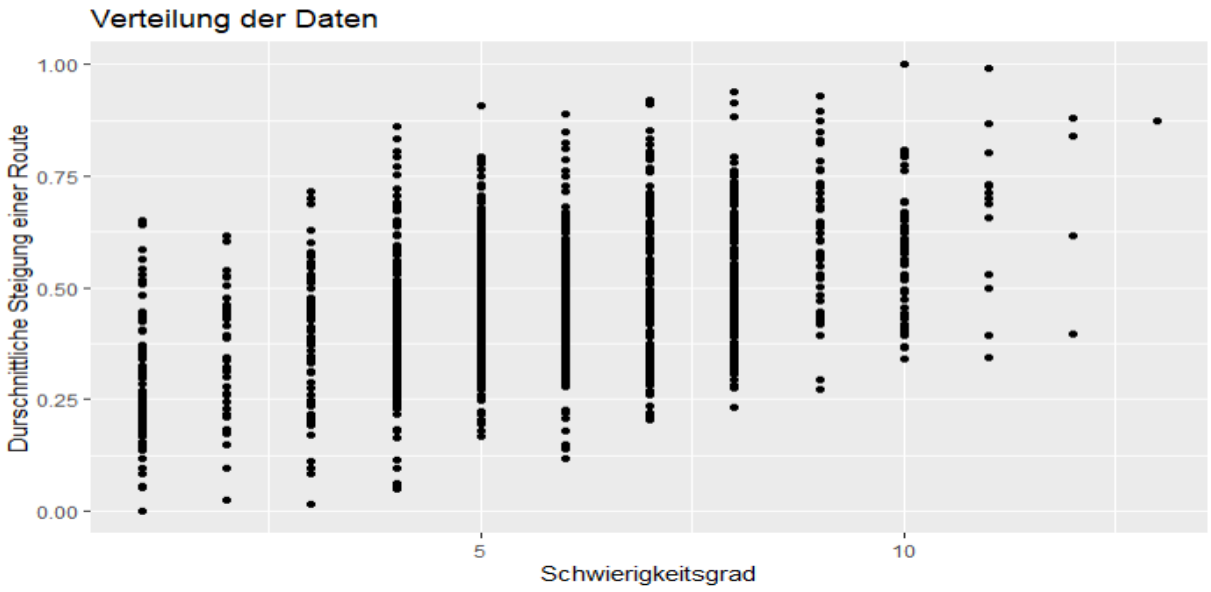


Abbildung 32 Diagramm Verteilung der Daten in Bezug auf die durchschnittliche Steigung (normalisiert)

Da den schlussendlich ausgewählten Modellen dies jedoch nicht benötigen, oder intern bereits selbst vornehmen, ist diese Umwandlung nicht im finalen Produkt enthalten.

6 Modellauswahl

Im nachfolgenden Abschnitt werden die ausgewählten Modelle beschrieben und wie diese optimiert wurden. Folglich werden alle Modelle spezifischen Anpassungen hier beschrieben.

6.1 Multiple Lineare Regression

Aus dem Grundgedanken her, dass jedes Feature in gewisser Weise eine lineare Beziehung zum Schwierigkeitsgrad aufweist, scheint die lineare Regression eine gute Wahl. Beispielsweise ist ein Trend erkennbar, bei welchem die Steigung einen linearen Einfluss auf den Schwierigkeitsgrad hat. (Je steiler, je schwerer). Gleiches gilt für Planc, Absturz Gefahrenwerte, Walddichte usw. (Siehe Lineare Charakteristika)[7]

6.2 Random Forest

Random Forest gehört zu der Gruppe der **Error! Reference source not found.** Methoden. (Siehe Supervised Learning) Dabei wird ein Wald von Entscheidungsbäumen erstellt. Dieses Konzept wird auch oft in Ablaufdiagrammen verwendet. Ein Produkt wird Anhand seiner diversen Features bewertet.
Dies führ Anschliessend zu einer Reihe von positiven und negativen Aussagen. Bei einer schwergewichts Bildung wird danach das entsprechende Resultat gewählt.[8]

In Abbildung 33 Beispiel eines Entscheidungsbaumes wird der Aufbau eines Entscheidungsbaums erklärt. Es wird die Entscheidungsfindung beim Lebenmittel Einkauf aufgezeit. Die blauen Rechtecke symbolisieren Features der Lebensmittel. Die anderer Rechtecke symbolisieren Antworten. Bein den Verbindung sind die jeweiligen Auswahlkiterien exemplarisch aufgelistet.[7]

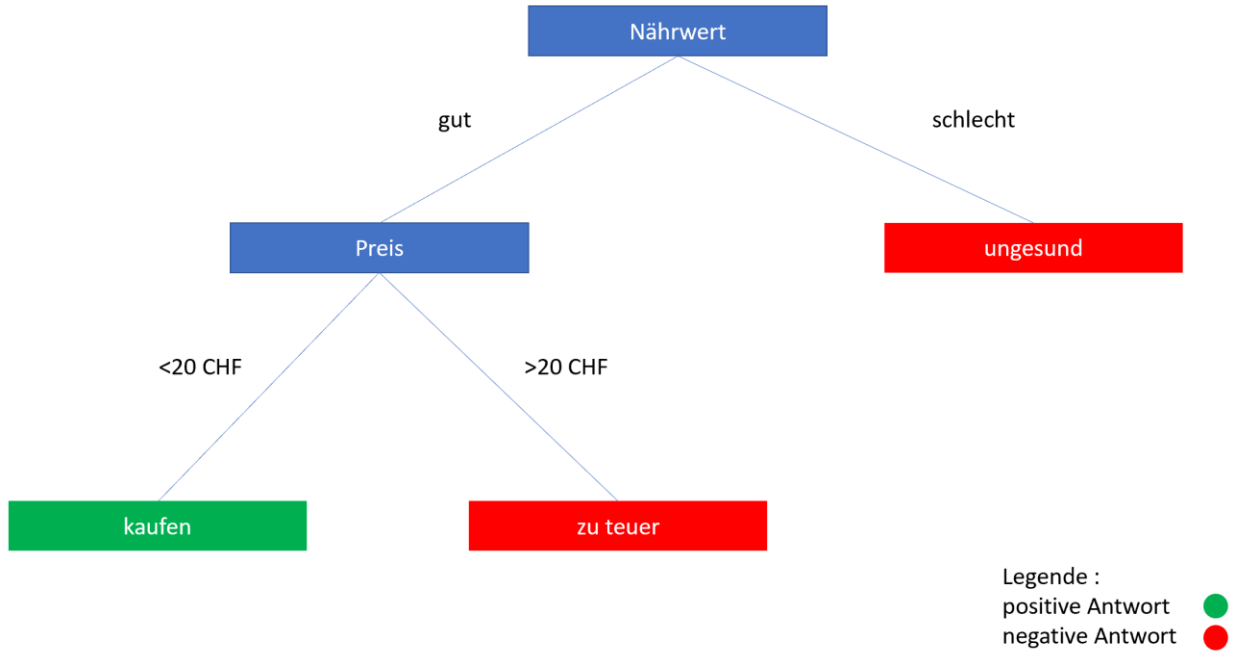


Abbildung 33 Beispiel eines Entscheidungsbaumes

Bei Random Forest werden meherer solcher Entscheidungsbäume erstellt. Anschliessend werden die verschiedenen Outputs der Bäume mit einander verglichen. Das Resulat mit der höchsten Frequenz wird nacher als Output gewählt. Diese Methode wird auch Bagging gennant.[8]

In Abbildung 34 sieht man das Random Forest mehrere Entscheidungsbäume kombiniert, welche so unabhängig wie möglich von jeglicher Korrelation sein sollten. Dies erhöht die Genauigkeit massiv. Diese angestrebt unabhängikeit erreicht Random Forest durch das gezielte hinzufügen von Zufälligkeit. Die geschieht in zwei Formen zum einen werden die Features immer zufällig in sogenannte Bags unterteilt, welche danach für die Entscheidungsfindung verwendet werden. Dies ist in der Grafik durch die unterschiedliche Anzahl blauer Kreise vertreten, welche Features symbolisieren. Zuätzlich zu diesem Schritt können die Schwellenwerte auch so angepasst werden, dass diese nicht immer nach dem bestmöglichsten Schwellenwert suchen. Diese Faktoren erhöhen die Diversität der Resulate der Bäume und verbessern entsprechend die Genauigkeit.[8]

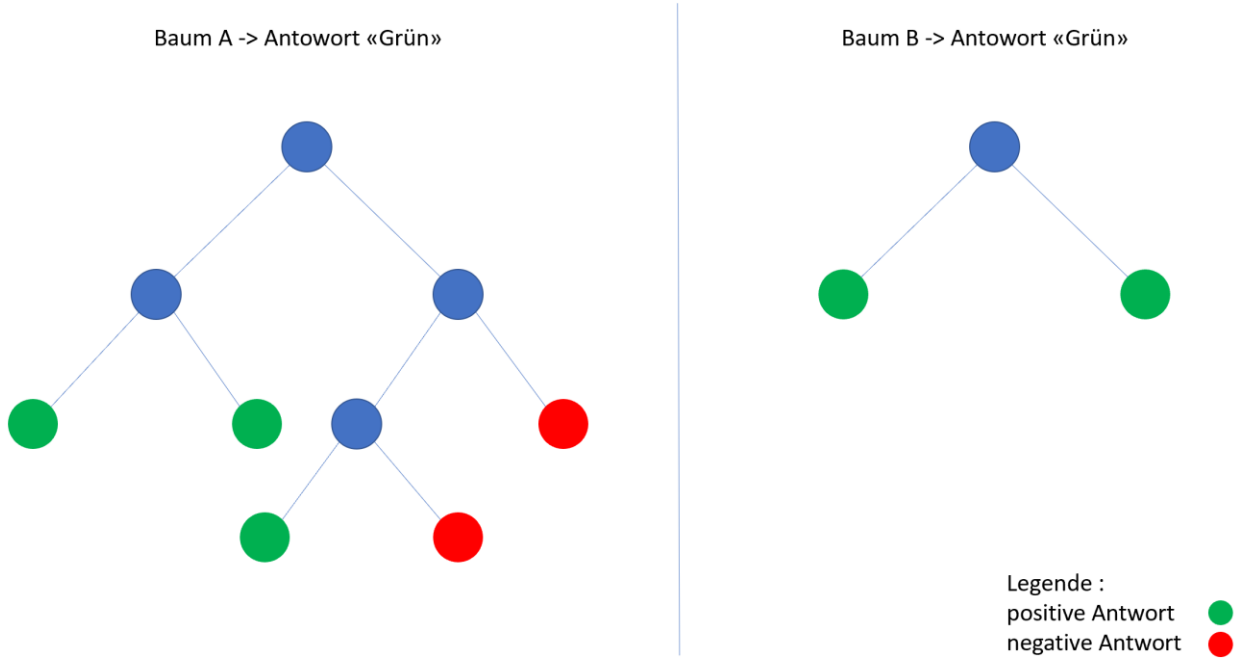


Abbildung 34 Beispiel eines Random Forest

Als Analogie könnte man das Buchen einer Reise in einer Gruppe betrachten. Jede Stimme wird dabei gleichgewichtet. Somit kann man jetzt jede Person als Entscheidungsbaum betrachten die nach verschiedenen Kriterien ein passendes Reiseziel aussuchen. Sofern die Mehrheit der Leute zur selben Entscheidung gelangt, ohne von den Anderen beeinflusst worden zu sein. Steigt die Chance statistisch gesehen das Sie ein gutes Reiseziel ausgewählt haben.

7 Auswertung

Nachfolgend werden die resultierenden Zahlen dargestellt und erläutert. Auch deren Einfluss und Bedeutung wird genauer betrachtet.

7.1 Genauigkeit (Accuracy)

Die Genauigkeit der Modelle wird dabei wie folgt definiert: Anzahl getroffener Schwierigkeitsgrade / Anzahl aller Daten = Genauigkeit in %.

Da die Rohdaten jedoch bereits Schwankungen unterliegen, haben wir eine Abweichung von einem Schwierigkeitsgrad zugelassen. Heisst also, wenn eine Route als 3 eingestuft wird, die ursprüngliche Wertung jedoch eine 2 war, so zählen wir dies trotzdem als korrekte Schätzung.

Diese Genauigkeit ist nachfolgend als «Verbesserte Genauigkeit» aufgeführt.

7.2 Regression

Die Regression erzielt mit den gegebenen Daten folgende Ergebnisse:

Datensatz	Genauigkeit	Verbesserte Genauigkeit	Durchschnittlicher Error
Randklassen abgeschnitten	22.3%	58.3%	1.56
Randklassen gruppiert	21.1%	56.2%	1.61
Alle Levels	16.9%	53.1%	1.87
Randklassen abgeschnitten (ohne Fusspassagen)	24.5%	61.3%	1.47
Randklassen gruppiert (ohne Fusspassagen)	21.6%	56.0%	1.63
Alle Levels (ohne Fusspassagen)	21.4%	57.2%	1.72

Tabelle 15 Resultate der linearen Regression

Dies in der Abbildung 35 grafisch verdeutlicht:

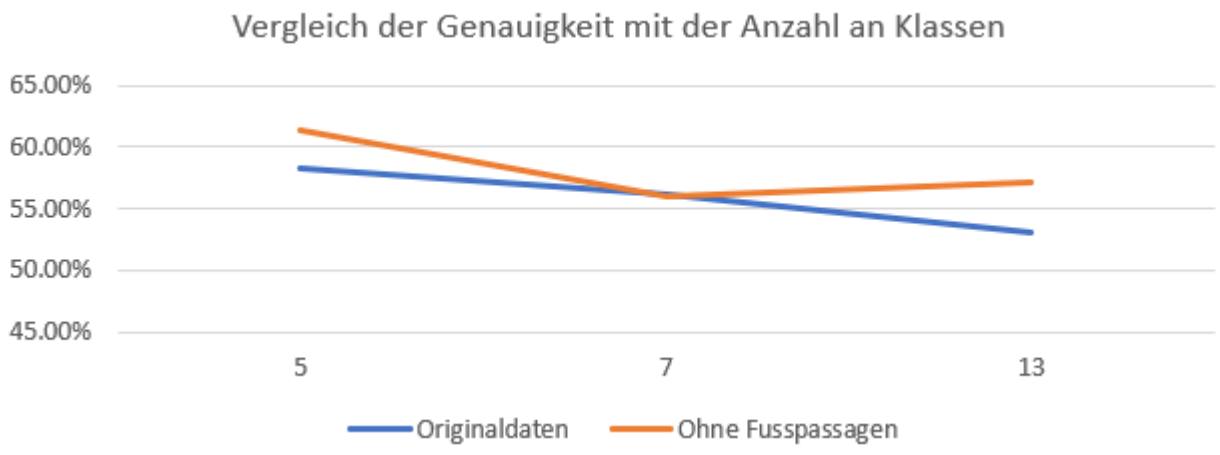


Abbildung 35 Diagramm Genauigkeit der Regression abhängig der Anzahl Levels

Das Entfernen der Fusspassage verbessert das Modell im Schnitt durchaus um ein paar Prozent. Ebenfalls deutlich: Je mehr von den Randklassen mit einbezogen werden, desto ungenauer wird das Modell.

7.3 Random Forest

Die Regression erzielt mit den gegebenen Daten folgende Ergebnisse:

Datensatz	Genauigkeit	Verbesserte Genauigkeit	Durchschnittlicher Error
Randklassen abgeschnitten	29.0%	69.1%	1.11
Randklassen gruppiert	37.1%	69.0%	1.05
Alle Levels	26.4%	61.2%	1.39
Randklassen abgeschnitten (ohne Fusspassagen)	32.3%	74.3%	0.99
Randklassen gruppiert (ohne Fusspassagen)	31.6%	68.1%	1.20
Alle Levels (ohne Fusspassagen)	25.6%	62.5%	1.31

Tabelle 16 Resultate der linearen Regression

Dies in der Abbildung 36 grafisch verdeutlicht:

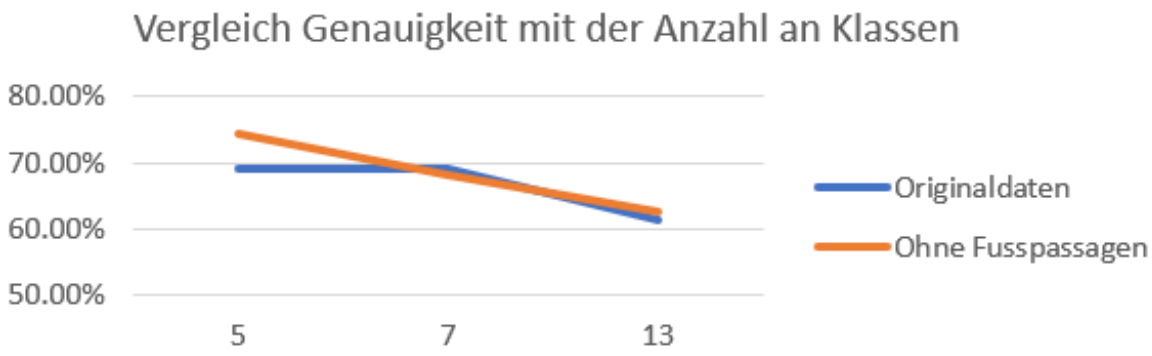


Abbildung 36 Diagramm Genauigkeit der Regression abhängig der Anzahl Levels

Interessanterweise verbessert das Entfernen der Fusspassagen die Resultate nur bei der reduzierten Anzahl Klassen. Bei den restlichen bleibt die Genauigkeit etwa die gleiche. Auch die Streuung wird nicht einheitlich besser.

7.4 Lineare Regression vs. Random Forest

Im Schnitt liefert das Modell mit dem Random Forest die besseren Ergebnisse. Auch ist die Streuung deutlich geringer, was ein wichtiger Aspekt für das einschätzen eines Schwierigkeitsgrades ist. Liegt der Schwierigkeitsgrad 1-2 Stufen daneben ist dies noch vertretbar, doch wird eine Route des Schwierigkeitsgrads 5, plötzlich als 1 eingeschätzt, würde dies für den Skitourer eine unangenehme bis gefährliche Erfahrung werden.

Doch wieso liefert die Regression eher bescheidene Resultate, wenn doch ein linearer Zusammenhang innerhalb der Daten erkennbar ist?

Auch wenn die Daten im Schnitt linear sind, ist die Streuung der Daten doch äusserts stark. Im Kapitel Lineare Charakteristika wurde für die wichtigsten Feature der Durschnitt über alle Routen gruppiert nach Schwierigkeitsgrad aufgezeigt. Wie in Abbildung 37 ersichtlich für die Steigung.[7]

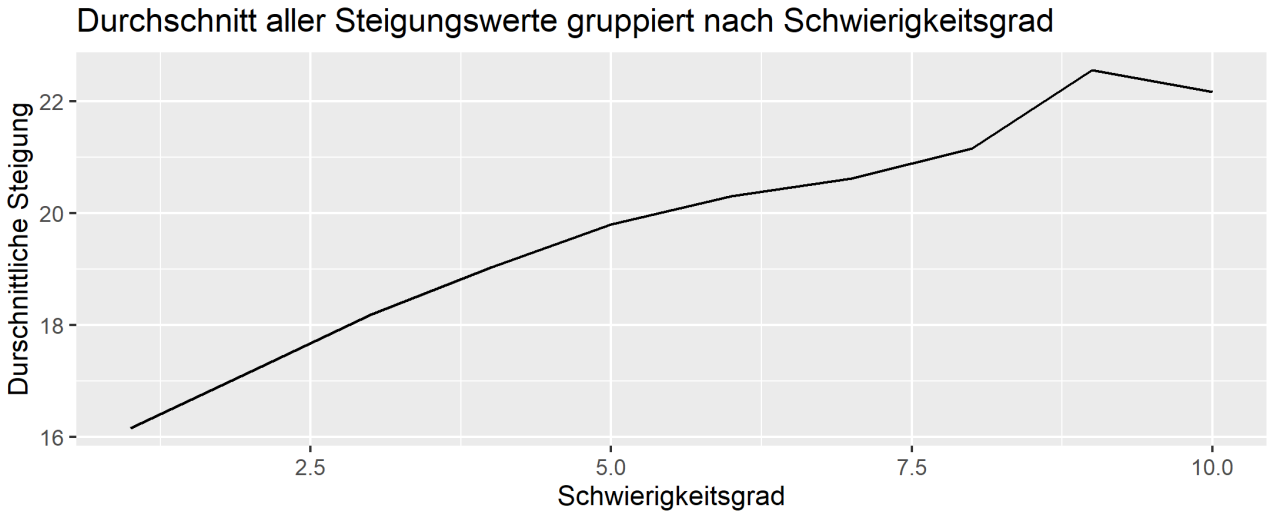


Abbildung 37 Diagramm Durschnitt aller Steigungswerte gruppiert nach Schwierigkeitsgrad

Werden die einzelnen Durchschnittswerte jedoch pro Route dargestellt, wie in Abbildung 38 ersichtlich, so ergibt sich ein deutlich anderes Bild.

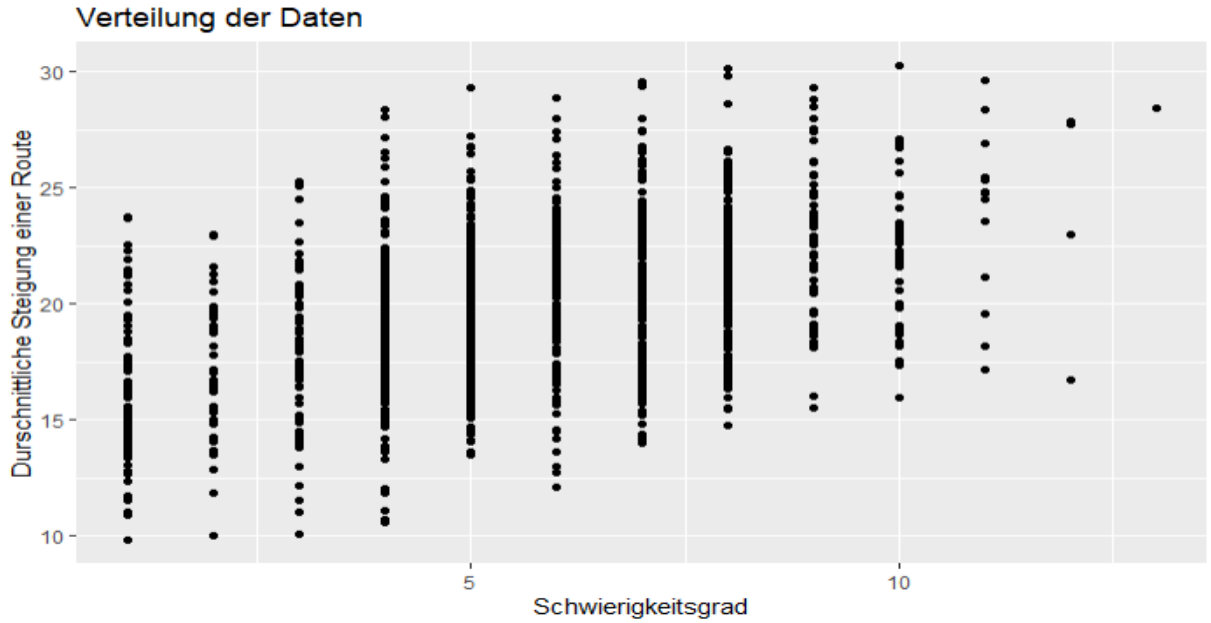


Abbildung 38 Diagramm Verteilung der Daten in Bezug auf die durchschnittliche Steigung

Obwohl die Werte im Schnitt linear sind, ist die Streuung enorm. Wird durch die Daten nun eine lineare Kurve erzeugt, so ist diese im Durschnitt korrekt, ergibt jedoch bei Routen an den Extremwerten eine falsche Vorhersage.

7.5 Data Augmentation

Aus den vorherigen Resultaten entstand ein weiterer Datensatz, welcher durch künstliche Erweiterung der Datensätze eine gute Balance zwischen den Klassen aufwies. Ausserdem wurden die Klassen über dem Schwierigkeitsgrad 9 entfernt, da nach Absprache mit Experten die unteren Schwierigkeitsgrade für den Service Anbieter eine deutlich höhere Gewichtung aufweist. Dies in Kombination mit dem Random Forest als Modell, einem neuem Seed und dem Abschneiden der Fusspassgen führte zu den bisher besten Resultaten.

Genauigkeit:	50.16%
Erweiterte Genauigkeit:	77.14%
Durchschnittlicher Fehler:	0.86
Quadrierter durchschnittlicher Fehler:	1.84

Vorhergesagter Wert \ Tatsächlicher Wert	1	2	3	4	5	6	7	8	9
1	42	6	7	5	1	0	0	0	0
2	1	30	1	0	0	0	0	0	0
3	0	0	25	3	0	1	0	0	0
4	1	0	5	18	7	3	1	1	0
5	1	1	2	8	4	5	4	1	1
6	0	0	0	3	4	3	6	0	0
7	0	0	0	12	8	12	30	8	8
8	0	0	1	2	1	6	12	6	7
9	0	0	0	0	0	0	1	0	0

Abbildung 39 Konfusionsmatrix des finalen Ergebnisses

Wie in Abbildung 39 ersichtlich ist, beweget sich die Grosszahl der Ergebnisse genau auf der mittleren Achse oder liegen innerhalb des Bereichs der erweiterten Genauigkeit (gelb). Die auffällige Ausnahme stellt hierbei der Schwierigkeitsgrad 4 dar, welcher nach unten wie oben starke Ausreisser aufweist.

Diese Ausreisser waren in allen vorgenommenen Analysen erkennbar, wobei die Ursache unbekannt blieb. Die Untersuchung der Ursache wäre ein primäres Thema für eine nachfolgende Arbeit zu diesem Thema und dazu eine äusserts vielversprechende.

8 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse der Arbeit noch einmal zusammenfassend präsentiert.

8.1 Installationsanleitung

In der Installationsanleitung wird das Vorgehen, wie man das Modell auf seinem Rechner laufen lassen kann. n Folge dessen werden auch das verwendete Paket und Bibliotheken noch einmal zusammenfassend aufgelistet.

1. Installation von RStudio: https://rstudio.com/products/rstudio/download/#download
2. Das vorliegende GIT Repository lokal klonen: https://gitlab.fhnw.ch/skitourenguru/bachelorthesis/tree/master
3. Das Projekt im RStudio öffnen
4. Die Datei IP6_Final.Rmd öffnen
5. Die Codeabschnitte ausführen. Entweder über den grünen «Ausführen» Pfeil oder über Ctrl+Shit+Enter für den aktuellen Abschnitt, oder mit Ctrl+Alt+N für den jeweils nächsten Abschnitt.

Im dazugehörigen Readme sind weitere Informationen zum Projektaufbau, der Code Formatierung und Hilfestellungen zur Einrichtung des Projekts enthalten.

8.2 Regionaler Bias

Eine der Fragen lautete zudem, ob ein regionaler Trend bei der Vergabe der Schwierigkeitsgrade erkennbar ist. Also, ob beispielsweise Routen im Tessin tendenziell leichter bewertet werden, als z.B. im Wallis.

Dabei sind die Werte wie folgt zu interpretieren:

- +: Routen werden im Schnitt um so viel schwerer geschätzt, als das Modell dies tut
- -: Routen im Schnitt um so viel einfacher geschätzt, als das Modell dies tut

Westschweiz	-0.67
Wallis West	-0.53
Berner Oberland Ost	-0.35
Berner Oberland West	-0.13
Wallis Ost	0.23
Graubünden Süd	0.25
Zentralschweiz	0.38
Glarus	0.39
Graubünden Nord	0.54
Tessin	0.97

Tabelle 17 Übersicht der regionalen Bias Werte

Dies zeigt, dass die Westschweiz Routen im Schnitt einfacher einschätzt, als das Tessin. Dies soll jedoch nur ein Richtwert sein. Es schwer ein Bias präzise zu belegen, da es die gleiche Route nicht in zwei verschiedenen Regionen gibt. Ausserdem ist, wie bei den Grunddaten bereits, eine grosse Streuung vorhanden. Die Werte sind also mehr als ein Indiz zu verstehen.

9 Schlusswort

Der letzte Abschnitt fasst die erzielten Ergebnisse und Resultate zusammen. Ausserdem zeigt er die Zukünftigen Entwicklungsmöglichkeiten auf.

9.1 Fazit

Im Verlauf der Arbeit konnten wir uns vertieft mit den Daten beschäftigen. Anfänglich schien es äusserst düster, mit der geringen Menge an Daten mit den grossen Streuungen eine genaue Vorhersage erreichen zu können.
Über mehrere Iterationen wurde mehr und mehr Verständnis über die Bedeutung und Zusammenhänge der einzelnen Eigenschaften einer Route erarbeitet und nach und nach wurde ein Weg klarer. Auch die Herausforderungen wurden dadurch ersichtlicher. So fiel die ungleichmässige Verteilung auf. Die kritischen Randklassen mit nur wenig Routen. Die Fusspassagen, durch welche Werte zustande kamen, die gemäss SAC nicht zulässig waren für den gegebenen Schwierigkeitsgrad.

Mit dieser Erkenntnis wurden die einzelnen Herausforderungen angegangen und Lösungen gesucht. Abtrennen der Fusspassagen, Sampling anhand der Klassenverteilung, Entfernen von Randgruppen beim Sampling usw.

Dies resultierte am Ende in einem vertretbaren Vorhersagen durch das Modell.

9.1.1 *Projektabschluss*

Die genannten Projektziele wurden erfüllt, jedoch wäre eine höhere Genauigkeit natürlich immer wünschenswert, was mit mehr Zeit bestimmt auch machbar wäre. Im nachfolgenden Kapitel (Siehe: Ausblick) ist beschrieben, welche Ansätze dabei helfen könnten.

Die Fragestellungen wurden weitestgehend beantwortet. Es wurden leider keine Beziehungen innerhalb der Daten gefunden und die Länge wies auch keinen Einfluss auf den Schwierigkeitsgrad auf. Hier könnte durchaus noch weiter geforscht werden, um die Anzahl oder Gewichtung der Features zu anzupassen.

Die Frage des Bias wurde beantwortet, jedoch mit der Weisung, die Zahlen mit Vorsicht zu geniessen. Mit mehr Daten könnte auch diese Frage deutlich genauer beantwortet werden. Doch es gibt einen Eindruck von der aktuellen Situation, was auch die ursprüngliche Intention der Fragestellung war.

9.1.2 *Reflexion*

Die anfänglich einfache Methode, in welcher beispielweise die Steigung als Durchschnitt verwendet wurde, lieferte schlechte vorhersagen, da dadurch wichtige Eigenschaften einer Route komplett verloren gingen. Eine Route kann durchgehend eher einfach sein, doch reicht bloss an einer Stelle eine schwierige Passage, um dieser Route einen höheren Schwierigkeitsgrad zu vergeben. Gleiches gilt für Dinge wie Maximum, oder eine Verzerrung nach oben. Durch dieses Massnahmen gehen Informationen verloren. Erst durch die Summe all dieser Charakteristika (TSFresh), konnten Daten zusammengefasst werden, ohne die wichtigen Eigenschaften einer Route zu verlieren.

9.1.3 *Ausblick*

Eine der vielversprechendsten Aspekte wäre die Erweiterung der Datenmenge. Besonders in den Randgruppen. Mit mehr Daten können genauere Modelle erstellt werden und weitere Analysieren vorgenommen werden. Auch würde es erlauben, die grosse Streuung anzugehen, indem

beispielsweise ein Limit gesetzt wird und alle Routen ausserhalb dieses Bereichs werden für das Training des Modells ignoriert.

Die grösste Chance sehen wir aber in der Überarbeitung der Skala für die Bewertung der Skitouren. Damit ein besseres gemeinsames Verständnis der Schwierigkeitsgrade zwischen Menschen und Computer entstehen kann. Es sollten nur noch Faktoren in die Bewertung einfliessen, welche Messbar und reproduzierbar sind. Eine «Umschreibung» einer Sturzgefahr in Kombination mit der Übersetzung in Deutsch, Italienisch und Französisch, führt unweigerlich dazu, dass unterschiedliche Interpretationen entstehen, was wiederum zu unterschiedlichen Einschätzungen führt. Dies müsste vom SAC angegangen werden, um die Bewertungen einheitlicher zu gestalten.

Des Weiteren bringen womöglich aktuelle Forschungen zur Klassifizierung von Zeitreihendaten neue Möglichkeiten und Modell für eine Validierung zu Tage, welche für dieses Thematik von Nutzen sein können.

Eine weitere grosse Möglichkeit würde in dem Benutzen von 3D-Daten liegen bei welchem die Berge als 3D Modell erfasst sind. Danach könnte man von jedem Punkte die Einsehbarkeit der weiteren Routen berechnen.

Der letzte Schritt wäre eine Bildverarbeitung von Satellitenbildern, bei welcher versucht wird anhand der Bilddaten identifizierungsmerkmale zu erkennen, welche einen gewissen Schwierigkeitsgrad rechtfertigen. Dies ist wird aber durch die sehr lasche Definition der aktuellen Bewertung fasst verunmöglicht.

10 Literatur- und Quellenverzeichnis

[1] «Skitourenguru». [Online]. Verfügbar unter: https://www.skitourenguru.ch/. [Zugegriffen: 17-März-2020].

[2] *I4DS01: Automatische Bestimmung des Schwierigkeitsgrades von Skitouren*. FHNW.

[3] «Skibergsteigen», *Wikipedia*. .

[4] F. MARTINESCU-BĂDĂLAN und R. STĂNCIULESCU, «HISTORY AND DEBUT OF THE SKI-MOUNTAINEERING». 13-Jan-2020.

[5] K. Winkler, H.-P. Brehm, und J. Haltmeier, *Bergsport Winter*, 4. Aufl. SAC-Verlag.

[6] G. Sanga, *Die klassischen Skitouren*, First Edition. 2015.

[7] A. Ng, «Coursera», *Coursera*. [Online]. Verfügbar unter: https://www.coursera.org/learn/machine-learning/home/week/1. [Zugegriffen: 17-März-2020].

[8] H. Wickham und G. Grolemund, *R for Data Science*. O'Reilly Vlg. GmbH & Co., 2017.

[9] «Snapshot». .

[10] W. Koehrsen, «Statistical Significance Explained», *Medium*, 05-Feb-2018. [Online]. Verfügbar unter: https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687. [Zugegriffen: 18-März-2020].

[11] «SMOTE explained for noobs - Synthetic Minority Over-sampling TEchnique line by line · Rich Data». [Online]. Verfügbar unter: http://rikunert.com/SMOTE_explained. [Zugegriffen: 20-März-2020].

[12] J. Brownlee, «SMOTE Oversampling for Imbalanced Classification with Python», *Machine Learning Mastery*, 16-Jan-2020. [Online]. Verfügbar unter: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. [Zugegriffen: 22-Feb-2020].

11 Abbildungsverzeichnis

12 Tabellenverzeichnis

13 Abkürzungen

Abkürzung	Bedeutung
SAC	Schweizer Alpen Club
FHNW	Fachhochschule Nordwestschweiz
L	Leicht
WS	Wenig schwierig
ZS	Ziemlich schwierig
S	Schwierig
SS	Sehr schwierig
AS	Ausserordentlich schwierig
EX	Extrem schwierig
tbd	To be defined
SMOTE	Synthetic Minority Oversampling Technique

A Anhang

A1. SAC Bewertungsskala

A2. Die Einstufung erfolgt in 7 verschiedenen Stufen, diese können teilweise mit den Modifikatore, Plus und Minus noch verfeinert werden. Diese sollten Schätzungweise einem Drittel entsprechen. Die Bewertung entspricht dem Spitzenwert der Hauptkriterien einer Tour. Die Schwierigkeitsangaben sollen Richtwerte bei guten Schnee-, Witterungs- und Sichtverhältnissen darstellen. Die Bewertung bezieht sich aussschliesslich auf den skifahrerischen Teil der Touren. Aplintechnische Schwierigkeiten werden ausseracht gelassen und meist separat zu den entsprechenden Touren mit einer eigenen Skala angegeben.

	Hauptkriterien					Hilfskriterien		
Grad	Steilheit	Ausgesetztheit	Geländeform Aufstieg und Abfahrt	Engpässe in der Abfahrt	Beispiele	Erschwerte Orientierung in Aufstieg und Abfahrt;	Routenverlauf nicht einsehbar	Routenfehler sind kaum oder gar nicht mehr korrigierbar
L (+)	bis 30°	keine Ausrutschgefahr	weich, hüglig, glatter Untergrund	keine Engpässe	Niderhorn von Boltigen Steghorn von Lämmerenhütte Faulhorn von Süden Grünhornlücke			
WS (- +)	ab 30°	kürzere Rutschwege, sanft auslaufend	überwiegend offene Hänge mit kurzen Steilstufen. Hindernisse mit Ausweichmöglichkeiten (Spitzkehren nötig)	Engpässe kurz und wenig steil	Bunderspitz Arpelistock von Geltenhütte Sattelhorn (Kandertal) Sattelhorn (Driest)			
ZS (- +)	ab 35°	längere Rutschwege mit Bremsmöglichkeit	kurze Steilstufen ohne Ausweichmöglichkeiten, Hindernisse in	Engpässe kurz, aber steil	Männliflue von Süden			

		en (Verletzungsgefahr)	mässig steilem Gelände erfordern gute Reaktion (sichere Spitzkehren nötig)		Rinderhorn Normalweg Bundstock von Kandersteg Grosshorn von Süden			
S (- +)	ab 40°	lange Rutschwege, teilweise in Steilstufen abbrechend (Lebensgefahr)	Steilhänge ohne Ausweichmöglichkeiten. Viele Hindernissen erfordern eine ausgereifte und sichere Fahrtechnik	Engpässe lang und steil. Kurzschwingen für Könner noch möglich	Winterhore N-Flanke Vorder Lohner SW-Flanke Altels NW-Flanke Dreispitz Wyssi Frau NW-Rücken			
SS (- +)	ab 45°	Rutschwege in Steilstufen abbrechend (Lebensgefahr)	allgemein sehr anhaltend steiles Gelände. Oft mit Felsstufen durchsetzt. Viele Hindernissen in kurzer Folge	Engpässe lang und sehr steil. Abrutschen und Quersprünge nötig	Märe N-Couloir Balmhorn N-Wand direkt Dündenhorn S-Seite Lauteraarhorn Mönch S-Wand			
AS (- +)	ab 50°	äusserst ausgesetzt	äusserst steile Flanken oder Couloirs. Keine Erholungsmöglichkeit in der Abfahrt	Engpässe lang und sehr steil, mit Stufen durchsetzt, nur Quersprünge und Abrutschen möglich	Mönch NE-Wand			
EX	ab 55°	extrem ausgesetzt	extreme Steilwände und Couloirs	evtl. Abseilen über	Eiger NE-Wand			

				Felsstufen nötig				

A3. Projektvereinbarung

Projektvereinbarung

Kontaktdaten

Autoren:

Fabian Brunner
Bellevuestrasse 26
2540 Grenchen
fabian.brunner@windowslive.com
+41 79 266 20 63

Michel Bittel
Meisenweg 2
3422 Kirchberg
michel.bittel@bluewin.ch
+41 79 653 17 13

Betreuer:

Csillaghy Andre
Bahnhofstrasse 6
5210 Windisch
andre.csillaghy@fhnw.ch
+41 56 202 76 85

Kunde:

Günter Schmudlach
Im Sydefädeli 19
8037 Zürich
schmudlach@gmx.ch
+41 77 471 90 39

Ausgangslage

Die Web-Plattfrom «www.skitourenguru.ch» unterstützt Wintersportler bei der Auswahl einer Skitour mit tiefem Lawinenrisiko. Dabei hat setzt die Plattform auf künstliche Intelligenz um diese Einschätzungen vorzunehmen. Nach anfänglicher Skepsis von Skitouren Experten und Forschungsämter, gelang es der Web-Plattform sich zu etablieren und sich laufend zu verbessern, heute wird Sie aktiv von Sponsoren umworben und wird offiziell vom Schweizer Alpen-Club für die Skitourenplanung empfohlen. Nun möchte die Plattform den nächsten Schritt wagen und auch die Einstufung des Schwierigkeitsgrades mittels Machine Learning lösen. Bisher wurde dies manuell durch Experten in sogenannten Tourguides abgedeckt. In einer Zeit aufreibenden Arbeit wurde der Schwierigkeitsgrad von 1200 Skitouren von Hand bestimmt. Nun soll ein entsprechendes Modell entwickelt werden um den Schwierigkeitsgrad aus Daten abzuleiten. Daher ist die Web-Plattform an die FHNW herangetreten.

Die Daten können mit QGIS dargestellt werden. Diese können auch in ein csv oder SQLite exportiert werden. Die Daten zur Beschaffenheit beinhalten dabei die folgenden Informationen:

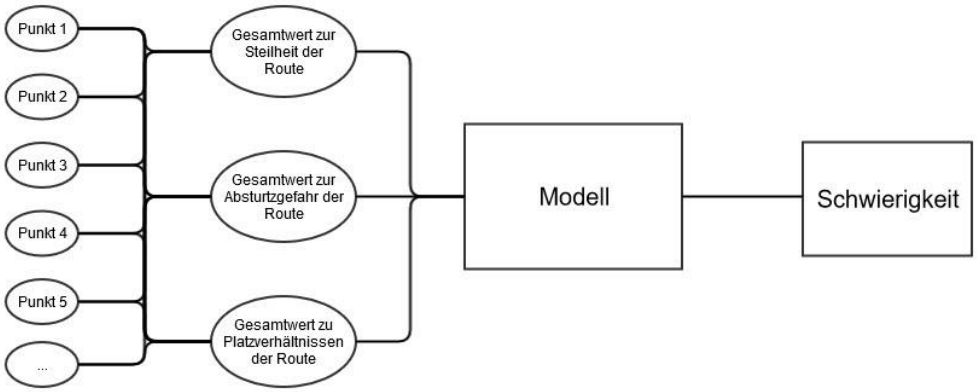
- Steilheit
- Absturzgefahr
- Platzverhältnisse

Aufgabenstellung

“Ziel diese Arbeit ist die Entwicklung eines Modells, das aus Geländedaten den Schwierigkeitsgrad einer Skitour ableiten kann.” (FHNW 2019)

Primärer Fokus dieser Arbeit liegt im erarbeiten eines Modells, womit aus den vorhandenen Daten eine einheitliche Schwierigkeitsabschätzung erzeugt werden kann. Dazu müssen die Daten erst untersucht und beurteilt werden, um die Auswahl der möglichen Modelle einzugrenzen.

Für das Modell haben wir uns in einem ersten Schritt darauf geeinigt, dass die Daten erst vorbereitet werden und damit 3 Variablen (Prädiktoren) zu erzeugen, mit welchem das Modell trainiert wird. Optional wird ein zweites Modell aufgesetzt, welches die Rohdaten direkt verarbeitet. Erster Schritt:



Input

Als Input stehen 1200 Skitouren zur Verfügung. Die Skitouren selbst besitzen folgende Eigenschaften:

Feldinhalt	Datentyp	Skalierung
Schwierigkeitsgrad	Zahl	0-18
Start- und Endpunkt	Text	-
RegionsID	Zahl	Autogen
Wildlife	Zahl	nicht Interpretierbar bis jetzt

Diese Routen besitzen alle 10 Meter einen Messpunkt. Dieser enthält wiederum folgende Eigenschaften:

Feldinhalt	Datentyp	Skalierung
Neigungswinkel	Zahl	
Planare Krümmung	Zahl	
Summe der Beschleunigung beim Fall	Zahl	
Maximale Beschleunigung beim Fall	Zahl	
Summe der Geschwindigkeit beim Fall	Zahl	
Maximale Geschwindigkeit beim Fall	Zahl	
Walddichte	Zahl	0-100 (0 entspricht keinem Wald)
Korridorbreite	Zahl	

Output

Das Modell soll eine Bewertung nach dem Standard des Schweizer Alpen Clubs zurückgegeben.

Grad	**Steilheit**	**Ausgesetztheit**	**Geländeform Aufstieg und Abfahrt**	**Engpässe in der Abfahrt**
L (+)	bis 30°	keine Ausrutschgefahr	weich, hügelig, glatter Untergrund	keine Engpässe
WS (- +)	ab 30°	kürzere Rutschwege, sanft auslaufend	überwiegend offene Hänge mit kurzen Steilstufen. Hindernisse mit Ausweichmöglichkeiten (Spitzkehren nötig)	Engpässe kurz und wenig steil
ZS (- +)	ab 35°	längere Rutschwege mit Bremsmöglichkeiten (Verletzungsgefahr)	kurze Steilstufen ohne Ausweichmöglichkeiten, Hindernisse in mässig steilem Gelände erfordern gute Reaktion (sichere Spitzkehren nötig)	Engpässe kurz, aber steil
S (- +)	ab 40°	lange Rutschwege, teilweise in Steilstufen abbrechend (Lebensgefahr)	Steilhänge ohne Ausweichmöglichkeiten. Viele Hindernissen erfordern eine ausgereifte und sichere Fahrtechnik	Engpässe lang und steil. Kurzschwingen für Könner noch möglich
SS (- +)	ab 45°	Rutschwege in Steilstufen abbrechend (Lebensgefahr)	allgemein sehr anhaltend steiles Gelände. Oft mit Felsstufen durchsetzt. Viele Hindernissen in kurzer Folge	Engpässe lang und sehr steil. Abrutschen und Quersprünge nötig
AS (- +)	ab 50°	äusserst ausgesetzt	äusserst steile Flanken oder Couloirs. Keine Erholungsmöglichkeit in der Abfahrt	Engpässe lang und sehr steil, mit Stufen durchsetzt, nur Quersprünge und Abrutschen möglich

Ziel

Das Ziel der Arbeit kann in einem Satz folgendermassen formuliert werden.

Anhand der gelieferten Daten soll ein Modell erstellt werden, welches einer Skitour den entsprechend Schwierigkeitsgrad gemäss obiger Tabelle zuordnet.

$$\text{Schwierigkeitsgrad} = f(\text{Neigungswinkel}, \text{Absturzgefahr}, \text{Platzverhältnisse})$$

Zielkriterien

Für die Arbeit stehen zwei Ziele im Fokus:

Kundenorientiertes Ziel einer möglichst hohen Genauigkeit

Der Kunde möchte eine möglichst hohe Genauigkeit erreichen. Darauf soll im Rahmen der Möglichkeiten hingearbeitet werden. Am Ende muss die erzielte Genauigkeit begründet werden können.

Ziel eines möglichst geeigneten und angepassten Modells

Mit den gegebenen Daten und Mitteln ein gutes Modell entwerfen. Gut wird hierbei mittels Qualitätsprüfungen gemessen und ausgewertet. Abweichungen müssen begründbar sein.

Problemstellungen

Nach der ersten Analyse des Auftrages wurden folgende Probleme identifiziert. Es können zu einem späteren Zeitpunkt noch unbekannte Probleme auftreten, welche es zu bewältigen gilt. Diese werden aber immer einen Bezug zu einem der unten aufgeführte Punkte haben.

1. Ski Tauglichkeit eines Punktes anhand der Daten bestimmen
2. Identifikation von der Beziehung der Daten zum Schwierigkeitsgrad
3. Aufbereitung und Analyse der Daten
4. Auswahl eines geeigneten Modells
5. Training und Validierung des Modelles
6. Anpassung der Modellparameter
7. Prüfung der Qualität des Modelles
8. Analyse der Resultate
9. Identifikation von Erweiterungen

Fragestellungen

Im Verlauf des Projektes werden sich immer wieder auch ander Fragen aufdrängen, welche auch beantwortet werden müssen. Diese Fragen sind als Schwergewichts Bildung zu betrachten. //Mehr like: folgendes sind die drei Hauptfragen der Arbeit.

1. Mit welchem Modell erreichen wir die höchste Genauigkeit? (Kundenziel)

2. Sind noch andere Daten vorhanden, welche bei der Bewertung von Skitouren hilfreich sind? (Akademisch)

3. Sind zwischen den Daten bisher noch unbekannte Beziehungen vorhanden? (Akademisch)

Optionale Fragestellung:

4. Welchen Bias haben die einzelnen Gebietsführer des SAC (Kundenziel)

Abgrenzung

Die Vermessung und Erfassung von Daten ist nicht bestandteil dieser Arbeit. Der Fokus ist auf der Auswahl und dem Training eines geeigneten Modells gelegt.

Die Arbeit befasst ausschließlich mit der Bewertung des ski technischen Teiles einer Skitour. Abschnitte, welche man kletternd oder zu Fuss überwinden muss werden nicht berücksichtigt, sofern die sich am Zielpunkt befinden. Fusspassagen zwischen der restlichen Route fliessen in der Schwierigkeitsgrad mit ein. Werde jedoch gesondert behandelt.

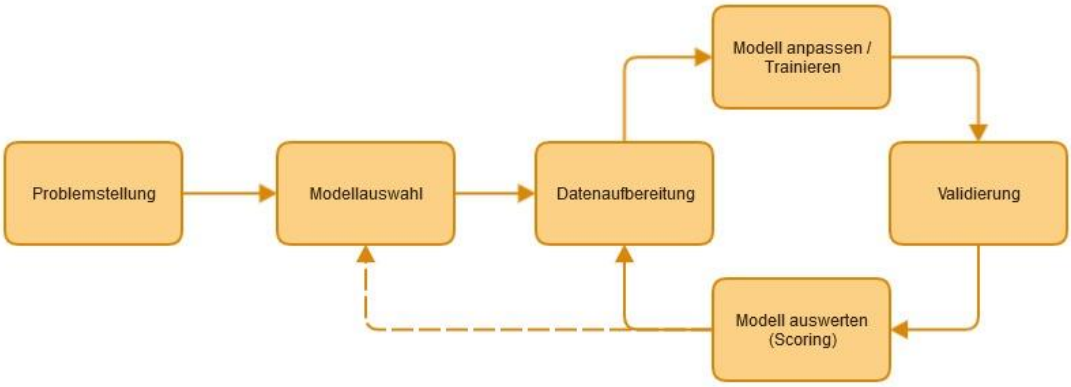
Technisch

Realisierung erfolgt mit R-Statistics, welches sich in C# integrieren lässt. Dies damit der Kunde die Umsetzung nach Abschluss des Projekts bei Bedarf selber erweitern oder anpassen kann.

Iterationsplanung

Das Projekt wird in zwei grössere Teile aufgeteilt. Der erste dient der Entwicklung eines eher einfachen Modells. Dieses wird in mehreren Iterationen immer weiter verbessert.
Im zweiten Teil wird versucht ein komplexeres aber hoffentlich genaueres Modell zu erstellen. Auch hier wird mit Iterationen gearbeitet. Sollte dieser zweite Teil jedoch kein besseres Ergebnis liefern, ist durch den ersten Teil trotzdem ein brauchbares Modell für den Kunden vorhanden.

In beiden Teilen verläuft die Iteration wie folgt:



Eine Iteration läuft über eine Dauer von schätzungsweise **zwei Wochen**. Sollte sich ein Modell als absolut unpassend erweisen, kann nach Absprache das Modell komplett ausgetauscht werden. (gestrichelte Linie)

Vorbereitung

Start	30.09.2019	Ende	28.10.2019	(4 Wochen)

Fokus	In der Vorbereitungsphase werden wir uns in das Thema einarbeiten und die Situation genauer Analysieren. Dabei wählen wir auch die Modelle, für den ersten und zweiten Teil. (Beim zweiten Teil kann sich dies nachträglich noch ändern, je nach Feedback zum ersten Teil)
Ziel	Ein grundlegendes Verständnis in Machine Learning. Auswahl eines Modell für den ersten und zweiten Teil.

Teil 1: Einfaches Modell

Start	28.10.2019	Ende	23.12.2019	(8 Wochen)

Fokus	Erstellen eines einfaches Modells. Simpel beginnen und dann nach und nach eine bessere Genauigkeit erreichen über die verschiedenen Iterationen.
Ziel	Fertiges Modell basierend auf einer einfacheren Basis, welches über eine akzeptable Genauigkeit verfügt.

Teil 2: Komplexeres Modell

Start	23.12.2019	Ende	17.02.2019	(8 Wochen)

Fokus	Erstellen eines komplexeren Modells. Auch hier wird wieder mit einer möglichst kleinen Variante begonnen und über mehrere Iteration stetig verbessert und erweitert.
Ziel	Fertiges Modell basierend auf komplexerer Basis, welches über eine akzeptable Genauigkeit verfügt.

Teil 3: Optimierung des vielversprechenden Modells

Start	17.02.12.2019	Ende	09.03.2019	(3 Wochen)

Fokus	Nach Absprache mit dem Kunden, wird das Modell, welches als besser empfunden wurde, nochmals optimiert um eine möglichst hohe Genauigkeit zu erzielen.
Ziel	Fertiges Modell basierend auf gewählter Basis, welcher über eine möglichst gute Genauigkeit verfügt.

Lieferobjekte

Folgende Objekte werden am Ende der Arbeit übergeben:

Artefakt	Medium
Quellcode	GIT - Repository (GitLab)
Modell	Lauffähiger Prototyp
Bachelor Thesis Dokument	Als PDF Dokument sowie ausgedruckte und unterschriebene Version.

Kommunikation

Es werden nach Bedarf, Meetings mit dem Kunden zusammen durchgeführt. Ausserdem sollte die Kommunikation über Email sichergestellt werden. Im Notfall kann man auch jemanden telefonisch erreichen.

Mit dem Coach wird einmal pro Woche ein Meeting stattfinden. Dies ist als Hilfestellung der Studierenden anzusehen.

Unterschriften

Beide Partien haben die Vereinbarung gelesen und sind damit einverstanden.

Projektteam

Michel Bittel

Ort, Datum	Unterschrift

Fabian Brunner

Ort, Datum	Unterschrift

Betreuer

Csillaghy Andre

Ort, Datum	Unterschrift

Auftraggeber

Günter Schmudlach

Ort, Datum	Unterschrift

Quellen

FHNW (2019): I4DS01: Automatische Bestimmung des Schwierigkeitsgrades von Skitouren

A4. Ehrlichkeitserklärung

«Wir versicheren, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der im Literaturverzeichnis angegebenen Quellen und Hilfsmittel angefertigt haben. Die wörtlich oder inhaltlich den im Literaturverzeichnis aufgeführten Quellen und Hilfsmitteln entnommenen Stellen sind in der Arbeit als Zitat bzw. Paraphrase kenntlich gemacht. Diese Bachelor Thesis «Künstliche Intelligenz & Skitouren» ist noch nicht veröffentlicht worden. Sie ist somit weder anderen Interessierten zugänglich gemacht noch einer anderen Prüfungsbehörde vorgelegt worden.»

Fabian Brunner

_____ _____

Ort,Datum Unterschrift

Michel Bittel

_____ _____

Ort,Datum Unterschrift