FH Westküste
Wirtschaft und Technik

Hausaufgabe

Alireza Abbasi, Darius Bonk, Kiran Raj Krishnakumar, Shittheswaran Selvakumar Kalaivani, Marc Täuber

22023309, 22213311, 22213304, 22123301

22123306,

Bericht der Hausaufgabe im Studienfach „Maschinelles Lernen“

bei

Dr.-Ing. John-Harry Wieken, Fachhochschule Westküste

vorgelegt am 12.06.2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Alle Abbildungen sind in eigener Darstellung entstanden

Tabellenverzeichnis

Alle Tabellen sind in eigener Darstellung entstanden

1. Aufgabe 1

a) Verwenden Sie im RapidMiner ein Multilayer-Perzeptron (NeuralNet) mit zwei Layern und drei Neuronen im ersten und 2 Neuronen im zweiten Layer für die Vorhersage des Kreditrisikos auf Basis von Bonität, Verschuldungsgrad und Kredithöhe. Verwenden Sie ein Stratified Sampling mit 70% Trainingsdaten.
Verwenden Sie zur Vergleichbarkeit wieder einen Random Seed von 42.
Geben Sie eine Confusion Matrix und einen Scatterplot für die Bonität und den Verschuldungsgrad mit dem Kreditrisiko(predict) als Markierung der Punkte (Farbe) an.
Beurteilen Sie die Performanz des Netzes insbesondere unter dem Aspekt der Accuracy und der Precision aus Sicht der Bank.
Exportieren Sie den Prozess als rmp-Datei Aufgabe 1a.
Machen Sie zwei Vorschläge zur Verbesserung des Modells bei der Accuracy und insbesondere bei der Sicherheit der Vorhersage des tatsächlichen Kreditrisikos eines Kunden für den die Klassen „Kreditrisiko durchschnittlich“ vorhergesagt wird.

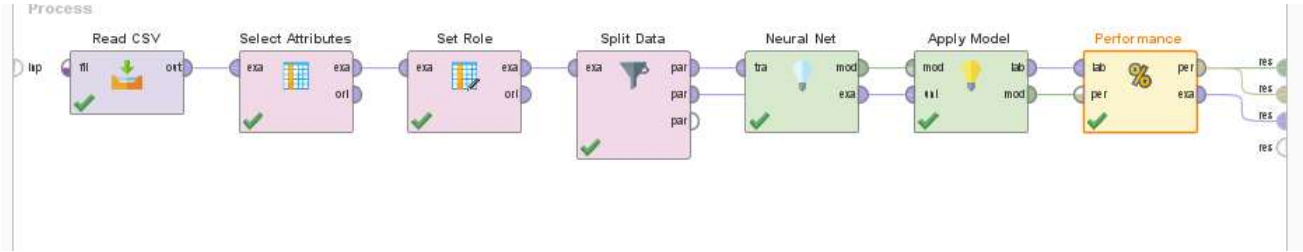

Abbildung 1: Prozessverlauf des Multilayer-Perzeptron-Modells (NeuralNet)

Vorgegeben war, dass mit Hilfe eines neuronalen Netzwerks ein Modell erzeugt wird, um eine Vorhersage zu treffen. Abbildung 1 zeigt, wie der Prozessverlauf in Rapidminer aufgebaut wurde, um am Ende die Confusion Matrix und den Scatter-plot zu erhalten. Dabei sind die Einstellung des neuronalen Netzwerkes so eingestellt (Anhang „Aufgabe 1a Ohne Optimierung“):

- Training Cycle auf 200
- Learning rate auf 0.1
- Momentum auf 0.9
- Shuffle wurde aktiviert
- Normalize wurde aktiviert
- Local Random Seed auf 42

Mit Hilfe der Confusion Matrix, zu sehen in Tabelle 1 und dem Scatterplot, dargestellt in Abbildung 2, können wir das Model bewerten.

Tabelle 1: Confusion Matrix des Modelles (Ohne Optimierung)

accuracy: 88.61%

	true durchschnittl...	true hoch	true gering	true nicht kreditw...	true sehr gering	class precision
pred. durchschnit...	53	1	0	0	0	98.15%
pred. hoch	0	60	0	3	0	95.24%
pred. gering	0	0	66	0	19	77.65%
pred. nicht kredit...	0	0	0	0	0	0.00%
pred. sehr gering	0	0	0	0	0	0.00%
class recall	100.00%	98.36%	100.00%	0.00%	0.00%	

- Zu sehen ist, dass die Genauigkeit bei 88.61% liegt, was für ein gutes Ergebnis spricht.
- Die Vorhersage der Stufe „nicht Kreditwürdig“ falsch und als Stufe „hoch“ eingeordnet wird
    - Daraus resultiert eine große Gefahr für den Anwender des Modelles
- Die Vorhersage der Stufe „Sehr gering“ falsch und als Stufe „gering“ eingeordnet wird
    - dieses könnte akzeptiert werden, da die beide Gruppen allgemein gesagt, nicht kreditwürdig sind.
- Das Modell hat sowohl die Stufe „durchschnittlich“ als auch „gering“ mit einem Recall von 100% und die Stufe „hoch“ mit einem Recall von 98.36% vorhergesagt wird. Dabei ist zu berücksichtigen, dass deren Class Precision bei 98.15%, 77.65% und 95.24% liegt
- Es ist auch zu beachten, dass das Model die Vorhersage nur in drei Classen (durchschnittlich, hoch, gering) einsortiert und die Classen „nicht Kreditwürdig“ in „hoch“ und „sehr gering“ in „gering“ eingepackt hatte
    - Der Grund des Problems ist die Anzahl der Neuronen, wobei im ersten Layer 3 Neuronen verwendet wurden

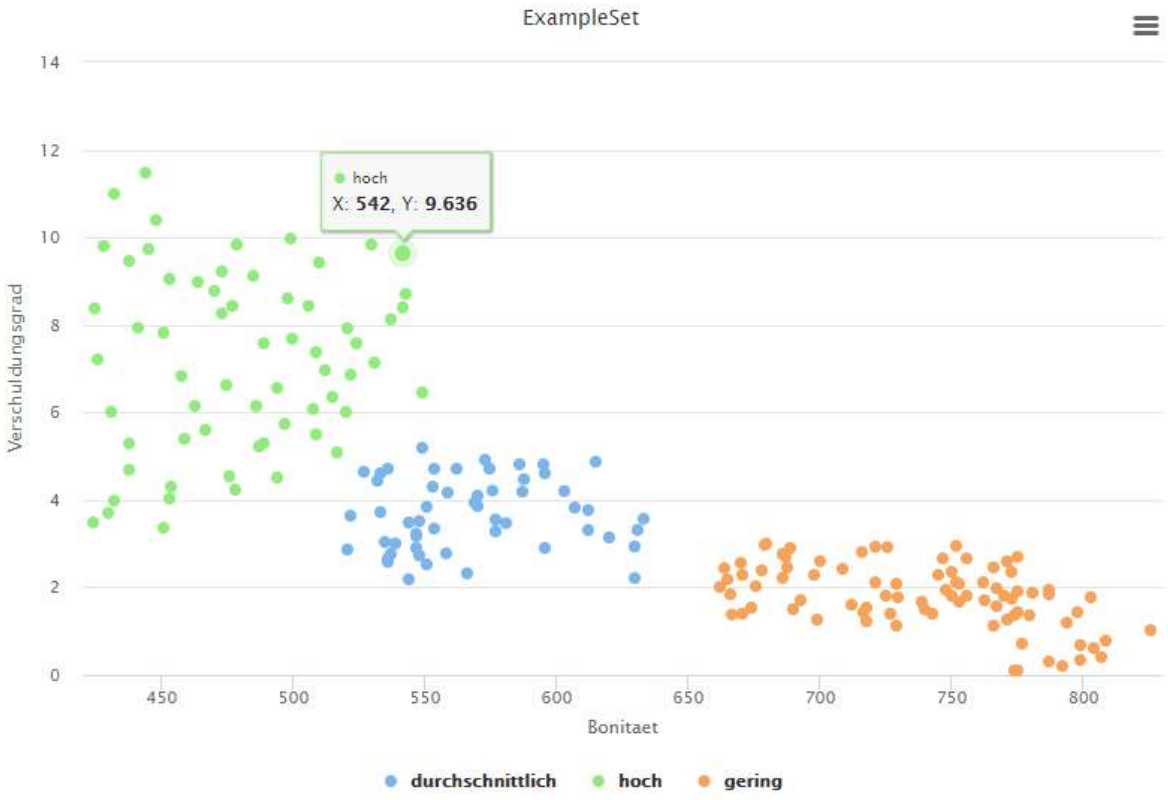

Abbildung 2: Scatterplot basieret auf Bonität und Verschuldungsgrad (Ohne Optimierung)

Um das Modell zu verbessern bzw. zu optimieren wurden folgende Maßnahmen ergriffen (Anhang „Aufgabe 1a mit Optimierung“):

- Die Neuronen im ersten und zweiten Layer wurden auf 5 erhöht
- Training Cycle wurde auf 6500 erhöht
- Ein weiteres Attribut wurde hinzugefügt: „Liquidität“
- Am Ende wurde versucht mit der Änderung der Learning rate und Momentum eine maximale Genauigkeit zu erhalten. (Learning rate auf 0.01 und Momentum auf 0.85 ergaben das beste Ergebnis)

Aus dem Scatter Plot des optimierten Modelles in Abbildung 3 und der entsprechenden Confusion Matrix in Tabelle 2 ist zu entnehmen, dass durch diese Maßnahmen eine bessere Genauigkeit von 97.52% und eine bessere Vorhersage der Stufe „sehr gering“ erreicht werden konnte. Das Problem der Einschätzung „nicht kreditwürdig“ bleibt bestehen und liegt vermutlich daran, dass der Anteil dieses Labels im Datensatz sehr klein ist und das Modell nicht in der Lage ist gut zu lernen. Hier ist eine größere Datengrundlage notwendig.

Tabelle 2: Confusion Matrix des Modelles (mit Optimierung)

accuracy: 97.52%

	true durchschnittl...	true hoch	true gering	true nicht kreditw...	true sehr gering	class precision
pred. durchschnit...	53	0	0	0	0	100.00%
pred. hoch	0	61	0	3	0	95.31%
pred. gering	0	0	65	0	1	98.48%
pred. nicht kredit...	0	0	0	0	0	0.00%
pred. sehr gering	0	0	1	0	18	94.74%
class recall	100.00%	100.00%	98.48%	0.00%	94.74%	

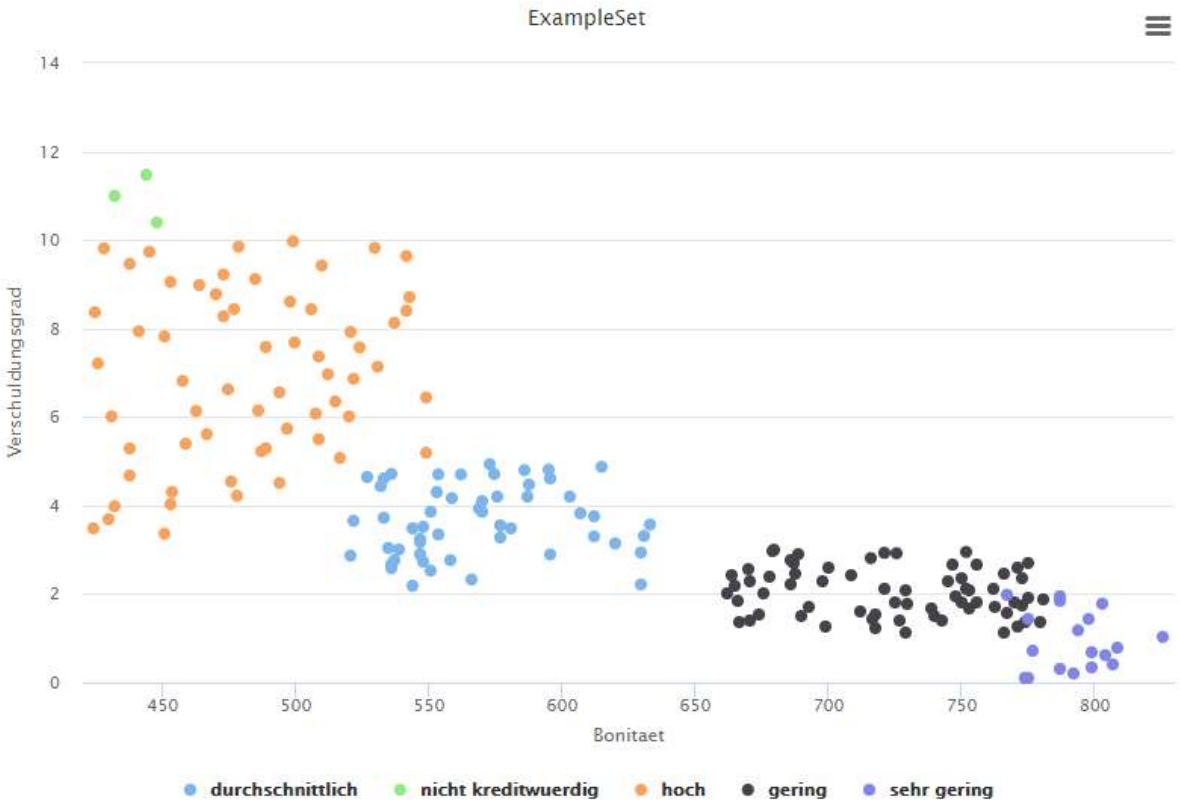

Abbildung 3: Scatterplot basieret auf Bonität und Verschuldungsgrad (mit Optimierung)

b) Erstellen Sie mittels des RapidMiners jetzt ein Perzeptron-Modell zur Vorhersage des Kreditrisikos mit einem Klassifizierer. Fassen Sie das Kreditrisiko zusammen (gering, sehr gering und nicht kreditwürdig wird zu Nein, die anderen Werte werden zu Ja) Verwenden Sie dieselben Features wie in 1a) und nutzen Sie für die übrigen Angaben zum Sampling und Random Seed ebenfalls die Werte aus Aufgabe 1a). Lassen Sie das Modell mit einer Lernrate von 0,05 und 100 Iterationen lernen. Geben Sie in Ihrer Lösung den Bias und die Gewichte für die Verwendung der drei Features an.
Bestimmen Sie die Klassifikation des Kreditrisikos für eine Bonität von 600 und einen Verschuldungsgrad von 3 bei Anfrage für einen Kredit über 300000 Euro. Exportieren Sie den Prozess als rmp-datei Aufgabe 1b.

Wie in der Aufgabe vorgegeben wurde, sollte ein Perzeptron-Modell erzeugt werden, um die Gewichte und den Bias von drei Features (Bonität, Verschuldungsgrad, Kredithöhe) zu bestimmen. Um das zu ermöglichen, wurde der Prozessverlauf erstellt, welcher in Abbildung 4 zu sehen ist. (zusätzlich: Anhang Aufgabe_1b).

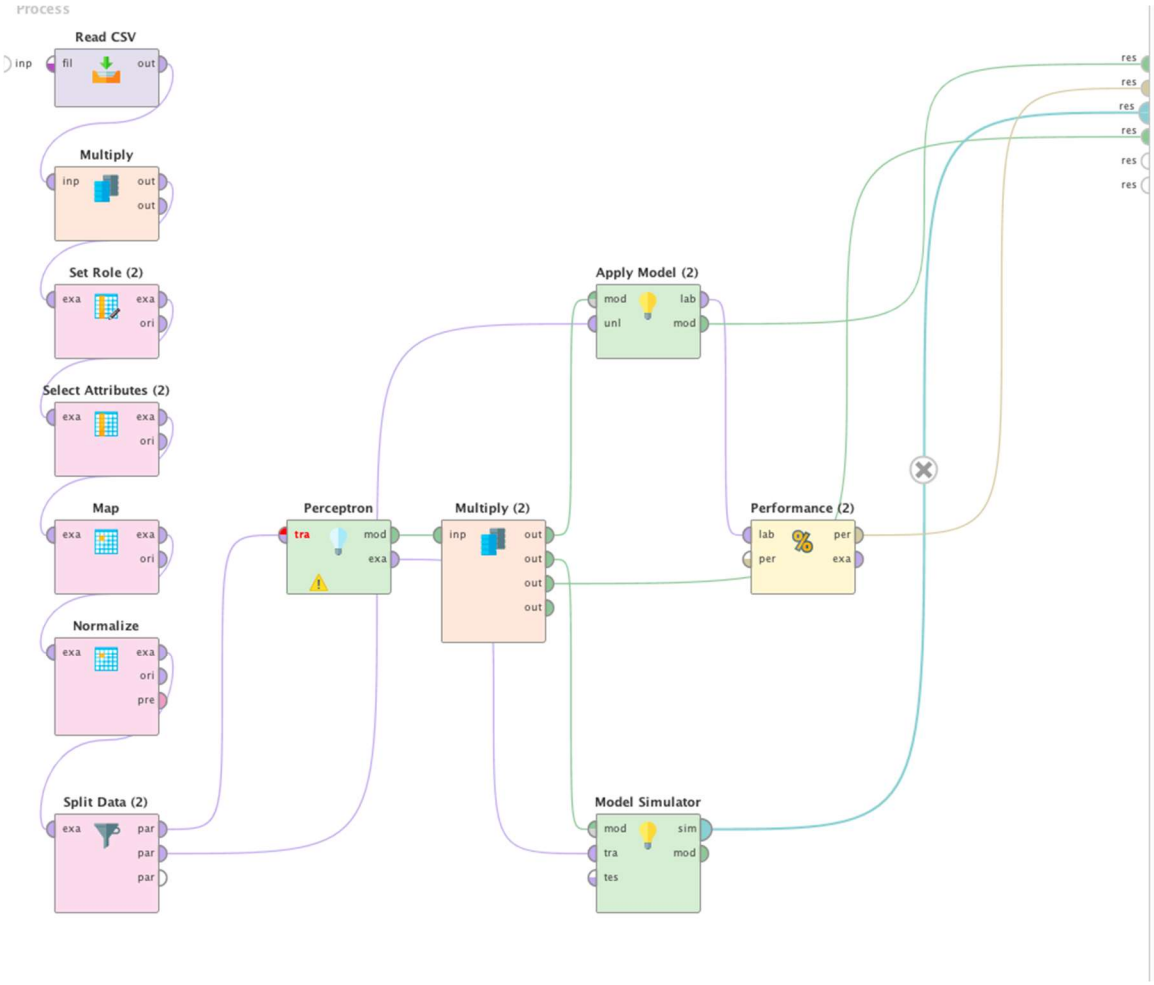

Abbildung 4: Prozessverlauf des Perzeptron-Modells

Mit Hilfe der Hyperplane in Abbildung 5 wurden die Werte des Bias und der gewichte abgelesen:

- Bias: 0.05324519491152731
- Bonitätsgewicht: -0.365
- Verschuldungsgradsgewicht: 0.159
- Kredithöhesgewicht: 0.013

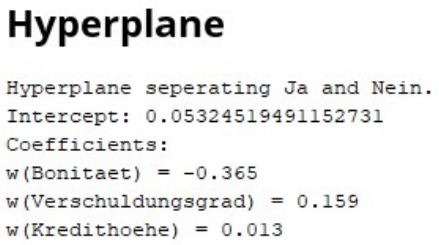

Abbildung 5: Hyperplane

Tabelle 3: Confusion Matrix des Perzeptron-Modelles mit zwei Klassen

accuracy: 98.03%

	true Nein	true Ja	class precision
pred. Nein	137	2	98.56%
pred. Ja	2	62	96.88%
class recall	98.56%	96.88%	

Um herauszufinden zu welcher klasse die eingegebenen Werte am besten passten könnten, wird der Model Simulator verwendet. Zuerst sollten die gewünschte Werte normalisiert werden und dafür wird eine „Range Transformation“ angewendet:

$$Feature\ Wert_{norm} = \frac{\text{Feature Wert} - \text{FeutureStichprobe}_{min}}{\Delta\ Feature\ Stichprobe}$$

Mit einem Break vor dem Operator Normalize wird der maximale und minimale Wert von jedem Feature ausgelesen und aus diesen die Normalisierten Werte berechnet:

$$Bonität = \frac{600-397}{826-397} = 0.473193$$

$$Verschuldungsgrad = \frac{3-0.050}{12.523-0.050} = 0.236510$$

$$Kredithöhe = \frac{300000-39236}{449485-39236} = 0.236510$$

In Abbildung 6 ist zu sehen, dass die eingegeben Werte der Features mit 50.37% Sicherheit der Klasse „Nein“ zugeordnet wird.

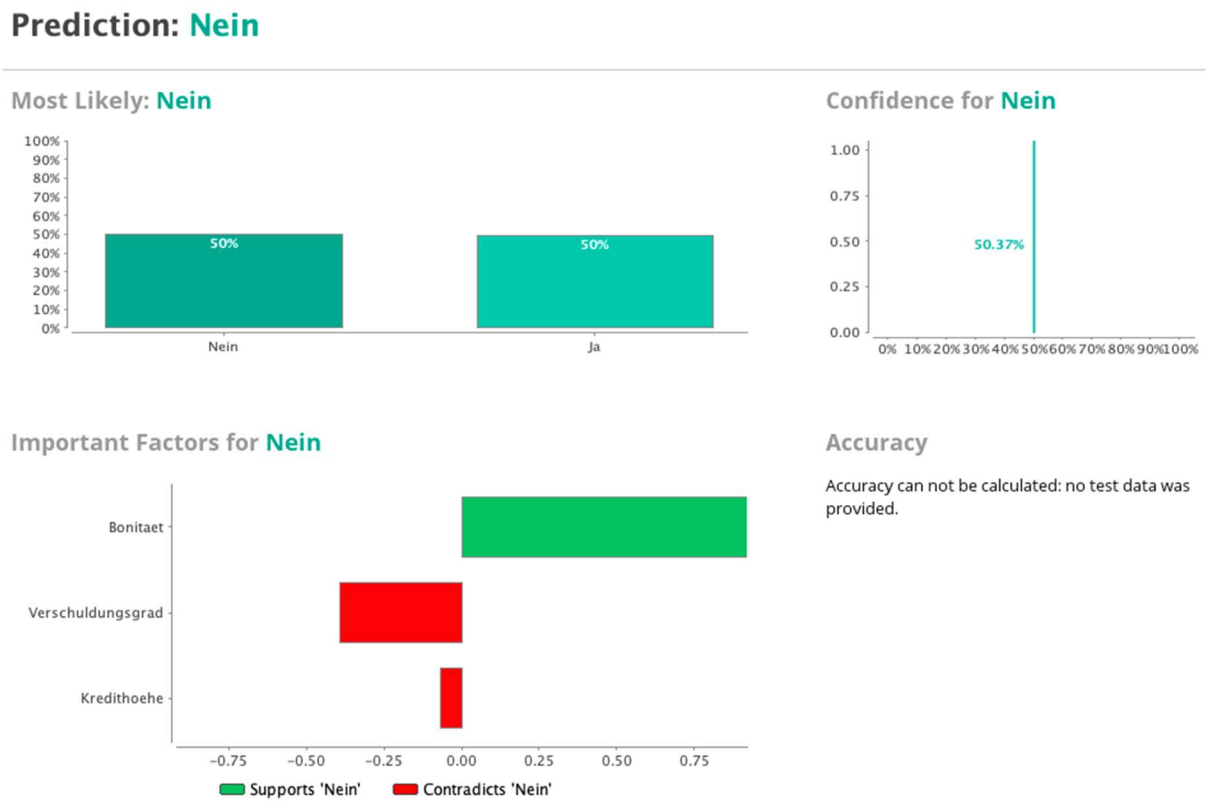

Abbildung 6: Simulationsergebnisse des Modells auf Grundlage der vorgegeben Werte

2. Aufgabe 2

a) Analysieren Sie in einem neuen Prozess die Qualität der Features sowie den Zusammenhang zwischen allen Features untereinander und dem Label unabhängig von einem Modell.
Wählen Sie begründet 5 Features aus, die Sie für die weitere Analyse der Kreditwürdigkeit verwenden wollen.
Speichern Sie den Prozess als Aufgabe 2a.rmp

Abbildung 7 zeigt den Prozessverlauf für die Auswahl von fünf Merkmalen, wobei “Quality Measures”, “Weight by Information gain” und "Correlation Matrix” von Rapidminer für die Analyse der Merkmale verwendet werden.

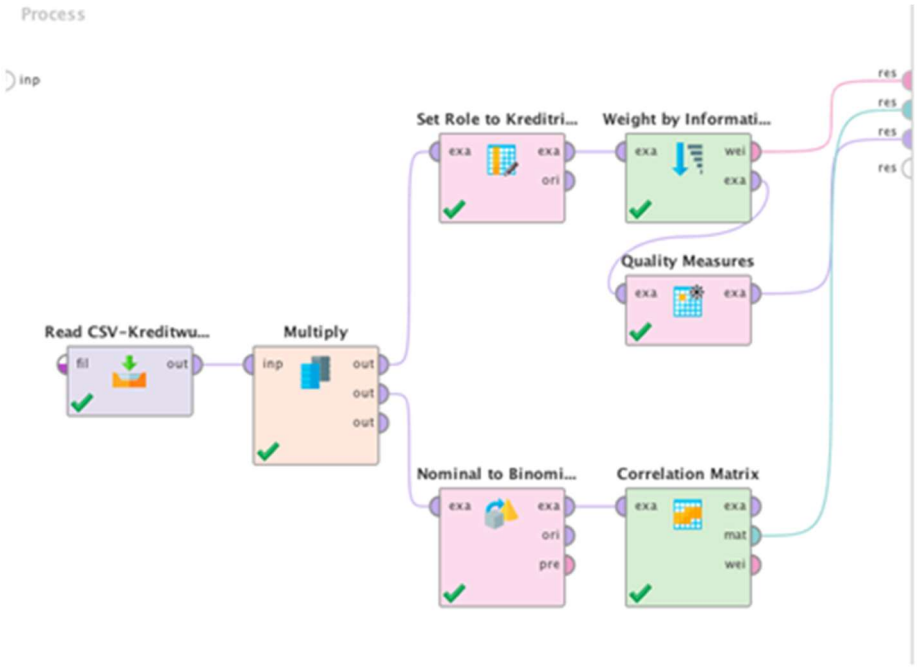

Abbildung 7: Prozessverlauf in Rapid Miner für Aufgabe 2a

Analyse der Ergebnisse:

Der oben beschriebene Prozess liefert uns eine Tabelle mit der Gewichtung der Informationsgewinne (siehe Tabelle 4), eine Tabelle mit den Qualitätsmaßen (siehe Tabelle 5) und eine Korrelationsmatrix (siehe Tabelle 6).

Aus der Tabelle 4 und 5 zu lesen ist, dass “ID” und “Kredithöhe” keinen großen Einfluss auf “Kreditrisiko” haben, da der ID-Wert von beiden 1 bzw. 0,997 ist und die Werte für die “Weight by information gain” 0,009 bzw. 1,28 betragen.

Tabelle 4: “Quality Measure“ Tabelle

Row No.	Attribute	Correlation ↓	ID-ness	Stabillity	Missing	Text-ness
2	Bonitaet	0.518	0.466	0.010	0	0
5	Verschuldungsgrad	0.506	0.004	0.009	0	0
3	Zahlungsverzuege	0.475	0.033	0.142	0	0
4	beschaeftigteMonate	0.299	0.123	0.038	0	0
7	Liquiditaet	0.286	0.979	0.003	0	0
8	AnzahlKreditlinien	0.230	0.018	0.179	0	0
6	Kredithoehe	0.057	0.997	0.003	0	0
1	ID	0.004	1	0.001	0	0

Aus der Tabelle 5 geht hervor, dass “Bonität”, “Verschuldungsgrad” und “Zahlungsverzuege” die wichtigsten Merkmale für die Vorhersage des “Kreditrisikos” sind. Aus der Tabelle 4 und 6 wird deutlich, dass diese drei Merkmale stark miteinander korrelieren, so dass es wichtig ist ein Merkmal zu entfernen.

Tabelle 5: “The weight by information gain” Tabelle

attribute	weight ↓
Bonitaet	0.983
Verschuldungsgrad	0.780
Zahlungsverzuege	0.778
Liquiditaet	0.608
AnzahlKreditlinien	0.376
beschaeftigteMonate	0.322
Kredithoehe	0.128
ID	0.009

Tabelle 6: Correlation Matrix

Attribu...	Kreditr...	Kreditr...	Kreditr...	Kreditr...	Kreditr...	ID	Bonitaet	Zahlungsverzuege	beschaeftigteMonate	Verschuldungsgrad	Kredithoehe	Liquiditaet	AnzahlKreditlinien
Kreditri...	1	-0.391	-0.413	-0.069	-0.194	-0.061	-0.204	0.007	-0.166	-0.043	-0.018	-0.284	0.194
Kreditri...	-0.391	1	-0.460	-0.077	-0.215	0.038	-0.720	0.689	-0.418	0.712	-0.199	-0.534	0.479
Kreditri...	-0.413	-0.460	1	-0.081	-0.227	-0.003	0.630	-0.553	0.247	-0.508	0.063	0.535	-0.473
Kreditri...	-0.069	-0.077	-0.081	1	-0.038	0.045	-0.187	0.297	-0.106	0.331	0.000	-0.109	0.222
Kreditri...	-0.194	-0.215	-0.227	-0.038	1	0.018	0.497	-0.322	0.547	-0.366	0.239	0.447	-0.370
ID	-0.061	0.038	-0.003	0.045	0.018	1	-0.017	0.064	0.026	0.031	-0.075	0.015	0.035
Bonitaet	-0.204	-0.720	0.630	-0.187	0.497	-0.017	1	-0.779	0.619	-0.792	0.232	0.763	-0.680
Zahlung...	0.007	0.689	-0.553	0.297	-0.322	0.064	-0.779	1	-0.496	0.751	-0.206	-0.622	0.578
bescha...	-0.166	-0.418	0.247	-0.106	0.547	0.026	0.619	-0.496	1	-0.516	0.208	0.531	-0.451
Verschu...	-0.043	0.712	-0.508	0.331	-0.366	0.031	-0.792	0.751	-0.516	1	-0.199	-0.610	0.589
Kredith...	-0.018	-0.199	0.063	0.000	0.239	-0.075	0.232	-0.206	0.208	-0.199	1	0.212	-0.146
Liquidit...	-0.284	-0.534	0.535	-0.109	0.447	0.015	0.763	-0.622	0.531	-0.610	0.212	1	-0.563
AnzahlK...	0.194	0.479	-0.473	0.222	-0.370	0.035	-0.680	0.578	-0.451	0.589	-0.146	-0.563	1

Da die Werte für den Informationsgewinn für die drei oben genannten Merkmale fast gleich sind, haben wir das Modell dreimal ausgeführt. Bei jedem Durchlauf wurde eines der oben genannten Merkmale weggelassen und die Ergebnisse in Tabelle 7 zusammengefasst.

Tabelle 7: Genauigkeitstabelle für verschiedene Szenarien

	Features	MODEL DATA ACCURACY	SAMPLE DATA ACCURACY
SCENARIO 1	Verschuldungsgrad, Kredithöhe, ID	96.62%	97.54%
SCENARIO 2	Bonität, Kredithöhe, ID	95.775	95.07%
SCENARIO 3	Zahlungsverzeuge, Kredithöhe, ID	**97.25%**	**99.015**

Aus der Tabelle 7 geht hervor, dass Szenario 3 die höchste Genauigkeit aufweist, weshalb wir die folgenden Merkmale auswählen:

- Bonität
- Verschuldungsgrad
- Beschäftigte Monate
- Liquidität
- Anzahlkreditlinien

b) Erstellen Sie eine geeignete Regressionsfunktion für das Label Kreditrisiko auf Basis der fünf in 2a) ausgewählten Features. Fassen Sie das Kreditrisiko zusammen (gering, sehr gering und nicht kreditwürdig als Nein, den Rest als Ja). Schätzen Sie das Kreditrisiko eines Kunden mit Bonität: 767, Zahlungsverzüge: 0, beschäftigte Monate: 40, Verschuldungsgrad: 2.70, Kredithöhe 270000, Liquidität: 18820, Kreditlinien:3 mittels des Modells ein. Speichern Sie den Prozess als Aufgabe 2b.rmp

Die Abbildung 8 zeigt das Prozessdiagramm für ein “Logistic Regression” Model, um herauszufinden, ob ein Kunde kreditwürdig ist oder nicht. Entsprechend der Fragestellung haben wir Hoch und nicht kreditwürdig in "ja" und die restlichen Werte des Kreditrisikos in "nein" mit Hilfe des Map-Operators umgewandelt. Da wir ein Klassifizierungsproblem haben, verwenden wir die logistische Regression für die Vorhersage des Wertes. Mithilfe des Modellsimulationsoperators können wir die Kreditwürdigkeit eines Kunden ermitteln. Wir haben die Leistung des entworfenen Modells sowohl für die Test- (30%), als auch die Modelldaten (70%), mithilfe des Performance-Operators, berechnet.

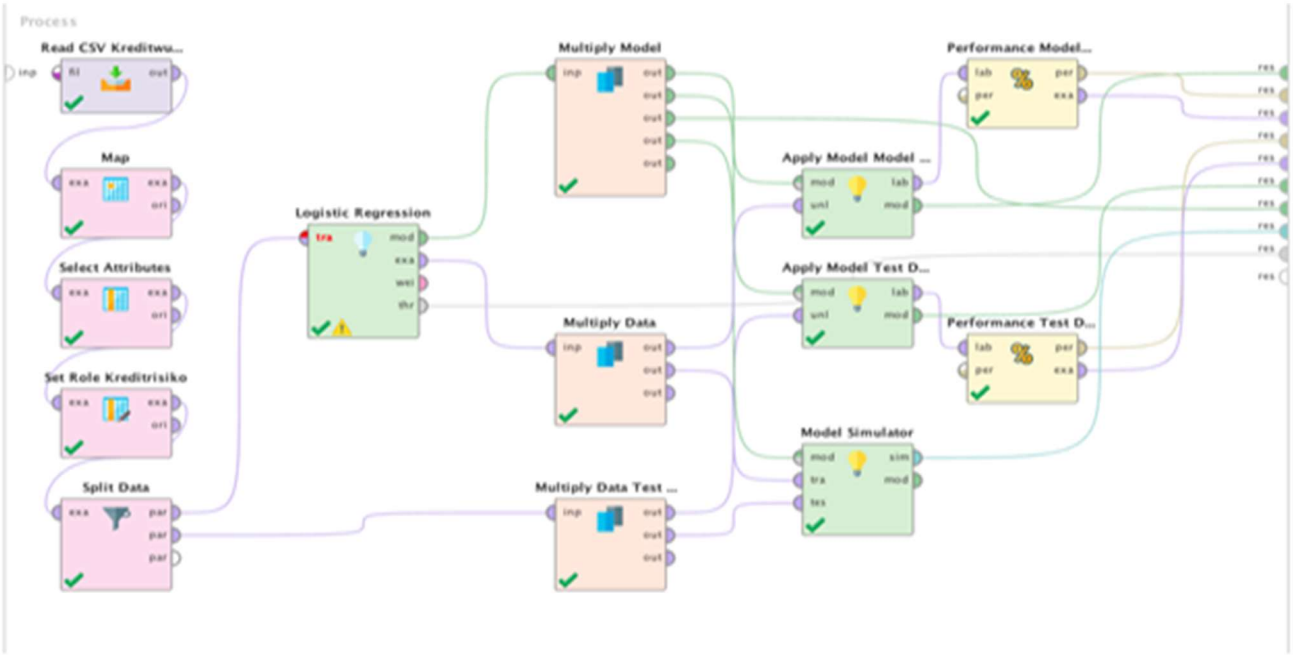

Abbildung 8: Entworfener Prozessverlauf in Rapid Miner für Aufgabe 2b

Die Tabelle 8 zeigt die Koeffizienten des logistischen Regressionsmodells für die verschiedenen Merkmale. Der Verschuldungsgrad-Koeffizient hat den höchsten Wert, so dass wir sagen können, dass er das einflussreichste Merkmal in unserem Modell ist. Liquidität weist den niedrigsten Wert und den geringsten Einfluss auf. Der „Threshold“ unseres Regressionsmodells ist 0,574. Unter diesem Wert wird das Modell "nein", oberhalb „Ja“ vorhersagen.

Tabelle 8: Logistic Regression Coefficient Tabelle

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Bonitaet	-0.083	-9.795	0.019	-4.410	0.000
beschaeftigteMonate	-0.061	-1.157	0.042	-1.470	0.141
Verschuldungsgrad	1.626	3.973	0.404	4.026	0.000
Liquiditaet	-0.001	-3.616	0.000	-2.889	0.004
AnzahlKreditlinien	0.271	0.682	0.176	1.539	0.124
Intercept	38.690	-11.463	10.044	3.852	0.000

Tabelle 9 und 10 zeigen die Ergebnisse unseres logistischen Regressionsmodells der Test- und Modelldaten. Die Genauigkeit mit Testdaten ist 99,01%, bei welchem das Modell 137 "nein" mit 100% Präzision und 66 "ja" mit 96,97% Präzision vorausgesagt hat. Mit der Grundlage der Testdaten hat das Modell 326 "nein" mit 97,55% Präzision und 147 "ja" mit 96,60% Präzision vorausgesagt.

Tabelle 99: Confusion Matrix des Modelles mit Testdaten (30%)

accuracy: 97.25%

	true nein	true ja	class precision
pred. nein	318	8	97.55%
pred. ja	5	142	96.60%
class recall	98.45%	94.67%	

Tabelle 10:Abbildung:Confusion Matrix des Modelles mit Trainingsdaten (70%)

accuracy: 99.01%

	true nein	true ja	class precision
pred. nein	137	0	100.00%
pred. ja	2	64	96.97%
class recall	98.56%	100.00%	

In Abbildung 9 sind die Simulationsergebnisse der folgenden Parameter zu finden:

- Bonität 767
- Beschäftige Monate 40
- Liquidität 18820
- Anzahl Kredit Linien 3
- Verschuldungsgrad 2.7

Abbildung 9: Simulationsergebnisse des Modells auf Grundlage der vorgegebenen Parameter

Wie in Abbildung 9 zu sehen ist, wurde mit einer Wahrscheinlichkeit von 100% „nein“ vorausgesagt. Die Genauigkeit liegt hier bei 99%.

C) Erläutern Sie kurz den Unterschied

a. zwischen einem Klassifikator und einem Regressor
b. einem überwachten und einem unüberwachten Lernprozess
c. einem Recall und der Precision

Klassifikator	**Regressor**
Nur diskrete Werte wie Wahr oder Falsch, Spam oder nicht Spam können hier gefunden werden.	Kontinuierliche Werte wie Preisbereich, numerischer Bereich=1,1.2 oder Menge (von 0 bis 10)
Die Bewertung erfolgt anhand von Accuracy	Die Bewertung erfolgt anhand des mittleren quadratischen Fehlers (Root mean square error)
Die Art der vorhergesagten Daten ist ungeordnet	Die Art der vorhergesagten Daten ist geordnet
Beispiele für Algorithmen sind Entscheidungsbaum, logistische Regression, usw.	Beispielalgorithmen sind Randomforest, lineare Regression, etc.
Die Ausgaben sind Klassen	Die Ausgaben sind reelle Zahlen
Versuch, eine Entscheidungsgrenze zu finden	es versucht, die beste Linie zu finden.
• Beispiel: Vorhersage des Sieges einer Mannschaft? Beispiel: Vorhersage, ob der Patient überleben wird oder nicht?	• Beispiel: Wettervorhersage (kontinuierlicher Wert) Beispiel: Vorhersage des Aktienkurses (kontinuierlicher Wert)

Überwachter Lernprozess	**Unüberwachter Lernprozess**
Sie dient dem Training des Algorithmus, indem sie den gelabelten Datensatz bereitstellt.	Der Algorithmus wird nicht trainiert und das System muss selbst lernen, indem es die strukturellen Merkmale der Eingabedaten durchgeht.
Ausgabemuster sind bekannt	Ausgabemuster sind unbekannt
Die Ausgabe ist genauer und zuverlässiger	Die Ausgabe ist mäßig genau und verlässlich
Es ist weniger komplex, da es das Muster der Ausgangs- und Eingangsdaten bereitstellt	Es ist komplexer, weil es kein Training wie beim überwachten Lernen gibt.
Beispiel: Klassifizierungs- und Regressionsprobleme werden durch überwachtes Lernen gelöst.	Beispiel: Clustering und Assoziationsregeln sind die Probleme, die durch unüberwachtes Lernen gelöst werden.
Es nutzt die Offline-Analyse	Es nutzt die Echtzeit-Analyse

Precison	**Recall**
Er hilft dabei, die Fähigkeit zu messen, positive Proben im Modell zu klassifizieren.	Sie hilft zu messen, wie viele positive Proben durch das ML-Modell richtig klassifiziert werden.
Die Berechnung erfolgt über die Anzahl der gefundenen relevanten Elemente/Gesamtzahl der Elemente	Die Berechnung erfolgt anhand der Anzahl der gefundenen relevanten Elemente/Anzahl der relevanten Elemente insgesamt
Hier wird die Vorhersage als Basislinie verwendet	Hier wird die Wahrheit als Grundlinie verwendet.

3. Aufgabe 3

Setzen Sie die Aufgabe 1a wahlweise in R oder in Python um (nur den technischen Teil, eine erneute Beurteilung oder Optimierung ist nicht notwendig) und geben Sie das Skript/Programm ebenfalls ab.

Der technische Teil der Aufgabe 1a wurde hier mithilfe von Python erneut umgesetzt.

Um dem Neuralen Netz die Möglichkeit zu geben eine Vorhersage für die Kreditwürdigkeit aufzustellen wurden die Strings der Kreditwürdigkeitsspalte durch ganze positive Zahlen ersetzt. Dabei wurden die Labels Durchschnittlich, Gering, Hoch, Nicht kreditwürdig und Sehr gering durch die Zahlen von 0-4 ersetzt. Des Weiteren wurden die Features, Bonität, Verschuldungsgrad, Kredithöhe normalisiert damit alle drei Feature eine gemeinsame Skala erhalten.

Über eine Schleife, die über alle fünf Labels läuft, werden mehrere Streudiagramme erzeugt. Diese stellen die Bonität gegen den Verschuldungsgrad, die Bonität gegen die Kredithöhe und den Verschuldungsgrad gegen Kredithöhe gegeneinander auf. Somit kann die Zuordnung der Label in Abhängigkeit der Features erfolgen.

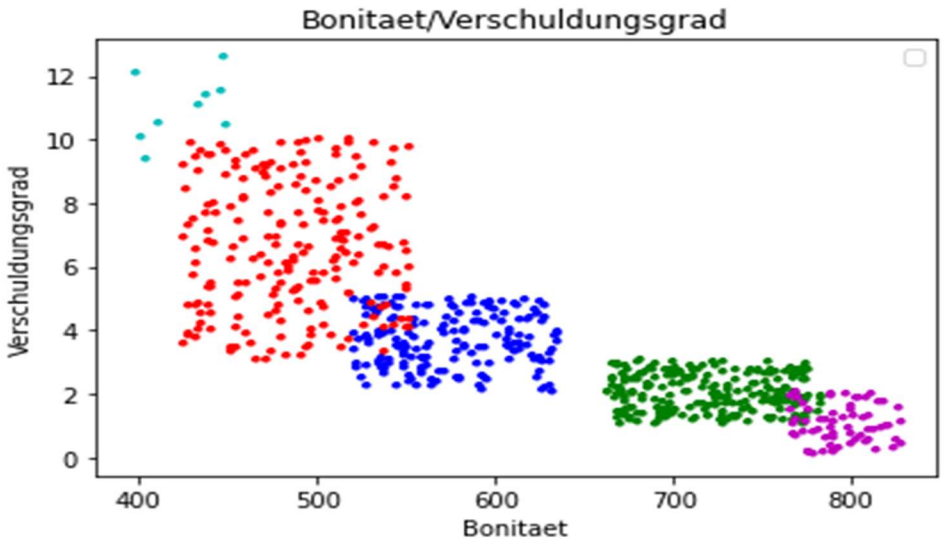

Abbildung 10: Streudiagramm des Kreditrisikos auf Grundlage des Verschuldungsgrads und der Bonität

In Abbildung 10 ist beispielhaft das Streudiagramm für den Verschuldungsgrad auf der Abszisse und die Bonität auf der Ordinate zusehen. Die lila Punkte stehen hierbei für ein sehr geringes Kreditrisiko, grün für geringes Kreditrisiko, dunkelblau für ein durchschnittliches Kreditrisiko, rot für ein hohes und die hellblauen Punkte als nicht kreditwürdig eingestuft.

Beim Modell-Training wurden die Daten außerdem mit “Stratified Sampling” versehen. Dabei wurden der Datensatz auf 70 % reduziert, 474 Datenpunkte, reduziert um Rechenzeit zu sparen. Durch das “Stratified Sampling” wird die Rate, in der die einzelnen Labels auftreten berechnet und bei der Reduzierung beibehalten.

Zuletzt wird anhand der vorhergesagten und wahren Ergebnisse eine Confusion-Matrix erstellt. Diese wird erstellt um eine schnelle Übersicht zu geben wie gut die Vorhersagen mit den wahren

Ergebnissen übereinstimmen. Dies wird in Abbildung 11 dargestellt.

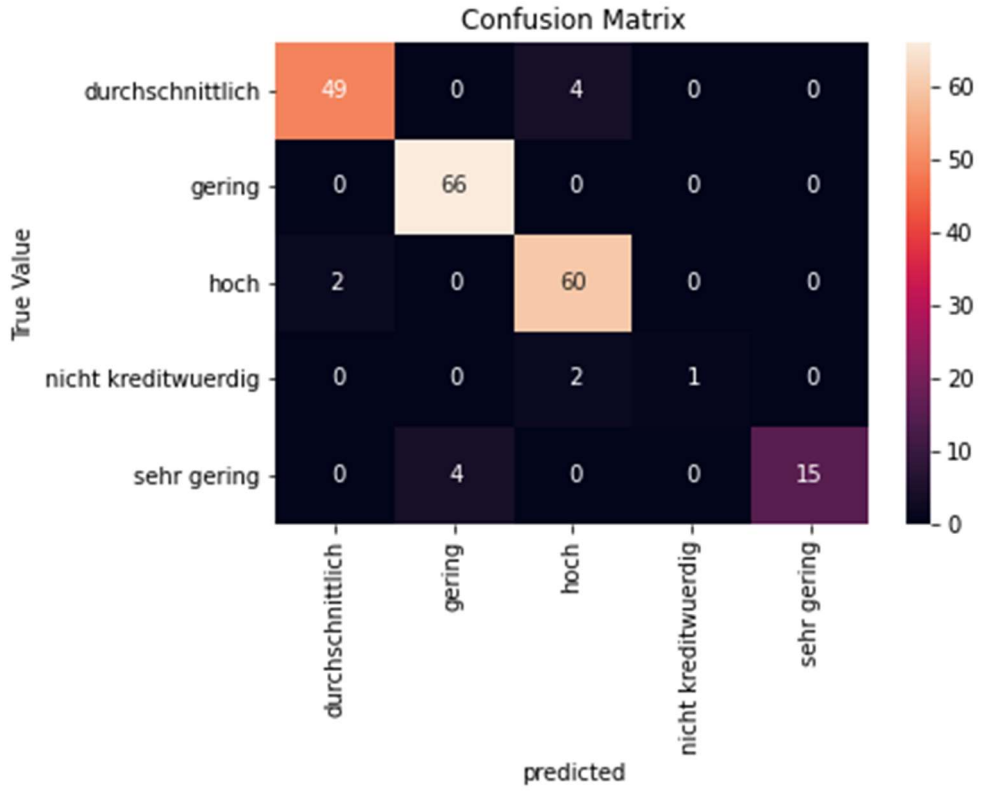

Abbildung 11: Confusion-Matrix in absoluten Zahlen

Durch diese Confusion-Matrix wird außerdem ersichtlich wie häufig die Vorhersagen zutreffen und die Häufigkeit, in der die Labels aufgetreten sind.

4. Aufgabe 5

a) Erstellen Sie mit einem Werkzeug Ihrer Wahl ein k-Means Model für den Datensatz Kreditwuerdigkeit (ohne das Label Kreditrisiko). Wählen Sie k=5.
b) Beschreiben Sie die inhaltliche Bedeutung der Kundenklassifizierung für die 5 Klassen.
c) Ordnen Sie den Klassen einen Wert des Labels sinnvoll zu und bestimmen Sie die Accuracy dieser Zuordnung gegenüber dem tatsächlichen Label.

Die Abbildung 12 zeigt das Prozessdiagramm für das Clustermodell, in dem wir mit Hilfe des Select Attribute Operators das Kreditrisiko gemäß der Frage und der ID weggelassen haben, um die Genauigkeit des Modells zu erhöhen, da alle Werte der ID's zu unterscheiden sind. Wir verwenden den Clustering Operator, Input k = 5 und „equclideanDistance“ als numerische Maße und mit Hilfe des Performance Operators haben wir die Performance der Cluster bewertet. (Anhang Aufgabe 5)

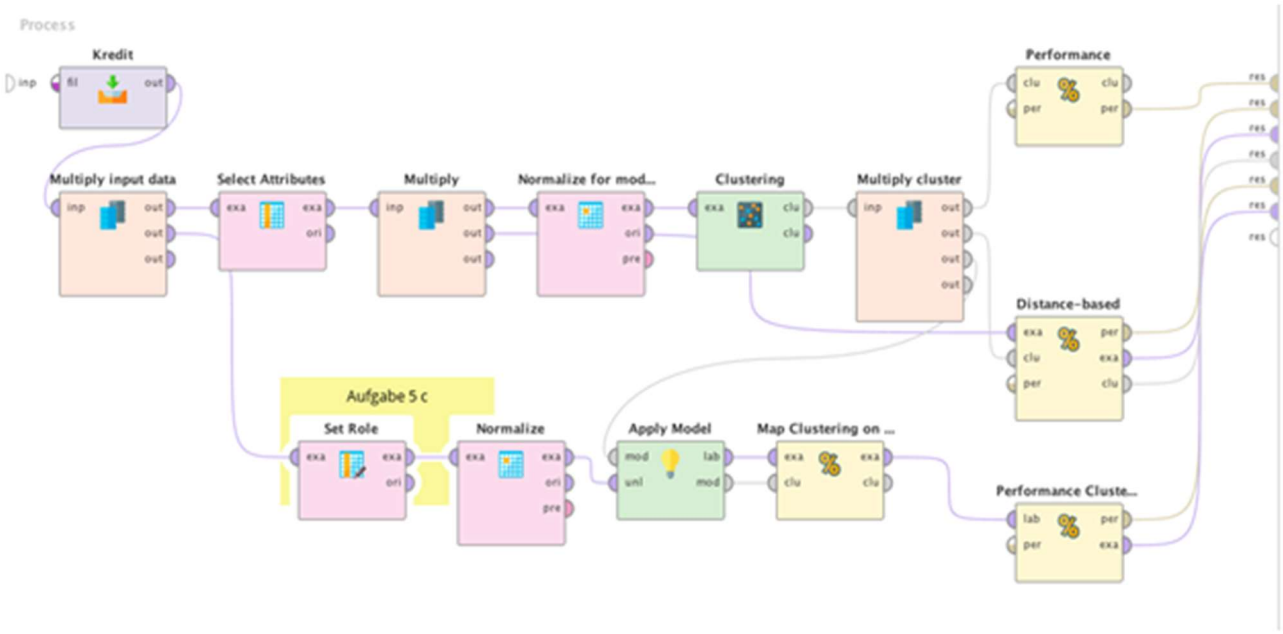

Abbildung 12: Entworfener Prozessverlauf in Rapid Miner für Aufgabe 5a und 5c

Die insgesamt 676 Items des Modells sind in 5 Cluster unterteilt:

- Cluster 1 hat 202 Items
- Cluster 1 hat 96 Items
- Cluster 2 hat 88 Items
- Cluster 3 hat 104 Items
- Cluster 4 hat 186 Items

Durch die Analyse der Tabelle 11 können wir sagen, dass die Label Liquidität und beschaftigeMonate eine große Rolle bei der Bildung von Clustern gespielt haben. Die restlichen Merkmale haben drei ausgeprägte centroid-Werte. Dies beruht vermutlich darauf, dass die Cluster 1,2 und 3 stark

mit den übrigen Merkmalen in der gleichen Weise korrelieren.

Tabelle 11: Centroid Tabelle

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Bonitaet	0.385	0.797	0.828	0.750	0.194
Zahlungsverzuege	0.229	0.053	0.042	0.062	0.463
beschaeftigteMonate	0.195	0.383	0.481	0.285	0.129
Verschuldungsgrad	0.291	0.133	0.125	0.150	0.564
Kredithoehe	0.360	0.219	0.692	0.366	0.295
Liquiditaet	0.153	0.640	0.573	0.284	0.084
AnzahlKreditlinien	0.461	0.207	0.184	0.212	0.570

Die Abbildung 13 zeigt Cluster 4 und Cluster 2 als Extremwerte unserer Klassifizierungen und Cluster 0 bleibt als Mittelwert. Die Zentren von Cluster 1, 2 und 3 sind fast ununterscheidbar, was zu Ungenauigkeiten führt.

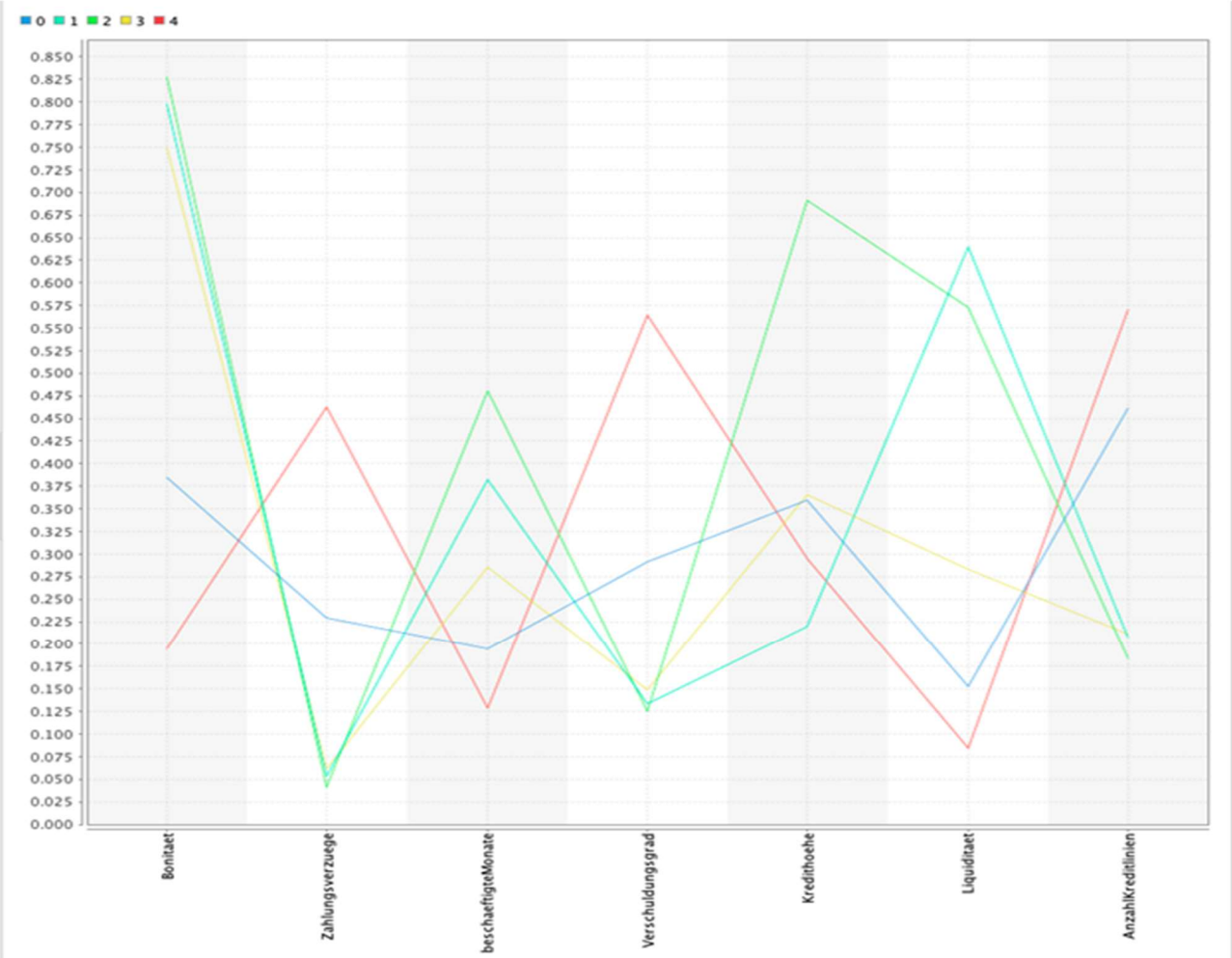

Abbildung 13: Centroid Plot

Anhand der beiden Balken Diagramme (Abbildung 14 und Abbildung 15) kann die Klassifizierung für die fünf Klassen erklärt werden. Zum einen führt eine hohe Bonität bei geringem Zahlungsverzug und geringem Verschuldungsgrad dazu, dass die Einstufung in ein sehr geringes oder geringes Kreditrisiko erfolgt. Zusätzlich ist für diese Einstufungen auch eine hohe Anzahl an beschäftigten Monaten nötig.

Dem gegenüber steht eine Einstufung für die Kreditunwürdigkeit (hier: nicht Kreditwürdig). Dabei ist die Anzahl der Kreditlinien sehr hoch sowie die Zahlungsverzüge und der Verschuldungsgrad. Ein hohes Kredit Risiko zeichnet sich durch dieselben Merkmale wie die Klasse nicht kreditwürdig aus. Ein Unterschied, wodurch eine andere Einstufung entsteht, ist das die Bonität und die Anzahl an beschäftigten Monaten höher ist.

Durch die beiden Abbildungen wird außerdem deutlich, dass eine geringe Anzahl der beschäftigten Monate einen Einfluss auf die Liquidität hat. Daraus folgt auch ein erhöhter Verschuldungsgrad.

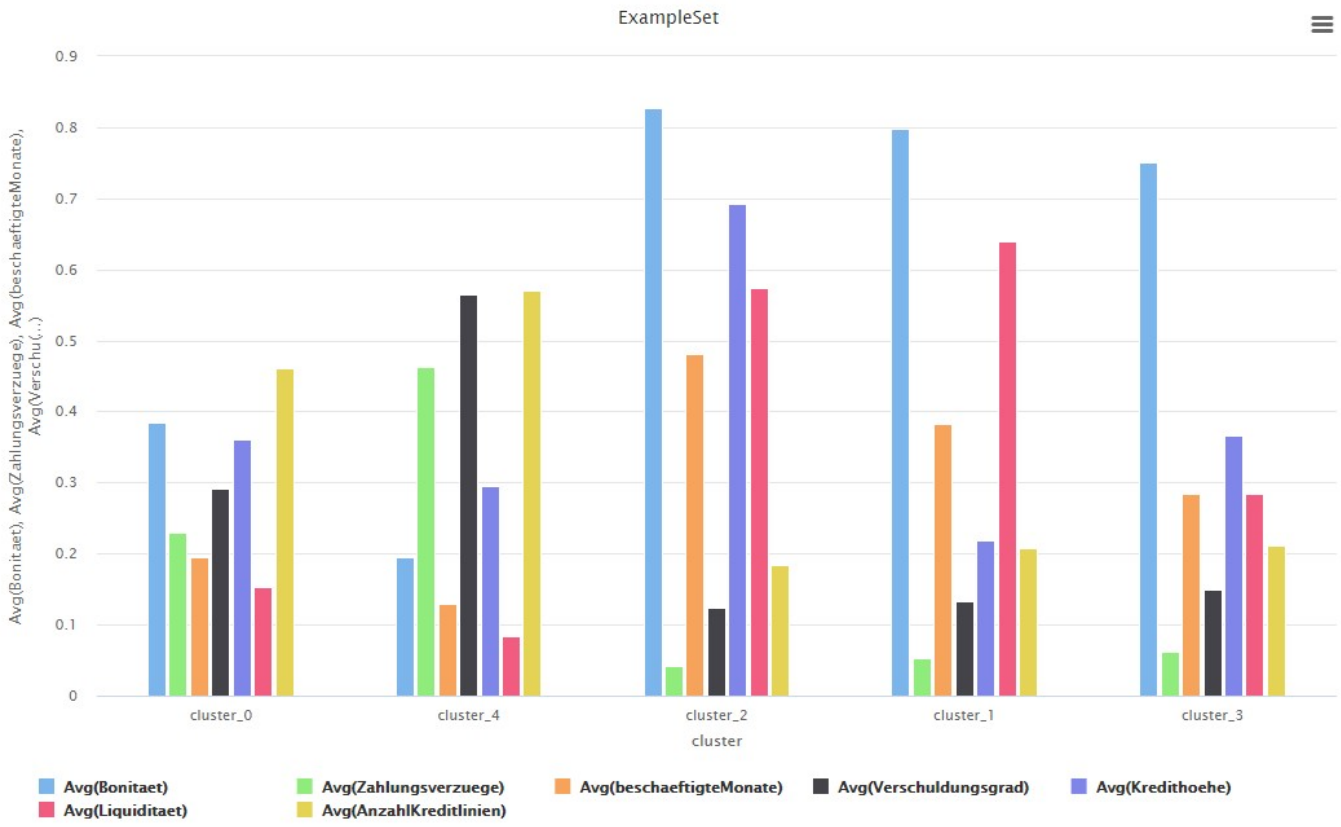

Abbildung 14: Darstellung der Cluster in Bezug auf Merkmale

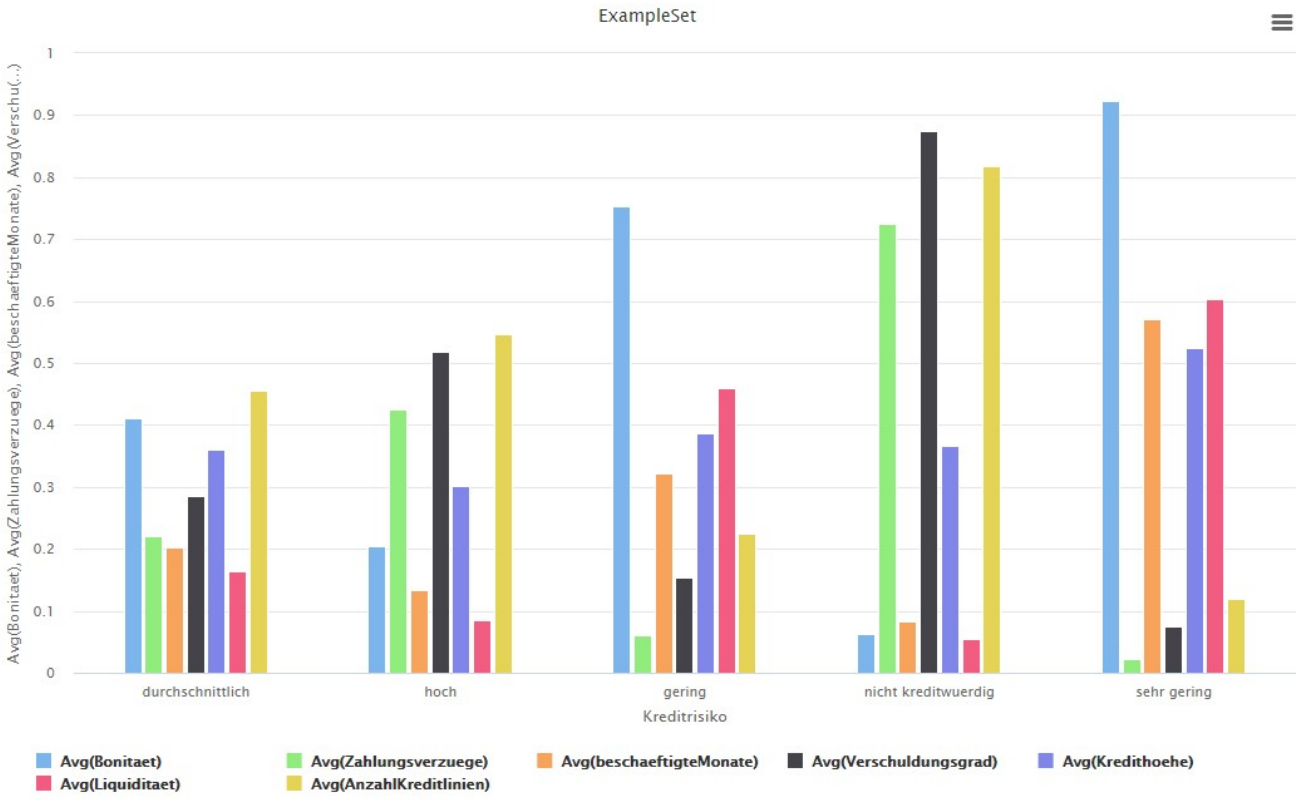

Abbildung 15: Darstellung der Labels in Bezug auf Merkmale

Die Tabelle 12 zeigt die Genauigkeit, Präzision und den Recall des K-Means clustering im Vergleich zu den wahren Werten. Insgesamt stellt sich dabei eine Genauigkeit von 71,30 % dar. Auf der rechten Seite der Tabelle wird die Klassen-Präzision dargestellt. Dabei wird erkennbar, dass zu 95,16 % der Fälle ein hohes Kreditrisiko korrekt festgestellt werden kann. Der niedrigste Wert der Präzision liegt bei nicht kreditwürdig. Bei diesem Fall würde das Modell, unter den gegebenen Daten, immer falsch liegen. Dabei wurde festgestellt, dass die gelabelten Daten und die Cluster mit den untenstehenden Werten übereinstimmen:

- Cluster 0 = durchschnittlich mit 98.86% Genauigkeit
- Cluster 1 = nicht kreditwürdig mit 0% Genauigkeit (vermutlich aufgrund von geringer spezifischer Datenmenge)
- Cluster 2 = sehr gering mit 55.38% Genauigkeit
- Cluster 3 = gering mit 42.99% Genauigkeit
- Cluster 4 = hoch mit 86.34% Genauigkeit

Tabelle 12: Confusion Matrix des Cluster Modells

accuracy: 71.30%

	true durchschnittl...	true hoch	true gering	true nicht kreditw...	true sehr gering	class precision
pred. durchschnit...	174	28	0	0	0	86.14%
pred. hoch	0	177	0	9	0	95.16%
pred. gering	2	0	95	0	7	91.35%
pred. nicht kredit...	0	0	74	0	22	0.00%
pred. sehr gering	0	0	52	0	36	40.91%
class recall	98.86%	86.34%	42.99%	0.00%	55.38%	