Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Modulschlussprüfungen 2024 ☐ HS ☒ FS
BSc Business Artificial Intelligence

Modul: Maschinelles Lernen

Dozent/in: Hans Friedrich Witschel, Andreas Martin

Prüfungsdatum: 25.06.2024 **Dauer:** 60 Minuten

Standort: ☒ Olten ☐ Basel ☐ Brugg-Windisch

Prüfungsart: ☒ Papier

☐ Beaxi ☐ MooPad ☐ Moodle

☐ mit Beilage (bewertet) ☐ mit Beilage (nicht bewertet)

Erlaubte Hilfsmittel: ☐ Keine Hilfsmittel erlaubt ☒ Taschenrechner TI-30

☐ Laptop ☒ Taschenrechner TI-84

☒ Andere: 1 (ein) A4 Blatt (beidseitig) mit eigenen Notizen

Hinweise:

- Bitte versehen Sie zum **Prüfungsbeginn alle Blätter** mit Ihrem **Namen**.
- Schreiben Sie Ihre Lösungen **nur** auf die **ausgeteilten Prüfungsblätter**.
- Falls der für die Lösung vorgesehene Platz nicht reicht, nutzen Sie die Rückseite. Falls auch dieser Platz nicht reicht, können Sie weitere Blätter bekommen.
- Schreiben Sie **nicht** mit **Bleistift**. (Einzige Ausnahme: Grafiken)

Punkte:

Prüfungsteil	Max. Punkte	Erreicht
Hans Friedrich Witschel	48	
Andreas Martin	12	
Punkte Klausur	**60**	
Teilnote Klausur (70%)		
Teilnote LIM (Gruppenarbeit, 30%)		

Gesamtnote:

1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 1 (10 Punkte)

a) (6P) Eine Bank möchte maschinelles Lernen nutzen, um Kreditkartenbetrug vorherzusagen. Es soll ein Modell gelernt werden, welches jedes Mal, wenn eine Kreditkarte genutzt wird, prüft, ob ein Missbrauch der Karte vorliegt.
Bitte formalisieren Sie dies als Klassifikationsaufgabe, indem Sie Instanzen, 2 Attribute und das Klassenattribut festlegen. Bitte begründen Sie für jedes der zwei Attribute kurz, warum Sie es für hilfreich für die Vorhersage halten.

Instanzen: Kreditkartentransaktionen

Attribut 1 + Begründung: Distanz zwischen der aktuellen und der letzten mit der gleichen Karte getätigten Transaktion. Weite Entfernungen sind verdächtig, insbesondere bei kurzen Abständen (siehe Attribut 2)

Attribut 2 + Begründung: Zeit seit der der letzten Transaktion; sinnvoll in Kombination mit Attribut 1

Klassenattribut: Betrug (ja / nein)

Punkteverteilung: 1P für Instanzen, 2P je Attribut / Begründung, 1P für Klassenattribut; andere Attribute sind natürlich möglich

b) (4P) Bitte geben Sie ein Beispiel für ein (plausibles) multivariates Muster, welches von dem Klassifikationsmodell erlernt werden könnte!

Lösung: z.B. Distanz > 2500 km *und* Zeit < 5 Minuten (siehe Attributdefinitionen in Aufgabenteil a))

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 2 (4 Punkte)

Die Firma Teleflow bietet 1-Jahres-Verträge für Telefon- und Internet-Services an. Für ein Modell zur Vorhersage von Kundenabwanderung wird eine Datenaufbereitung durchgeführt. Nach einigen Aufbereitungsschritten liegen die zwei folgenden Tabellen vor:

contractId	Service
0004-TLHLJ	DeviceProtection
0013-EXCHZ	TechSupport
0013-EXCHZ	StreamingTV
0011-IGKFF	OnlineBackup
0011-IGKFF	DeviceProtection
0011-IGKFF	StreamingMovies
0011-IGKFF	StreamingTV

Cancellation_in_period	contractID	PeriodStart	PeriodEnd	EndDate	gender	MonthlyCharges
no	0011-IGKFF	27.11.2021	27.11.2022	22.12.2022	Male	98
yes	0004-TLHLJ	14.12.2021	13.04.2022	13.04.2022	Male	73.9
yes	0013-EXCHZ	17.08.2023	15.11.2023	15.11.2023	Female	83.9
yes	0011-IGKFF	27.11.2022	22.12.2022	22.12.2022	Male	98

Nun werden weitere Aufbereitungsschritte durchgeführt. Danach sehen die (gleichen) Daten so aus:

StreamingTV	DeviceProtection	OnlineBackup	StreamingMovies	TechSupport	Cancellation_in_period	contractID	PeriodStart	PeriodEnd	EndDate	gender	MonthlyCharges
1	1	1	1	*NULL*	no	0011-IGKFF	27.11.2021	27.11.2022	22.12.2022	Male	98
NULL	1	*NULL*	*NULL*	*NULL*	yes	0004-TLHLJ	14.12.2021	13.04.2022	13.04.2022	Male	73.9
1	*NULL*	*NULL*	*NULL*	1	yes	0013-EXCHZ	17.08.2023	15.11.2023	15.11.2023	Female	83.9
1	1	1	1	*NULL*	yes	0011-IGKFF	27.11.2022	22.12.2022	22.12.2022	Male	98

Welche Operationen wurden durchgeführt? (mehrere Antworten können richtig sein, falsche Antworten führen zu Abzug!)

- ☐ Pivotierung
- ☐ Diskretisierung von numerischen Attributen
- ☐ One-Hot-Encoding
- ☐ Skalierung
- ☐ Verknüpfung von Tabellen mittels Join

Punkteverteilung: 2P für jede korrekte Auswahl, 1P Abzug für jede falsche
Ausnahme: wenn «One-Hot-Encoding" angekreuzt ist, nicht aber «Pivotierung», dann werden für diese Kombination nur 1.5P abgezogen statt 3.

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 3 (3 Punkte)

Die Don't Worry-Versicherung (DWV) verliert immer wieder Kunden, die all ihre bestehenden Policen auflösen. Mit Hilfe eines Datensatzes mit 20'000 Kunden soll ein Modell gelernt werden, welches Kundenabwanderung vorhersagen kann. Der Datensatz wurde im csv-Format bereitgestellt und in Orange geladen. Dort sehen die Daten wie folgt aus:

ANZ_ST	N numeric	feature	
SEX_CD	C categorical	feature	M, W
SPRACH_CD	C categorical	feature	D, E, F, I
VERTRIEBSKANAL	N numeric	feature	
AGE_ST	N numeric	feature	
NATION_CD_b	C categorical	feature	CH, other
TIME	N numeric	feature	
STATUS	**C categorical**	**target**	**0, 1**

Die Bedeutung der Attribute:

- ANZ_ST: Anzahl Policen bei Start über alle Branchen
- SEX_CD: Geschlecht
- SPRACH_CD: Sprache – D-Deutsch, F-Französisch, I-Italienisch, E-Englisch
- Vertriebskanal: 11000 (Aussendienst), 13000 (Broker), 14000 (Hauptsitz), 15000 (alternative Sales-Kanäle)
- AGE_ST: Alter bei Start
- NATION_CH: Nationalität – mögliche Ausprägungen (CH, other)
- TIME: wie viele Jahre war dieser Kunde im Bestand?
- STATUS: hat der Kunde gechurnt? (0: Kunde ist noch im Bestand, 1: Kunde ist nicht mehr im Bestand)

Betrachten Sie die Typen der Attribute. Bei welchem/n Attribut(en) sollte der Typ geändert werden? (mehrere Antworten können richtig sein, falsche Antworten führen zu Abzug!)

- ☐ ANZ_ST
- ☐ SEX_CD
- ☐ SPRACH_CD
- ☐ VERTRIEBSKANAL
- ☐ AGE_ST
- ☐ NATION_CD_b
- ☐ TIME
- ☐ STATUS

Punkteverteilung: 3P für die richtige Auswahl, 1P Abzug für jede falsche

Name, Vorname _____ _____
(in Blockbuchstaben)

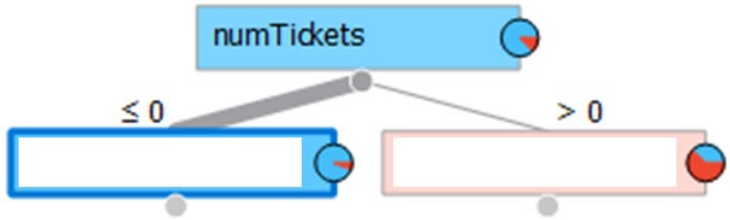
Matrikelnummer _____ Klasse _____

Aufgabe 4 (6 Punkte)

Wir betrachten wieder den Fall der Firma TeleFlow, welche die mögliche Abwanderung ihrer Kunden vorhersagen will. Aus den vorhandenen Daten wurde ein Entscheidungsbaum gelernt, welcher «Ja» vorhersagt, falls ein Kunde seinen Vertrag innerhalb einer Vertragsperiode kündigt, sonst nein. Der Baum nutzt unter anderem die Attribute «numTickets» (durchschnittliche Anzahl Support-Tickets, die vom Kunden innerhalb der Periode eröffnet wurden), sowie «MonthlyCharges» (monatlicher Rechnungsbetrag).

Aus dem Baum wurde ein Abwanderungsmuster herausgelesen, das sich so beschreiben lässt: «Kunden, welche ein oder mehr Support-Tickets hatten und monatlich mehr als 30 CHF bezahlen, wandern ab.»

Aufgabe: Ergänzen Sie das folgende Baum-Fragment so, dass es zu der Interpretation passt.
Hinweis: Sie müssen wirklich nur das ergänzen, was zu dem beschriebenen Muster passt!



Lösung: den rechten Knoten mit «MonthlyCharges» beschriften, dann von dort eine Verbindung zu einem mit «Ja» beschrifteten Knoten herstellen, welche ihrerseits mit «>30» beschriftet ist.
Punkteverteilung: 2P für die Beschriftung des vorhandenen Knotens, 2P für den mit «Ja» beschrifteten Blattknoten, 2P für die mit «>30» beschriftete Verbindung

Aufgabe 5 (5 Punkte)

Eine Smartphone-App misst über verschiedene eingebaute und angeschlossene Sensoren die Vitalfunktionen von Patienten mit erhöhtem Risiko für Herzinfarkte. Da die Sensoren oft günstig und daher unzuverlässig sind, können Werte oft nicht erhoben werden, d.h. sie fehlen.

Die App enthält ein Klassifikationsmodell, welches für eine gegebene Konstellation aus Parameterwerten einen Alarm auslöst («Ja») oder nicht («Nein»). Die Messungen erfolgen häufig und mögliche Warnungen sollen daraufhin sehr schnell erfolgen, um keine lebenswichtige Zeit zu verlieren.

Wenn Sie für die Konstruktion des Klassifikationsmodell die Wahl haben zwischen einem Entscheidungsbaum- und einem k-Nearest-Neighbour-Algorithmus, für welchen entscheiden sie sich? Bitte begründen Sie Ihre Wahl anhand der o.g. Informationen!

Lösung: relevant ist hier die Robustheit im Umgang mit fehlenden Werten, sowie die Schnelligkeit der Vorhersage. Beides ist beim Entscheidungsbaum deutlich besser, dieser ist demnach vorzuziehen.
Punkteverteilung: 2.5P für die Identifikation und Zuordnung jedes der beiden Kriterien

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 6 (8 Punkte)

a) (4P) Eine private Krankenversicherung nutzt ein lineares Regressionsmodell, um die verursachten Behandlungskosten für neue Kunden vorherzusagen. Folgendes Modell wurde gelernt:

	name	coef
1	intercept	-666.938
2	age	256.856
3	sex=female	65.6572
4	sex=male	-65.6572
5	bmi	339.193
6	children	475.501
7	smoker=no	-11924.3
8	smoker=yes	11924.3
9	region=northeast	587.009
10	region=northw...	234.045
11	region=southeast	-448.013
12	region=southw...	-373.042

Welche der untenstehenden Aussagen sind korrekt (mehrere Antworten können richtig sein, falsche Antworten führen zu Abzug!)?

- ☐ Weibliche Patientinnen verursachen tendenziell mehr Kosten als männliche
- ☐ Das Alter hat einen geringeren Einfluss auf die vorhergesagten Kosten als die Region, in der jemand lebt
- ☐ Jedes Lebensjahr lässt die vorhergesagten Kosten signifikant (um fast 257 CHF) ansteigen
- ☐ Die Attribute mit negativen Koeffizienten (z.B. männliches Geschlecht) haben keinen Einfluss auf die Vorhersage

Punkteverteilung: 2P für jede korrekt angekreuzte Aussage, 2P Abzug für jede falsch angekreuzte

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

b) (4P) Aufgrund schlechter Performance des ursprünglichen Modells wurde ein neues lineares Regressionsmodell gelernt, welches so aussieht:

	name	coef
1	intercept	8301.33
2	age	251.161
3	sex=female	-0
4	sex=male	0
5	bmi	299.342
6	children	0
7	smoker=no	-17677
8	smoker=yes	1.60286e-11
9	region=northeast	0
10	region=northw...	0
11	region=southeast	-0
12	region=southw...	-0

Welche Modifikation am Regressionsalgorithmus kann die gezeigte Veränderung am ehesten bewirkt haben? Welche Überlegung bzw. welches Problem könnten dazu geführt haben, diese Veränderung vorzunehmen?

Lösung: Lasso-Regression, mit dem Ziel die Komplexität des Modells zu reduzieren – da evtl. Overfitting vorlag…
Punkteverteilung: 2P für die Nennung von Lasso, 2P für die Begründung

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 7 (5 Punkte)

Eine Unfallversicherung nutzt ein Klassifikationsmodell, um unberechtigte Kosten in Rechnungen aufzudecken. Wenn das Modell eine Warnung ausgibt, wird die entsprechende Rechnungsposition manuell von der Schadensabteilung geprüft, was durchschnittlich 15 Minuten (~ 20 CHF) in Anspruch nimmt. Eine unberechtigte Schadensforderung (Rechnungsposition) kostet im Schnitt 100 CHF. Dies resultiert in der folgenden Kostenmatrix, welche für die Evaluation des Klassifikationsmodells eingesetzt wird:

		predicted fraud	
		yes	no
actual fraud	yes	20	100
	no	20	0

Eine aktuelle Version des Klassifikationsmodells resultiert in der folgenden Confusion Matrix auf einer Testmenge von 1000 Rechnungspositionen:

		predicted fraud	
		yes	no
actual fraud	yes	20	30
	no	100	850

Berechnen Sie die Genauigkeit (Accuracy) des Klassifikationsmodells, sowie dessen Kosten. Geben Sie etwaige Zwischenschritte Ihrer Berechnungen an!

Lösung: Accuracy = (20 + 850)/1000 = 87%
Kosten = 20*20 CHF + 100*20 CHF + 30*100 CHF = 400 + 2000 + 3000 = 5400 CHF
Punkteverteilung: 2P für die Berechnung der Accuracy, 3P für die Kosten

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 8 (4 Punkte)

Für die in der vorigen Aufgabe beschriebene Klassifikation wurden zwei Modelle, A und B, mittels der Masse Genauigkeit (Accuracy) und Area Under the Curve (AUC) bewertet. Zusätzlich wurde die Genauigkeit des Baseline-Klassifikators «Constant» gemessen:

	Genauigkeit	AUC
Modell A	91%	0.53
Modell B	76%	0.83
Constant	92%	-

Welchen der folgenden Aussagen stimmen Sie zu (mehrere Antworten können richtig sein, falsche Antworten führen zu Abzug!)?
Hinweis: wählen Sie eine Aussage nur aus, wenn insbesondere auch die *Begründung* stimmig ist!

- ☐ Modell B ist nicht geeignet, da seine Genauigkeit weit unter der der Baseline liegt
- ☐ Modell A ist nicht geeignet, da seine Genauigkeit nicht einmal so gut ist wie die der Baseline
- ☐ Modell B ist aufgrund seines AUC-Wertes besser geeignet als Modell A
- ☐ Das Modell «Constant» ist hier aufgrund seiner hohen Genauigkeit am besten geeignet
- ☐ Modell A ist nicht geeignet, da sein AUC-Wert anzeigt, dass es nur minimal besser ist als «zufälliges Raten»

Punkteverteilung: 2P für jede korrekte Wahl, 1P Abzug für jede falsche

Aufgabe 9 (3 Punkte)

Zum Vergleich zweier Klassifikationsmodelle wurde eine Evaluierung mittels Area Under the Curve (AUC) durchgeführt, sowohl auf dem gesamten verfügbaren Datensatz (Trainingsmenge) als auch per Holdout. Dabei ergaben sich folgende Werte:

	AUC auf Trainingsmenge	AUC auf Testmenge (Holdout)
Modell A	0.99	0.71
Modell B	0.63	0.62

Welchen der folgenden Aussagen stimmen Sie zu (mehrere Antworten können richtig sein, falsche Antworten führen zu Abzug!)?

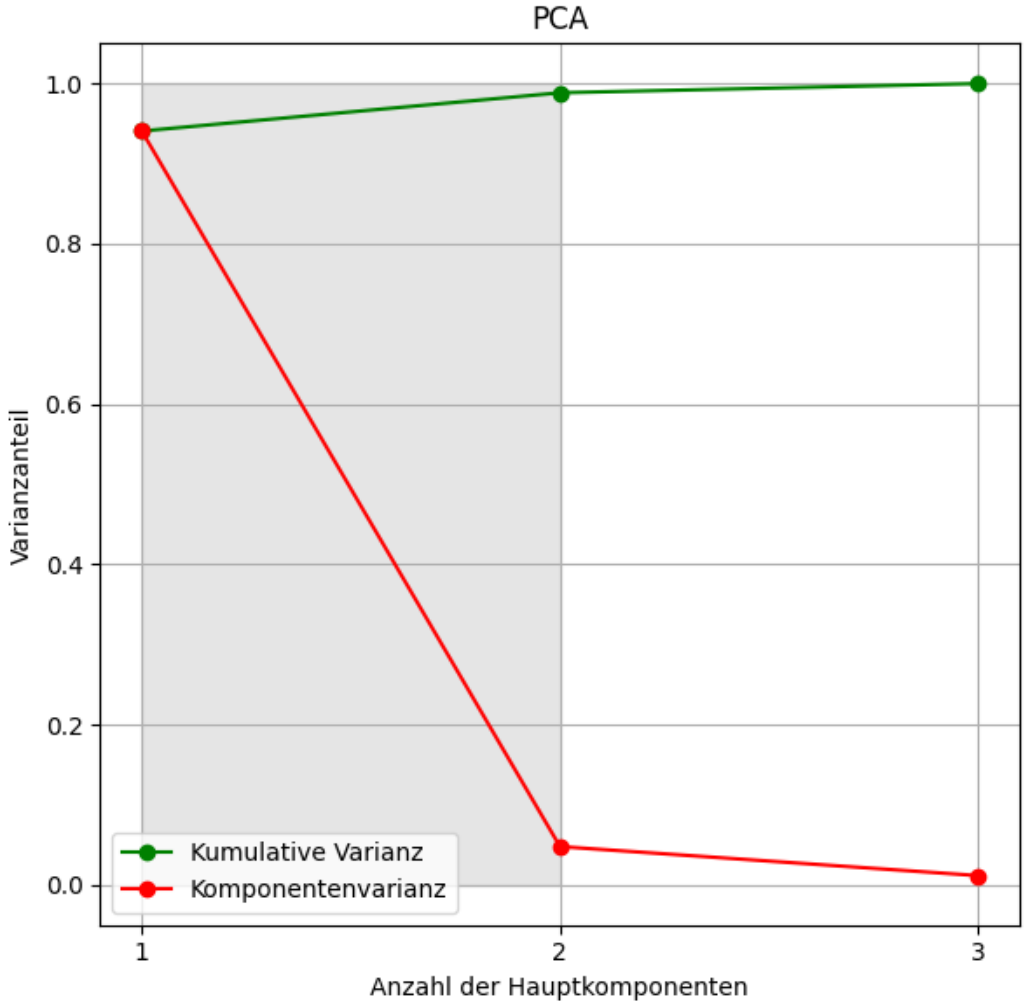
- ☐ Bei Modell A liegt Underfitting (hoher Bias) vor
- ☐ Bei Modell B liegt Underfitting (hoher Bias) vor
- ☐ Bei Modell A liegt Overfitting (hohe Variance) vor
- ☐ Bei Modell B liegt Overfitting (hohe Variance) vor
- ☐ Bei keinem der Modelle liegt Under- oder Overfitting vor
- ☐ Die Zahlen lassen keine der obigen Rückschlüsse zu

Punkteverteilung: 1.5P pro richtige Auswahl, 1P Abzug je falsche

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 10 (6 Punkte)



RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
☐	☐	Die erste Hauptkomponente (PC1) trägt über 80% zur Erklärung der Gesamtvarianz der Daten bei.
☐	☐	Die Summe der Varianzanteile aller betrachteten Hauptkomponenten bleibt unter 1.0 (unter 100 %).
☐	☐	Nach Einbeziehung der dritten Hauptkomponente (PC3) erreicht die kumulative Varianz der analysierten Daten einen Wert von 100%.
☐	☐	Obwohl die zweite Hauptkomponente (PC2) weitere Informationen zur Datenstruktur beiträgt, ist ihr Beitrag zur kumulativen Varianz im Vergleich zur ersten Hauptkomponente minimal.

- *- Für jede korrekt gewählte Aussage gibt es 0.5* ***Pluspunkte****.*
- *- Für jede nicht korrekt gewählte Aussage gibt es 0.5* ***Minuspunkte****.*
- *- Keine, einige oder alle Aussagen können RICHTIG oder FALSCH sein.*
- *- Es gibt ein Minimum von 0 Punkten.*

Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Name, Vorname _____ _____
(in Blockbuchstaben)

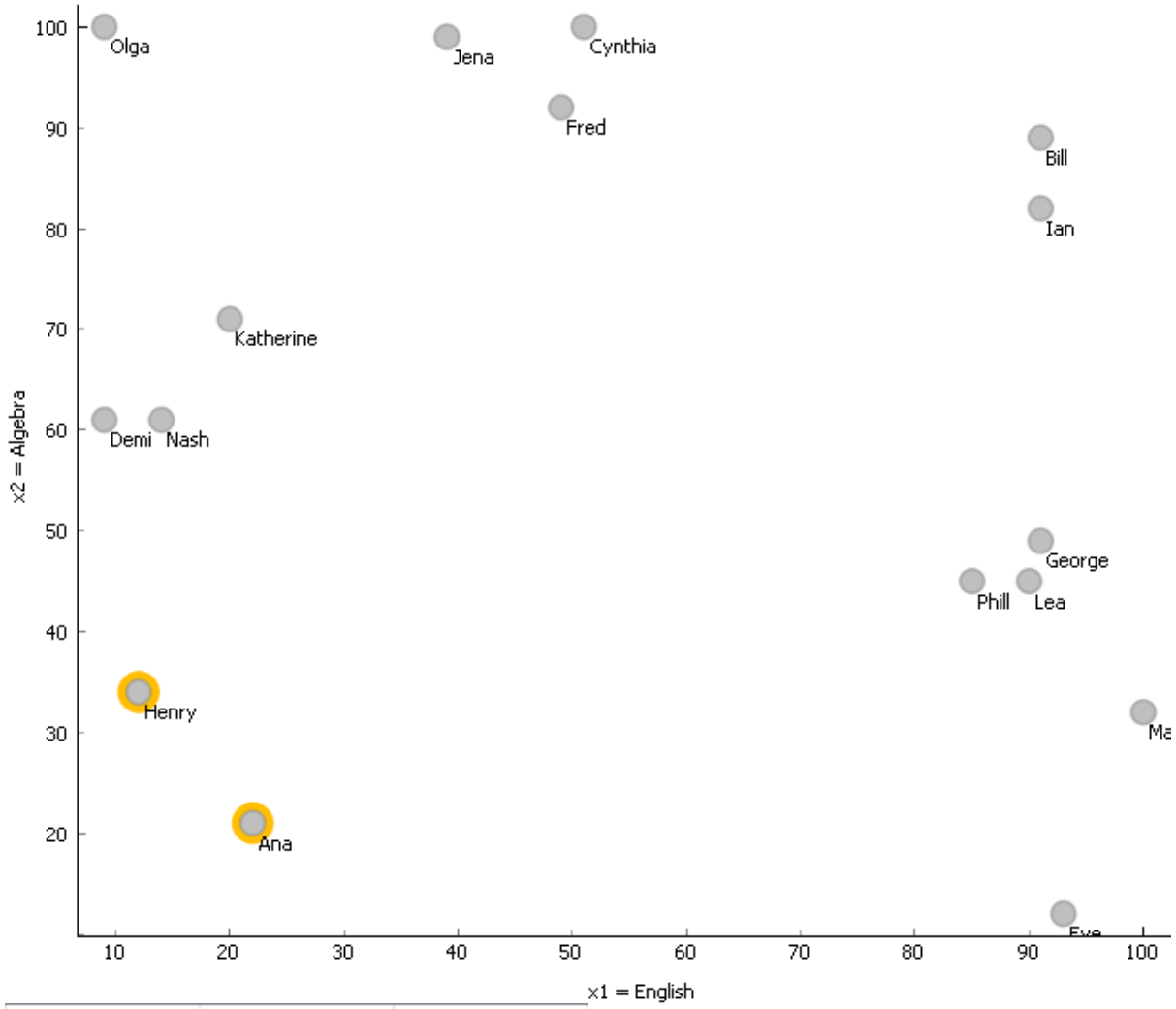
Matrikelnummer _____ Klasse _____

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
☐	☐	t-SNE tendiert dazu, Datenpunkte ähnlicher Typen enger und klarer zu gruppieren als MDS, besonders bei grösseren Datensätzen.
☐	☐	MDS ist besser geeignet als t-SNE, um die lokalen Nachbarschaftsbeziehungen zwischen Datenpunkten zu bewahren.
☐	☐	t-SNE ist besonders effektiv bei der Visualisierung von Daten mit komplexen, nicht-linearen Strukturen, indem es eine klare visuelle Trennung von Clustern ermöglicht.
☐	☐	Die Erstellung einer Distanzmatrix ist der erste Schritt im Prozess der multidimensionalen Skalierung (MDS).

- *- Für jede korrekt gewählte Aussage gibt es 0.5* ***Pluspunkte****.*
- *- Für jede nicht korrekt gewählte Aussage gibt es 0.5* ***Minuspunkte****.*
- *- Keine, einige oder alle Aussagen können RICHTIG oder FALSCH sein.*
- *- Es gibt ein Minimum von 0 Punkten.*

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____



Student	x1 = English	x2 = Algebra
Ana	22	21
Henry	12	34

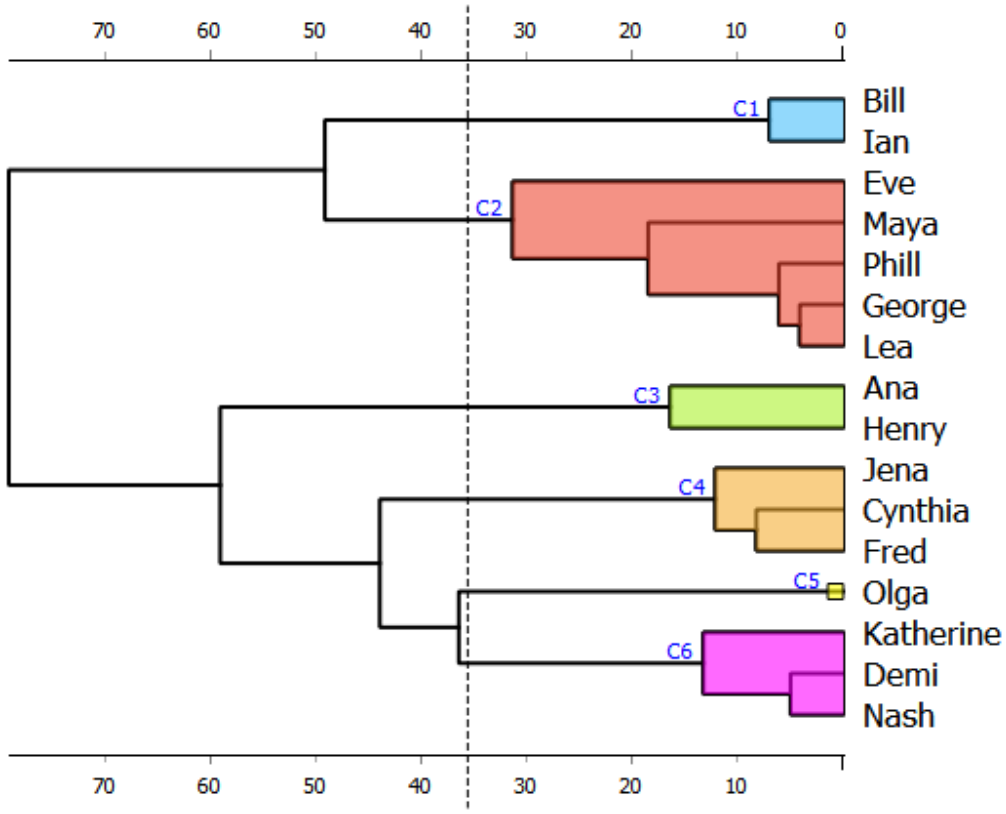
Die euklidische Distanz zwischen Henry und Ana ist:

_____

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Aufgabe 11 (6 Punkte)



RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
☐	☐	Hierarchisches Clustering beginnt mit jedem Datenpunkt als eigenständigem Cluster und verschmilzt schrittweise die nächstliegenden Cluster.
☐	☐	Ein Dendrogramm zeigt die Distanz der verschmolzenen Cluster und seine Linien können sich kreuzen, um komplexe Beziehungen zu verdeutlichen.
☐	☐	Die Wahl des Distanzmasses, wie z.B. die Euklidische Distanz, hat keinen signifikanten Einfluss auf die Struktur der durch hierarchisches Clustering gebildeten Cluster.
☐	☐	Durch das "Schneiden" des Dendrogramms an verschiedenen Punkten können unterschiedliche Anzahlen von Clustern erhalten werden, was eine flexible Anpassung an die Datenanalysebedürfnisse ermöglicht.

- *- Für jede korrekt gewählte Aussage gibt es 0.5* ***Pluspunkte****.*
- *- Für jede nicht korrekt gewählte Aussage gibt es 0.5* ***Minuspunkte****.*
- *- Keine, einige oder alle Aussagen können RICHTIG oder FALSCH sein.*
- *- Es gibt ein Minimum von 0 Punkten.*

Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
☐	☐	Der "Fluch der Dimensionalität" bezieht sich auf das Phänomen, dass mit zunehmender Anzahl an Dimensionen in einem Datensatz die Distanzen zwischen den Datenpunkten immer weniger aussagekräftig werden.
☐	☐	Die euklidische Distanz wird in hochdimensionalen Räumen zunehmend effektiver, da sie die direkte Linie zwischen zwei Punkten misst.
☐	☐	Die Cosinus-Distanz, die den Winkel zwischen zwei Vektoren misst, ist weniger anfällig für den Fluch der Dimensionalität und kann daher in hochdimensionalen Räumen besonders nützlich sein.
☐	☐	Eine Methode zur Minderung des Fluchs der Dimensionalität in der Datenanalyse ist die Normalisierung (bspw. 0 bis 1) der Daten vor der Anwendung von Clustering-Algorithmen.

- *- Für jede korrekt gewählte Aussage gibt es 0.5* ***Pluspunkte****.*
- *- Für jede nicht korrekt gewählte Aussage gibt es 0.5* ***Minuspunkte****.*
- *- Keine, einige oder alle Aussagen können RICHTIG oder FALSCH sein.*
- *- Es gibt ein Minimum von 0 Punkten.*

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
☐	☐	DBSCAN ist besonders geeignet für Datensätze mit variabler Clusterdichte und ist in der Lage, Ausreißer effektiv zu behandeln. Es ist nicht notwendig, die Anzahl der Cluster im Voraus festzulegen.
☐	☐	k-Means ist optimal für Datensätze, in denen die Cluster sphärisch und gut abgegrenzt sind. Dieser Algorithmus erfordert jedoch die Vorgabe der Clusteranzahl und ist empfindlich gegenüber der Platzierung der Anfangszentroide.
☐	☐	Hierarchisches Clustering ist ideal für große Datensätze, da es sehr skalierbar ist und keine Vorgabe der Clusteranzahl benötigt.
☐	☐	Hierarchisches Clustering bietet detaillierte visuelle Darstellungen der Datenstruktur durch ein Dendrogramm, was es besonders wertvoll für die explorative Datenanalyse macht.

- *- Für jede korrekt gewählte Aussage gibt es 0.5* ***Pluspunkte****.*
- *- Für jede nicht korrekt gewählte Aussage gibt es 0.5* ***Minuspunkte****.*
- *- Keine, einige oder alle Aussagen können RICHTIG oder FALSCH sein.*
- *- Es gibt ein Minimum von 0 Punkten.*

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

Lösung

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
X	☐	Die erste Hauptkomponente (PC1) trägt über 80% zur Erklärung der Gesamtvarianz der Daten bei.
☐	X	Die Summe der Varianzanteile aller betrachteten Hauptkomponenten bleibt unter 1.0 (unter 100 %).
X	☐	Nach Einbeziehung der dritten Hauptkomponente (PC3) erreicht die kumulative Varianz der analysierten Daten einen Wert von 100%.
X	☐	Obwohl die zweite Hauptkomponente (PC2) weitere Informationen zur Datenstruktur beiträgt, ist ihr Beitrag zur kumulativen Varianz im Vergleich zur ersten Hauptkomponente minimal.

Lösung

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
X	☐	t-SNE tendiert dazu, Datenpunkte ähnlicher Typen enger und klarer zu gruppieren als MDS, besonders bei grösseren Datensätzen.
☐	X	MDS ist besser geeignet als t-SNE, um die lokalen Nachbarschaftsbeziehungen zwischen Datenpunkten zu bewahren.
X	☐	t-SNE ist besonders effektiv bei der Visualisierung von Daten mit komplexen, nicht-linearen Strukturen, indem es eine klare visuelle Trennung von Clustern ermöglicht.
X	☐	Die Erstellung einer Distanzmatrix ist der erste Schritt im Prozess der multidimensionalen Skalierung (MDS).

Lösung

Student	x1 = English	x2 = Algebra
Ana	22	21
Henry	12	34

Die euklidische Distanz zwischen Henry und Ana ist: 16.401

Lösung:

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

X	☐	Hierarchisches Clustering beginnt mit jedem Datenpunkt als eigenständigem Cluster und verschmilzt schrittweise die nächstliegenden Cluster.
☐	X	Ein Dendrogramm zeigt die Distanz der verschmolzenen Cluster und seine Linien können sich kreuzen, um komplexe Beziehungen zu verdeutlichen.
☐	X	Die Wahl des Distanzmasses, wie z.B. die Euklidische Distanz, hat keinen signifikanten Einfluss auf die Struktur der durch hierarchisches Clustering gebildeten Cluster.
X	☐	Durch das "Schneiden" des Dendrogramms an verschiedenen Punkten können unterschiedliche Anzahlen von Clustern erhalten werden, was eine flexible Anpassung an die Datenanalysebedürfnisse ermöglicht.

Lösung:

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
X	☐	Der "Fluch der Dimensionalität" bezieht sich auf das Phänomen, dass mit zunehmender Anzahl an Dimensionen in einem Datensatz die Distanzen zwischen den Datenpunkten immer weniger aussagekräftig werden.
☐	X	Die euklidische Distanz wird in hochdimensionalen Räumen zunehmend effektiver, da sie die direkte Linie zwischen zwei Punkten misst.
X	☐	Die Cosinus-Distanz, die den Winkel zwischen zwei Vektoren misst, ist weniger anfällig für den Fluch der Dimensionalität und kann daher in hochdimensionalen Räumen besonders nützlich sein.
X	☐	Eine Methode zur Minderung des Fluchs der Dimensionalität in der Datenanalyse ist die Normalisierung (bspw. 0 bis 1) der Daten vor der Anwendung von Clustering-Algorithmen.

Lösung:

RICHTIG	FALSCH	*Welche Aussagen sind RICHTIG und welche sind FALSCH?*
X	☐	DBSCAN ist besonders geeignet für Datensätze mit variabler Clusterdichte und ist in der Lage, Ausreißer effektiv zu behandeln. Es ist nicht notwendig, die Anzahl der Cluster im Voraus festzulegen.
X	☐	k-Means ist optimal für Datensätze, in denen die Cluster sphärisch und gut abgegrenzt sind. Dieser Algorithmus erfordert jedoch die Vorgabe der Clusteranzahl und ist empfindlich gegenüber der Platzierung der Anfangszentroide.
☐	X	Hierarchisches Clustering ist ideal für große Datensätze, da es sehr skalierbar ist und keine Vorgabe der Clusteranzahl benötigt.

Name, Vorname _____ _____
(in Blockbuchstaben)

Matrikelnummer _____ Klasse _____

X	☐

Hierarchisches Clustering bietet detaillierte visuelle Darstellungen der Datenstruktur durch ein Dendrogramm, was es besonders wertvoll für die explorative Datenanalyse macht.