TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Informatik

Fachgebiet Knowledge Engineering

Maschinelles Lernen zur Hautkrebs-Vorhersage

Bachelorarbeit von Daniel Fischer

Betreuung durch

Prof. Dr. Johannes Fürnkranz

Dipl. Inf. Frederik Janssen

Dr. med. Matthias Herbst

Eidesstattliche Erklärung

Ich versichere an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich, oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Darmstadt, den 17.06.2011

(Daniel Fischer)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

„Von der ursprünglichen Wortbedeutung her (dia: durch, hindurch, auseinander, gnosis: Erkenntnis) ist Diagnostik Erkenntnisgewinnung zur Unterscheidung zwischen Objekten. […]“ (Hossiep & Wottawa, 1993)

Gemäß dieser Definition lassen sich große Parallelen zwischen einer medizinischen Diagnose und verbreiteten Methoden der Informatik ziehen. So erfolgt bspw. eine computergestützte „Diagnose“, bzw. eine Einstufung einer E-Mail automatisch durch den Spam-Filter, der anhand von festgelegten Charakteristika, wie etwa der Anzahl der Rechtschreibfehler, die E-Mail als (Spam-)Mail klassifiziert. Methoden wie diese entstammen allgemein dem Bereich des Maschinellen Lernens und finden in der heutigen Zeit in vielen Softwaresystemen Anwendung (Intrusion Detection, Anti-Viren Programme etc.). Maschinelles Lernen bezeichnet allgemein das Anwenden formaler Strukturen (Maschinen) zur Deduktion und Induktion. Im Gegensatz dazu beschäftigt sich das Data Mining mit der Generierung von Wissen aus Datensätzen und verwendet dafür Methoden des Maschinellen Lernens (Clarke et al., 2009). Dazu werden Algorithmen eingesetzt, die Muster in meist sehr großen Datensätzen erkennen und diese in verschiedenen Darstellungsformen (Regeln, Bäumen etc.) als Domänen-Wissen manifestieren. Damit lässt sich bspw. das Kaufverhalten von Kunden analysieren und eine Aussage darüber treffen, zwischen welchen Produkten gewisse Synergieeffekte bestehen. Die wohl populärste Erkenntnis, die aus der Anwendung von Data Mining resultiert, ist eine Synergie zwischen Windeln und Bier an Wochenendtagen (Clarke et al., 2009). Gehetzte Väter kaufen laut dieser Auswertung Windeln und Bier oft zusammen. Oder es kann eine Aussage darüber getroffen werden, welche Eigenschaften einer menschlichen Embryonalzelle die bestmögliche Überlebenschance für eine künstliche Befruchtung gewährleisten (Witten & Frank, 2005).

Einige Methoden des Data-Mining, die im weiteren Verlauf näher vorgestellt werden, werden in dieser wissenschaftlichen Arbeit auf den vorliegenden Datensatz angewandt. Ziel ist es dabei, Wissen über die unzureichend geklärte Entstehung von Hautkrebs und das damit verbundene Hautkrebsrisiko zu extrahieren, um eine Früherkennung und bestmögliche Heilungschance zu ermöglichen.

1.1 Einführung in das Thema Hautkrebs

Unter dem gängigen Begriff „Hautkrebs“ werden ganz allgemein alle bösartigen Veränderungen der Haut verstanden, die aus unterschiedlichen Zelltypen entstehen (Altmeyer & Bacharach-Buhles, 2002). Dabei ist in einer ersten groben Einteilung zwischen Krebsarten zu unterscheiden, die sich durch Melanozyten (pigmentbildende Zellen der Haut) entwickeln und solchen, die epithelial (nichtmelanozytär) entstehen. Der im Volksmund bekannte und äußerst gefährliche „schwarze Hautkrebs“, das ***maligne Melanom***, gehört dabei zur ersten Gattung. Es bezeichnet einen pigmentierten Hauttumor, der auf den unterschiedlichsten Hautflächen (u.a. auch Schleimhaut, Fuß- und Fingernägel) auftritt. Warnzeichen sind insbesondere die Neuentstehung, oder Veränderung von Pigmentmalen (s. ABCD-Regel, Abb. 2). Die besondere Gefahr des malignen Melanoms liegt dabei in der häufigen Metastasierung, also der „[…]Verschleppung maligner entarteter Zellen eines Primärtumors in andere Organe mit Ausbildung von Tochtergeschwülsten“ (Massalme, 2004) zu der die anderen Hautkrebsarten weniger neigen[1]. Breiten sich Tumorzellen bspw. in umliegende Lymphknoten aus, sinkt die 10-Jahres-Überlebensrate der betroffenen Melanompatienten auf 15 bis 30%. Viel häufiger als das maligne Melanom treten jedoch altersabhängige[2] Vertreter der zweiten Kategorie auf, zu der das ***Basaliom*** (Basaliomkarzinom) und das ***Spinaliom*** (Plattenepithelkarzinom) gezählt werden.

Basaliome sind die häufigste maligne Hauttumorart mit einer 75-80% relativen Häufigkeit. Basaliome sind Tumore, die zumeist im Kopf-Hals-Bereich als kleine, langsam wachsende Knoten auftauchen und lokal das Gewebe zerstören. Mit einer Sterblichkeit von 0,1% der Betroffenen und ihrer geringen Metastasierungswahrscheinlichkeit ist diese Tumorart vergleichsweise ungefährlich. Die gängige Therapie in Form einer vollständigen, operativen Entfernung des Tumors bereitet jedoch gelegentlich Probleme aufgrund der betroffenen Hautfläche (Augennähe etc.).

Unter dem Begriff Spinaliom werden Tumore zusammengefasst, die zu 90% als scharf begrenzte, gerötete Male an Hautregionen beginnen, die dem Sonnenlicht permanent

[1] Die Metastasierungswahrscheinlichkeit beträgt 0,003-0,5% (Basaliom) und 5-6% (Spinaliom). Beim malignen Melanom ist die Tumordicke entscheidend. (Breitbart, Wende, Mohr, Greinert, & Volkmer, 2004).

[2] Häufigkeitsgipfel: 65-69 Jahre (Basaliom) und 70-74 [männliche Patienten] / 75-79 [weibliche Patienten] Jahre (Spinaliom), während 50% der Melanompatienten jünger als 60 Jahre sind

ausgesetzt sind (Gesicht, Ohren, Lippen etc.). Im weiteren Krankheitsverlauf bildet sich eine Verhornung und ein Knoten auf dieser Fläche. Die Letalität des Spinalioms ist mit weniger als 5% ebenfalls eher gering. Bislang bekannte Risikofaktoren der melanozytären und nichtmelanozytären Hautkrebsarten sind in unten stehender Tabelle aufgelistet (Breitbart et al., 2004):

Tabelle 1 - Risikofaktoren

malignes Melanom	***Nichtmelanozytäre Hautkrebsarten***
Kurzzeitig starke UV-Strahlung	Langzeitige UV-Strahlung
Sonnenbrände in Kindheit und Jugend	Aktinische Keratosen
Hauttyp I oder II	Immunsupression[3]
Atypische Pigmentmale (dysplast. NZN)	Epithelialer Hautkrebs in der Eigenanamnese
Angeborenes, großes Pigmentmal	Strahlenschäden
Melanom in der Eigenanamnese[4]	Polyzyklische, aromatische Kohlenwasserstoffe
Mehr als 40 bis 50 gewöhnliche Pigmentmale	
Melanom in der Familienanamnese (1. Grades)	

Das Basaliom, Spinaliom und maligne Melanom stellen aufgrund ihrer kumulierten relativen Häufigkeit von etwa 95% aller Neuerkrankungen (Reinhold & Breitbart, 2007, S. 129-130) die drei Hauptvertreter des Hautkrebses dar. Weitere Arten sind das *Merkelzellenkarzinom*, das *Karposi-Sarkom* und das *kutane Lymphom*, die allesamt wegen ihres vergleichbar geringen Stellenwertes in der nachfolgenden Untersuchung keine Berücksichtigung finden. Einen umfassenden Überblick über die Krankheitsverteilungen im vorliegenden Untersuchungsdatensatz findet sich im Anhang (Kapitel 7) als Häufigkeitstabellen.

[3] Überbegriff für Verfahren bei denen „Immunsuppressiva“ eingesetzt werden. Diese Medikamente unterdrücken eine Immunantwort des Körpers [Bsp.: Antibiotika, Zytostatika (Krebsmittel) etc.] (Altmeyer & Bacharach-Buhles, 2002)

[4] „Das Wort ‚Anamnese‘ stammt aus dem Griechisch-Lateinischen, bedeutet ‚Erinnerung‘ und wird im medizinischen Kontext als Vorgeschichte einer Krankheit gesehen“ (Duden, 2001)

Die Dermatologie kategorisierte desweiteren eine Reihe von Gewebeänderungen, die den Ausgangspunkt zur Bildung genannter Hautkrebsarten darstellen können. Diese Krebs-Vorstufen werden unter dem Begriff *Präkanzerosen* subsumiert. Eine häufige Unterart der Präkanzerosen, die in dieser Untersuchung erhoben wurde, ist die ***aktinische Keratose***. Eine Existenz dieser Hauterkrankung geht mit einem Risiko von 20-25% auf Spinaliome einher (Reuter, 2004). Eine indirekte Vorstufe des malginen Melanoms sind ***dysplastische Nävi***, also „Muttermale", die eine entartete Form aufweisen und oft den Übgergang von melanozytären Nävi zum malignen Melanom darstellen (Stolz et al., 2001, S. 76 ff.). Basaliome hingegen besitzen keinerlei Vorläuferform (Breitbart et al., 2004, S. 23).

Auswertungen des Robert-Koch-Instituts der Daten des saarländischen Krebsregisters, das als einziges deutsches Krebsregister seit 1970 kontinuierlich Daten bereithält, ergaben einen signifikanten Trend, der in Abb. 1 zu verfolgen ist:

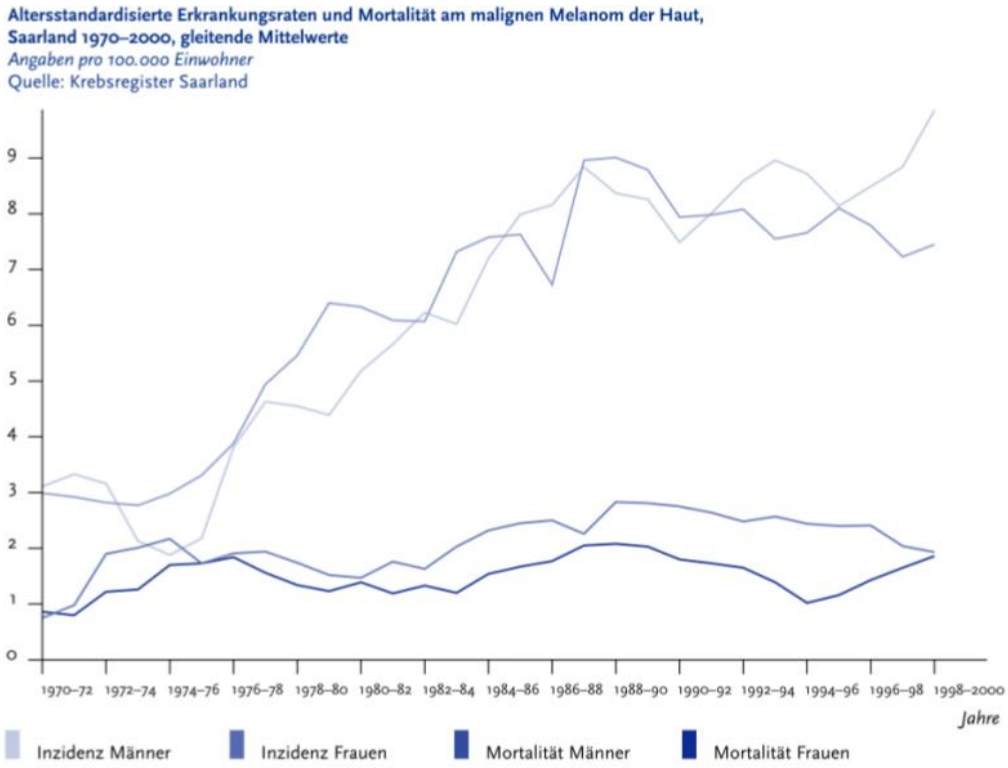


Abbildung 1 – Inzidenz und Mortalität beim malignen Melanom

Die Inzidenzen (Neuerkrankungsraten) aller drei Hautkrebsarten[5] steigen stetig an, während die Mortalitätsrate beim malignen Melanom nahezu stagniert und bei epithelialen Hautkrebsarten[6] (Basaliom, Spinaliom) sogar sinkt. Zurückzuführen ist dieser, auch international zu beobachtende, Sachverhalt (Breitbart et al., 2004) auf eine verbesserte Früherkennung und den Erfolg zahlreich durchgeführter Aufklärungskampagnen über Krebswarnzeichen. Die diagnostische Früherkennung von Hautkrebs gehört zur sekundären Prävention, während Aufklärungsarbeit in der Bevölkerung zur primären Prävention gezählt wird, da die darin aufgezeigten Möglichkeiten jederzeit und ohne ärztliche Aufsicht stattfinden können. Ein Beispiel für erfolgreiche Aufklärungsarbeit ist die *ABCD(E)*[7]*-Regel*, anhand derer Patienten gewöhnliche Nävi (Leberflecken) durch optische Erkennungszeichen in einer ersten, groben Selbsteinschätzung von Ausprägungen eines malignen Melanoms selbst abgrenzen können (Deutsche Krebshilfe e.V., 2008):

- ***A*** (Asymmetrie): Unregelmäßige Formen, Abgrenzungen eines Mals
- ***B*** (Begrenzung): Unebene, nicht klare Rand-Abgrenzung
- ***C*** (Colour): Die Farbe erscheint nicht einheitlich (heller und dunkler abwechselnd)
- ***D*** (Durchmesser): Male mit einem Durchmesser von ≥ 2mm

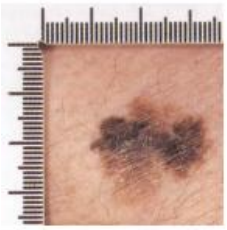


Abbildung 2 - ABCD

Die Attribute *ABCD(E)* ergaben sich aus einer multivariaten Analyse von 31 dermatoskopischen Kriterien (Stolz et al., 2001). Der Erfolg dieser Maßnahmen trägt maßgeblich zur Senkung der Mortalitätsraten bei. Über eines herrscht breite Gewissheit in der dermatologischen Praxis: Hautkrebs ist bei Erkennung in einem Frühstadium nahezu immer heilbar (Breitbart et al., 2004). In der Hoffnung an diese Präventionserfolge anzuknüpfen und sie weiter auszubauen wurde diese Arbeit verfasst. Das Verständnis des Risikos (Mortalitätsraten, Metastasierungswahrscheinlichkeiten) und des Zusammenspiels der einzelnen Hautkrebsarten (Praecancerosen bedingen Spinaliom / malgines

[5] Aus Platzgründen wurde auf die Darstellung der Inzidenzen und Mortalität von Basaliom und Spinaliom verzichtet. Ausführliche Statistiken, die gleichartige Entwicklungen (steigende Inzidenzen bei stagnierenden, oder rückläufigen Mortalitätsraten) belegen sind in (Breitbart, Wende, Mohr, Greinert, & Volkmer, 2004) zu finden

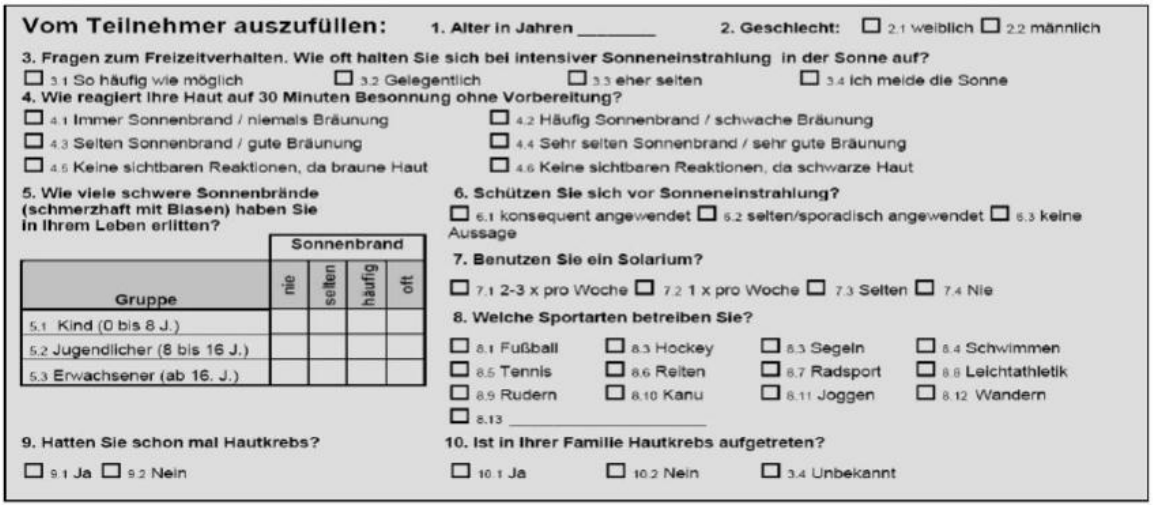
[6] Epithelialer Hautkrebs ist auch als *'weißer Hautkrebs'* bekannt

[7] Das „E“ steht für Erhabenheit (rasche Veränderung bzw. Vergrößerung) und wird in einigen Umsetzungen der ABCD-Regel noch mitgeführt (Garbe, 2006)

Melanom) ist dabei für spätere Modellierungszwecke im Data-Mining (Kapitel 5) elementar. Zusätzlich verschafft der Überblick über bisherig bekannte Risikofaktoren (Tabelle 1) die Möglichkeit die Ergebnisse der Experimente zu verifizieren.

1.2 Herkunft der Daten

Die notwendigen Daten zur Untersuchung auf Zusammenhänge zwischen Hautkrebserkrankungen und Merkmalen bzw. Verhaltensweisen der Betroffenen, lieferte ein „Hautcheck-Programm“ der Qualitätsgemeinschaft südhessischer Dermatologen e.V., das im Januar 2006 startete. Ziel dieses Programms war die Verbesserung der Früherkennung von Hautkrebs. Hierzu wurden bundesweit etwa 7.000 Fragebögen (Abb. 1) und durch Patienten ausgefüllt und ärztliche Untersuchungen (Abb. 2) an jedem teilnehmenden Patient durchgeführt. Der Fragebogen erhob Informationen über das Alter, Geschlecht, Krankheitsgeschichte und Freizeitverhalten der Teilnehmer. Neben einer Ganzkörperuntersuchung wurden Patienten im Zuge der Hautkrebsvorsorge über Präventionsmaßnahmen, Hauttyp, Lichtschutz und Pflegemaßnahmen durch den behandelnden Dermatologen aufgeklärt. Patienten mit einem konkreten Hautkrebs-Risiko, oder bereits vorliegendem Hautkrebs, wurden zusätzlich über die erforderlichen Folgeschritte unterrichtet. In akuten Fällen fanden deshalb im Laufe des „Hautchecks“ mehrere (Folge-)Untersuchungen eines Patienten durch teils verschiedene Ärzte statt.



Vom Teilnehmer auszufüllen: 1. Alter in Jahren _____ 2. Geschlecht: ☐ 2.1 weiblich ☐ 2.2 männlich

3. Fragen zum Freizeitverhalten. Wie oft halten Sie sich bei intensiver Sonneneinstrahlung in der Sonne auf?
☐ 3.1 So häufig wie möglich ☐ 3.2 Gelegentlich ☐ 3.3 eher selten ☐ 3.4 Ich meide die Sonne

4. Wie reagiert Ihre Haut auf 30 Minuten Besonnung ohne Vorbereitung?
☐ 4.1 Immer Sonnenbrand / niemals Bräunung
☐ 4.2 Häufig Sonnenbrand / schwache Bräunung
☐ 4.3 Selten Sonnenbrand / gute Bräunung
☐ 4.4 Sehr selten Sonnenbrand / sehr gute Bräunung
☐ 4.5 Keine sichtbaren Reaktionen, da braune Haut
☐ 4.6 Keine sichtbaren Reaktionen, da schwarze Haut

5. Wie viele schwere Sonnenbrände (schmerzhaft mit Blasen) haben Sie in Ihrem Leben erlitten?

	Sonnenbrand			
Gruppe	nie	selten	häufig	oft
5.1 Kind (0 bis 8 J.)				
5.2 Jugendlicher (8 bis 16 J.)				
5.3 Erwachsener (ab 16. J.)				

6. Schützen Sie sich vor Sonneneinstrahlung?
☐ 6.1 konsequent angewendet ☐ 6.2 selten/sporadisch angewendet ☐ 6.3 keine Aussage

7. Benutzen Sie ein Solarium?
☐ 7.1 2-3 x pro Woche ☐ 7.2 1 x pro Woche ☐ 7.3 Selten ☐ 7.4 Nie

8. Welche Sportarten betreiben Sie?
☐ 8.1 Fußball ☐ 8.3 Hockey ☐ 8.3 Segeln ☐ 8.4 Schwimmen
☐ 8.5 Tennis ☐ 8.6 Reiten ☐ 8.7 Radsport ☐ 8.8 Leichtathletik
☐ 8.9 Rudern ☐ 8.10 Kanu ☐ 8.11 Joggen ☐ 8.12 Wandern
☐ 8.13 _____

9. Hatten Sie schon mal Hautkrebs?
☐ 9.1 Ja ☐ 9.2 Nein

10. Ist in Ihrer Familie Hautkrebs aufgetreten?
☐ 10.1 Ja ☐ 10.2 Nein ☐ 3.4 Unbekannt

Abbildung 1 - Patientenfragebogen

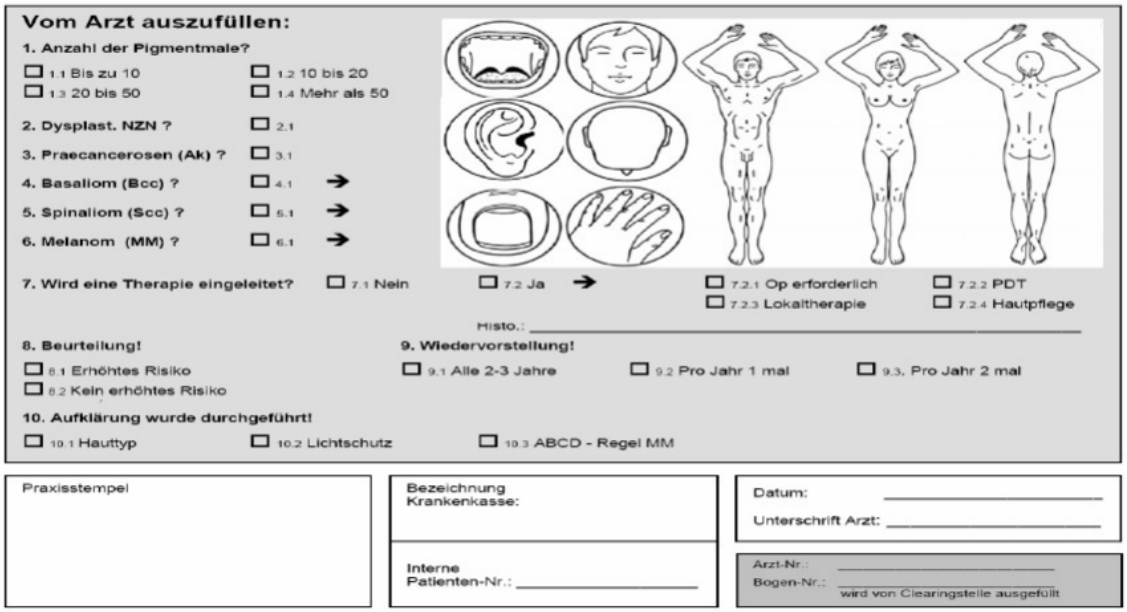


Abbildung 2 - Ärztefragebogen

Finanziert wurde der „Hautcheck“ über die Berufskrankenkassen der Merck KG, sowie des HEAG Konzerns, um ihre Kunden- / Patientenzufriedenheit zu steigern.

Im Anschluss an die Befragungen und Untersuchungen wurden die Daten durch die Iatrocon GmbH als Dienstleister in ein Excel-Sheet überführt und erste statistische Kennzahlen wie Häufigkeiten, Varianzen und Standardabweichungen des Datensatzes ermittelt. Ausgangspunkt dieser Arbeit ist der Datensatz nach seiner Aufbereitung durch die Iatrocon GmbH als .xslx-Datei *(Microsoft Excel 2007).*

1.3 Ziel und Aufbau der Arbeit

Das Ziel dieser Arbeit ist die Ermittlung und anschließende Verifikation von Modellen zur Klassifikation eines Hautkrebs-Risikos. Wie eingangs erwähnt, übernimmt die Prävention durch Früherkennung und Aufklärung eine Schlüsselrolle in der Bekämpfung von Hautkrebs. Bei nahezu fehlerfreier Klassifikation wäre ein Einstufungssystem für Patienten in Risikoklassen der nächste, aufbauende Schritt. Dieses System würde einerseits die ärztliche Arbeit unterstützen und zudem die Patienten für ein eventuelles Risiko sensibilisieren.

Hierfür werden zunächst die *Grundlagen des Data Mining (Kapitel 2)* abgedeckt, um Begrifflichkeiten zu definieren und das System des Data Mining näher zu erläutern. Im Anschluss daran werden die erforderlichen Schritte des *Data Preprocessing (Kapitel 3)*, der Daten-Vorbereitung zur weiteren Verwendung, behandelt. Kapitel 3 beinhaltet desweiteren die Kurzbeschreibung eines Programms zur Konvertierung von *.csv*-Dateien in das *.arff*-Format. Nachfolgend wird auf die verwendeten Algorithmen der Lern-Methoden im Kapitel *Algorithmen des Data Mining (Kapitel 4)* eingegangen. Die Ergebnisse und Effizienz dieser Algorithmen für den Datensatz des „Hautchecks“ werden im nächsten Kapitel, *Experimente (Kapitel 5)*, festgehalten. Abschließend wird in *Diskussion und Ausblick (Kapitel 6)* der Ablauf der Arbeit resümiert und daraus mögliche, zukünftige Schritte geschlussfolgert.

2 Grundlagen des Data-Mining

"We are drowning in information, but starving for knowledge." (Rutherford D. Roger)

Im jetzigen, digitalen Zeitalter scheint unser Gedächtnis schon längst durch die überall gegenwärtige Datenflut überholt. Daten hinterlassen bei jedwedem Einkauf, beim Internet-Surfen und vielen anderen Aktivitäten unsere Fußspuren in diversen Datenbanken. Begünstigt wird diese Digitalisierungs-Entwicklung durch zunehmend günstigeren Speicher und leistungsfähigere Hardware im Allgemeinen. Riesige Datenmengen alleine schaffen jedoch keinerlei Mehrwert, wenn nicht Informationen und Wissen daraus extrahiert werden können. Mit der Datenmenge wächst offensichtlich jedoch nicht das Verständnis über selbige. Dieser Wunsch, Sachverhalte zu analysieren und Erkenntnisse aus ihnen zu gewinnen, ist so alt wie die Menschheit. Ohne derartige Fähigkeiten wäre ein Lernprozess schlichtweg nicht möglich gewesen. In der Vergangenheit wurden Daten überwiegend manuell analysiert und ausgewertet, wie bspw. in den Anfängen der Statistik. Bei Datenbeständen geringer Größe stellt eine Auswertung per Hand noch keine Herausforderung dar. Die heutigen Datenbanken mit z.T. mehreren millionen Einträgen sind jedoch durch manuelle Bearbeitung nicht mehr zu bewältigen. Data Mining befasst sich deshalb mit der Automatisierung des Lernprozesses. Eine gute Definition für Data-Mining liefern Hand, Mannila und Smyth:

„Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner." (Hand et al., 2001)

Data Mining ist die Analyse von (häufig großen) beobachteten Datenmengen, um unerwartete Beziehungen zu finden und die Daten in einer neuartigen Weise zusammenzufassen, die sowohl verständlich, als auch nützlich für den Datenbesitzer ist.

Moderne Algorithmen des Maschinellen Lernens analysieren Datensätze auf logische und funktionale Zusammenhänge und liefern in kürzester Zeit meist erstaunliche Ergebnisse.

2.1 Der „Knowledge-Discovery in Databases“ (KDD) Prozess

Um das weitere Vorgehen dieser Arbeit und den großen Rahmen der Wissensentdeckung aus Datensätze (KDD) zu erläutern, werden im folgenden Abschnitt KDD-Prozessmodelle eingeführt und eines dieser Modelle detaillierter beschrieben.

Der komplexe KDD-Prozess kann formal in mehrere, vereinfachte Teilschritte zerlegt werden. Hierzu wurden von verschiedenen Parteien Prozessmodelle definiert, um ein einheitliches Vorgehen für jedes Data-Mining-Projekt zu etablieren. Ziel dieser Prozessmodelle war neben einer Vereinfachung und Verbesserung der Planung, Ausführung und Kontrolle von Data-Mining-Anwendungen, das Etablieren von Standards bezüglich der einzelnen Ergebnisse der Prozessschritte, sowie ein Hinweis auf übliche Probleme in KDD-Projekten (Kietz, 2009). In diesem Sinne entwickelte bspw. ein Konsortium aus Mitarbeitern der Unternehmen *Daimler Chrysler*, *SPSS Inc.*, *OHRA Bank Groep B.V.* und *NCR Systems Copenhagen* in den Jahren 1996-1999 den *CRoss Industry Standard Process for Data Mining (CRISP)*, ein Quasi-Standard-Prozessmodell. Andere bekannte Prozessmodelle wurden durch Han (Han & Kamber, 2006) und Fayyad (Fayyad et al., 1996) vorgestellt. Trotz den ausgesetzten Entwicklungsarbeiten an *CRISP 2.0*, dessen Weiterentwicklung im Juli 2006 angekündigt wurde (CRISP-DM-Konsortium, 2007), findet CRISP im industriellen und privaten Sektor weitverbreitete Anwendung (KDnuggets, 2007). Aus diesem Grund wird im Folgenden das *CRISP-Modell* näher erläutert.

Das *CRISP-Modell* stellt ein hierarchisches Prozessmodell bezüglich vier Abstraktionsebenen dar: *Phases*, *Generic Tasks*, *Specialized Task* und *Process Instances.*

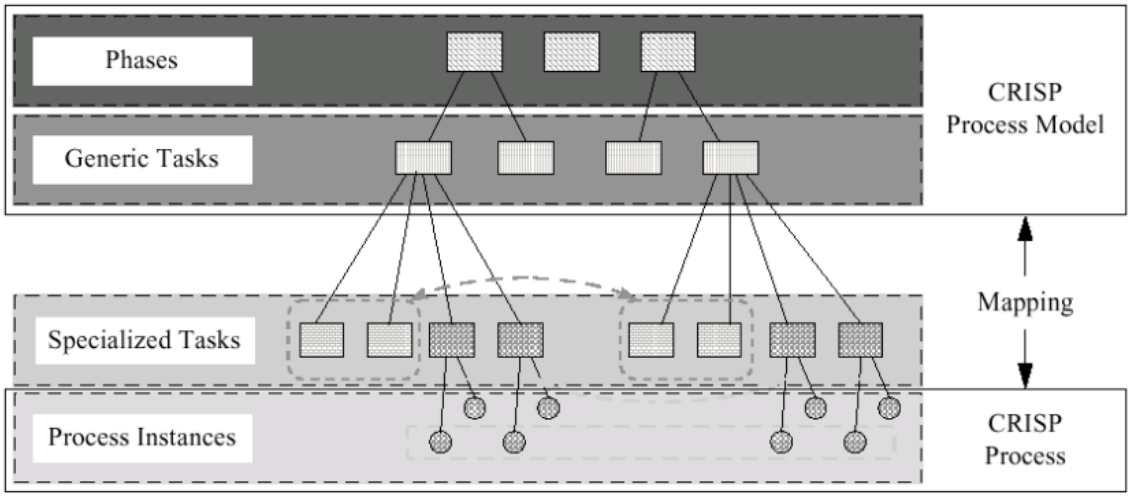


Abbildung 3 - CRISP Abstraktionsebenen (CRISP-DM-Konsortium, 2006)

Die Ebene *Phases* unterteilt den Prozess des Data Mining in sechs allgemeine Phasen, die im weiteren Verlauf noch erläutert werden. *Generic Tasks* bezeichnet im Anschluss daran die zweite Ebene, in der es um die Aufzählung und Beschreibung der möglichen Aufgaben jeder Phase geht. Diese Aufgaben sollten gemäß der Richtlinien vollständig (*complete*) und robust (*stable*) sein. Gemeint ist hiermit, dass eine vollständige Abdeckung des Data-Mining-Prozesses und aller Data-Mining-Anwendungen stattfinden muss und das bisherige Modell seine Validität auch bei zukünftigen Modellierungsänderungen beibehalten soll. In der dritten Ebene, *Specialized Tasks*, erfolgt schließlich die Abbildung der allgemeinen Aufgaben aus der vorigen *Generic-Tasks*-Ebene auf die individuellen, spezialisierten Aufgabengebiete. Beispielsweise spezifiziert sich der allgemeine Schritt des *Data-Cleaning* je nach Attributtyp der konkreten Anwendung zu einem numerischen oder nominalem *Data-Cleaning*. Abschließend enthält die letzte Phase, *Process Instances*, zu Dokumentationszwecken Aufzeichnungen der Aktionen, Entscheidungen und Ergebnisse einer realen Prozess-Instanz (Chapman et al., 2000).

Die zentralen Abschnitte der *Phases*-Ebene sind: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* und *Deployment*.

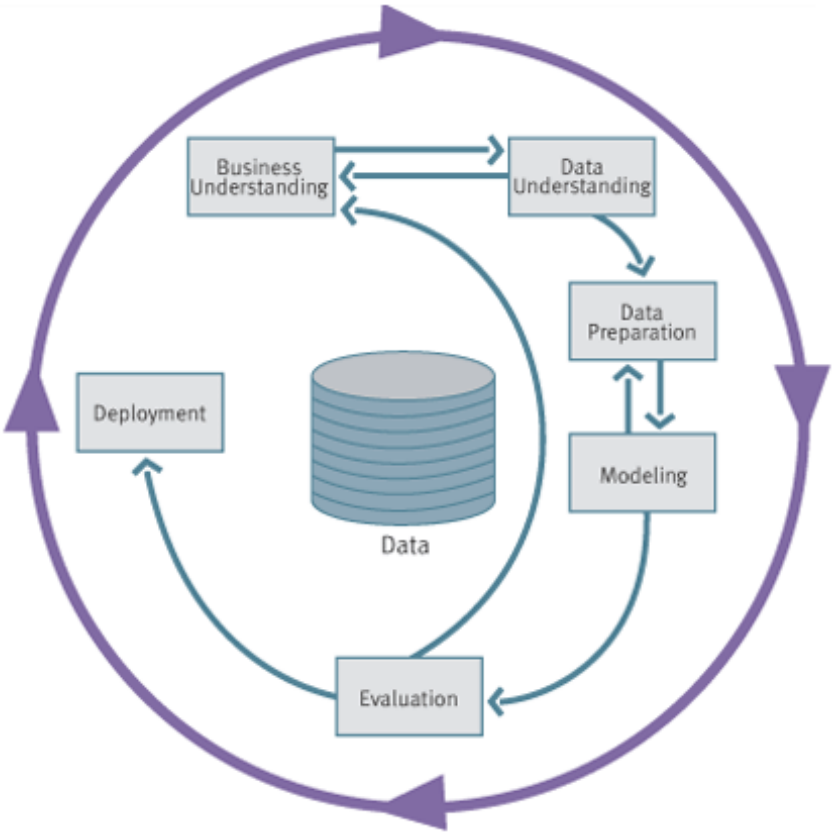


Abbildung 4 - CRISP Phasen (CRISP-DM-Konsortium, 2006)

Nachstehend werden die Hauptaufgaben jedes Abschnitts kurz zusammengefasst. Im Schritt *Business Understanding* soll der Data-Mining-Anwender sich das Hintergrundwissen über den realen (Geschäfts-)Prozess aneignen und in Anlehnung an die Projektziele- und Anforderungen eine Data-Mining Problemstellung ausarbeiten. Im Anschluss daran erfolgt während des *Data Understanding* eine Analyse der Ausgangsdaten im Hinblick auf mögliche Qualitätsprobleme, erste Hypothesen und versteckte Informationen. Diese Vorbereitung dient der nächsten Aufgabe, der *Data Preparation*, in der Datensätze und Attribute ausgewählt, bereinigt und ggf. transformiert werden. In der Modellierungsphase *Modeling* werden diverse Data-Mining-Techniken auf den Datensatz angewandt und optimale Parameter bestimmt. Typischerweise werden mehrere Methoden zur Modellbildung genutzt, die oft bestimmte Anforderungen an die Daten stellen. Methoden, die auf einer linearen Regression basieren, verlangen bspw., dass alle Attribute in numerischer Form vorliegen. Um diesen Ansprüchen gerecht zu werden ist in solch einem Fall deshalb eine Rückkehr zur *Data Preparation* notwendig (Chapman et al., 2000).

Das *CRISP*-Modell, in seiner vorgestellten Form als Phasenmodell, definierte den weiteren Ablauf der Arbeit. Die Phase *Business Understanding* wurde hierbei durch Kapitel 1.1 umgesetzt, um einen groben Überblick über Hautkrebsarten, Relevanz und bisherig bekannte Risikofaktoren zu verschaffen. Die darauffolgenden beiden Phasen, *Data Understanding* und *Data Preparation*, finden sich nach einer Einführung in die Data-Mining Grundlagen in Kapitel 3 wieder. Kapitel 3 geht dabei detailliert auf Qualitätsprobleme der Daten und qualitätssteigernde Maßnahmen ein. *Modelling* und *Evaluation* sind abschließend durch Experimentreihen an drei Modellen in Kapitel 5 dargestellt.

2.2 Klassifikation

Es lassen sich hauptächlich vier Arten des Lernens in Data Mining-Anwendungen unterscheiden: Die *Klassifikation*, die *Assoziation*, das *Clustering* und die *numerische Vorhersage*. Assoziatives Lernen beschäftigt sich mit den Zusammenhängen aller Attribute untereinander und erweitert damit die Klassifikation, die sich auf einzelne Attribute der Relation fokussiert und Zusammenhänge zwischen der Ausprägung dieser Attribute und dem Rest darstellt. Im Clustering werden Beispieldatensätze anhand von Ähnlich-

keiten in Gruppen zusammengefasst und eine numerische Vorhersage stellt bei einer Relation, die ausschließlich auf numerischen Werten basiert, ein Funktionsterm auf, mit dem ein Attributswert eines neuen Beispiels berechnet werden kann. Nachfolgend wird die Klassifikation näher betrachtet, da diese essenziell für die Problemstellung der Einstufung eines Hautkrebsrisikos ist.

Den Ausgangspunkt für die *Klassifikation* bildet ein Datensatz mit klassifizierten Beispielen[8]. Die Klassifikation bezieht sich meist auf ein Attribut. Die Ausprägung dieses Attributs muss in dem Trainingsdatensatz enthalten sein. Anschließend versucht ein Algorithmus Muster in den Beispielen zu finden, die ein Klassifikationsmodell ermöglichen. Diese Muster oder Regeln werden schließlich verwendet, um neue, unklassifizierte Beispiele einzuordnen. Ein stark vereinfachtes, fiktives Beispiel ist der unten stehenden Tabelle *(Tabelle 1)* zu entnehmen. Hierbei wird versucht anhand einiger Umweltzustände wie der Fitness der Spieler, der Stärke des Gegners etc. den Ausgang der nächsten Spiels eines Fußballvereins zu schätzen. Dazu wurden an vergangenen Spieltagen die entsprechenden Beobachtungen in der Tabelle Fußballspiel dokumentiert. Auf Basis dieser Beobachtungen wurde ein Klassifikationsmodell bestimmt, das in Tabelle 2 zu finden ist.

Tabelle 2 - Fußballspiel

Fitness der Spieler	Stärke des Gegners	Heimspiel	Sieg im letzten Spiel	Ergebnis
fit	stark	nein	nein	Niederlage
ausgewogen	schwach	ja	ja	Sieg
erschöpft	stark	nein	nein	Niederlage
fit	stark	nein	ja	Sieg
erschöpft	schwach	nein	nein	Unentschieden
erschöpft	stark	ja	nein	Niederlage
ausgewogen	schwach	ja	nein	Unentschieden
erschöpft	schwach	ja	ja	Sieg
fit	schwach	nein	ja	Sieg
fit	stark	nein	nein	Unentschieden
fit	schwach	ja	ja	Sieg
ausgewogen	stark	nein	nein	Niederlage
ausgewogen	stark	nein	ja	Unentschieden
erschöpft	stark	nein	nein	Niederlage
ausgewogen	stark	ja	ja	?

[8] Das Lernen anhand von klassifizierten Beispielen wird auch als *„supervised learning“* bezeichnet, da die Ausprägung der Werte in der Trainingsmenge 'überwacht' wird. *„Unsupervised learning“* bildet demzufolge das Gegenstück mit unbekannten Merkmalswerten, wird aber in dieser Arbeit nicht näher erläutert. Eine gute Einführung bietet (Witten & Frank, 2005, S. 254-271)

Zeile eins der Tabelle 1 enthält die ***Attribute*** der Objekte, in diesem Fall die Fitness der Spieler, die Gegnerstärke usw. Die darauffolgenden Zeilen bestehen aus Beispieldatensätzen mit konkreten Ausprägungen der Attribute. Im weiteren Verlauf der Arbeit wird jedoch von ***Instanzen*** gesprochen. Die Menge von Instanzen, die zur Klassifikation benutzt wird, wird als ***Trainingsmenge*** bezeichnet (Witten & Frank, 2005, S. 45). Das Attribut, dessen Konzept erlernt werden soll (in diesem Fall *Ergebnis*) wird als ***Klassenattribut*** bezeichnet. Gesucht wird im Zuge der Klassifikation eine Zuordnung von neuen, unbekannten Instanzen zu einer Ausprägung des Klassenattributs. Im Beispiel des *„Fußball-Problems"* lässt sich dieses Zuordnungsproblem in der letzten Zeile erkennen. Es stellt sich nun die Frage, ob die unbekannte Mannschaft unter den gegebenen Zuständen (ausgewogen, stark, ja, ja) ihr nächstes Spiel gewinnt, verliert, oder unentschieden spielt. Eine beispielhafte Repräsentation, in Form eines Entscheidungsbaums, eines erlernten Modells ist unten abgebildet:

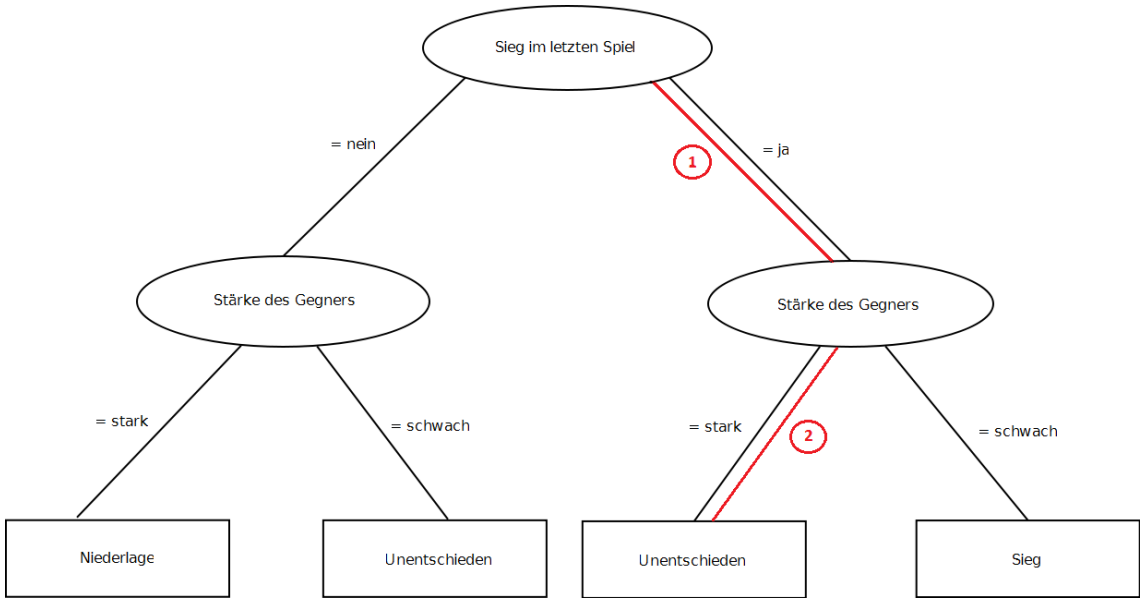


Abbildung 5 - Entscheidungsbaum Fußball-Beispiel

Bei Anwendung des neuen, unklassifizierten Beispieldatensatzes auf das Modell, ist ersichtlich, dass der Ausgang des Spiels „Unentschieden" lauten müsste. Die Traversierung des Baums ist rot markiert und in Teilschritten dargestellt. Der erste Schritt zur Klassifikationsentscheidung ergibt sich durch das Attribut *Sieg im letzten Spiel*. Da die Ausprägung der neuen Instanz in diesem Attribut „ja" entspricht, wird für die weitere

Traversierung der rechte Teilbaum ausgewählt. Die *Stärke des Gegners*, in diesem Fall „stark“, führt schließlich zur Klassifikationsentscheidung „Unentschieden“.

Allgemeiner gefasst lässt sich der Klassifikationsprozess gemäß Han und Kamber (Han & Kamber, 2006, S. 286-288) in 2 Teilabschnitte gliedern:

1. Lernschritt:

   Innerhalb des Lernschritts wird ein Klassifikationsmodell auf Basis der Trainingsdaten $\boldsymbol{T}$ erstellt. Eine Instanz $\boldsymbol{X} \in \boldsymbol{T}$ aus einem Trainingsmenge mit $\boldsymbol{N}$ Instanzen wird dabei durch einen *n*-dimensionalen Attributvektor $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{y}, \boldsymbol{x_n})$ dargestellt. Die zugehörigen Attribute sind $\boldsymbol{A_1}, \boldsymbol{A_2}, \dots, \boldsymbol{A_n}$. Das zu erlernende Attribut $\boldsymbol{y}$ heißt Klassenattribut und muss beim Supervised Learning bei jeder Instanz vorhanden sein. Ziel ist es, ein Mapping, bzw. eine Funktion, $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X})$ zu finden, die eine Vorhersage des Klassenattributs für jede neue Instanz ermöglicht und damit $\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_N}$ bestmöglich in die Ausprägungen des Klassenattributs unterteilt. Übliche Formen eines solchen Modells sind Klassifikationsregeln, Entscheidungsbäume, oder mathematische Formeln, wie sie bei Regressionsmethoden verwendet werden.[9]
2. Klassifikationsschritt:

   Im Klassifikationsschritt wird das erstellte Klassifikationsmodell zur Vorhersage des Klassenattributes genutzt. Zur Einschätzung der Güte dieser Klassifikation wird die Genauigkeit verwendet.

Die ***Genauigkeit*** $\boldsymbol{g}$ eines Modells entspricht der Rate der korrekt klassifizierten Instanzen auf den Trainingsdaten. Wenn $\boldsymbol{E}$ die Anzahl erfolgreich klassifizierter Instanzen bei einer Gesamtmenge von $\boldsymbol{N}$ Instanzen bezeichnet, gilt folglich: $\boldsymbol{g} = \frac{\boldsymbol{E}}{\boldsymbol{N}}$ Zur Evaluation sollte jedoch nicht die Menge an Trainingsinstanzen herangezogen werden, da auf diesen bereits das Modell gelernt wurde und eine Vorhersage daher regelmäßig sehr optimistisch ausfällt und lediglich als obere Schranke anzusehen ist. Vielmehr wird versucht, die aus den Trainingsdaten gewonnenen Erkenntnisse auf neue, unbekannte Daten anzuwenden und die daraus resultierende Genauigkeit als Gütekriterium des Modells festzulegen. Hierfür existieren zwei Ansätze:

[9] In Kapitel 4 werden die Arten und Darstellungen näher behandelt

Falls die Menge an verfügbaren Daten ausreichend groß ist, lässt sich eine Aufteilung der Daten in Trainings- und Testdaten vornehmen. Das auf der Trainingsmenge erlernte Modell kann somit auf einer unabhängigen, neuen Testmenge getestet werden. In der Praxis üblich ist eine Trennung im Verhältnis von 2:1 in Bezug auf die Trainingsdaten (Witten & Frank, 2005, S. 149).
Ein anderes Vorgehen beschreibt die ***k-fache Kreuzvalidierung***, die geeignet ist, falls der Datensatz keine hinreichende Größe erreicht, oder ein Lernen auf der kompletten Menge erwünscht ist.

- Die Menge an Trainingsdaten $\boldsymbol{T}$ wird dazu in $\boldsymbol{k}$ gleich große, zufällige Teile unterteilt
- Es beginnen $\boldsymbol{k}$ Durchgänge:
    - Auf den Trainingsdaten $T_1, T_2, \dots, T_{i-1}, T_{i+1}, \dots, T_k$ wird ein Modell M gelernt
    - M wird durch die Restmenge, bzw. die Testdaten T_i evaluiert und die Genauigkeit g_i wird bestimmt
    - i wird inkrementiert
- Als finales Performanzmaß des Modells wird die durchschnittliche Genauigkeit $\bar{g} = \frac{\sum_{i=1}^{k} g_i}{k}$ ermittelt

Als geeignete Wahl für *k* stellte sich durch unzählige Experimente die Zahl zehn heraus. Statistische Beweise für diese Empfehlung existieren jedoch nicht (Witten & Frank, 2005, S. 150). Die Zufälligkeit der Aufteilung von $\boldsymbol{T}$ beeinflusst das Ergebnis, weshalb es in der Praxis üblich ist, eine *k-fache Kreuzvalidierung* zehnmal hintereinander mit jeweils neuer Zufallsaufteilung durchzuführen und wiederum das Mittel aus diesen Versuchen als Genauigkeit zu wählen. Als Performanzmaß für die Experimentreihen in Kapitel 5 wurde deshalb eine *10-malige 10-fache Kreuzvalidierung* gewählt.

Eine annähernd 100%-ige Performanz auf den Trainingsdaten lässt meist auf eine negative Problematik schließen, die sich ***Overfitting*** nennt. Es bezeichnet die Überanpassung des Modells an die Trainingsdaten und ist meist durch ein komplexes Klassifikationsmodell und schlechte Ergebnisse auf unabhängigen Testdaten gekennzeichnet. Mit diesen Modellen lassen sich die Trainingsdaten, jedoch ausschließlich diese, sehr genau klassifizieren. Bei unbekannten Instanzen sinkt die Klassifikationsgenauigkeit signifi-

kant, da das erlernte Konzept zu stark durch die Trainingsdaten und deren mögliches, zufälliges Rauschen geprägt wurde. In den Kapiteln 3 und 4 werden durch die *Feature-Subset Selection* und das *Pruning* Techniken vorgestellt um Overfitting zu vermeiden (Falkenauer, 1998).

3 Data Preprocessing

Die Ergebnisse eines Data-Mining Projekts hängen signifikant von der Qualität der Ausgangsdaten ab. In den meisten Fällen werden diese Daten jedoch in einer mangelhaften Form vorliegen. Ziel des Data Preprocessing ist es, die Daten, die häufig unvollständig (fehlende Attributwerte), verzerrt (enthaltene Ausreißer), oder inkonsistent (widersprüchliche Attributausprägungen, keine einheitlichen Schreibweisen etc.) sind, so aufzubereiten, dass ihre Qualität steigt und damit auch effizientere Modelle schneller gefunden werden. Dafür stehen eine Reihe von Methoden zur Auswahl, die im weiteren Verlauf aufgezeigt werden (Han & Kamber, 2006).

3.1 Aufbereitung und Kodierung der Daten

Die Grundlage dieser Arbeit bildete eine *.csv*-Datei, die die Auswertungen der Patienten-Fragebögen und ärztlichen Untersuchungen beinhaltete. Die Darstellung und Formatierung der Daten erfolgte in Anlehnung an die grundsätzliche Struktur eines Fragebogens und war für eine weitere Verarbeitung, insbesondere zum Zwecke des Data Mining, unvorteilhaft. Aus diesem Grund waren einige Schritte der Vorverarbeitung nötig, um die Dateien in eine geeignete Form zu überführen. Eine grobe Skizzierung des *.csv*-Aufbaus wird in folgender Darstellung veranschaulicht.

Tabelle 3 - .csv im Ursprung

	A	B	C	D	E	F	G	H
1	26.10.2009	102	10010750	55		x		x
2	26.10.2009	102	10010752	52	x			x
3	26.10.2009	102	10010747	40		x		x
4	26.10.2009	102	10010749	58	x			x
5	26.10.2009	102	10010748	65	x		x	
6	26.10.2009	102	10010746	62	x			x
7	26.10.2009	102	10010751	60	x			x
8	26.10.2009	10	10010753	52		x		
9	26.10.2009	10	10010754	49	x			x
10	26.10.2009	10	10010755	23	x			x
11	26.10.2009	10	10010756	64	x			
12	26.10.2009	10	10010757	54		x		x
13	26.10.2009	10	10010758	37		x		
14	26.10.2009	10	10010759	53		x		
15	24.10.2009	19	10010744	41		x		
16	24.10.2009	19	10010738	61	x			x
17	24.10.2009	25	10010735	66	x			x

Ein „x“ repräsentiert in dieser Form eine zutreffende Aussage durch einen Patient, oder Arzt. Dabei kann eine Antwort einer von drei Kategorien zugeordnet werden:

1. Felder enthalten vollständigen (intervallskalierten / verhältnisskalierten) Wert
    a. Vorwiegend bei numerischen Merkmalen (Alter, Arzt-Nummer, Datum etc.)
2. Felder ('x') kodieren nominalen / kategorialen / ordinalen Merkmalswert
    a. Üblicherweise für Fragen mit vielen Antwortmöglichkeiten (Häufigkeit der Solarium-Nutzung etc.)
3. Felder ('x') kodieren binären, nominalen Merkmalswert
    a. Typische Ja-/Nein-Fragen (Wird Sport ausgeübt etc.)

Zunächst mussten die Bezeichnungen der Attribute und Merkmalswerte den Platzhaltern ('x') zugeordnet werden. Dies lässt sich, ähnlich der weiteren Vorverarbeitung, sehr schnell durch die Verwendung von Formeln in *MS Excel 2007* realisieren. Um diese Ersetzung und eine direkte Konvertierung in das .arff-Format zu automatisieren, wurde das Programm .csv2arff entwickelt, das im *Kapitel 3.3* näher vorgestellt wird.

Bei der Darstellung der Attributausprägungen muss zusätzlich darauf geachtet werden, die natürliche Ordnung der Daten möglichst exakt abzubilden (Pyle, 1999), denn es gilt:

"If you torture data sufficiently, it will confess to almost anything." (Fred Menger)

Grundsätzlich kategorisiert die Statistik vier verschiedene Skalenniveaus zur Messung eins Merkmals / Attributs (Fahrmeier et al., 2007):

- **Nominalskala:** Ein Merkmal ist nominalskaliert, wenn seine Ausprägungen Namen, oder Kategorien sind. (Bsp.: Farben, Geschlecht etc.)
- **Ordinalskala:** Ein Merkmal ist ordinalskaliert, wenn seine Ausprägungen geordnet werden können, aber ihre Abstände nicht interpretierbar sind. (Bsp.: Schulnoten, Hauttyp etc.)
- **Intervallskala:** Ein Merkmal ist intervallskaliert, wenn es eine Ordnung besitzt und ihre Abstände interpretiert werden können. Quotienten lassen sich jedoch nicht sinnvoll interpretieren. (Bsp.: Jahreszahlen, Temperaturen)
- **Verhältnisskala:** Ein Merkmal ist verhältnisskaliert, wenn es intervallskaliert ist und zusätzlich einen sinnvollen Nullpunkt definiert, der Quotientenbildungen zulässt. (Bsp.: Semesterzahl, Abstände etc.)

In vielen Softwaresystemen, so auch in WEKA, werden diese Skalenniveaus nur durch numerische und nominale Typen umgesetzt (Witten & Frank, 2005). Diese Beschränkung hat eine Zuordnung der ursprünglichen Attributtypen zu den Klassen der nominalen oder numerischen Werte zur Folge. Ziel ist es dabei, weder Informationen einzufügen, die ursprünglich nicht existierten, noch Details durch zu simple Darstellungen zu verbergen. In einem fiktiven Beispiel soll das Wachstum eines Bakteriums in Abhängigkeit der Temperatur prognostiziert werden. Eine Abbildung der Temperatur in Form *(heiß, warm, kalt)* eignet sich dafür wenig, da dem Lerner wertvolle Informationen über Zwischenschritte (Unterschied zwischen 35° und 30° [beide dem ordinalen Intervall *heiß* zugeordnet] nicht erkennbar) vorenthalten werden. Bei der Ersetzung ist es wichtig die Daten bestmöglich darzustellen. Für nominalskalierte und ordinalskalierte Attributwerte (bspw. Frage 5, Abb. 1) wurde deshalb ein nominaler Wertebereich verwendet, da eine Ordnung der Werte, falls diese besteht (*nie < selten < häufig < oft*), durch eine willkürlich gewählte, numerische Transformation (*nie → 1, selten → 2, häufig → 3, oft → 4*) zerstört würde (*häufig = 3*nie*).

"It is hard to imagine how more damage can be done to the natural ordering of the data than by arbitrary number assignment to categorical." (Pyle, 1999)

Verhältnisskalierte *(Alter)*, oder intervallskalierte Attribute *(Solariumnutzung, Pigmentzahl, Untersuchsungszyklus)*, bei denen ein Abstand messbar ausgedrückt werden konnte, wurden numerisch transformiert, indem für jedes Intervall einer Attributausprägung (bspw. Pigmentzahl: 0-10, s. Abb. 1) das arithmetische Mittel zur Kodierung verwendet wurde (0-10 → 5) (Tabelle 4):

Tabelle 4 - Attribute_Kodierung

Attribut	**Transformierte Darstellung**
Alter	Keine Veränderung
Solariumnutzung	{2,5 ; 1 ; 0,2 ; 0}
Pigmentzahl	{5 ; 15 ; 35 ; 70}
Untersuchungszyklus	{2,5 ; 1 ; 0,5}

3.1.1 Behandlung von Inkonsistenzen

Innerhalb des langen Befragungs- und Untersuchungszeitraums wurden an Patienten, je nach Schwere des gesundheitlichen Problems, bis zu sechs ärztliche Untersuchungen durchgeführt. Eine Übersicht über die Anzahl der durchgeführten Untersuchungen an den insgesamt 6938 Patienten lässt sich aus der unten aufgeführten Grafik entnehmen:

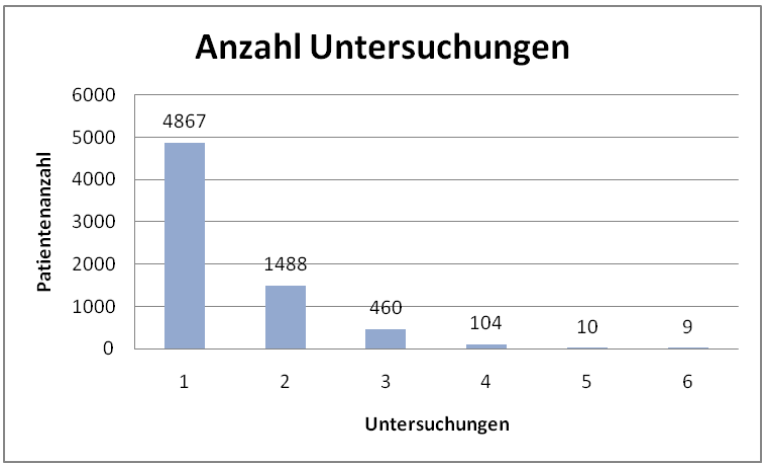


Abbildung 6 - Anzahl Untersuchungen pro Patient

Die Ergebnisse dieser Untersuchungen enthielten teils erhebliche Abweichungen zu vorherigen Diagnosen. Bei einigen Patienten wurde bspw. wechselhaft Risiko zu- oder abgesprochen, oder eine Pigmentanzahl von weniger als zehn und nach einigen Monaten über fünfzig festgestellt. Diese unterschiedlichen Ärzteberichte mussten zusammengeführt werden, um eine einheitliche, möglichst plausible Datenbasis pro Patient zu erhalten. Zu diesem Zweck wurde nach Rücksprache mit Dr. Michael Herbst anfänglich ein '*Mehrheitsvotum*' verwendet, gemäß dem die einzelnen Attributausprägungen durch das arithmetische Mittel (für numerische Attribute, wie der *Anzahl der Pigmente*), oder den Modus (nominale Attribute, wie *Erhöhtes Risiko*, s. Formel (3.2)) der Untersuchungen festgelegt wurde (Fahrmeier et al., 2007). Für unterschiedliche Werte x_i in den Attributen x der Untersuchungen berechnen sich die neuen, einheitlichen Werte x wie folgt:

Arithmetisches Mittel:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{3.1}$$

Modus: $x_{mod} : Ausprägung\ mit\ größter\ Häufigkeit.$ (3.2)

Den Grundgedanken des Mehrheitsvotums bildete die Annahme der gleichen Kompetenz jedes Arztes. Da der Modus bei einer identischen Anzahl von Ausprägungen zweier, oder mehrerer Merkmale kein eindeutiges Ergebnis liefert, musste für diesen Fall auch eine Regel festgelegt werden. Bei einem solchen „Patt" zwischen den Untersuchungsergebnissen in den nominalen / kategorialen Attributen, wurde der zeitlich jüngste Untersuchungswert als Referenz gesetzt, da die Mehrheit der erhobenen Werte zeitlich abhängig ist. Hintergrund ist die Überlegung, dass ein Patient, der ein Jahr zuvor kein Anzeichen eines Basalioms aufzeigt, im darauffolgenden Jahr davon betroffen ist. Deshalb wird im direkten Vergleich eine jüngere Untersuchung bevorzugt. Die Methodik des *„Mehrheitsvotums"* erwies sich jedoch im späteren Verlauf als wenig praktikabel, da viele der erhobenen Attribute eine große zeitliche Abhängigkeit aufweisen und durch eine simple Aggregation aller Ergebnisse weitere Inkonsistenzen entstanden. So ist es bspw. vorstellbar, dass in einer ersten, ärztlichen Diagnose Basaliome, Spinaliome, maligne Melanome etc. diagnostiziert wurden und eine darauffolgende Behandlung der Patienten diese Befunde wiederum beseitigte. Werden nun alle Untersuchungsergebnisse dieser Patienten aggregiert, so könnten einige Attribute der ersten Diagnose, die auf einen Befund deuten, mit Werten einer zweiten, oder dritten Untersuchung ohne Befunde zusammentreffen. Um derartige Konstellationen zu verhindern, wurde entschieden, das *„Mehrheitsvotum"* lediglich auf das Attribut *Anzahl der Pigmente* als relativ zeitstabiles Attribut (Traupe & Hamm, 2006, S. 121 ff.) anzuwenden und für weitere Zwecke eine einheitliche Untersuchung pro Patient zu wählen, die jeweils die Anzahl der Befunde maximiert, um die Datenbasis zu verbessern.

Neben den ärztlichen Attributen mussten noch weitere Inkonsistenzen im Datensatz bereinigt werden. Frage 8 des Patientenfragebogens (Abb. 1) behandelte detailliert eine Frage zu ausgeübten Sportarten. Zwölf vorformulierte Sportvarianten, sowie ein freies Textfeld, waren als Antwortmöglichkeiten gegeben. Eine Bereinigung der „sonstigen"

sportlichen Tätigkeiten um Indoor-Sportarten erwies sich aufgrund zahlloser unterschiedlicher Schreibvarianten („*Badminton“*, „*Badmington“*; „*Bowling“*, „*Boling*“ etc.) als schwierig. Das eigentliche Problem war jedoch die Aussagekraft des Sport-Attributs. Sonnenexposition, bspw. in Form von Outdoor-Sport, ist de facto ein Risikofaktor für Hautkrebserkrankungen (s. Tabelle 1), jedoch auch und insbesondere in Abhängigkeit des zeitlichen Rahmens. Diese Dauer wurde nicht erfasst und kann durch simple Klassifikationsmodelle kaum sinnvoll geschätzt werden. Anhand der vorliegenden Daten lässt sich bspw. der Unterschied zwischen einem Hobby-Fußballer (ca. eine Stunde wöchentlich) und einem Profi- oder Vereins-Fußballer (mind. sechs Stunden wöchentlich) nicht erkennen. Aus diesem Grund und der relativ großen Anzahl an einzelnen, möglichen Sportarten, bzw. Attributen, und der daraus resultierenden Gefahr von Overfitting wurde diese Frage zu einem einzelnen, binären Attribut *Outdoor-Sport* zusammengefasst.

Eine fehlende, einheitliche Definition des Attributs *Beurteilung* erschwerte zusätzlich die spätere Klassifikation. Das nominale Attribut (Ausprägungen: erhöhtes Risiko, kein erhöhtes Risiko) diente ursprünglich als Einstufung der Patienten in Gruppierungen, die aufgrund ihrer körperlichen Veranlagung und einem entsprechenden Risiko regelmäßiger einen Dermatologen aufsuchen sollten, und solchen, die dies nicht tun müssen. Eine visuelle Darstellung der Daten in Form einer Kreuztabelle zeigt die Problematik der fehlenden, einheitlichen Begriffsabgrenzung auf:

Tabelle 5 - Kreuztabelle: Melanom-Beurteilung

		Beurteilung			
		erhöhtes Risiko	kein erhöhtes Risiko	fehlender Wert	
malignes Melanom	ja	26	3	2	31
	nein	2505	4189	213	6907
	fehlender Wert	0	0	0	0
		2531	4192	215	

Etwa 9,6% (3/31) der betroffenen Melanom-Patienten wurde demzufolge kein erhöhtes Risiko in der Beurteilung zugesprochen. Ähnliche Werte lassen sich bei den anderen Hautkrebsarten finden (16% bei Spinaliompatienten, 9,7% für Basaliom-Erkrankte). Eine Interpretation dieser Größe lässt den Schluss zu, dass das Attribut *„Beurteilung“* anstelle eines Präventionsrisikos (Malignes Melanom liegt vor, daher müsste *„erhöhtes Risiko“* gegeben sein) die Chancen bzw. die Aussicht auf Erfolg bei einer Behandlung eines vorliegenden Hautkrebs bezeichnen. Gegen dieses Begriffsschema des Attributs *Beurteilung* sprechen jedoch Diagnosen, die bei keinem erkennbaren Risikofaktor (kei-

ne erhöhte Anzahl an Nävi [Muttermalen], kein Hautkrebs in der Eigen- oder Familienanamnese etc.[10]) und keinen Hautkrebsbefunden (malignes Melanom, Spinaliom, Basaliom), dennoch *„erhöhtes Risiko“* enthalten. Eine Analyse des Datensatzes[11] wies 40 Patienten mit *„erhöhtem Risiko“* bei fehlenden Risikofaktoren auf. Da jedoch hinter der Ausprägung des Attributs *Beurteilung* eine medizinisch fundierte, kompetente Diagnose steht, die Umstände einschließt, die ggf. nicht durch die Datenerhebung ersichtlich sind, wurden Überlegungen über das Löschen, bzw. Reklassifizieren der betroffenen Instanzen nicht weiter verfolgt. Eine unbedachte Bearbeitung des Konzepts verfälscht sowohl die Datenbasis, als auch die Ergebnisse. Insbesondere spielte die Hoffnung auf das Auffinden neuer, unbekannter Risikofaktoren in dieses Vorgehen mit. An dieser Stelle wird jedoch darauf hingewiesen, dass eine solche, nicht einheitliche Definition und Bewertung die Performanz der Lerner signifikant reduziert.

Nach Rücksprache mit Dr. Herbst wurde in einem letzten Schritt ein Attribut für zukünftige Erhebungen entfernt (Frage 9 a) + b) bezüglich des Hautkrebs in der Eigenanamnese), das für die aktuelle Erhebung nicht vorgesehen und auf dem Fragebogen nicht enthalten war.

3.1.2 Behandlung fehlender Werte

Datensätze enthalten nahezu immer fehlende Werte aus unterschiedlichen Ursachen. Im vorliegenden Fall der Datenerhebung über einen Fragebogen sind die wahrscheinlichsten Gründe die simple Verweigerung der Beantwortung durch den Patienten, Verunsicherung über die Fragestellung, oder schlicht das Fehlen der entsprechenden Information. So könnte es bspw. vorkommen, dass Fragen über Hautkrebsvorfälle in der Familie des Betroffenen aus Gründen der Privatsphäre nicht beantwortet werden, Sonnenbrände im Kindesalter mit hohem Alter vergessen wurden (*72,05% der Patienten, die diese Frage nicht beantworteten waren mind. 40 Jahre alt. Das Durchschnittsalter aller Patienten in der Grundgesamtheit beträgt 42,247 Jahre*), oder Jugendliche nicht die Frage beantworten können, ob sie als Erwachsener einen schweren Sonnenbrand haben wer-

[10] S. Tabelle 1 für eine Übersicht der bekannten Risikofaktoren einzelner Hautkrebsarten
[11] Hierfür wurde die Excel-Funktion ZÄHLENWENN (bei nicht vorliegenden Risikofaktoren, bspw. Für Pigmentmale #: 5, oder 15, oder Hautkrebs in der Eigenanamnese: nein) und eine Sortierung nach *Beurteilung* genutzt

den. Eine individuelle Berücksichtigung der Ursache des Fehlens von Instanzwerten ist entscheidend für die weitere Bearbeitung. Sind die Werte bspw. schlicht nicht zugänglich, wie im Beispiel der Jugendlichen, oder Kinder, wurden fehlende Werte mit der Ausprägung *„n.a.“*[12] ersetzt. Eine Ersetzung dieser Werte durch die im folgenden vorgestellten Verfahren ist nicht sinnvoll, da diese Daten nicht fehlen im engeren Sinne, sondern eine Angabe nicht möglich ist.

Fehlende Daten sollten aus diversen Gründen beseitigt werden. Einige Algorithmen sind bspw. nicht im Stande mit fehlenden Werten umzugehen und ignorieren somit im schlechtesten Fall eine komplette Instanz aufgrund eines einzelnen fehlenden Wertes, was eine signifikante Verringerung der Klassifikations-Genauigkeit zur Folge haben kann. Andere Tools nutzen Default-Methoden zur Ersetzung dieser Werte und fügen damit oft erhebliche Verzerrungen ein. Um also weder Datensätze zu ignorieren, noch zu verfälschen, sollten diese in Anbetracht der jeweiligen Data-Mining-Situation sinnvoll ersetzt werden. Hierfür existieren mehrere Ansätze (Saar-Tsechansky & Provost, 2007):

- *Löschen aller unvollständigen Instanzen*
- *Beschaffung der fehlenden Werte*
- *Einheitliche Ersetzung durch einen Globalwert (Mittelwert, Median etc.)*
- *Klassifikation von unvollständigen Instanzen (bspw.):*
    - *Lineare Regression*
    - *Multiple Lineare Regression*
    - *Nearest Neighbor Schätzer*

Die erste Methode, das Löschen derjenigen Instanzen mit fehlenden Werten, erweist sich als sehr drastisch und wenig vorteilhaft. Trotz fehlender Werte enthalten diese Instanzen oft wichtige Informationen, die durch diese Vorgehensweise unzugänglich werden. Eine Beschaffung der nicht vorhandenen Werte stellt die genaueste, jedoch aufwändigste Lösung dar und ist bei einer solch umfassenden Datenmenge[13] schlichtweg nicht praktikabel. Eine Globalwert-Ersetzung wird durch eine Vielzahl von Data-Mining Programmen unterstützt. *WEKA* bspw. ermöglicht über die Methode

[12] not applicable – entfällt (angewandt auf alle Patienten, die jünger als 18 Jahre zum Zeitpunkt der Untersuchung waren)
[13] Betroffen sind 3003 Instanzen (~ 43,2%) mit einem bis mehreren fehlenden Werten

weka.filters.unsupervised.attribute.ReplaceMissingValues eine Ersetzung mit dem entsprechenden Modus (bei nominalen Attributen, s. Formel (3.2)) oder arithmetischen Mittel (bei numerischen Attributen, s. Formel (3.1)). Für wenige fehlende Werte ist diese Methode durchaus vorteilhaft aufgrund ihrer geringen Komplexität. Im Falle vieler fehlender Werte wird jedoch durch eine globale Ersetzung eine erhebliche Verzerrung eingeführt, da die vorhandene Verteilung der Daten und deren Beziehungen zu anderen Attributen keinerlei Berücksichtigung finden. Einige der Methoden, die diese Zusammenhänge zwischen den Attributen zur Ersetzung nutzen, sind unter Punkt drei aufgeführt. Lineare Regression und multiple lineare Regression sind Schätzmethoden für Datensätze, die ausschließlich numerische Attribute enthalten. Hierbei werden spezifische Gewichte zu allen Attributen durch die *Kleinste-Quadrate-Methode* ermittelt und durch einen Funktionsterm eine spätere Klassifikation ermöglicht[14].

Die wohl geeignetste Methode im vorliegenden Fall, mit überwiegend nominalen Attributen ist die Klassifikation der fehlenden Werte durch einen *Nearest Neighbor Schätzer*. Diese Methode hat ihren Ursprung im Jahr 1967 durch die Arbeit von Cover & Hart (Cover & Hart, 1967) und wurde seitdem in vielen darauffolgenden Arbeiten überarbeitet. Der *k Nearest Neighbor (KNN)* Algorithmus gehört zu der Klasse der *Instance-based learning algorithms (IBL)* oder *lazy learning algorithms*. Diese Bezeichnung steht den *Eager Algorithms* (Sammut & Webb, 2010) gegenüber, deren Aufgabe der Entwurf eines generalisierenden Klassifikationsmodells darstellt. IBL hingegen arbeiten ausschließlich auf den Trainingsdaten (Instanzen) und verwenden nur diese, anstelle eines Modells, zur Klassifikation von zukünftigen Instanzen. Als IBL nutzt KNN die ihm verfügbaren Trainingsinstanzen, um durch ein Mehrheitsvotum der *k* nächsten Nachbarn die Klassifikationsentscheidung zu fällen. KNN kann dabei sowohl mit numerischen, als auch mit nominalen Attributen umgehen. Zur Klassifikation werden neben den Trainingsdaten eine Metrik zur Distanzberechnung und eine Ganzzahl *k* benötigt. *k* legt die Anzahl der nächsten Nachbarn fest, die betrachtet werden. Dabei sollte *k* weder zu kleine Werte annehmen[15], da die Klassifikation ansonsten sehr stark auf Ausreißer in den Trainingsdaten reagiert, noch zu große, weil durch stetig steigendes *k* sich die Vorhersage automatisch der häufigsten Ausprägung innerhalb der Klasse anpasst (Witten & Frank, 2005). Ein optimaler Parameter *k* lässt sich durch eine mehrfache Kreuzvalidie-

[14] Wird aufgrund der Aufgabenstellung (nominale statt numerische Attribute) nicht näher behandelt. Eine gute Einführung findet sich in (Fahrmeier et al., 2007, S. 475-515)
[15] Im Falle von k=1 spricht man auch von einer Nearest-Neighbor-Methode

rung ermitteln (Hastie et al., 2009). Das *k* mit maximaler Genauigkeit bei einer Kreuzvalidierung *(Kapitel 2)* auf den Trainingsdaten wird schließlich gewählt.

Als Distanzmetriken zur Ermittlung des Abstands zweier Instanzen x und y mit den jeweiligen Attributen $x_1, x_2, \dots, x_n$, bzw. $y_1, y_2, \dots, y_n$, stehen der *WEKA*-Umsetzung des *KNN*-Algorithmus (*weka.classifiers.lazy.IBk*) die folgenden Verfahren zur Verfügung: Levenshtein-, Chebyshev-, Manhattan- und euklidische Distanz, deren Formeln untenstehender Grafik zu entnehmen sind (Rieck et al., 2006).

Tabelle 6 - KNN_Distanzmetriken

Chebyshev-Distanz	$D(x,y) \coloneqq max_i(\lvert x_i - y_i \rvert)$
Manhattan-Distanz	$D(x,y) = \sum_i \lvert x_i - y_i \rvert$
Euklidische Distanz	$D(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2$

Zur Auswahl einer geeigneten Metrik wurde im vorliegenden Fall eine Auswertung der Klassifikationsgenauigkeit von KNN in Bezug auf das Attribut *„schwere Sonnenbrände in der Kindheit“* durchgeführt. *k* wurde zur Vergleichbarkeit jeweils auf einen konstanten Wert gesetzt[16]. Wie der nachstehenden Grafik zu entnehmen ist, existieren kaum Unterschiede zwischen der Klassifikations-Genauigkeit der Manhattan- und der euklidischen Metrik. Die Chebyshev-Metrik hingegen ist mit 13% geringerer Genauigkeit nicht konkurrenzfähig. Die Levenshtein-Distanz lässt sich ausschließlich für Text-Klassifikationen verwenden und konnte deshalb nicht angewandt werden.

[16] Drei Variationen wurden getestet: $k \in \{5, 10, 15\}$. Die Ergebnisse wiesen jeweils denselben Trend auf, der in Abb. 5 zu erkennen ist. Auszugsweise wird deshalb nur k = 5 grafisch in Abb. 5 dargestellt.

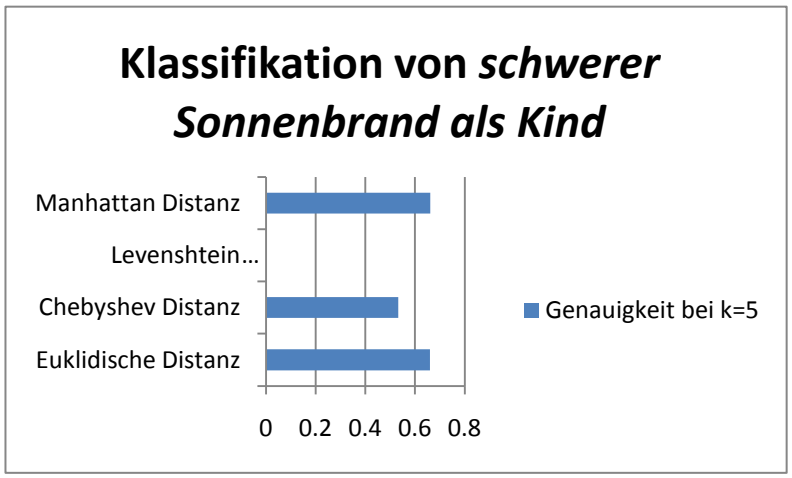


Abbildung 7 - Klassifikationsgenauigkeit KNN_Metriken

Für die Zwecke der Ersetzung fehlender Werte wird im Folgenden KNN mit der euklidischen Distanz gewählt.

Numerische Variablen können und sollten zuvor normiert werden, um die Auswirkungen der unterschiedlichen Maßstäbe zu minimieren. WEKA ermöglicht eine automatische Normalisierung in den *IBk*-Optionen (*'dontNormalize'* bei der Auswahl einer Distanzmetrik). Die Normalisierung transformiert numerische Variablen v mit den Variablenwerten v_i dabei bspw. wie folgt:

$$\hat{v}_i = \frac{v_i - \min v}{\max v - \min v} \quad \hat{v}_i \in [0,1] \tag{3.3}$$

Nominale Attribute v_i und v_j werden meist durch eine einfache, binäre Distanz verglichen:

$$d_A(v_i, v_j) = \begin{cases} 0 & wenn \quad v_i = v_j \\ 1 & wenn \quad v_i \neq v_j \end{cases} \tag{3.4}$$

Da KNN keine Beziehungen (Korrelation) der Attribute untereinander mit in die Klassifikation einbezieht, sondern jedes Attribut identisch handhabt, kann es passieren, dass die Klassifikation durch Attribute mit häufigen Werten (bspw. binäre Attribute) dominiert und verfälscht wird. Um den Einfluss dieser Attribute zu verringern, empfiehlt sich

eine Feature Subset Selection (Kapitel 3.4) vorzuschalten, oder eine Attribut-Gewichtung vorzunehmen. Zur Bestimmung eines ersten, optimalen Parameters *k* wurde eine Klassifikation des Attributs *Sonnenbrand als Kind* mit einer 10-fachen Kreuzvalidierung durchgeführt:

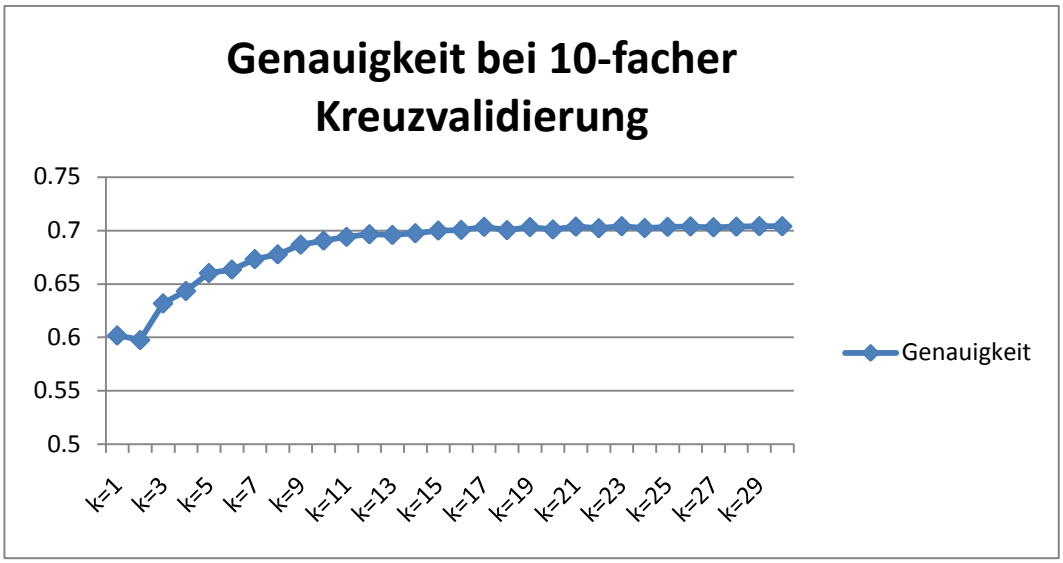


Abbildung 8 - Nearest Neighbor (Wahl von k für „Sonnenbrand als Kind“)

In der Abbildung ist ein asymptotischer Verlauf zu beobachten. Mit steigendem *k* nähert sich die Genauigkeit einem Grenzwert von etwa 71%. Ab $k = 17$ stellt sich lediglich ein langsames Wachstum ein. Daher wurde dieser Wert für das weitere Verfahren gewählt.

Analog (mit einer optimalen Parameterbestimmung für k) wurde für andere unvollständige Attribute verfahren. Im Anschluss wurde KNN mit der entsprechenden Konfiguration (Wahl von *k* und euklidische Distanz) zur Ersetzung nachfolgender Werte verwendet:

- *Sonnenbrand als Kind* : 2.035 fehlende Werte
- *Sonnenbrand als Jugendlicher* : 1.754 fehlende Werte
- *Sonnenbrand als Erwachsener* : 1.064 fehlende Werte
- *Outdoorzeit bei intensiver Sonneneinstrahlung* : 93 fehlende Werte
- *Solariumnutzung* : 239 fehlende Werte
- *Hautreaktion* : 180 fehlende Werte
- *Hautkrebs in der Eigenanamnese* : 158 fehlende Werte

Um die Auswirkungen dieser Ersetzungen nachzuvollziehen und mögliche, manuell hinzugefügten Verzerrungen[17] zu kontrollieren, wurden im weiteren Verlauf vier Varianten, zwei mit einer Ersetzung und zwei ohne, getestet:

I. Datensatz mit allen Ersetzungen
II. Datensatz mit wenigen Ersetzungen (ohne eine Ersetzung Attribute *Sonnenbrand als Kind*, *Sonnenbrand als Jugendlicher*, *Sonnenbrand als Erwachsener* aufgrund der erheblichen Anzahl an fehlenden Daten)
III. Datensatz mit fehlenden Werten
IV. Datensatz mit gelöschten Instanzen

3.3 Konvertierung der Daten: Von CSV zu ARFF

Zur späteren Analyse der Daten mussten diese in ein Format überführt werden, das von gängiger Data-Mining Software unterstützt wird. Da in dieser Arbeit die Open-Source Data-Mining Workbench *WEKA*[18] (Waikato Environment for Knowlegde Analysis) verwendet wird, handelte es sich hierbei um eine *.arff*-Datei (Attribute-Relation File Format). WEKA unterstützt nativ den direkten *.csv*-Import und eine Konvertierung in das Zielformat *.arff*, jedoch nicht in der zugrunde liegenden Fragebogen-Form. Zur Konvertierung dieses speziellen *.csv*-Aufbaus wurde im Rahmen dieser Arbeit das Programm *csv2arff* in der plattformunabhängigen Programmiersprache Java entwickelt, das im Folgenden vorgestellt wird. Um die Arbeitsweise des Programms zu verdeutlichen, muss zunächst der allgemeine Aufbau einer *.arff*-Datei aufgezeigt werden. Diese Dateien setzen sich stets aus zwei Hauptkomponenten zusammen: Dem Header, in dem Informationen über den Relationsnamen ('@relation') und die enthaltenen Attribute ('@attribute') inklusive ihres Attributtyps festgehalten werden, und einem Daten-Teil, der die einzelnen Datensätze und Merkmalsausprägungen der Relation darstellt. Die Werte einer Instanz werden durch ein Komma abgetrennt (Paynter et al., 2008).

[17] Durch eine fortlaufende Ersetzung, die sich der bereits ersetzten Werte bedient, wird der Fehler, bzw. die Verzerrung durch die Klassifikation möglicherweise fortgeführt und vervielfacht (Zur Klärung dieser These wird auf die Auswertung der Experimente in Kapitel 5 verwiesen)
[18] http://www.cs.waikato.ac.nz/ml/weka/

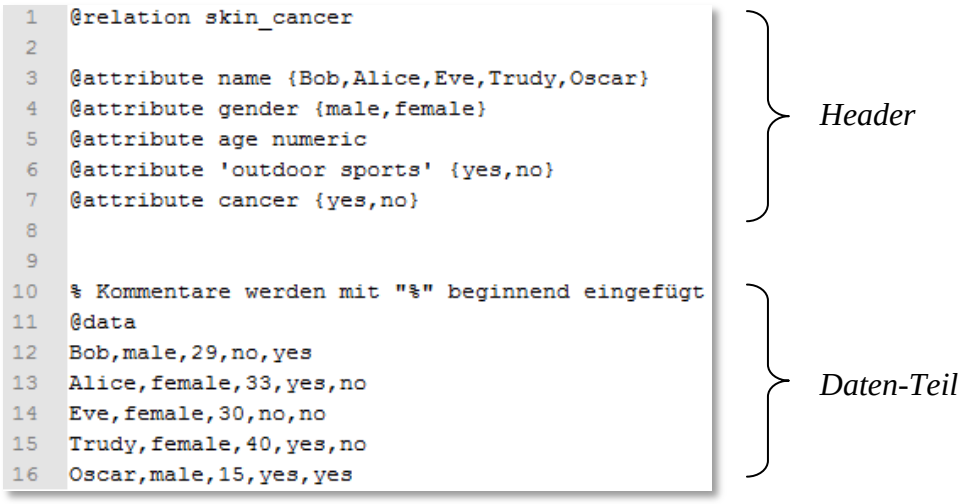


Abbildung 9 - Aufbau einer .arff-Datei

WEKA unterscheidet vier Attributtypen: *Nominal*, *numeric*, *String* und *date*, wobei *String* und *date* jeweils Sonderfälle der beiden anderen Typen bilden. *Csv2aff* bestimmt demzufolge für jedes Attribut aus der Relation dessen Typ und gibt im Falle eines nominalen Attributs eine duplikatsfreie Liste von dessen Merkmalsausprägungen wieder. Fehlende Informationen in Datensätzen müssen nach der WEKA-Deklaration mit einem Fragezeichen gekennzeichnet werden.

Gemäß der in Kapitel 3.1 vorgestellten Merkmals- bzw. Fragen-Kategorien muss bei der Konvertierung die Art der Ersetzung berücksichtigt werden. Die Art jedes Attributs wird hierfür in die zweite Zeile unter das entsprechende Merkmal des *.csv*-Dokuments geschrieben. Eine Übersicht der Typen findet sich in nachstehender Tabelle:

Tabelle 7 - CSV2ARFF_Kodierung

Attributtyp	Kodierung in 2. Zeile der .csv-Datei
Numerisch (Merkmalsausprägung enthalten)	-
„Singuläre“ / binäre Fragen (yes/no)	S (x → yes)
Nominale Fragen	„merkmalswert“ (bspw.: männlich, oft, selten)

Nach Auswahl einer Datei, muss das verwendete Trennzeichen angegeben werden, um die Daten später korrekt auslesen zu lassen. Das Trennzeichen lässt sich manuell über die Regions- und Sprachoptionen in Microsoft Windows Betriebssystemen ändern, oder

indirekt über die Wahl der Systemsprache bei *CSV*-exportierender Software wie MS Excel. Der Standard hierfür ist ein Komma in der englischen[19] Form und ein Semikolon in der deutschen Fassung.

Im Anschluss an die Wahl der Sprachoptionen wird durch einen Klick auf den Konvertieren-Button eine gültige *.arff*-Datei und eine vereinfachte *.csv*-Datei für ein mögliches Data-Preprocessing in MS Excel[20] erzeugt.

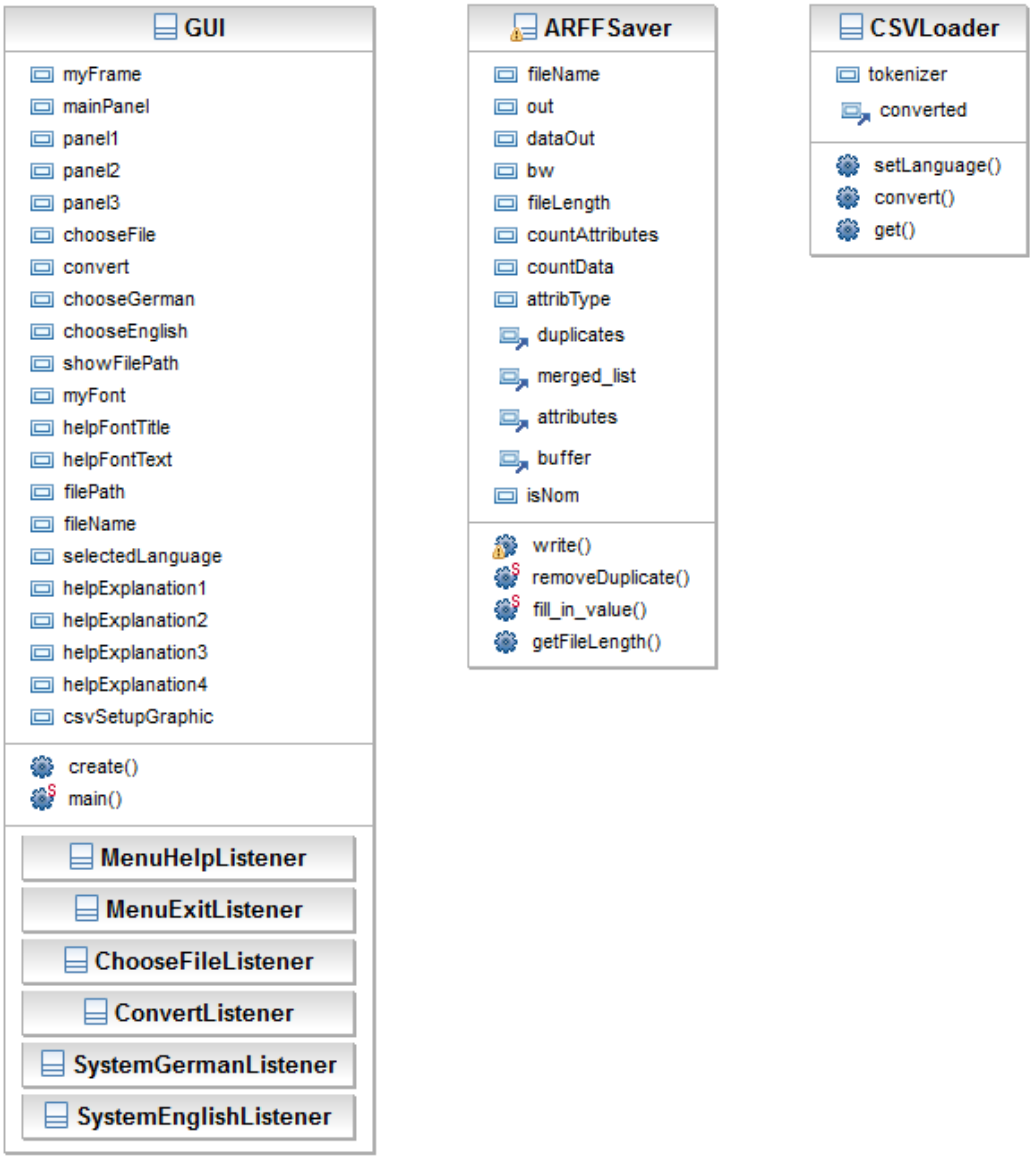


Abbildung 10 - Klassendiagramm csv2arff

[19] Der Grund liegt in den unterschiedlichen Dezimaltrennzeichen
[20] MS Excel eignete sich aufgrund der Daten-Visualisierung, Filter-, Sortier- & Formelmöglichkeiten sehr gut für verschiedene Aufgaben des Data Preprocessing und fand auch in dieser Arbeit Anwendung.

Abbildung 11 - csv2arff UI

3.4 Feature Subset Selection

Die in Kapitel 4 vorgestellten Lernalgorithmen zielen darauf ab, in jeder Iteration die bestmöglichen Attribute zur Klassifikation auszuwählen. Attribute, die irrelevant bezüglich der Klassifikationsentscheidung sind, sollten demzufolge durch den Lernalgorithmus ausgeschlossen werden. Zahlreiche Experimente[21] belegen jedoch die Erkenntnisse, die auch in dieser Arbeit beobachtet wurden. Bei der Anwendung des Entscheidungsbaum-Lerners C4.5[22] auf die ursprünglichen Daten, ohne vorherige manuelle, oder automatisierte *Feature Subset Selection* (FSS), wurde bereits auf der vierten Ebene ein Knoten eingefügt, der das Attribut *'Arzt#'* für eine weitere Entscheidung über das Krebsrisiko des Patienten auswählt. Sind genügend positive Trainings-Beispiele durch andere Entscheidungen abgedeckt, erscheint eine Unterteilung durch die Ärzte-Nummer, oder anderen, irrelevanten Attributen dem Lernalgorithmus im weiteren Verlauf zufällig[23] als sinnvoll. Darunter leidet meist die Qualität der Vorhersage. Methoden der *Feature Subset Selection* beheben dieses Problem, indem sie die relevanten[24] Attribute als

[21] Experimentreihen mit einem C4.5-Lerner, bei denen ein zufälliges, binäres Attribut (bspw. das Ergebnis eines Münzwurfs) hinzugefügt wurde. Die Performanz des Lerners sank infolge dessen jeweils um 5-10% (Witten & Frank, 2005, S. 288)
[22] Entspricht der WEKA-Implementierung J48 (weka.classifiers.trees.J48)
[23] Bei einer kleinen Restmenge von Trainingsbeispielen, wie sie bspw. in Entscheidungsbäumen ab einer gewissen Tiefe erreicht wird, erscheint bspw. ein zufälliges, binäres Attribut (Ergebnis eines Münzwurfs) für eine Teilung oft durch den Zufall der größten Abdeckung am geeignetsten (Witten & Frank, 2005).
[24] Eine allgemein gültige Definition der „Relevanz“ von Attributen existiert nicht. (John et al., 1994) liefert jedoch einen guten Überblick über vorhandene Definitionen.

Schritt des Data Preprocessing ermitteln. Ziel ist neben der Qualitätssteigerung der Vorhersage auch kürzere, verständlichere Repräsentationen der Konzepte zu schaffen (bspw. weniger Regeln oder kleinere Bäume bei nahezu gleicher Vorhersagegenauigkeit). Die Attribut-Auswahl kann dabei manuell, über das Wissen eines Domänenexperten, sowie automatisiert erfolgen. Eine manuelle Auswahl ist automatisierten Verfahren stets vorzuziehen, da hierfür langjährige und zahlreiche Erfahrungswerte, sowie semantische Zusammenhänge berücksichtigt werden. Automatisierte Verfahren lassen sich in folgende Kategorien einteilen:

- *Wrapper-Verfahren*: Die Wrapper-Methode evaluiert in Abhängigkeit eines Lernalgorithmus (entspricht dem Algorithmus zur späteren Modellbildung) jede Untermenge des Attributraums und wählt schließlich die Untermenge mit den geringsten Vorhersagefehlern[25].
- *Filter-Verfahren*: Die Filter-Methode geht von gleicher Relevanz aller Attribute im Ziel-Konzept aus. Durch Suchheuristiken wird im Attributraum eine minimale Menge ermittelt, die alle Instanzen unterscheiden kann und eine Messgröße M maximiert. Der nachträglich angewandte Lernalgorithmus wird nicht berücksichtigt. Filter-Verfahren erzielen deshalb meist geringwertigere Ergebnisse als Wrapper-Verfahren, arbeiten jedoch effizienter (Liu & Yu, 2005).

In dieser Arbeit wurden die manuelle und automatisierte Feature Subset Selection kombiniert, um Vorteile beider Verfahren zu nutzen. Zu Beginn eines jeden Experiments wurden manuell Attribute entfernt, die offensichtlich keinen kausalen Zusammenhang mit dem zu erlernenden Konzept aufwiesen. In nachfolgender Tabelle sind sämtliche Attribute mit entsprechender Begründung aufgelistet, die global, über alle Experimente hinweg, wegen ihres fehlenden Bezugs zum Klassenattribut gelöscht wurden.

Im Anschluss an diese manuelle Vorauswahl wurden für die durchgeführten Experimente jeweils zwei Varianten jedes Modells evaluiert, eines mit einer automatisierten FSS[26] und eines ohne. Bei nahezu gleicher Performanz beider Varianten wurde schließlich das Ergebnis der FSS benutzt, um einerseits Overfitting zu vermeiden und andererseits eine einfachere Darstellung und Interpretation des Modells zu ermöglichen.

[25] Einen guten Überblick über die Verfahren der FSS inkl. Pseudocode bietet (Liu & Yu, 2005)
[26] Wrapper-Verfahren mit jeweiligem Lern-Algorithmus

Tabelle 8 - Manuelle FSS

Attribut	**Erklärung**	**Grund des Ausschlusses**
Datum	Datum der ärztlichen Untersuchung	Kein kausaler Zusammenhang
Ärzte#	Primärschlüssel für jeden Arzt (numerischer Code)	Kein kausaler Zusammenhang
Bogen#	Primärschlüssel des Befragungsbogens (numerischer Code)	Kein kausaler Zusammenhang
Kommentar	Kommentar des Arztes	Keine Angaben verfügbar
Therapie	Wurde eine Therapie eingeleitet? Falls ja, in welcher Form? (Operation, PDT, Lokaltherapie, Hautpflege)	Keinerlei Informationsgehalt für ein Präventionssystem, da ein Risiko bereits vorlag
Untersuchungszyklus	Numerischer Wert für die Anzahl der Hautarztbesuche pro Jahr	Kausaler Zusammenhang eher: Risiko → Untersuchungszyklus

4 Algorithmen des Data Mining

Es existiert eine Vielzahl von unterschiedlichen Algorithmen im Bereich des Maschinellen Lernens, die je nach Problemstellung und Datenstruktur individuelle Stärken und Schwächen aufweisen. Um bestmögliche Ergebnisse zu erzielen, empfiehlt es sich deshalb mehrere, vom Ansatz verschiedene, Lernalgorithmen auf eine Datenbasis anzuwenden und zu vergleichen (Witten & Frank, 2005, S. 84). Die in dieser Arbeit genutzten Methoden decken daher ebenfalls eine möglichst große Bandbreite bezüglich des Vorgehens ab. Im Folgenden werden die Algorithmen in ihrer Grundidee dargestellt.

4.1 Entscheidungsbaum-Lerner

Entscheidungsbäume stellen aufgrund ihrer zahlreichen Vorzüge eine in der Praxis oft genutzte Klassifikationsart dar. Sie bieten eine leicht nachvollziehbare Darstellung der Daten, die keinerlei Vorkenntnisse des Maschinellen Lernens erfordert, erzielen überwiegend gute Ergebnisse und arbeiten mit einer Gesamtkomplexität von $O(mn \log n) + O(n(\log n)^2)$ [27] vergleichsweise effizient. Die heutigen Verfahren basieren auf dem *ID3-Algorithmus* von Ross Quinlan (Quinlan, 1986). Nachfolger dieses Algorithmus sind zum einen der *C4.5-Algorithmus*, der in vielen Open-Source-Implementierungen verfügbar ist und zum anderen der kostenpflichtige Quasi-Industriestandard *See5/C5.0*[28]. In Kapitel 5 wird für die Experimentreihen mit Entscheidungsbäumen der in WEKA verfügbare C4.5-Algorithmus benutzt. C4.5 bietet im Vergleich zu ID3 folgende Verbesserungen an: Den Umgang mit kontinuierlichen Daten (bspw. einer Temperatur), mit fehlenden Werten [werden zur Entropie- und Information Gain-Berechnung (nachfolgend erläutert) nicht verwendet], mit gewichteten Attributen, sowie die Möglichkeit des Prunings und die Umwandlung von Entscheidungsbäumen zu einer Regelmenge (Witten & Frank, 2005, S. 105).

[27] Komplexität des C4.5 Algorithmus bei m Attributen und n Instanzen, inkl. Baum-Erstellung und Pruning. Siehe (Witten & Frank, 2005, S. 198) für weitere Ausführungen.
[28] Weitere Informationen sind unter http://www.rulequest.com/see5-info.html zu finden

Ein Entscheidungsbaum besteht im Allgemeinen, wie in der Informatik üblich, aus Knoten, Kanten und Blättern. Wie in Abb. 5 bereits zu sehen war, repräsentieren die Baum-Knoten die Attribute des Datensatzes und die jeweils anliegenden Kanten deren Attributausprägungen. Die Blätter enthalten schließlich die Werte des Klassenattributs.

Die grundlegende Divide & Conquer-Arbeitsweise der Entscheidungsbaum-Algorithmen wird im Folgenden vorgestellt:

1. Wenn alle verbleibenden Trainingsdaten $T = (X_1, X_2, \dots, X_N)$ (nahezu) derselben Klasse angehören, füge ein Blatt mit dieser Klassenausprägung zum Entscheidungsbaum hinzu,
2. Sonst:
    a. Für jedes Attribut A_i
        i. Berechne das Split-Kriterium $s(A_i)$
    b. Wähle das Attribut A_{max} aus, das den maximalen Wert des Split-Kriteriums erzielt: $A_{max} = \max_{A_i} s(A_i)$
    c. Erzeuge mit A_{max} einen neuen Knoten im Entscheidungsbaum
    d. Bestimme die Menge von Datensätzen T_i, die jeweils dieselbe Ausprägungen von A_{max} besitzen
    e. Fahre rekursiv mit den verbleibenden Trainingsdaten T_i fort und füge weitere Knoten als Kindsknoten an

Das Split-Kriterium $s(A_i)$ entspricht einer Heuristik und wird von C4.5 entweder durch den normalisierten ***Information-Gain***, oder die ***Gain-Ratio*** realisiert.

$$Information\ Gain\ (T, A_i) = I(T) - \sum_{i=1}^{N} \frac{|T_i|}{|T|} I(T_i) \tag{4.1}$$

$$I(T) = -\sum_{j=1}^{x} RH(C_j, T) \log(RH(C_j, T)) \tag{4.2}$$

Formel (4.2), auch *Entropie* genannt, entstammt der Arbeit von Shannon: *„The Mathematical Theory of Communication“* (Shannon & Weaver, 1963). Dabei entspricht

RH der relativen Häufigkeit der Klassenausprägungen C_j im Trainingsset T. $I(T)$ bezeichnet demnach den Informationsgehalt einer Nachricht, die die Klassenzugehörigkeit einer Instanz identifiziert. Nach einer Aufteilung des Trainingssets T in die Teilmengen $T_1, T_2, \dots, T_N$ durch einen Split an einem spezifischen Attribut A_i wird der Information Gain schließlich durch $Information\ Gain\ (T, A_i)$ ermittelt und das Attribut A_i mit einem maximalen Information Gain zum Split gewählt. Der Nachteil des Information Gains ist, dass ein Attribut mit vielen Ausprägungen für einen Split bevorzugt wird. So erreicht der $Information\ Gain(T, A_i)$ einen maximalen Wert, wenn alle Instanzen bezüglich A_i eine unterschiedliche Ausprägung besitzen.

Die Gain-Ratio umgeht dieses Problem, indem sie den potentiellen Zugewinn eines Splits an dem jeweiligen Attribut A_i berücksichtigt.

$$P(T, A_i) = -\sum_{i=1}^{N} \frac{|T_i|}{|T|} \log\left(\frac{|T_i|}{|T|}\right) \tag{4.3}$$

Dabei wird das Attribut gewählt, das den Term $Information\ Gain\ (T, A_i)\ / P(T, A_i)$ maximiert.

Um Overfitting zu vermeiden, sollte der Baum nach / während[29] seiner Erstellung zusätzlich durch ***Pruning*** gekürzt werden. Durch eine statistische Heuristik auf den Trainingsdaten, die die echte Fehlerrate q an jedem Knoten durch einen Schätzer e approximiert (e wird hergeleitet durch ein Bernoulli-Experiment mit der Fehlerrate f auf den Trainingsdaten. e entspricht dabei der oberen Intervallsgrenze des Konfidenzintervalls aus diesem Experiment), erfolgt eine Abschätzung, ob der Knoten durch ein Blatt mit der Mehrheitsklasse ersetzt wird[30]. Wenn N die Anzahl von Instanzen, z die Standardabweichung und $f = \frac{E}{N}$ den Anteil erfolgreich klassifizierter Instanzen bezeichnet, berechnet sich eine pessimistische Abschätzung der Fehlerrate e durch folgenden Term:

[29] Unterscheidung zwischen Pre- und Post-Pruning (s. Kapitel 2)

[30] In dieser stark gekürzten Erklärung kann nicht auf unterschiedliche Pruning-Arten (Post-Pruning [Subtree-Replacement] und Pre-Pruning [Subtree-Raising]) und deren Hintergrund eingegangen werden. Eine gute Darstellung findet sich in (Witten & Frank, 2005, S. 193-196)

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \tag{4.3}$$

C4.5, bzw. J48, ersetzt einen Knoten demnach, wenn seine geschätzte, echte Fehlerrate größer ist als die des darunter liegenden Teilbaums. Die Fehlerrate eines Teilbaums errechnet sich dabei durch die gewichtete Summe der Fehlerrate seiner Blätter. Der weiter unten liegende Teilbaum ersetzt in diesem Fall den oberen Knoten *(Subtree Replacement)* (Witten & Frank, 2005, S. 192 ff.). Durch eine Variation des Parameters C (gibt das Konfidenzintervall an) kann das Pruning gesteuert werden.

4.2 Regel-Lerner

Schriftlich formulierte Regeln sind für den Menschen eine der verständlichsten Darstellungsformen von Wissen. Bei vielen Aufgabenstellungen des Maschinellen Lernens erzielen Regel-Lerner dabei bessere Ergebnisse als Entscheidungsbaumlerner (Cohen, 1995, S. 1). Eine beispielhafte Regelsammlung aus einem *WEKA*-Experiment des Fußball-Problems aus dem Klassifikations-Kapitel 2.2 sieht folgendermaßen aus[31]:

- *(Sieg im letzten Spiel = ja)* ➔ *Ergebnis=Sieg (7.0/2.0)*
- *(Stärke des Gegners = schwach)* ➔ *Ergebnis=Unentschieden (2.0/0.0)*
- ➔ *Ergebnis=Niederlage (6.0/1.0)*

Der WEKA-Output ist hierbei nahezu selbsterklärend. Die Regelmenge wird von oben nach unten interpretiert. Im ersten Fall, bei einem Sieg im letzten Spiel, sagt der Regel-Lerner bspw. einen Sieg voraus. Die Zahlen in den Klammern hinter jeder Regel geben die Anzahl abgedeckter Instanzen mit diesem Attributwert, sowie die Anzahl der jeweiligen Fehlklassifikationen durch die einzelne Regel an. Ein „Sieg im letzten Spiel“ wurde demzufolge sieben Mal vorhergesagt, in zwei dieser Fälle hatte dies jedoch keinen Sieg im aktuellen Spiel zur Folge.

[31] JRip-Durchlauf mit Default-Werten

Der in dieser Arbeit verwendete Regel-Lerner wird als *„RIPPER“*[32] (Cohen, 1995) bezeichnet. RIPPER entstand aus einer Modifikation des IREP-Algorithmus aus dem Jahr 1994 (Fürnkranz & Widmer, 1994) und bot u.a. Änderungen bezüglich des Abbruchkriterums, der Heuristik zum Pruning und einer zusätzlichen Optimierungs-Phase an[33]. Seine grundsätzliche Vorgehensweise wird nachfolgend stichpunktartig erläutert:

I. **Aufbau-Phase:** Die Trainingsmenge $T = \{Pos, Neg\}$, die aus positiven und negativen Beispielen besteht, wird in eine Growing-$Grow = \{GrowPos, GrowNeg)$ und eine Pruning-Menge $Prune = \{PrunePos, PruneNeg\}$ unterteilt. Die Schritt 1.a und 1.b werden solange wiederholt bis entweder:
    i. Keine positiven Beispiele mehr vorhanden sind, oder…
    ii. Die Description Length (DL) der Regelmenge und der verbleibenden Beispiele ≥ Bisherige DL + 64 bits ist, oder…
    iii. Die Fehlerrate ≥ 50% erreicht hat

    a) **Grow-Phase:** Auf der Growing-Menge werden durch Anwendung des FOIL-Algorithmus[34] Regeln $R_1, R_2, \dots, R_n$ mit einer 100%igen Genauigkeit gefunden und zur Regelmenge hinzugefügt.

    b) **Pruning-Phase:** Jeder Attributtest einer eben gelernten Regel der Regelmenge wird durch Anwendung auf der Pruning-Menge mit der Heuristik $\frac{p-n}{p+n}$[35] in einer last-to-first Reihenfolge getestet. Dabei wird eine endliche Anzahl an Tests aus einer Regel gelöscht bis die Heuristik maximiert wird.

II. **Optimierungs-Phase:** Für jede gefundene Regel R_i aus der initialen Regelmenge $\{R_1, R_2, \dots, R_n\}$ werden in dieser Phase zwei neue, alternative Regeln gelernt. Die erste heißt *Replacement* R_i^{rep} und wird auf einer neuen Growing-Menge durch eine Neuaufteilung gelernt. Dabei werden sämtliche Instanzen aus der

[32] RIPPER steht für Repeated Incremental Pruning to Produce Error Reduction und wird, in WEKA durch JRip mit geringfügigen Änderungen implementiert (weka.classifiers.rules.JRip)
[33] Detaillierte Änderungen von IREP zu RIPPER sind der Arbeit von Cohen (Cohen, 1995) zu entnehmen.
[34] FOIL steht für „First Order Inductive Learner“ und wurde 1990 durch J.R. Quinlan vorgestellt (Quinlan, 1990)
[35] JRip verwendet die leicht abgewandelte Form $\frac{p+1}{p+n+2}$, da hierbei im Falle von $p + n = 0$ die Heuristik zu 0,5 evaluiert

Menge entfernt, die durch andere Regeln bereits abgedeckt sind. Die zweite Regel nennt sich *Revision* R_i^{rev} und entsteht durch eine weitere Verfeinerung von R_i auf der neuen Growing-Menge, indem zusätzliche Attributtests angefügt werden. Beide Regeln R_i^{rep} und R_i^{rev} werden durch die Pruning-Metrik $\frac{p+N-n}{P+N}$ geprunt, wobei P und N der Anzahl der positiven, bzw. negativen Instanzen der jeweiligen Klasse und p, n die Anzahl positiver und negativer Abdeckungen der spezifischen Regel entsprechen. Im Anschluss wird durch das Kriterium der MDL (Minimum Description Length) die beste der drei Regeln, also R_i, das *Replacement*, oder die *Revision* als finale Regel zur Regelmenge hinzugefügt. Durch diese Optimierungs-Phase soll die Genauigkeit des Modells weiter verbessert werden, indem Regel für Regel verbessert wird (Witten & Frank, 2005, S. 205-207).

Sollten nach diesem Ablauf noch positive Beispiele verbleiben, wird die Prozedur wiederholt.

4.3 Naive Bayes

Naive Bayes[36] repräsentiert einen der simpelsten und ältesten Algorithmen im Repertoire des Maschinellen Lernens. Aufgrund seiner Einfachheit bei gleichzeitig soliden Ergebnissen findet er in zahlreichen Data-Mining-Projekten immer noch Anwendung. Sein einfacher Aufbau ohne zusätzliche Optimalitätsparameter, seine geringe Komplexität und leicht interpretierbare Ergebnisse sind weitere Gründe für seine Popularität trotz stark vereinfachter Grundannahmen (Wu et al., 2008, S. 24-27). Die Methode erstellt auf Basis des durch den englischen Mathematiker Thomas Bayes aufgestellten Bayestheorems (Bayes, 1763) einen Klassifikator. Neuen Datensätzen wird damit die „wahrscheinlichste“ Klasse bezüglich der Trainingsdaten zugewiesen. Formal gliedert sich der Prozess wie folgt:

Für Trainingsdaten mit n Instanzen, m Attributen und k Klassen werden zunächst relative Häufigkeiten jeder einzelnen Attributausprägung $a_i(h)\ f\ddot{u}r\ i = 1, \dots, m$ mit Aus-

[36] In WEKA in der Klasse *weka.classifiers.bayes.NaiveBayes* implementiert

prägungen h bezüglich der Klassenwerte C_j $für$ $j = 1, \dots, k$ und der jeweiligen Anzahl, ausgedrückt durch „#“, bestimmt (Han & Kamber, 2006, S. 310-315):

$$Likelihood\left(a_i(h), C_j\right) = \frac{\#\, a_i(h)}{\#\, C_j} \tag{4.3.1}$$

Gemeinsam mit dem sogenannten *„Prior“* des Klassenwertes, der relativen Häufigkeit der Klassenausprägung im gesamten Trainingsset T, lässt sich durch Anwendung der bedingten Wahrscheinlichkeit (Bayes-Theorem: $P(a|b) = P(b|a) * P(a)$) die *„Posterior“*-Wahrscheinlichkeit P berechnen:

$$P\left(C_j \mid a_i(h)\right) = \frac{Likelihood\left(a_i(h), C_j\right) *\ Prior\ (C_j)}{\sum_{i=1}^{m} Likelihood\ (a_i(h), C_j)} \tag{4.3.2}$$

Der Prior lässt sich durch die Wahrscheinlichkeit einer einzelnen Klassenausprägung C_j unter Gewissheit des Attributs $a_i(h)$ interpretieren. Ein auf diese Arbeit zutreffendes Beispiel bildet der Posterior $P(erhöhtes\ Risiko = "yes" \mid Anzahl\ Pigmente = " > 50")$, also die Wahrscheinlichkeit, dass bei einem Patienten mit mehr als 50 Pigmentmalen ein erhöhtes Krebsrisiko diagnostiziert wurde. Da eine Instanz x im Normalfall mehrere Attribute besitzt, muss ein Weg gefunden werden, diese einfachen Wahrscheinlichkeiten für einzelne Attributwerte zu kombinieren, um die Wahrscheinlichkeit eines Klassenwertes für eine ganze Instanz, bzw. einen Patienten, zu errechnen. Im Folgenden geht Naive Bayes zur namensgebenden Vereinfachung von Rechenschritten fälschlicherweise von einer statistischen Unabhängigkeit aller Attribute aus, die im Realfall nie vorliegt, aber eine gute Approximation bietet:

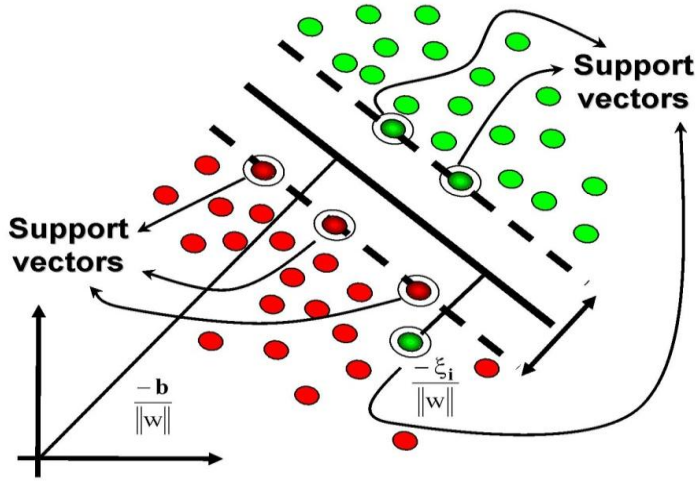
$$Likelihood\left(x, C_j\right) = \prod_{i=1}^{m} Likelihood\ (a_i(h), C_j) \tag{4.3.3}$$

Der Posterior $P\left(C_j \mid x\right)$ errechnet sich analog, durch Ersetzung der Likelihood-Terms in (4.3.2) mit (4.3.3). Der Klassenwert einer neuen Instanz x wird nach Berechnung aller

Posterior-Terme schlussendlich durch deren Maximum, also den wahrscheinlichsten Wert auf Basis der Trainingsdaten, determiniert.

4.4 Support-Vector-Machines

Support-Vector-Machines (SVM) sind aus heutigen Data-Mining-Projekten nicht mehr wegzudenken und gelten als besonders robust und genau.[37] Die Grundidee der SVMs entstammt einer Arbeit von Frank Rosenblatt aus dem Jahr 1958 (Rosenblatt, 1958). Rosenblatt beschrieb als Erster die Trennung durch eine Hyperebene, worauf schließlich Wapnik und Chervonenkis 1979 die Idee für die Entwicklung einer SVM nutzten (Wapnik & Chervonenkis, 1979).



- b: Bias
- w: Normalenvektor der Ebene
- ξ: Schlupfvariable für individuelle Verletzung der Nebenbedingung

Abbildung 12 - Support Vector Machine (Üstün, 2003)

Der Grundgedanke hinter SVMs ist es, eine Klassifikation über eine mathematische, grafische Trennung der Daten durch eine Hyperebene zu realisieren. Mit Hilfe dieser Ebene ist es anschließend möglich durch Bestimmung der relativen Lage neuer Instanzen im Hinblick auf die Hyperebene die Klassenzugehörigkeit zu ermitteln. Dazu muss zunächst die beste aller möglichen Ebenen gefunden werden. SVMs maximieren hierfür den Abstand der Klassen voneinander durch eine Maximierung des Abstands der örtlich nächsten Punkte auf beiden Seiten. Nur diese Punkte werden für eine Beschreibung der Hyperebene genutzt und ihre Vektoren verleihen der Methode ihren Namen als „Stütz-

[37] SVMs erreichten in einer Abstimmung der ICDM (International Conference on Data Mining) durch die IEEE im Dezember 2006 den dritten Platz in einer Top10-Rangliste (Wu, et al., 2008)

vektoren“ (Support Vectors) der Ebene. Eine vollständige, mathematische Erläuterung der komplexen Vorgehensweise von SVMs kann im Rahmen dieser Arbeit nicht geboten werden. Eine sehr gute Einführung und einen Überblick über die Einsatzmöglichkeiten von SVMs im Data-Mining bietet Steinwart und Christmann (Steinwart & Christmann, 2008). Eine grobe Skizzierung der Arbeitsweise bietet Abb. 12.

Der in dieser Arbeit genutzt Algorithmus trägt die Bezeichnung ***SMO***[38] (sequential minimal optimization algorithm for support vector classification) und ist in WEKA durch die Klasse weka.classifiers.functions.SMO implemetiert.

4.5 Bagging

Der Meta-Lerner Bagging[39] beschreibt ein Ensemble-Verfahren[40] aus dem Jahr 1994 (Breiman, 1996) zur Optimierung eines Klassifikationsmodells (hauptsächlich für Entscheidungsbaumlerner). Dazu werden, wie in Abb. 13 dargestellt ist, k Stichproben (mit Zurücklegen) $T_1, T_2, \dots, T_k$ mit derselben Größe n, von der Trainingsmenge T erzeugt. Auf allen Stichproben wird anschließend ein entsprechendes Klassifikationsmodell (Entscheidungsbaum) durch einen Lernalgorithmus generiert.

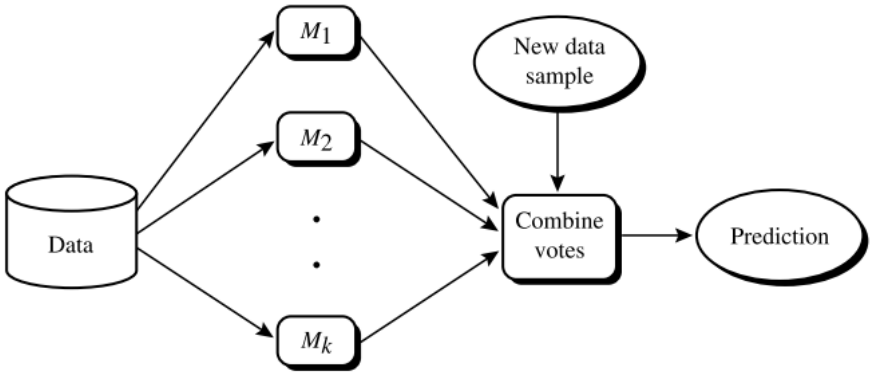


Abbildung 13 - Ensemble-Verfahren
(Han & Kamber, 2006, S. 366)

[38] Eine Einführung in die Arbeitsweise von SMO kann im Rahmen dieser Arbeit nicht gegeben werden, findet sich jedoch in (Platt, 1999)
[39] Ursprünglich *„Bootstrap Aggregating“*, wird durch die Klasse weka.classifiers.meta.Bagging in WEKA implementiert
[40] Beschreibt allgemein die Kombination verschiedener Basislerner

Bei der finalen Evaluation auf den Testdaten, oder der Klassifikation von neuen Instanzen, dienen diese Modelle dann als eine Art „Expertengremium“ (Ähnlichkeit zum vorgestellten *„Mehrheitsvotum“*, Kapitel 3, bei dem die unterschiedlichen ärztlichen Diagnosen aggregiert wurden) und bestimmen, in Abhängigkeit des Attributtyps des Klassenattributs, gemeinsam eine Vorhersage. Im Falle eines numerischen Zielkonzepts werden die Vorhersagen aller Modelle durch eine Mittelwertbildung aggregiert (s. Formel 3.1, Kapitel 3). Bei nominalen Attributen entscheidet das häufigste Votum der Modelle die Vorhersage.

Hintergrund des Verfahrens ist die Instabilität von Klassifikationsmodellen, insbesondere der Entscheidungsbaumlerner. Bereits bei geringfügigen Änderungen der Trainingsdaten sind oft große, strukturelle Unterschiede im Entscheidungsbaum zu erkennen. Mit dem Bagging wird dieser Schwäche Rechnung getragen, indem die Varianz reduziert wird. Begründet wird diese Methode mit einem Verfahren namens „Bias-Variance-Decomposition“, bei dem der Klassifikationsfehler in seine drei Bestandteile „Noise“, „Bias“ und „Variance“ theoretisch aufgeteilt wird. Der Gesamtfehler wird durch eine Varianz-Reduktion damit reduziert[41].

In vielen Fällen stellt sich durch den Einsatz von Bagging neben einer erheblichen Varianzreduktion ebenfalls ein signifikanter Performanzzuwachs in der Klassifikation ein. In sehr wenigen Anwendungen kann das aggregierte Modell jedoch auch schlechter abschneiden (Witten & Frank, 2005, S. 317). Die Versuchsreihen im folgenden Kapitel enthalten jeweils einen Performanzvergleich zwischen Experimenten mit und ohne Bagging.

[41] Eine ausführliche, mathematische Erläuterung ist u.a. in (Hastie, Tibshirani, & Friedman, 2009) zu finden

5 Experimente

Um das gegebene Ziel einer Hautkrebsvorhersage zu erreichen, werden in diesem Kapitel, analog zu den Phasen *Modeling* und *Evaluation* im CRISP-Modell (s. Kapitel 2.1), mehrere Modelle vorgestellt und mit den in Kapitel 4 eingeführten Algorithmen getestet. Datengrundlage hierfür sind die aufbereiteten Trainingsdaten (Kapitel 3) in mehreren Versionen. Hierfür wurden aus dem ursprünglichen Trainingsset die Inkonsistenzen der ärztlichen Untersuchung und sonstigen Attributen beseitigt (Kapitel 3.1.2), irrelevante Attribute manuell gelöscht (manuelle *Feature-Subset Selection*, Kapitel 3.4) und fehlende Werte durch eine *k-nearest-neighbor* Klassifikation ersetzt (Kapitel 3.1.3). Instanzen ohne Ausprägung des jeweiligen Klassenattributs[42] wurden nicht berücksichtigt. Um die Auswirkungen der zahlreichen Ersetzungen auf die Performanz der Modelle zu verfolgen, wurden jeweils vier Trainingsmengen getestet (s. Kapitel 3.1.3):

- T_1: Mit fehlenden Werten
- T_2: Ohne fehlende Werte[43]
- T_3: Volle Ersetzung aller fehlenden Werte mit KNN
- T_4: Halbe Ersetzung der fehlenden Werte mit KNN[44]

Ein Überblick über die Performanzmaße der einzelnen Kombinationen aus den Trainingsmengen $T_1, \dots, T_4$ sowie den Algorithmen[45] J48 (Entscheidungsbaum-Lerner C4.5), JRip (Regel-Lerner RIPPER), Naive Bayes und SMO (Support-Vector-Machine) wird aus Platzgründen auf Tabelle 10-12 geboten. Fett formatiert ist pro Modell und Trainingsmenge der jeweils beste Algorithmus, sowie blau eingefärbt die beste Kombination aus Algorithmus und Trainingsmenge T_i. Zur Vergleichbarkeit und Nachweisbarkeit eines Lerneffekts wurde pro Modell und Trainingsmenge ein Baseline-Wert angegeben, der dem Erwartungswert der häufigsten Klassenausprägung entspricht. Interpretierbar

[42] 167 Instanzen ohne *Beurteilung*
[43] Löschungen aller Instanzen ohne fehlende Ausprägung in den relevanten Attributen: 4.006 vollständige Instanzen (ca. 57,74% der vorhandenen Datenmenge) genutzt
[44] 4.629 genutzte Instanzen (623 unvollständige Datensätze durch Klassifizierung ergänzt)
[45] Jeweils mit Default-Werten

ist der Baseline-Wert dabei als eine untere Schranke für die Performanz der Lerner. Erreicht ein Lernalgorithmus demnach bessere Werte als der Baseline-Wert, ist er einem simplen „Raten“ auf Basis der Verteilung des Klassenattributs überlegen und ein Lerneffekt damit evident.

Desweiteren lässt sich die Vermutung der fortgeführten, oder multiplizierten Verzerrung (Kapitel 3.1.3) durch die Resultate über alle Modelle hinweg verifizieren. Trainingsmenge T_3, mit einer vollen Ersetzung aller fehlenden Werte, belegt beim Ampel- und Ärztemodell durchschnittlich jeweils den letzten Platz und ist der Variante T_4 lediglich beim Patientenmodell überlegen. T_4 generiert mit Ausnahme des Patientenmodells erwartungsgemäß die besten Ergebnisse, da ein guter Mittelweg zwischen einer hinreichenden Datenmenge zur Erstellung eines genauen Modells und einer Verzerrung der Daten gefunden werden konnte.

Auffällig war zusätzlich die unterschiedliche Performanz der Lernalgorithmen für die unterschiedlichen Trainingsmengen. Obwohl JRip bei der Trainingsmenge T_4 für das Ampel- und Ärztemodell die beste Performanz erzielte, war J48 für T_1 geeigneter (durchschnittlich um 0,519% besser als JRip). Dieser Sachverhalt lässt den Schluss zu, dass J48 im vorliegenden Fall besser mit fehlenden Werten arbeiten kann als JRip.

5.1 Patientenmodell

Eine computergestützte Hautkrebsvorhersage ist besonders nützlich in Form einer sekundären Beratungsfunktion für den Patienten. Da eine individuelle, ärztliche Diagnose nie substituiert werden kann und soll, platziert sich ihr Einsatzgebiet vor allem im Aufklärungs- und Warnbereich. Ist sich bspw. ein bisher unauffälliger Patient (seltenes Aufsuchen eines Facharztes) bezüglich seiner Anfälligkeit für Hautkrebs unsicher, kann er durch Angabe von persönlichen Daten wie *Alter*, *Geschlecht*, *Freizeitverhalten* etc. eine Tendenz bestimmen lassen. Eine mögliche Form der Eingabe wäre dabei ein Online-Formular als Umsetzung des bisher benutzten Fragebogens (Abb. 1), der in Kapitel 1.2 vorgestellt wurde. Als erklärende, unabhängige Attribute wurden deshalb für diese

Versuchsreihe eingangs alle Patientendaten (12 Attribute) des ersten Fragebogens gewählt:

- *Alter*
- *Geschlecht*
- *Outdoor-Zeit*
- *Haut-Reaktion (Sonnenbrand)*
- *Sonnenbrand als Kind*
- *Sonnenbrand als Jugendlicher*
- *Sonnenbrand als Erwachsener*
- *Schutz vor Sonneneinstrahlung*
- *Solariumnutzung*
- *Outdoor-Sport*
- *Hautkrebs in Vergangenheit*
- *Hautkrebs in Familie*

Als abhängiges, oder zu erklärendes Attribut eignet sich insbesondere die ärztliche *Beurteilung* (Frage 8 des 2. Fragebogens, Abb. 2). Individuelle Hautkrebsarten wie das maligne Melanom, Spinaliom und Basaliom können ebenfalls gewählt werden, sind jedoch als einzelne, meist unabhängige Attribute[46] ungeeignet um eine umfassende Risikoanalyse zu einem generellen Hautkrebsrisiko zu ermöglichen.

Die insgesamt besten Resultate für das Patientenmodell ergab eine Kombination aus einer Support-Vector-Machine mit der Trainingsmenge T_4. Aus Gründen der Verständlichkeit für spätere Anwender wurde jedoch für eine *Feature Subset Selection* (s. Kapitel 3.4, Wrapper-Verfahren) der Entscheidungsbaum-Lerner J48 (mit Pruning) mit dessen performantester Trainingsmenge (T_1) gewählt. Ausgewählt wurde dabei ausschließlich das Attribut *Hautkrebs in der Vergangenheit*[47]. Die Performanz konnte dadurch um 0,536% auf 65,031% gesteigert werden. Der entsprechende, stark vereinfachte Baum ist im Folgenden abgebildet:

[46] Große Differenzen zwischen bisherig bekannten Risikofaktoren der einzelnen Hautkrebsarten (s. Tabelle 1: Vergleich malignes Melanom und nichtmelanozytärer Hautkrebs)
[47] Die Anzahl der Blätter wurde damit von 65 auf 2 reduziert

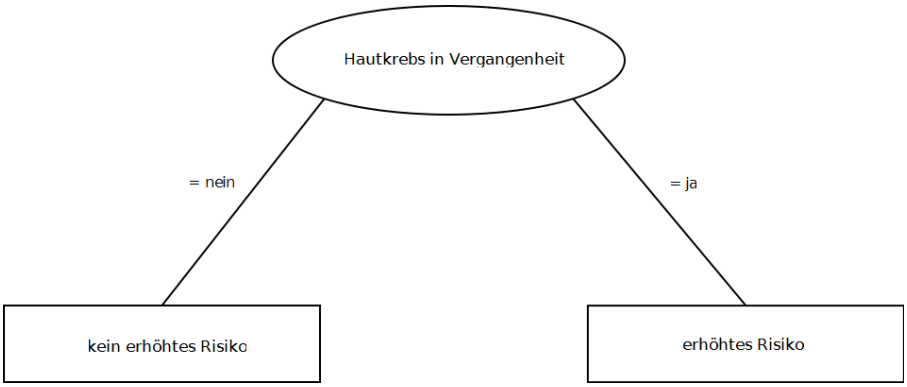


Abbildung 14 - Patientenmodell Entscheidungsbaum

Der unzureichende Informationsgewinn, (Performanz des Lerners – Baseline-Wert = 2,67%), die stark vereinfachte Attributmenge (ein einzelner Risikofaktor), sowie die bedeutsame Anzahl an *False-Negative* Klassifikationen weisen jedoch darauf hin, dass dieses Modell wenig praktische Bedeutsamkeit erlangen kann. Zur Skizzierung der Klassifikationsfehler kann eine *Konfusionsmatrix* verwendet werden (Witten & Frank, 2005, S. 160-165). Hierbei werden die klassifizierten Werte den tatsächlichen Ausprägungen in einer Tabellenform gegenüber gestellt (s. Abb. 14). Die jeweils richtigen Klassifikationen befinden sich auf der Diagonalen der Matrix. Im Fall eines binären Klassenattributs (trifft für die *Beurteilung* zu) kann zudem die Unterscheidung zwischen *False-Negative* und *False-Positive* bei den Fehlklassifikationen getroffen werden. *False-Negative* bezeichnet einen Klassifikationsfehler, bei dem fälschlicherweise ein negatives („kein erhöhtes Risiko“) Ergebnis prognostiziert wird. Analog steht *False-Positive* für einen Fall bei dem irrtümlich ein positives („erhöhtes Risiko“) Resultat vorhergesagt wird.

Abbildung 15 - Konfusionsmatrix Patientenmodell

a	b	$\overleftarrow{klassifiziert\ als}$
247	2.284	$a = erhöhtes\ Risiko$
67	4.125	$b = kein\ erhöhtes\ Risiko$

Gemäß Abb. 14 liegt beim Patientenmodell eine bedeutende Anzahl von *False-Negatives* vor. Dabei wurden 2.284 Patienten (90,24%) mit realem Risiko durch „kein erhöhtes Risiko“ klassifiziert. Für einen medizinischen Einsatz, sei er auch nur präventiv, eignet sich dieses Modell daher in keinem Fall, da bei einer solch sicherheitskritischen Anwendung Klassifikationsfehler im schlimmsten Fall die Gesundheit eines

Menschen ernsthaft gefährden. Mögliche Wege zur Verbesserung des Ergebnisses werden in Kapitel 5.3 aufgegriffen.

5.2 Ärztemodell

Kapitel 5.1 beschrieb ein Einsatzgebiet des Modells zur Patienten-Aufklärung und Sensibilisierung. Eine alternative Anwendungsmöglichkeit bildet die Unterstützung ärztlicher Arbeit durch eine erste, tendenzielle Rückmeldung neben einer individuellen, aufwändigen Diagnose. Da hierfür erste „ärztliche" Attribute genutzt werden konnten, ließ sich die Performanz des Modells im Vergleich zum Patientenmodell durch die Attributhinzunahme erheblich steigern. Hierfür wurden Attribute gewählt, die keine Laborergebnisse und zeitaufwändige Diagnosen zur Auswertung benötigen[48], um das Ziel der Prävention, bzw. einer ersten Einschätzung, weiter zu verfolgen. Gewählt wurden deshalb zusätzlich die Attribute:

- *Anzahl der Pigmentmale*
- *Dysplastische Nävi*
- *Aufklärung Hauttyp*
- *Aufklärung Lichtschutz*
- *Aufklärung ABCD-Regel*

Die besten Resultate dieses Modells erzielte, wie bereits im Ampelmodell, der Regellerner JRip auf der Trainingsmenge T_4. Diese Ergebnisse werden im Folgenden nach einer Feature Subset Selection (8 anstelle von 17 Attributen) dargestellt:

- *(Dysplast. NZN = ja)* ➔ *Beurteilung=erhöhtes Risiko (709.0/132.0)*
- *(Hautkrebs in Vergangenheit = ja)* ➔ *Beurteilung=erhöhtes Risiko (176.0/38.0)*
- *(Alter ≥ 62) und (Aufklärung - ABCD-Regel (MM) = nein)* ➔ *Beurteilung=erhöhtes Risiko (162.0/59.0)*

[48] Die gängigsten, diagnostischen Verfahren sind Gewebeproben, die in einem Labor untersucht werden müssen, oder Methoden der Auflichtmikroskopie [durch optische Hilfsmittel (bspw. hochauflösende Kameras, oder Körper-Scans) werden Befunde digital, oder analog vergrößert dargestellt und analysiert] (Berger, 2009, S. 66-71)

- *(Anzahl Pigmente ≥ 70) und (Alter ≤ 41) und (Geschlecht = männlich)* ➔ *Beurteilung=erhöhtes Risiko (91.0/34.0)*
- *(Anzahl Pigmente ≥ 70) und (Solariumsnutzung ≤ 0) und (Alter ≥ 21) und (Alter ≤ 42)* ➔ *Beurteilung=erhöhtes Risiko (52.0/19.0)*
- *(Anzahl Pigmente ≥ 35) und (Hautkrebs in der Familie = ja) und (Alter ≥ 53)* ➔ *Beurteilung=erhöhtes Risiko (18.0/4.0)*
- *(Anzahl Pigmente ≥ 35) und (Hautkrebs in der Familie = ja) und (Alter ≥ 34) und (Alter ≤ 39)* ➔ *Beurteilung=erhöhtes Risiko (24.0/7.0)*
- ➔ *Beurteilung=kein erhöhtes Risiko (3397.0/792.0)*

Die Performanz des Lerners ließ sich dabei um 0,043% auf 74,919% steigern. Die Konfusionsmatrix in Abb. 15 lässt erkennen, dass auch in diesem Modell sehr viele *False-Negatives* existieren. Jedoch sank die Fehlerquote signifikant um 41,829% auf 48,411% in Bezug auf die Klassifikationsfehler im Risikobereich.

Abbildung 16 - Konfusionsmatrix Ärztemodell

a	b	$\xleftarrow{klassifiziert\ als}$
893	838	$a = erhöhtes\ Risiko$
323	2.575	$b = kein\ erhöhtes\ Risiko$

Die Hinzunahme der „ärztlichen“ Attribute bewirkte damit eine erhebliche Verbesserung der Performanz.

5.3 Ampelmodell

Der ursprüngliche Fokus der Arbeit lag auf der Einstufung von Patienten in Risikoklassen. Eine von vielen möglichen Einstufungen wurde durch die vorangegangen Kapitel 5.1 und 5.2 vorgestellt. Eine alternative Formulierung der Risikoklassen wird im Folgenden durch Modellierung eines neuen Attributs getestet. Dabei wurde das bisherige Klassenattribut *Beurteilung* aufgrund seiner Inkonsistenz nicht weiter verwendet. Stattdessen wurde ein neues Attribut generiert, das die verschiedenen Hauterkrankungen entsprechenden Risikostufen zuordnet. Der Grundgedanke des Ampelmodells ist eine Risikobeurteilung der Hautkrebsarten anhand von Mortalitätsraten und Behandlungs-

chancen. Den individuellen Ausprägungen des Hautkrebs (*Basaliom*, *Spinaliom*, *malignes Melanom* etc.) werden anhand dieser Daten Risikostufen zugeordnet. Der Aufbau der Ampel wird untenstehend erläutert:

Tabelle 9 - Risiko der Hautkrebsarten

	Basaliom	***Spinaliom***	***malignes Melanom***
Metastasierungswahrsch.	0,2515	5,5	∞
Letalität	0,1	5	17,5[49]

Gemäß Tabelle 9, die die Informationen aus Kapitel 1.2 verkürzt darstellt, wurden Patienten in die Risikoklassen *rot*, *gelb* und *grün* unterteilt. Dabei ist ein erheblicher Anstieg der Metastasierungswahrscheinlichkeit und Letalität um die Faktoren 21,8 bzw. 50 im Vergleich der nicht-melanozytären Hautkrebsarten erkennbar. Das Spinaliom wurde deshalb gemeinsam mit dem malignen Melanom auf die höchste Risikostufe (*rot*) gesetzt. Stufe *gelb* teilten sich mit einem vergleichbar geringen Risiko das Basaliom und die aktinischen Keratosen als Vorstufe des Spinalioms. Waren keine der Befunde bei einem Patienten ersichtlich, wurde ihm die Risikostufe *grün* zugeordnet.

- *Rote Stufe*: Erhebliches Risiko (Malignes Melanom + Spinaliom)
- *Gelbe Stufe*: Moderates Risiko (Basaliom + aktinische Keratosen)
- *Grüne Stufe*: Kein erkennbares Risiko (Kein Hautkrebsbefund)

Wird der durchschnittliche Informationszuwachs (durchschnittliche Lernleistung pro Trainingsmenge T_i für alle Lerner – Baseline-Wert der Trainingsmenge) als Evaluationskriterium des Modells betrachtet, erzielt das Ampelmodell die schlechtesten Werte der drei getesteten Modelle. Bei keinem der getesteten Trainingsmengen $T_1, \dots, T_4$ konnte durchschnittlich ein positiver Lerneffekt verzeichnet werden. Ein simples Raten der Werte für neue Instanzen auf Basis der Verteilung der Klassenausprägungen in der Trainingsmenge wäre demzufolge den Klassifikationsmodellen vieler Algorithmen überlegen. Dieser Wert wird jedoch stark durch das konkurrenzunfähige Ergebnis des J48-Lerners (ohne Pruning) verzerrt. Wird dieser Wert isoliert und ein Durchschnitt der restlichen Ergebnisse gebildet, ist ein positiver durchschnittlicher Informationszuwachs von 0,083% (für Trainingsmenge T_1) bis 0,16% (für Trainingsmenge T_3) zu verzeichnen.

[49] Grundlage ist die Schätzung des Robert-Koch-Instituts der absoluten Inzidenz und Mortalität in Deutschland im Jahr 2000 (Breitbart, Wende, Mohr, Greinert, & Volkmer, 2004, S. 7 ff.).

Das beste Resultat erzielte für dieses Modell der Regellerner JRip, dessen Ergebnis im Folgenden nach einer *Feature Subset Selection* (automatisiertes Wrapper-Verfahren, s. Kapitel 3.4) detailliert dargestellt wird:

- *(Alter ≥ 57) und (Aufklärung - ABCD-Regel (MM) = nein) und (Geschlecht = männlich)* ➔ *Risikoklasse=gelb (119.0/41.0)*
- *(Alter ≥ 72) und (Aufklärung - ABCD-Regel (MM) = nein)* ➔ *Risikoklasse=gelb (42.0/16.0)*
- *(Alter ≥ 67) und (Alter ≤ 69) und (Sonnenbrand als Kind = selten) und (Geschlecht = männlich)* ➔ *Risikoklasse=gelb (24.0/8.0)*
- *(Alter ≥ 84)* ➔ *Risikoklasse=gelb (23.0/11.0)*
- *(Alter ≥ 56) und (Geschlecht = männlich) und (Sonnenbrand als Kind = häufig) und (Haut-Reaktion (Sonnenbrand) = häufig)* ➔ *Risikoklasse=gelb (10.0/1.0)*
- ➔ *Risikoklasse=grün (4411.0/359.0)*

Neben der erheblichen Reduktion der Attribute (18 Attribute → 5 Attribute) und einer daraus resultierenden, leichteren Interpretierbarkeit des Modells, stieg die Performanz um 0,173% auf 89,674%. Als relevanteste Attribute zur Klassifikation wurden gemäß der oben abgebildeten Regelmenge die folgenden Attribute gewählt: *Alter, Aufklärung – ABCD-Regel (MM), Geschlecht, Sonnenbrand als Kind, Haut-Reaktion (Sonnenbrand).* Drei der fünf Attribute lassen sich durch die in Kapitel 1.1 dargestellten Risikofaktoren (Tabelle 1) verifizieren [*Haut-Reaktion (Sonnenbrand)* entspricht nach Aussage von Dr. Michael Herbst dem Hauttyp eines Menschen. Die Ausprägung „häufig" kann mit Hauttyp II gleichgesetzt werden]. Interessant ist neben der Aufnahme des Attributs *Geschlecht* vor allem die Hinzunahme von *Aufklärung – ABCD-Regel (MM).* Ein (fehlendes) Bewusstsein für die Entstehung und die Folgen einer Hautkrebserkrankung wird somit indirekt als weiterer Risikofaktor auf dieser Datenmenge erkannt. Die relativ geringe, positive Abdeckung der beiden betroffenen Regeln (65,546% für Regel 1 und 61,904% für Regel 2) lässt jedoch die Richtigkeit dieser Regeln anzweifeln.

Abbildung 17 - Konfusionsmatrix Ampelmodell

a	b	c	$klassifiziert\ als \leftarrow$
112	362	0	$a = gelb$
87	4.039	0	$b = grün$
4	25	0	$c = rot$

Als weitere Evaluationsmöglichkeit wurde auch für dieses Modell eine Konfusionsmatrix angegeben (Abb. 15). Erkennbar ist hierbei ebenfalls eine beträchtliche Anzahl von Klassifikationsfehlern bei Risikopatienten (76,37% Fehlklassifikation für die Klasse „gelb“ und 100% Fehlklassifikation für die Klasse „rot“).

Auffällig ist zudem, dass keine Regel zur Klassifikation der Klassenausprägung „rot“ gelernt wurde. Eine Erklärung hierfür ist die ungünstige Verteilung der Testdaten[50] („rot“ als Minderheitsklasse lediglich mit 29 Instanzen vertreten). Mögliche Wege zur Behebung dieses Problems sind bspw. (Debray, 2009):

- ***Sampling-Methoden:***
  Erzeugen Variationen der Verteilung (in WEKA z.B. durch *weka.filters.supervised.instance.SMOTE* implementiert)
- ***Ensemble-Verfahren*** (s. Kapitel 4.5)
- ***Kostensensitives Lernen:***
  Erlaubt die unterschiedliche Gewichtung von Klassifikationsfehlern durch Angabe einer Kostenmatrix durch den Benutzer (bspw. enthalten in *weka.classifiers.meta.CostSensitiveClassifier*)

Aus Zeitgründen können diese Verfahren leider nicht in dieser Arbeit behandelt werden, sind jedoch für anschließende Experimente sehr empfohlen. In der jetzigen Form kann das Modell nicht in der Praxis Anwendung finden.

[50] Die Split- bzw. Auswahl-Kriterien in vielen Algorithmen (JRip, J48 etc.) basieren auf einer Maximierung der Klassifikationsgenauigkeit. Bei vorliegender, ungleicher Verteilung (0,626% der Daten als „rot“ klassifiziert) erzielt der Lernalgorithmus insgesamt eine höhere Genauigkeit durch die Vorhersage der Mehrheitsklasse (in diesem Fall „grün“ mit 89,13% aller Instanzen)

Tabelle 11 - Ampelmodell Auswertung

Ampel-modell	Baseline (Erwartungswert)	JRip	J48 (mit Pruning)	J48 (ohne Pruning)	J48 (mit Pruning + Bagging)	SMO	Naive Bayes	Durchschnitt	durchschnittlicher Zuwachs
T1 (mit fehlenden)	87.712%	87.668%	87.963%	84.803%	**88.052%**	87.172%	87.579%	87.206%	-0.506%
T2 (Löschung)	88.872%	89.047%	**89.122%**	84.905%	89.072%	88.872%	88.997%	88.336%	-0.536%
T3 (volle Ersetzung)	87.669%	87.874%	**88.130%**	84.009%	88.026%	87.669%	87.446%	87.193%	-0.477%
T4 (halbe Ersetzung	89.134%	89.501%	89.177%	85.958%	89.479%	89.134%	89.134%	88.730%	-0.403%
Durchschnitt		88.522%	88.598%	84.919%	**88.657%**	88.212%	88.289%		

Tabelle 12 – Patientenmodell Auswertung

Patienten-modell	Baseline (Erwartungswert)	JRip	J48 (mit Pruning)	J48 (ohne Pruning)	J48 (mit Pruning + Bagging)	SMO	Naive Bayes	Durchschnitt	durchschnittlicher Zuwachs
T1 (mit fehlenden)	62,353%	64,257%	64,495%	58,337%	63,826%	**65,031%**	64,168%	63,352%	0,999%
T2 (Löschung)	61,252%	62,949%	63,124%	56,263%	61,078%	**63,922%**	63,323%	61,776%	0,524%
T3 (volle Ersetzung)	62,353%	**64,302%**	63,573%	57,995%	62,532%	65,035%	63,707%	62,857%	0,504%
T4 (halbe Ersetzung	62,605%	63,988%	63,686%	57,010%	61,482%	65,025%	63,794%	62,497%	-0,108%
Durchschnitt		63,874%	63,719%	57,401%	62,229%	**64,753%**	63,748%		

Tabelle 10 - Ärztemodell Auswertung

Ärztemodell	Baseline (Erwartungswert)	JRip	J48 (mit Pruning)	J48 (ohne Pruning)	J48 (mit Pruning + Bagging)	SMO	Naive Bayes	Durchschnitt	durchschnittlicher Zuwachs
T1 (mit fehlenden)	62,353%	74,163%	**74,148%**	67,187%	73,881%	73,554%	73,152%	72,681%	10,328%
T2 (Löschung)	61,252%	**74,376%**	74,027%	66,667%	73,254%	73,927%	72,954%	72,534%	11,282%
T3 (volle Ersetzung)	62,353%	73,881%	**74,253%**	66,444%	73,494%	73,554%	72,810%	72,406%	10,053%
T4 (halbe Ersetzung	62,605%	74,876%	74,249%	67,056%	73,536%	74,379%	73,493%	72,932%	10,326%
Durchschnitt		**74,324%**	74,169%	66,838%	73,541%	73,853%	73,102%		

6 Diskussion und Ausblick

In diesem Kapitel wird der Ablauf der Arbeit abschließend reflektiert und mögliche Folgeschritte erläutert. Die mangelhafte Qualität der Ergebnisse ist auf mehrere Ursachen zurückzuführen. Einer der wesentlichen Gründe hierfür war die Qualität der Datenerfassung, bzw. der Fragen des Fragebogens. Eine Vielzahl von Fragen ermöglichte eine individuelle Einschätzung der Testpersonen bezüglich der Beantwortung, die letztlich eine Verzerrung des Datensatzes herbeiführte. Beispielhafte Ausschnitte sind unten abgebildet:

- *Frage 3: Wie oft halten sie sich bei intensiver Sonneneinstrahlung in der Sonne auf?*
    - *So häufig wie möglich*
    - *Gelegentlich*
    - *Eher selten*
    - *Ich meide die Sonne*

Eine genaue Einschätzung dieser und anderer Fragen seitens der Patienten fällt schwer, da Abgrenzungen oft schwammig ausfallen und nicht allgemeingültig definiert werden („gelegentlich“, „eher selten“). Eine Erfassung im Hinblick auf konkrete quantitative, begreifbare und einheitlich interpretierbare Größen wie etwa eine zeitliche Dauer, erscheint für zukünftige Datenerfassungen demzufolge sinnvoller.

Die Problematik der qualitativen Diskrepanz der Umfrage findet sich u.a. in der Sport-Frage, Frage 8 des Patienten-Fragebogens (Abb. 1), ebenfalls wieder (s. Kapitel 3.1.2). Für zwei von drei Hautkrebsarten (Basaliom, Spinaliom)[51] ist die kumulierte Dauer der ungeschützten Sonnenexposition ausschlaggebend, die durch Frage 8 nicht abgedeckt wurde. Bei einer Konkretisierung und Quantifizierung dieser Frage entstünde ein deutlicher Informationszuwachs. Ein beispielhafter Vorschlag ist:

- *Welche der folgenden Sportarten übten Sie in den letzten Jahren regelmäßig für die angegebenen Wochenstunden x aus? (Antwortmöglichkeiten pro angekreuzter Frage)*

[51] S. Tabelle 1 – Risikofaktoren von nichtmelanozytären Hautkrebsarten

- $x \leq 1$ *Stunde*
- $1 < x \leq 2{,}5$ *Stunden*
- $2{,}5 < x \leq 5$ *Stunden*
- $x > 5$ *Stunden*

Ähnliche Schwierigkeiten lassen sich in Frage 5 erkennen. Erfragt wurde wie viele schwere Sonnenbrände (inkl. Blasenbildung) der Patient in den Lebensabschnitten Kindheit, Jugendalter und als Erwachsener erlitten hat (Abb. 1). Die kategorialen Ausprägungen „nie, selten, häufig, oft“ stellen dabei erneut eine lediglich grobe Einteilung dar. Insbesondere eine Unterscheidung zwischen „häufig“ und „oft“ scheint schwierig. Bei einer Analyse der Verteilung der Daten fällt u.a. auf, dass bei allen betroffenen Attributen (*Sonnenbrand als Kind, Sonnenbrand als Jugendlicher, Sonnenbrand als Erwachsener*) mindestens 82,4% (bei *Sonnenbrand als Kind* 92,1%) der Patienten angaben, entweder „nie“, oder „selten“ von Sonnenbrand betroffen gewesen zu sein. Quantitative Maßstäbe eignen sich deshalb meiner Meinung nach auch bei dieser Fragestellung für eine exaktere, vergleichbare Messung. In dieser Frage findet sich zudem die Erschwernis, besonders für Patienten in hohem Alter, sich die ungefähre Anzahl von Sonnenbränden in Kindheit, oder Jugendalter in Erinnerung zu rufen. Aus diesem Sachverhalt resultierten die erheblichen Fehlbestände dieser Attribute, deren Ersetzung eine zusätzliche Verzerrung des Datensatzes zur Folge hatte. (s. Kapitel 3.1.3). Die vermutlich einzige, jedoch recht aufwändige, Lösung hierfür wäre eine Form von Tagebuch als Dokumentation, das entweder auf Seiten des Patienten, oder des behandelnden Dermatologen geführt werden kann.

Ein weiteres Manko der Datenerfassung ist in Frage 3 des Patientenumfrage (Abb. 1) ersichtlich. Wie bereits erwähnt repräsentiert die kumulierte Sonnenexposition im Leben eines Menschen einen besonderen Risikofaktor (Tabelle 1). Frage 3 zielt auf diesen Faktor ab, setzt jedoch den Fokus lediglich auf das Freizeitverhalten der Patienten. Die Sonnenlichtaufnahme durch berufliche Tätigkeiten wird dabei streng genommen vollkommen vernachlässigt. Anhand der momentanen Datenbasis lässt sich folglich in diesem Attribut bspw. nicht differenzieren zwischen einem Angestellten in einem Bürogebäude, der eine Stunde täglich seine Freizeit draußen verbringt, und einem Bauarbeiter mit dem gleichen Freizeitverhalten, der jedoch zusätzlich während seiner Arbeitszeit der Sonne ausgesetzt ist.

Ein weiterer, qualitätsschmälernder Faktor war die Inkonsistenz in den ärztlichen Untersuchungsergebnissen. Aufgrund der großen Diskrepanz der Diagnosen konnte ein einheitlich gültiges Ergebnis pro Patient nur durch eine Zusammenführung der Einzelresultate jedes Arztes approximiert werden (s. Kapitel 3.1.2).

Scheinbar inkonsistente Angaben bezüglich eines der Klassenattribute (*Beurteilung* für das Patienten- und Ärztemodell) reduzierten ebenfalls merklich die Performanz der Lerner (s. Kapitel 3.1.2).

Ich denke desweiteren, dass sich die Performanz der Modelle, insbesondere des Ampelmodells, bedeutend steigern lässt durch eine größere Anzahl von Patienten mit Hautkrebsbefunden (29 Spinaliome und 31 maligne Melanome wurden insgesamt diagnostiziert), bzw. der Anwendung der in Kapitel 5.3 vorgestellten Verfahren.

Im weiteren Verlauf ist eine Bereinigung der False-Negative Klassifikationen (s. Kapitel 5) unumgänglich. Für ein Modell, das die Gesundheit des Menschen simuliert, wäre für zukünftige Arbeiten ein *kostensensitives Lernen*[52] weitaus besser geeignet, das jedoch im Rahmen dieser Arbeit nicht abgedeckt werden kann.

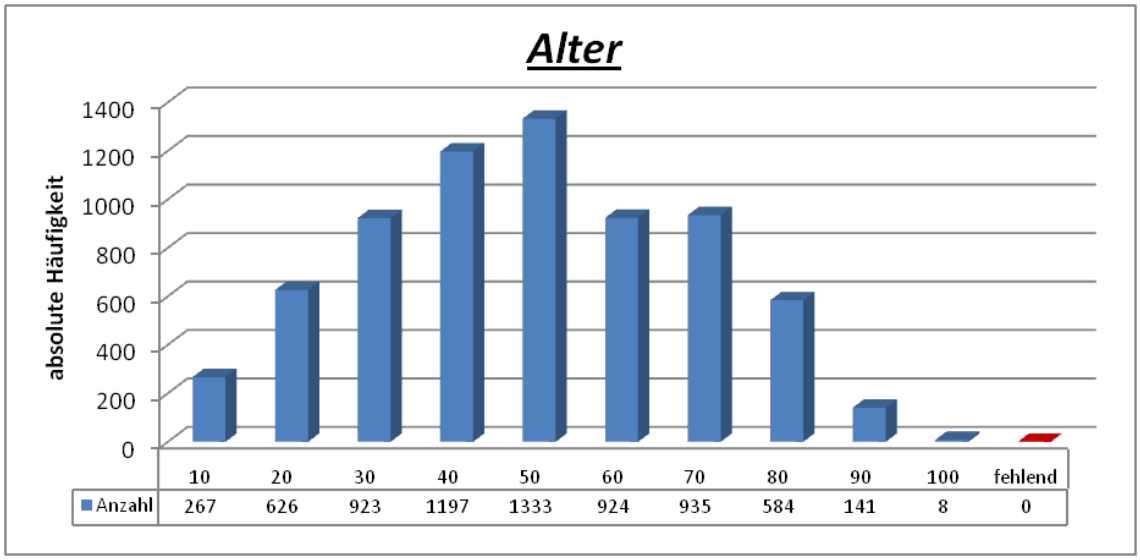
Da jedoch ein genereller Informationszuwachs[53] in den Experimenten zu verzeichnen war, besteht meiner Meinung nach die Annahme, dass Maschinelles Lernen auf Basis einer neuen Datenerhebung in der Lage sein wird, brauchbare Ergebnisse für diagnostische Früherkennung von Hautkrebs zu liefern. Ein nächster, wichtiger Schritt zur Anwendung von Maschinellem Lernen im Gebiet der Hautkrebsprävention ist meiner Ansicht nach die Konzeption eines neuen Fragebogens und dessen anschließende Datenerhebung. Wichtig wird zukünftig ebenfalls eine Abwägung zwischen der Genauigkeit der Vorhersage und der Beschaffbarkeit der Daten sein. Wie am Ärztemodell erkennbar war, ist eine Hautkrebsprävention durch die Hinzunahme zusätzlicher Attribute natürlich performanter. Ist jedoch ein Präventionssystem gefordert, dass nur auf Patientendaten arbeitet, muss eine Performanzverlust meiner Ansicht nach zwangsläufig akzeptiert werden.

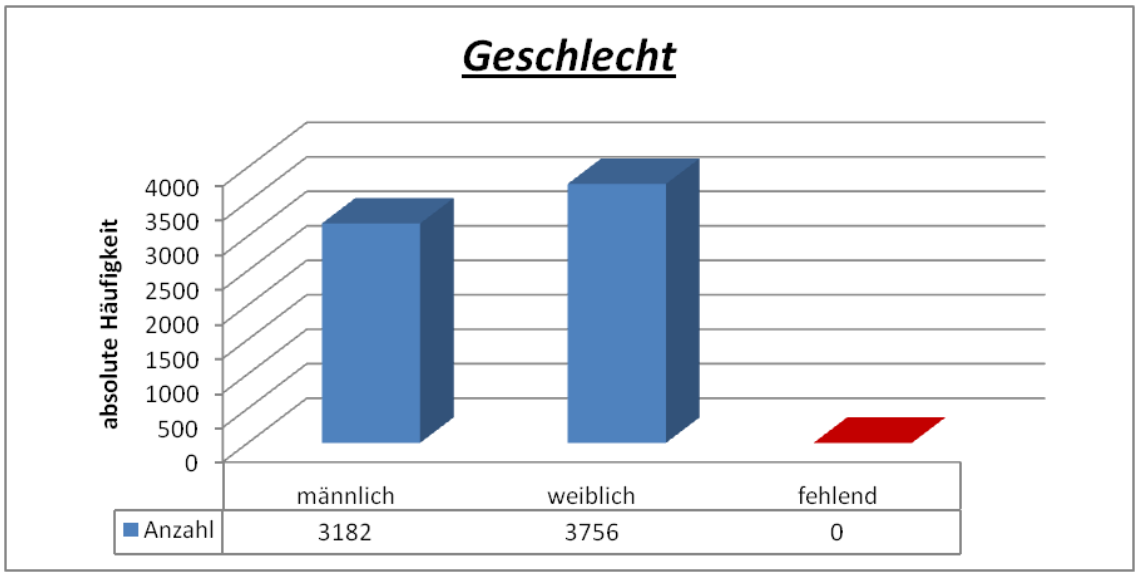
[52] Beim Lernprozess werden Klassifikationsfehler unterschiedlich gewichtet. Weitere Ausführungen können (Sammut & Webb, 2010, S. 231-235) entnommen werden

[53] S. Tabelle 10-12 für einen Vergleich mit dem entsprechenden Baseline-Wert der Trainingsmenge

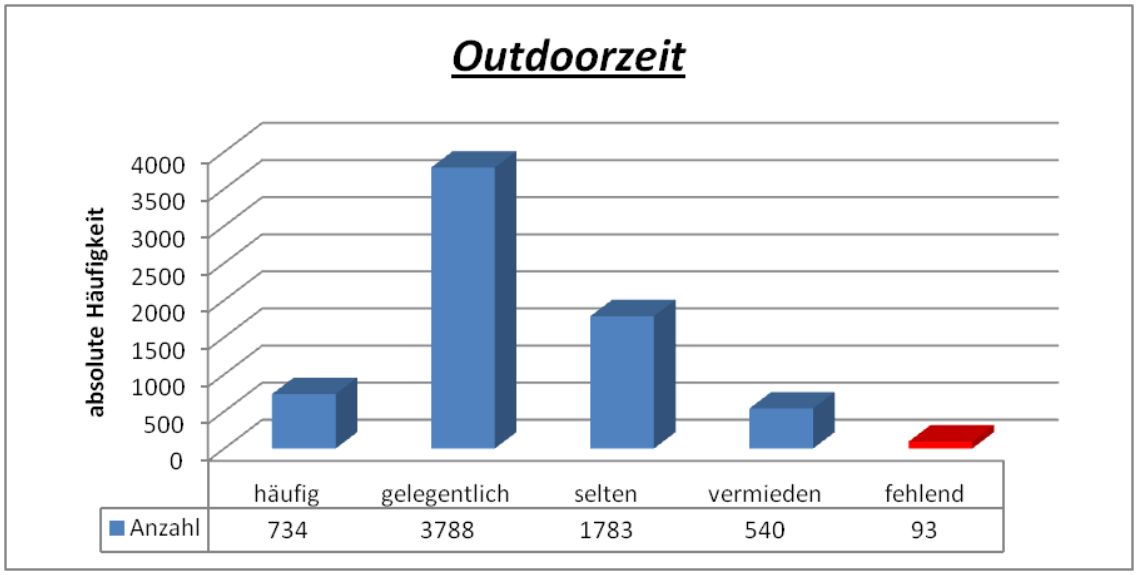
7 Anhang

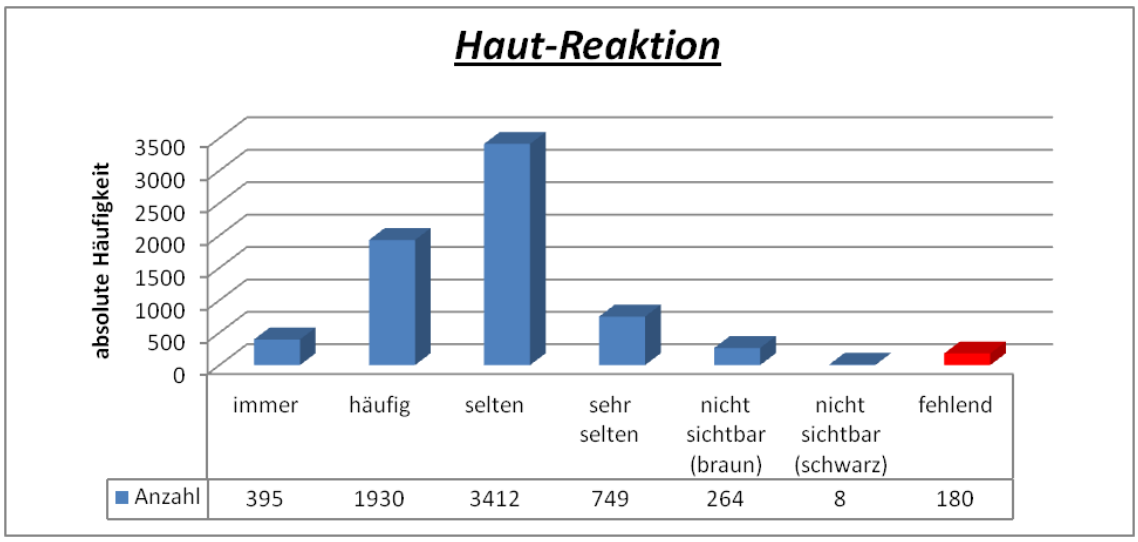
Im folgenden Kapitel sind Häufigkeits-Diagramme zu allen Attributen in ihrer ursprünglichen Form [vor Ersetzung (Kapitel 3.2), jedoch nach Aufbereitung (Kodierung und Behandlung von Inkonsistenzen (Kapitel 3.1)] zu finden. Aus Gründen der Vollständigkeit wurden Attribute, die nicht als relevant für die Modellbildung angesehen wurden, dennoch aufgeführt[54].

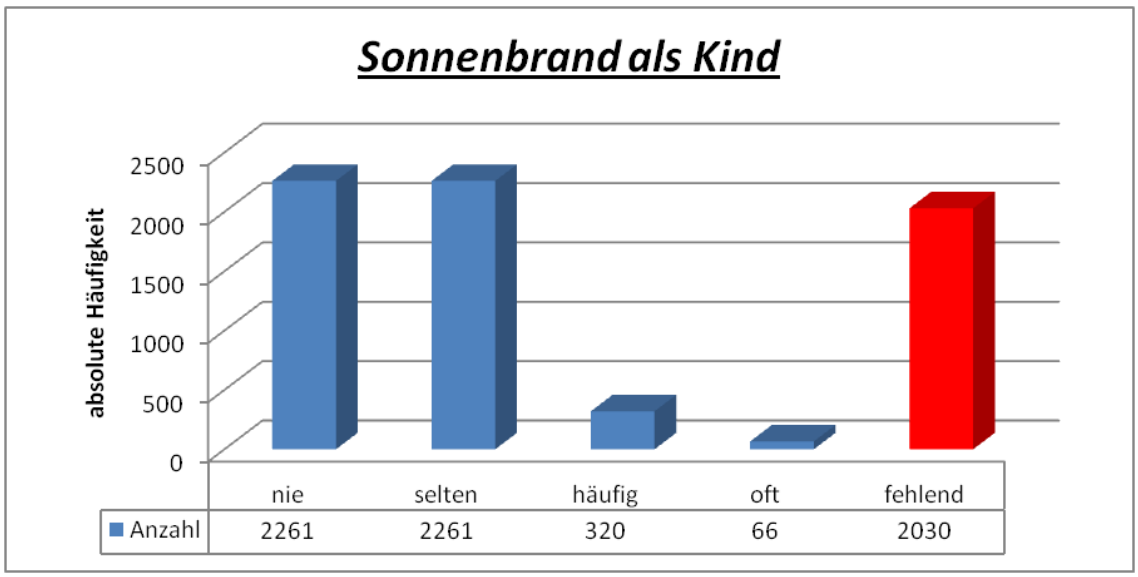


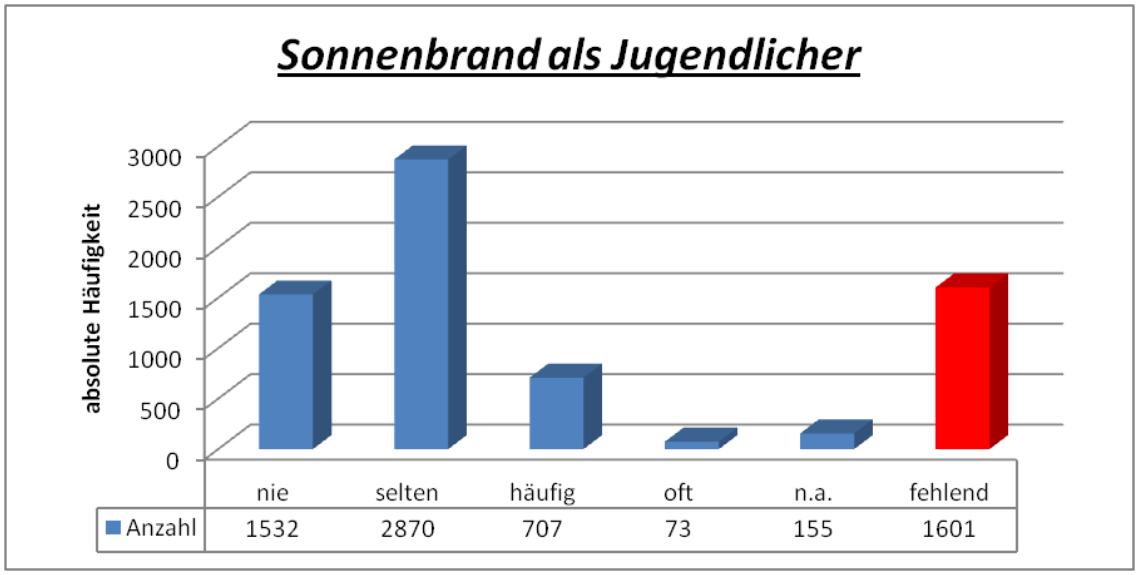


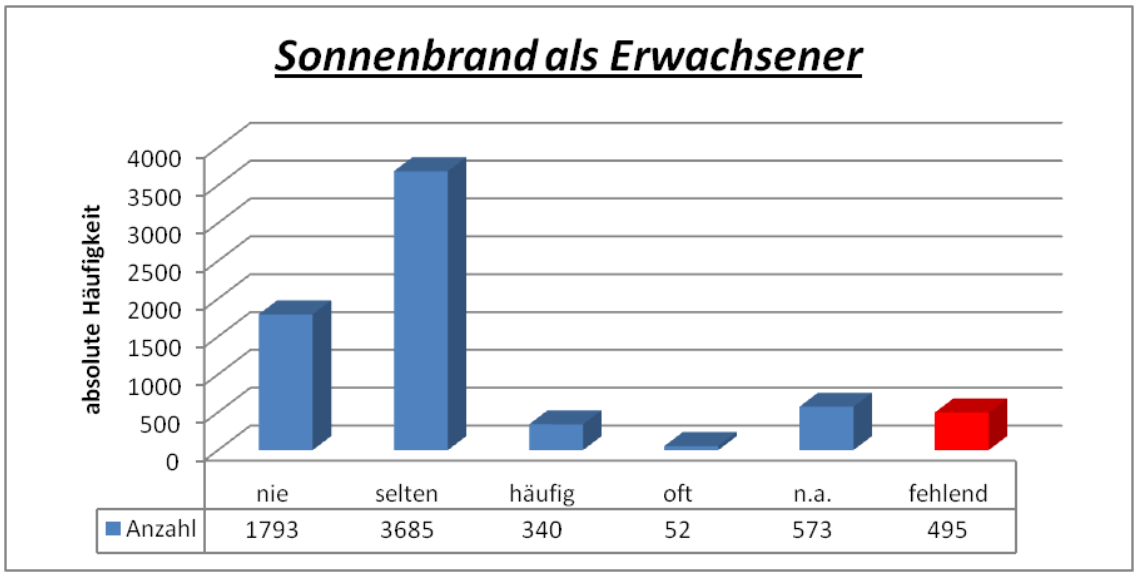
[54] Von dieser Regelung ausgenommen sind die (Primärschlüssel-)Attribute *Ärzte#*, *Bogen#* und *Datum der Untersuchung*

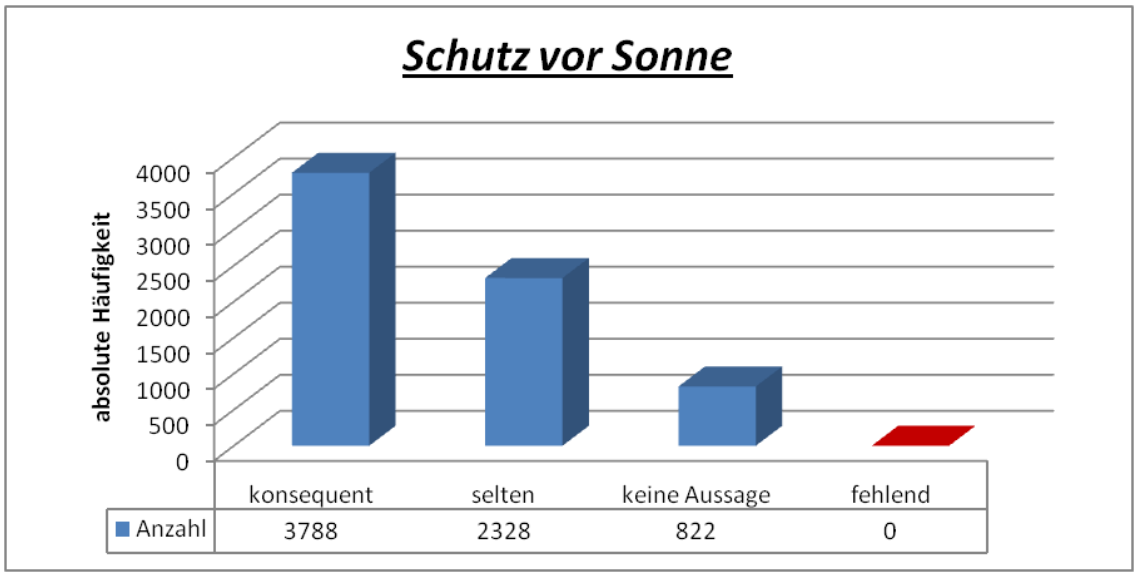

Outdoorzeit
absolute Häufigkeit
4000
3500
3000
2500
2000
1500
1000
500
0
häufig
gelegentlich
selten
vermieden
fehlend
Anzahl
734
3788
1783
540
93

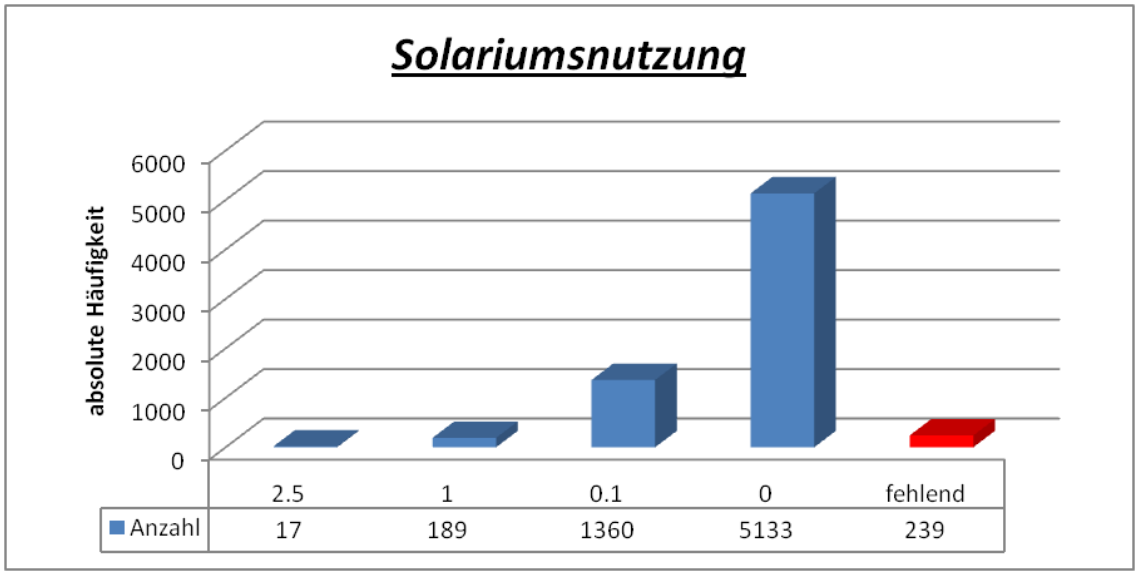

Haut-Reaktion
absolute Häufigkeit
3500
3000
2500
2000
1500
1000
500
0
immer
häufig
selten
sehr selten
nicht sichtbar (braun)
nicht sichtbar (schwarz)
fehlend
Anzahl
395
1930
3412
749
264
8
180

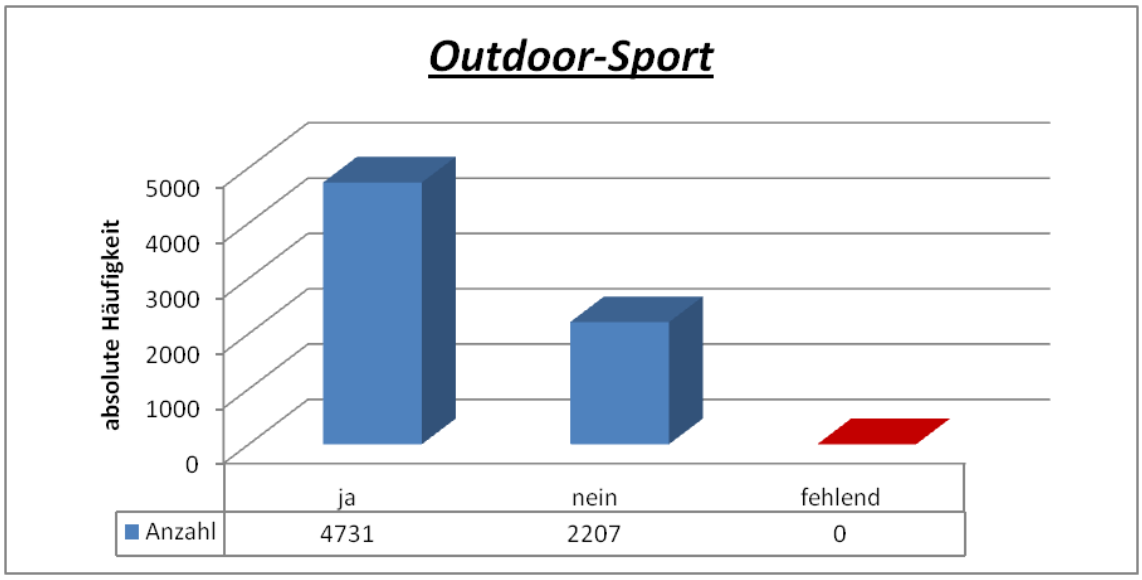

Sonnenbrand als Kind
absolute Häufigkeit
2500
2000
1500
1000
500
0
nie
selten
häufig
oft
fehlend
Anzahl
2261
2261
320
66
2030

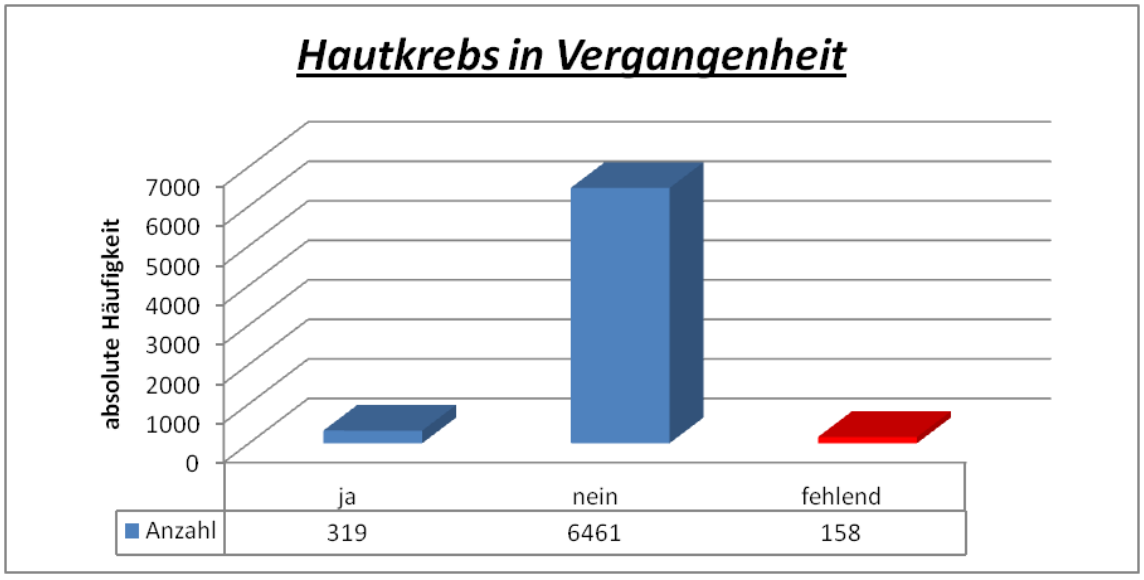

Sonnenbrand als Jugendlicher
absolute Häufigkeit
3000
2500
2000
1500
1000
500
0
nie
selten
häufig
oft
n.a.
fehlend
Anzahl
1532
2870
707
73
155
1601

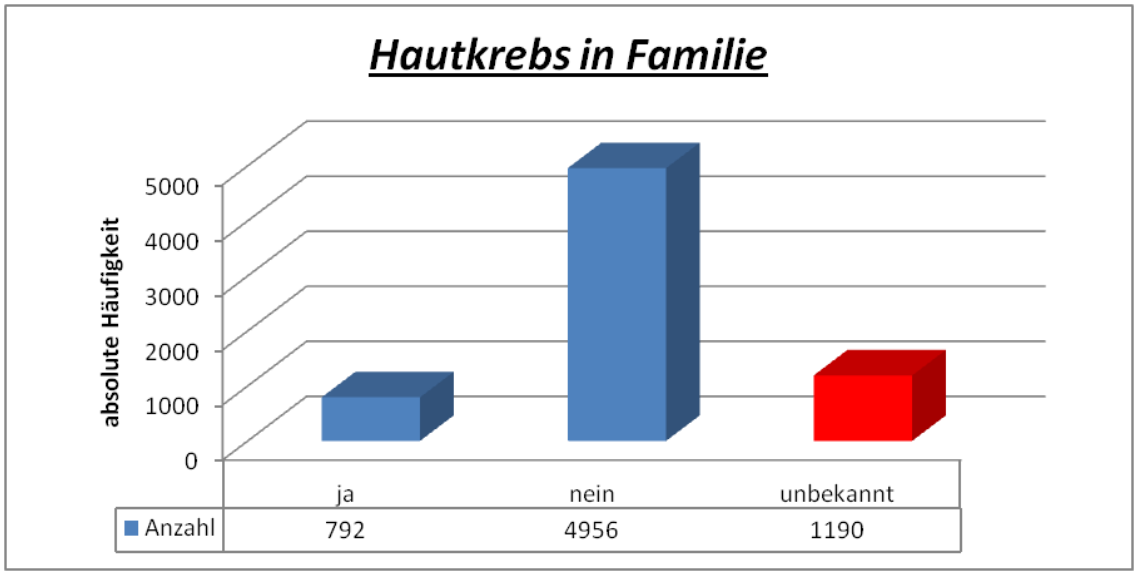

Sonnenbrand als Erwachsener
absolute Häufigkeit
4000
3500
3000
2500
2000
1500
1000
500
0
nie
selten
häufig
oft
n.a.
fehlend
Anzahl
1793
3685
340
52
573
495

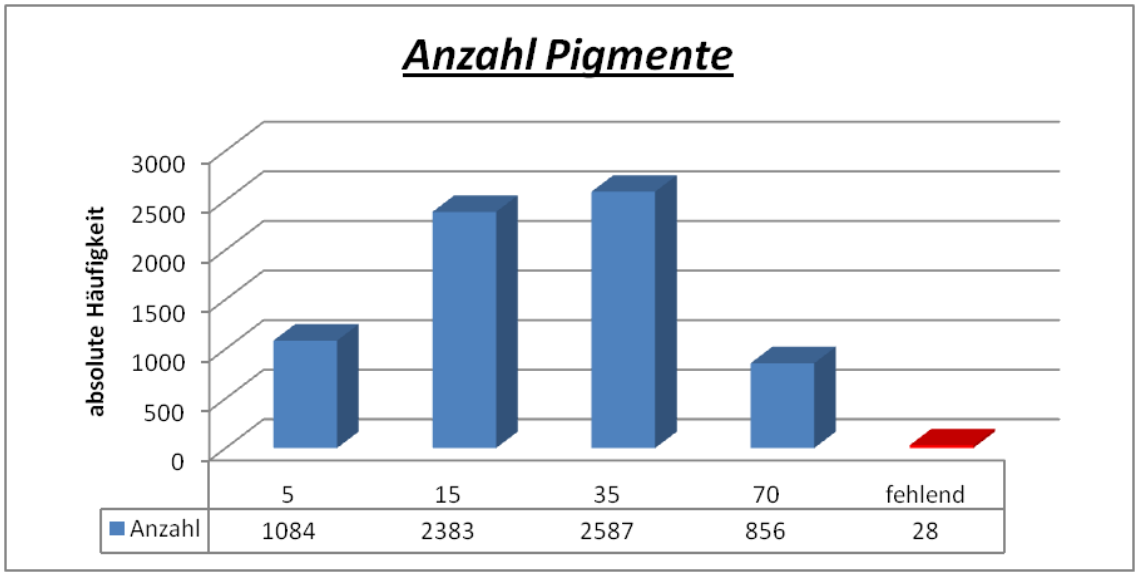

Schutz vor Sonne
absolute Häufigkeit
4000
3500
3000
2500
2000
1500
1000
500
0
konsequent
selten
keine Aussage
fehlend
Anzahl
3788
2328
822
0

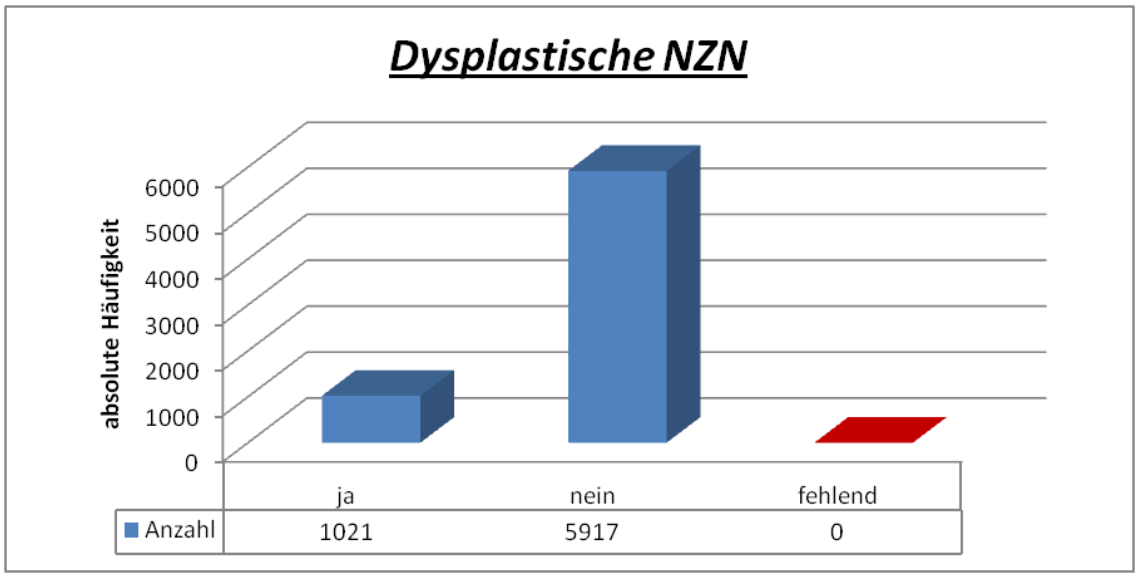

Solariumsnutzung
absolute Häufigkeit
6000
5000
4000
3000
2000
1000
0
2.5
1
0.1
0
fehlend
Anzahl
17
189
1360
5133
239

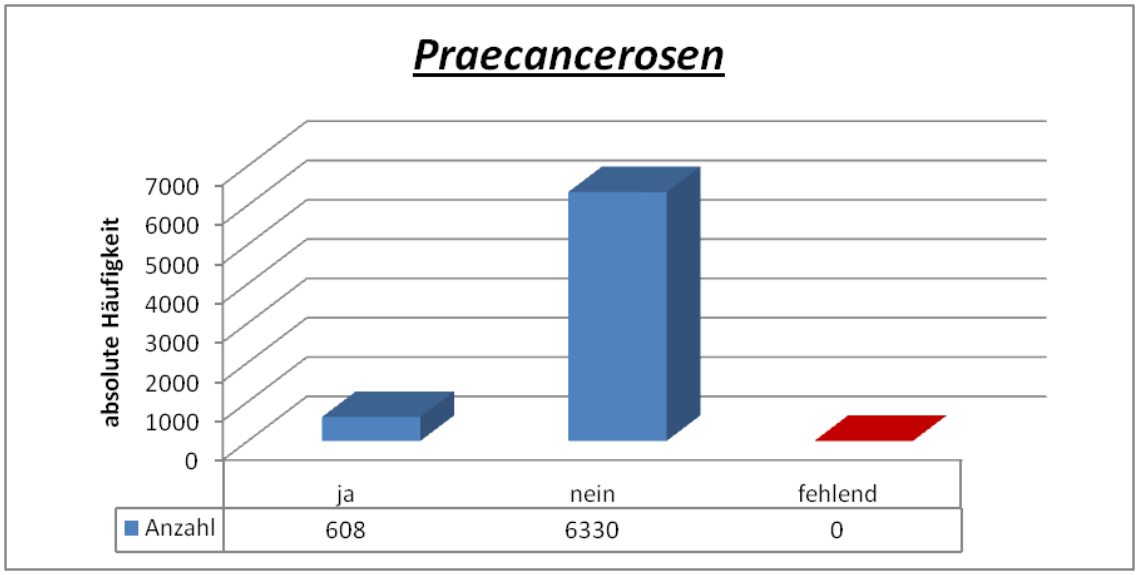

Outdoor-Sport
absolute Häufigkeit
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
4731
2207
0

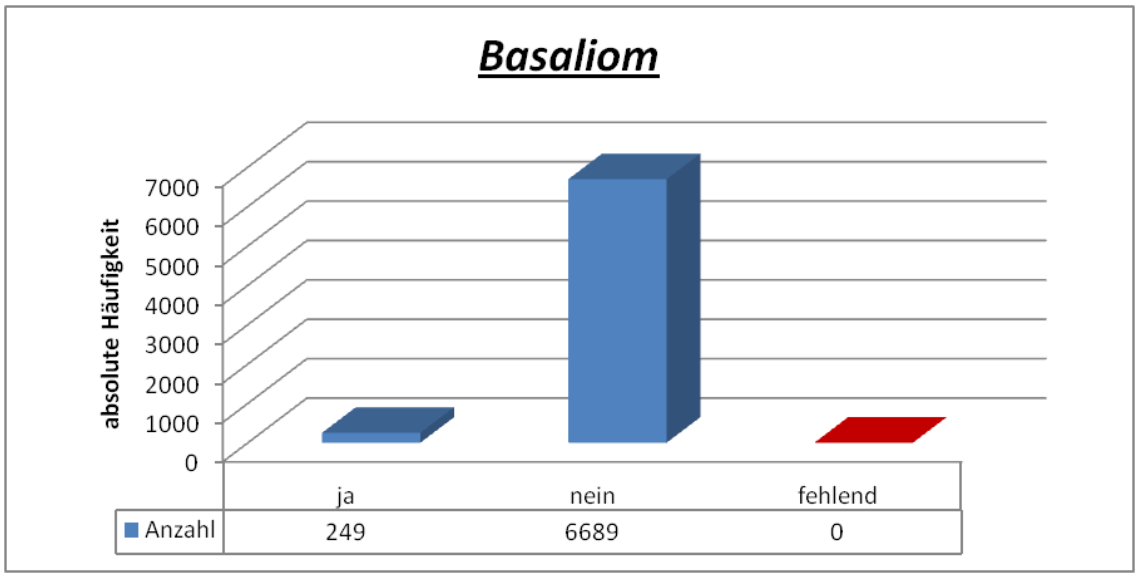

Hautkrebs in Vergangenheit
absolute Häufigkeit
7000
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
319
6461
158

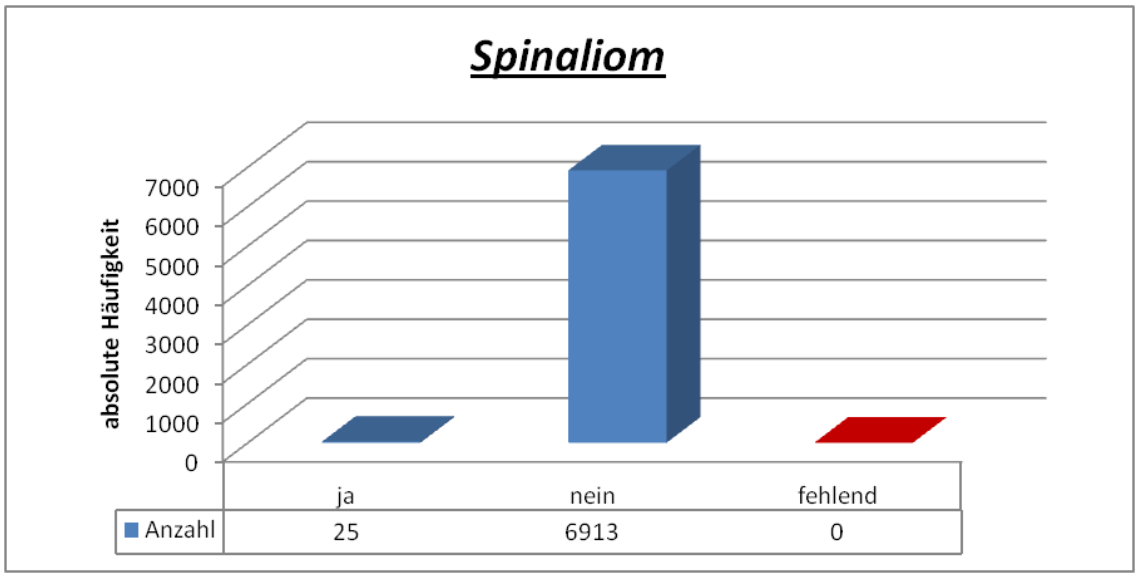

Hautkrebs in Familie
absolute Häufigkeit
5000
4000
3000
2000
1000
0
ja
nein
unbekannt
Anzahl
792
4956
1190

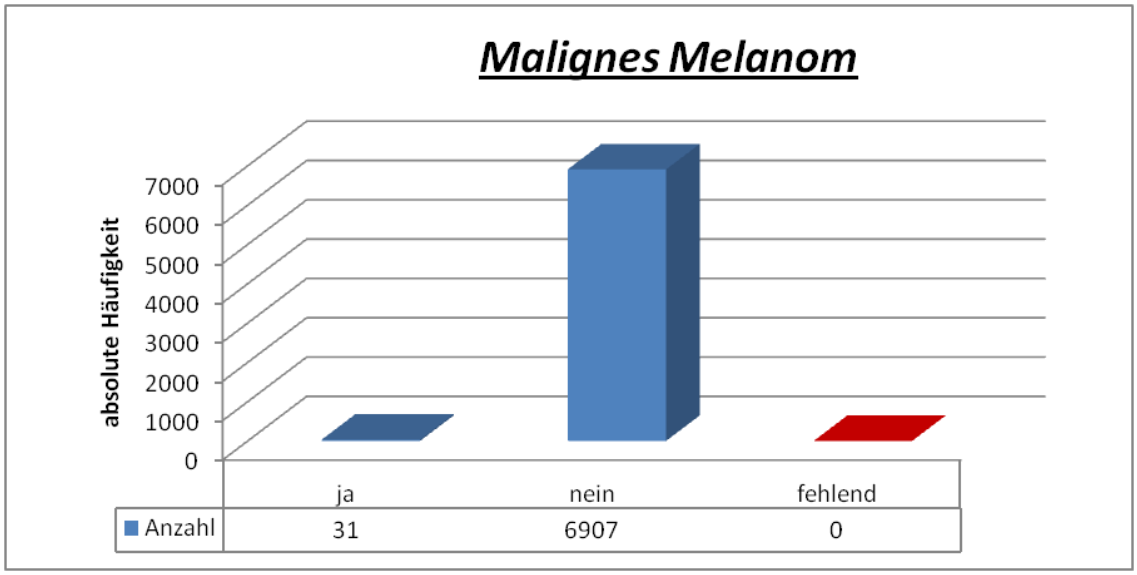

Anzahl Pigmente
absolute Häufigkeit
3000
2500
2000
1500
1000
500
0
5
15
35
70
fehlend
Anzahl
1084
2383
2587
856
28

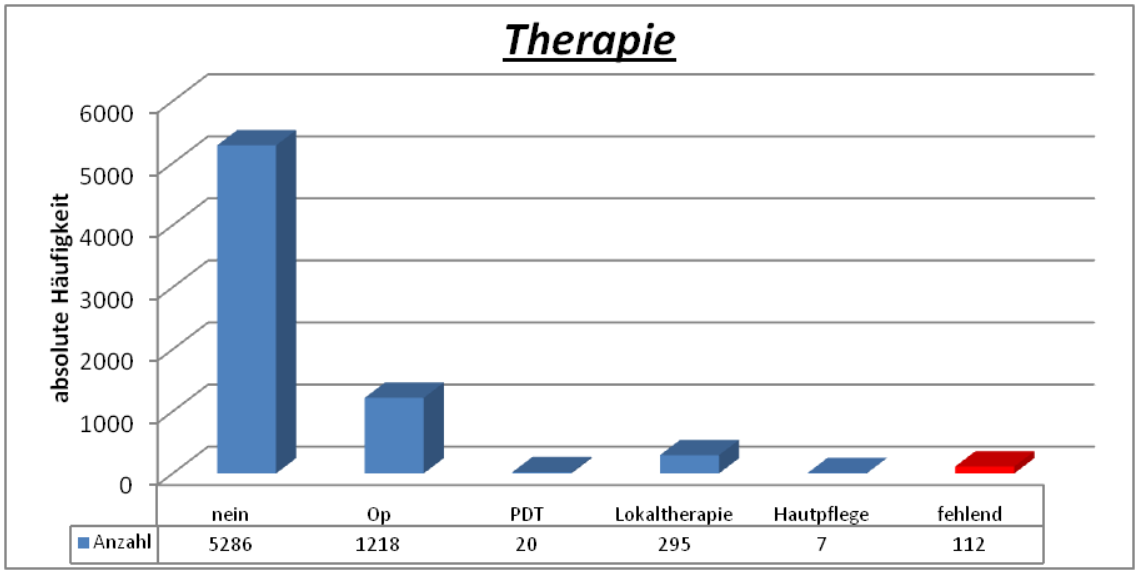

Dysplastische NZN
absolute Häufigkeit
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
1021
5917
0

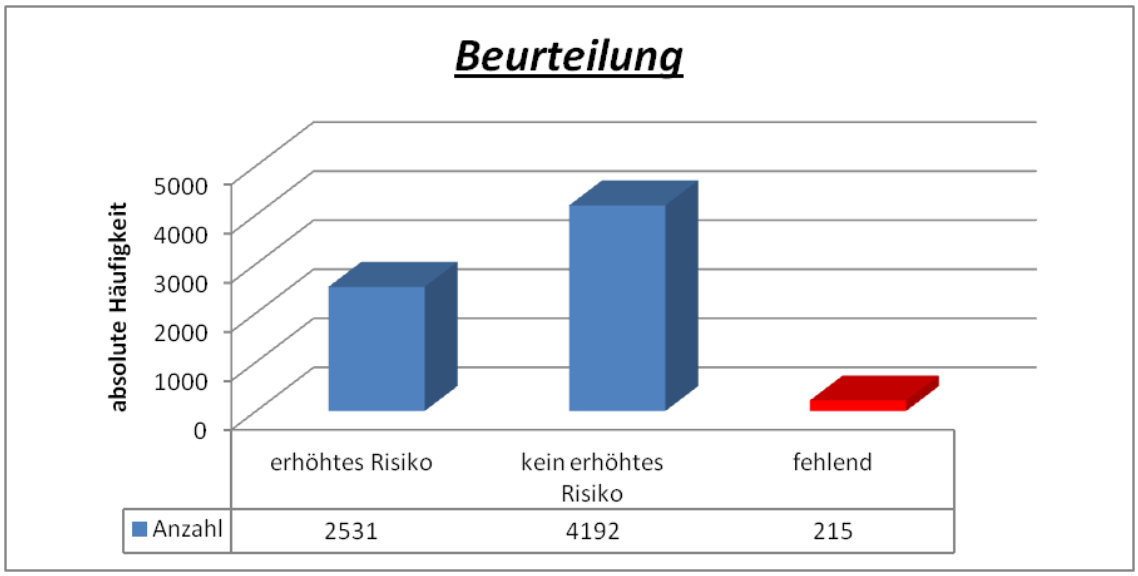

Praecancerosen
absolute Häufigkeit
7000
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
608
6330
0

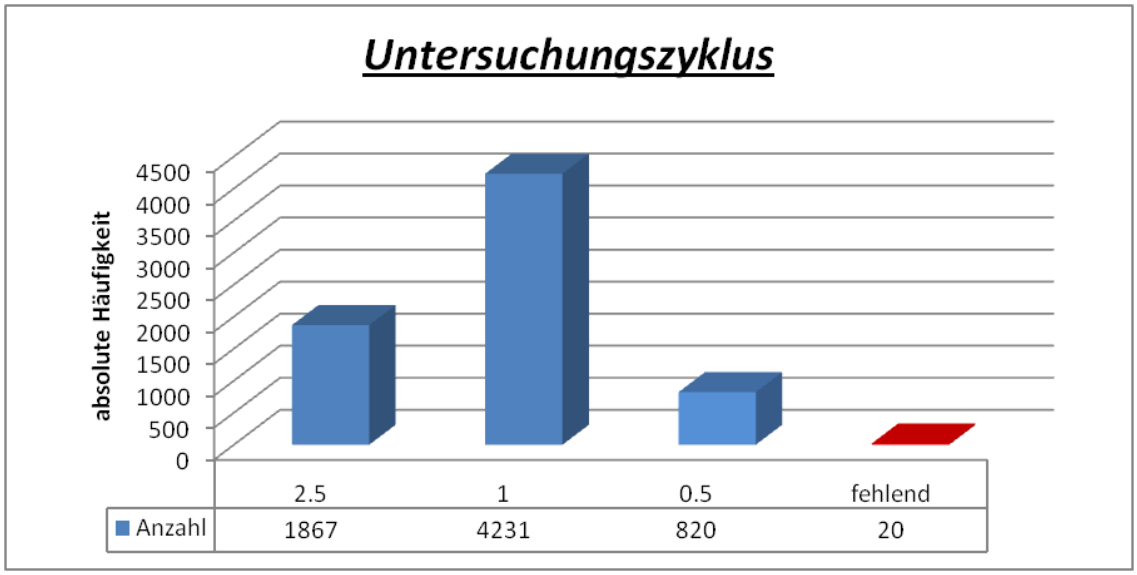

Basaliom
absolute Häufigkeit
7000
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
249
6689
0

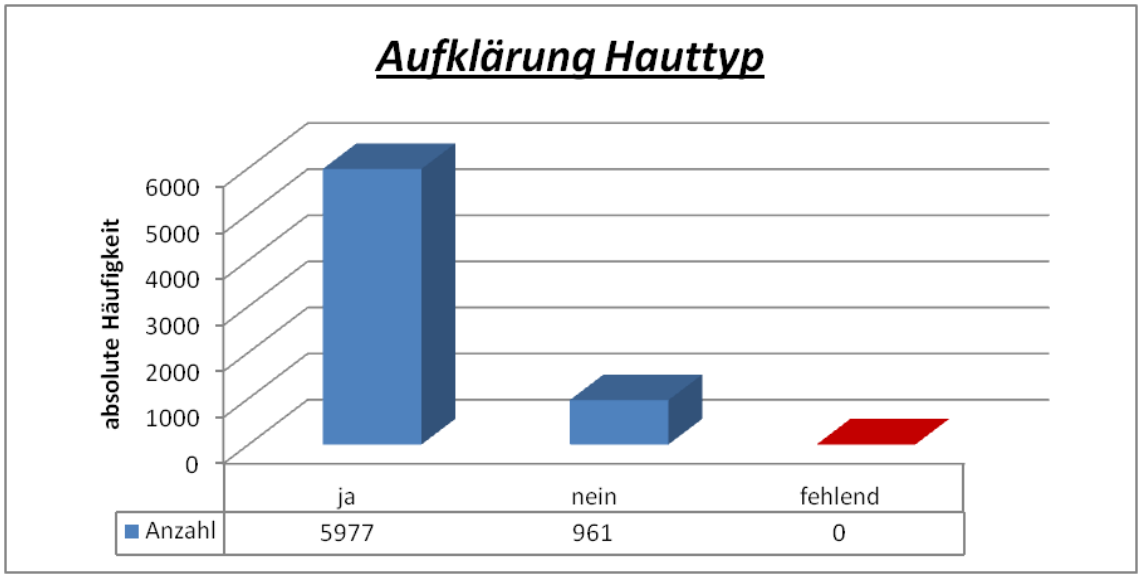

Spinaliom
absolute Häufigkeit
7000
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
25
6913
0

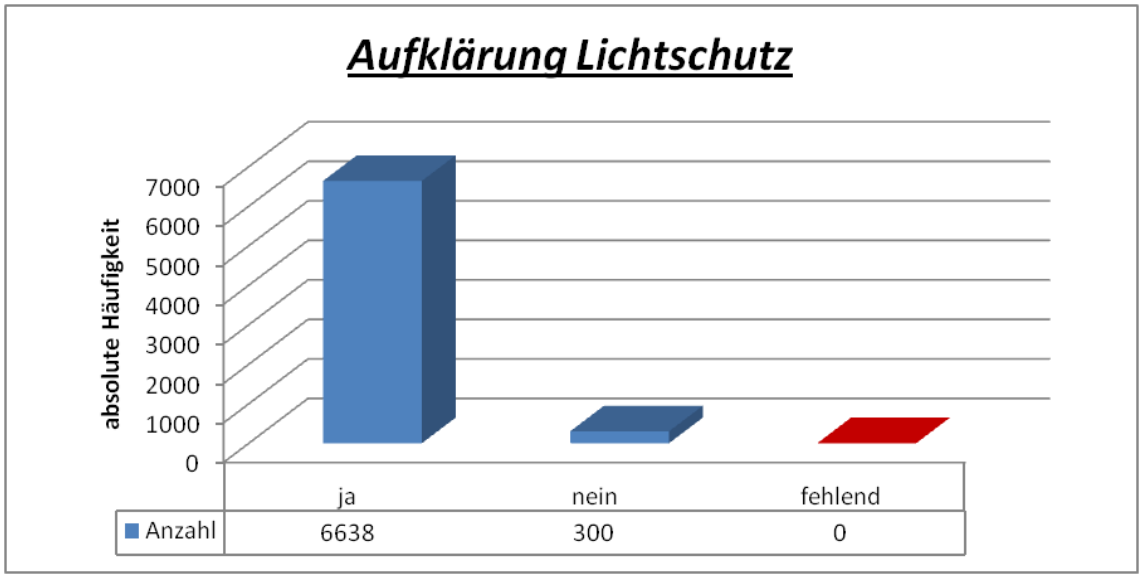

Malignes Melanom
absolute Häufigkeit
7000
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
31
6907
0

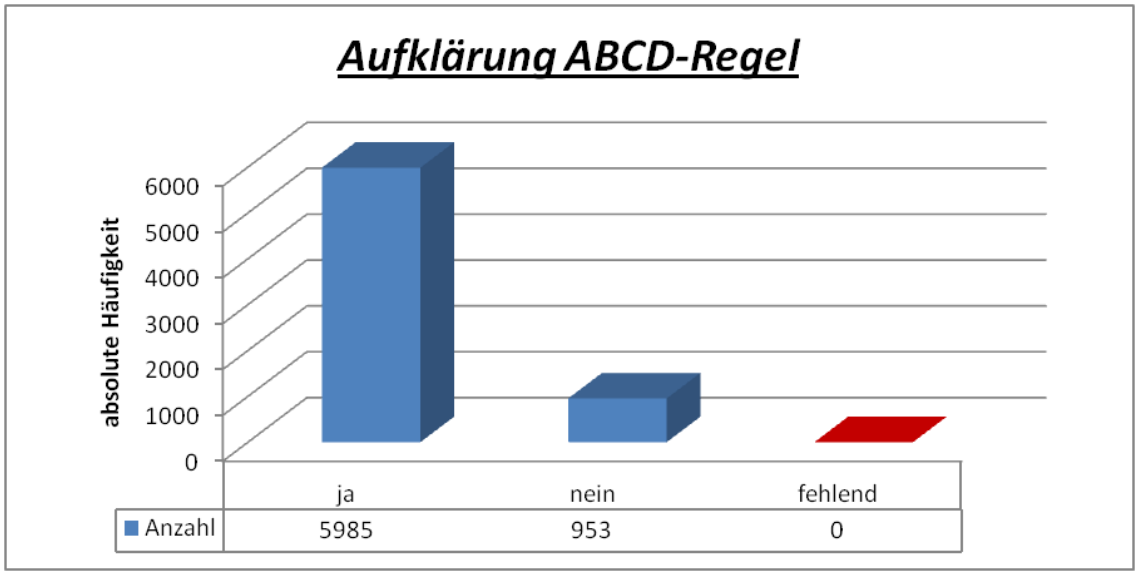

Therapie
absolute Häufigkeit
6000
5000
4000
3000
2000
1000
0
nein
Op
PDT
Lokaltherapie
Hautpflege
fehlend
Anzahl
5286
1218
20
295
7
112


Beurteilung
absolute Häufigkeit
5000
4000
3000
2000
1000
0
erhöhtes Risiko
kein erhöhtes Risiko
fehlend
Anzahl
2531
4192
215


Untersuchungszyklus
absolute Häufigkeit
4500
4000
3500
3000
2500
2000
1500
1000
500
0
2.5
1
0.5
fehlend
Anzahl
1867
4231
820
20


Aufklärung Hauttyp
absolute Häufigkeit
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
5977
961
0


Aufklärung Lichtschutz
absolute Häufigkeit
7000
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
6638
300
0


Aufklärung ABCD-Regel
absolute Häufigkeit
6000
5000
4000
3000
2000
1000
0
ja
nein
fehlend
Anzahl
5985
953
0

Literaturverzeichnis

Altmeyer, P., & Bacharach-Buhles, M. (2002). *Springer Enzyklopädie Dermatologie, Allergologie, Umweltmedizin.* Heidelberg: Springer.

Bayes, T. (1763). *An Essay towards solving a Problem in the Doctrine.*

Berger, U. (2009). *Diagnose Hautkrebs - die Krankheit, Ihre Ursachen und Behandlungsmethoden.* Noderstedt: Books on Demand GmbH.

Breiman, L. (August 1996). Bagging Predictors. *Machine Learning , 24* (2), S. 123-140.

Breitbart, E., Wende, A., Mohr, P., Greinert, R., & Volkmer, B. (2004). Hautkrebs. (R. Koch-Institut, Hrsg.) *Gesundheitsberichterstattung des Bundes* (22).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). *CRISP-DM 1.0*. Abgerufen am 23. November 2010 von Cross Industry Standard Process for Data Mining: http://www.crisp-dm.org/CRISPWP-0800.pdf

Clarke, B., Fokoué, E., & Zhang, H. H. (2009). *Principles and theory for data mining and machine learning.* Springer Verlag.

Cohen, W. W. (1995). Fast Effective Rule Induction. *In Proceedings of the Twelfth International Conference on Machine Learning* , S. 115-123.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory , 13* (1), S. 21-27.

CRISP-DM-Konsortium. (August 2006). *CRoss Industrial Standard Process for Data Mining*. Abgerufen am 24. November 2010 von http://www.crisp-dm.org/Images/Crisp-dmchartnew.gif

CRISP-DM-Konsortium. (18. Januar 2007). *CRoss Industrial Standard Process for Data Mining*. Abgerufen am 24. November 2010 von http://www.crisp-dm.org/index.htm

Debray, T. (2009). *Classification in Imbalanced Datasets.* Maastricht University, Faculty of Humanities and Sciences, Maastricht.

Deutsche Krebshilfe e.V. (1. Juli 2008). *Kurzinfos zur Früherkennung von Hautkrebs*. Abgerufen am 15. Januar 2011 von http://www.krebshilfe.de/fileadmin/Inhalte/Downloads/PDFs/Kurzinfos_zur_Frueherkennung_von_Hautkrebs.pdf

Duden. (2001). *Das Fremdwörterbuch.* Mannheim: Dudenverlag.

Fahrmeier, L., Künstler, R., Pigeot, I., & Tutz, G. (2007). *Statistik Der Weg zur Datenanalyse* (6. Ausg.). Berlin: Springer-Verlag.

Falkenauer, E. (September 1998). On Method Overfitting. *Journal of Heuristics , Vol. 4* (Nr. 3), S. 281-287.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Widener, T. (November 1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM , Vol. 39* (Nr. 11), S. 27-34.

Fürnkranz, J., & Widmer, G. (1994). Incremental Reduced Error Pruning. *In Proceedings the Eleventh International Conference on Machine Learning* , S. 70-77.

Garbe, C. (2006). *Management des Melanoms.* Heidelberg: Springer.

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques* (2. Ausg.). San Francisco: Morgan Kaufmann Publishers.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2. Ausg.). New York: Springer Verlag.

Hossiep, R., & Wottawa, H. (1993). Die Angewandete Psychologie in Schlüsselbegriffen. In R. Hossiep, H. Wottawa, & A. Schorr (Hrsg.), *Handwörterbuch*

der Angewandten Psychologie (S. 131-136). Bonn: Deutsche Psychologien Verlags GmbH.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In *Machine Learning: Proceedings of the Eleventh International Conference* (S. 121-129). San Francisco: Morgan Kaufmann.

KDnuggets. (August 2007). *Data Mining Methodology*. Abgerufen am 23. November 2010 von http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

Kietz, D.-U. (2009). *Data Mining zur Wissensgewinnung aus Datenbank Teil 2: Der KDD-Prozess*. Abgerufen am 23. November 2010 von Dr. Jörg-Uwe Kietz: http://www.kietz.ch/DataMining/Vorlesung/folien/02-Prozess.pdf

Liu, H., & Yu, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. (I. C. Society, Hrsg.) *IEEE Transactions on Knowledge and Data Engineering , 17* (4), S. 491-500.

Massalme, S. (2004). *Crashkurs Pathologie* (1. Ausg.). München: Urban & Fischer.

Paynter, G., Trigg, L., & Frank, E. (2008. November 2008). *Attribute-Relation File Format (ARFF)*. Abgerufen am 2010. Dezember 10 von The University of Waikato: http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Platt, J. (1999). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, C. Burges, & A. Smola, *Advances in Kernel Methods - Support Vector Learning* (S. 185-208). Cambridge, Massachusetts: MIT-Press.

Pyle, D. (1999). *Data Preparation for Data Mining* (1. Ausg.). San Francisco: Morgan Kaufmann.

Quinlan, J. (1990). Learning Logical Definitions from Relations. *Machine Learning , 5* (3), S. 239-266.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning , 1* (1), S. 81-106.

Reinhold, U., & Breitbart, E. (2007). *Hautkrebsprävention: Früherkennung und Vorbeugung*. Hannover: Schlütersche Verlag.

Reuter, P. (2004). *Springer Lexikon Medizin* (1. Ausg.). Florida, USA: Springer.

Rieck, K., Laskov, P., & Müller, K.-R. (2006). Efficient Algorithms for Similarity Measures over Sequential Data: A Look Beyond Kernels. In K. Franke, & e. al, *In Proceedings the Pattern Recognition, Proc. of 28th DAGM Symposium* (S. 374-383). Heidelberg: Springer.

Rosenblatt, F. (1958). The Perceptron, a Probabilistic Model for Information Storage and Organisation in the Brain. *Psychological Review , 65* (6), S. 386-408.

Saar-Tsechansky, M., & Provost, F. (2007). Handling Missing Values when Applying Classification Models. *Journal of Machine Learning Research* (8), S. 1625-1657.

Sammut, C., & Webb, G. I. (2010). *Encyclopedia of Machine Learning.* New York: Springer.

Shannon, C., & Weaver, W. (1963). *The Mathematical Theory of Communication.* University of Illinois Press.

Steinwart, I., & Christmann, A. (2008). *Support Vector Machines* (3. unveränderte Ausg.). New York: Springer Verlag.

Stolz, W., Braun-Falco, O., Bilek, P., Burgdorf, W. H., & Landthaler, M. (2001). *Farbatlas der Dermatoskopie* (3. unveränderte Ausg.). Berlin: Georg Thieme Verlag.

Traupe, H., & Hamm, H. (2006). *Pädiatrische Dermatologie* (2. Ausg.). Heidelberg: Springer.

Üstün, B. (2003). *Radboud University Nijmegen - Laboratory for Analytical Chemistry*. Abgerufen am 20. 1 2011 von http://www.cac.science.ru.nl/people/ustun/SVM.JPG

Wapnik, W., & Chervonenkis, A. (1979). *Theorie der Zeichenerkennung.* Berlin: Akademie-Verlag.

Witten, I. H., & Frank, E. (2005). *Data Mining Practical Machine Learning Tools and Techniques* (2. Ausg.). Waikato: Morgan Kaufmann Publishers.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems , 14* (1), S. 1-37.