Themenblock Prognosemodelle (Klassifikation und Regression): 02_05Daten _Datenexploration

Datenexploration

Erarbeitet von
Dr. Katja Theune

Lernziele 1
Inhalt 2
Einstieg 2
Datenvorverarbeitung 2
Visualisierung des Outputs 3
Visualisierung der Inputs 4
Visualisierung der Zusammenhänge von Output und Inputs 5
Zusammenhänge und Auswahl von Inputs 7
Abschluss 8
Quellen 8
Disclaimer 8

Lernziele

- Du kannst aus Datenvisualisierungen Informationen über die Inputs und Outputs ableiten
- Du kannst aus Datenvisualisierungen Informationen über die Zusammenhänge zwischen Inputs und Outputs ableiten

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Bevor wir mit den uns vorliegenden Daten Prognosemodelle bilden, müssen wir uns die Daten einmal genauer ansehen. Nur dann bekommen wir wichtige erste Eindrücke davon, ob sich die Daten für unsere Analysen überhaupt eignen. Darüber hinaus erhalten wir durch eine detaillierte Datenexploration schon Hinweise auf Zusammenhänge der Inputs untereinander und zwischen den Inputs und dem Output. Das hilft uns, die Daten passend für unsere weiteren Analysen aufzubereiten.

Datenvorverarbeitung

Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)	Unemployment rate	Inflation rate	GDP	Target
0	0	0.0	0	10.8	1.4	1.74	Dropout
6	6	13.666666666666666	0	13.9	-0.3	0.79	Graduate
0	0	0.0	0	10.8	1.4	1.74	Dropout
10	5	12.4	0	9.4	-0.8	-3.12	Graduate
6	6	13.0	0	13.9	-0.3	0.79	Graduate
17	5	11.5	5	16.2	0.3	-0.92	Graduate
8	8	14.345	0	15.5	2.8	-4.06	Graduate
5	0	0.0	0	15.5	2.8	-4.06	Dropout
7	6	14.142857142857142	0	16.2	0.3	-0.92	Graduate
14	2	13.5	0	8.9	1.4	3.51	Dropout
7	5	14.2	0	13.9	-0.3	0.79	Graduate
8	7	13.214285714285714	0	12.7	3.7	-1.7	Graduate
0	0	0.0	0	12.7	3.7	-1.7	Dropout
8	5	11.0	0	8.9	1.4	3.51	Graduate
5	5	12.0	0	10.8	1.4	1.74	Graduate
7	0	0.0	0	15.5	2.8	-4.06	Dropout
14	2	11.0	0	10.8	1.4	1.74	Enrolled
8	8	14.545	0	15.5	2.8	-4.06	Graduate
8	4	12.25	2	10.8	1.4	1.74	Graduate
8	6	13.5	0	16.2	0.3	-0.92	Enrolled
0	0	0.0	0	11.1	0.6	2.02	Graduate
9	8	11.425	0	12.7	3.7	-1.7	Enrolled
12	7	12.857142857142858	0	12.7	3.7	-1.7	Graduate
7	6	12.285714285714286	0	11.1	0.6	2.02	Graduate
9	7	14.114285714285714	0	11.1	0.6	2.02	Graduate
12	4	11.0	0	7.6	2.6	0.32	Enrolled
9	6	13.285714285714286	0	16.2	0.3	-0.92	Graduate
7	4	13.0	0	9.4	-0.8	-3.12	Enrolled
6	5	12.333333333333334	0	16.2	0.3	-0.92	Graduate
7	6	13.716666666666669	0	16.2	0.3	-0.92	Enrolled
17	5	10.571428571428571	0	16.2	0.3	-0.92	Enrolled
9	4	13.4	0	8.9	1.4	3.51	Graduate
8	2	13.5	0	8.9	1.4	3.51	Enrolled
8	6	14.375	0	12.4	0.5	1.79	Graduate
9	5	13.428571428571429	0	13.9	-0.3	0.79	Graduate
7	1	10.0	0	8.9	1.4	3.51	Dropout
0	0	0.0	0	7.6	2.6	0.32	Dropout
8	2	12.0	2	16.2	0.3	-0.92	Dropout

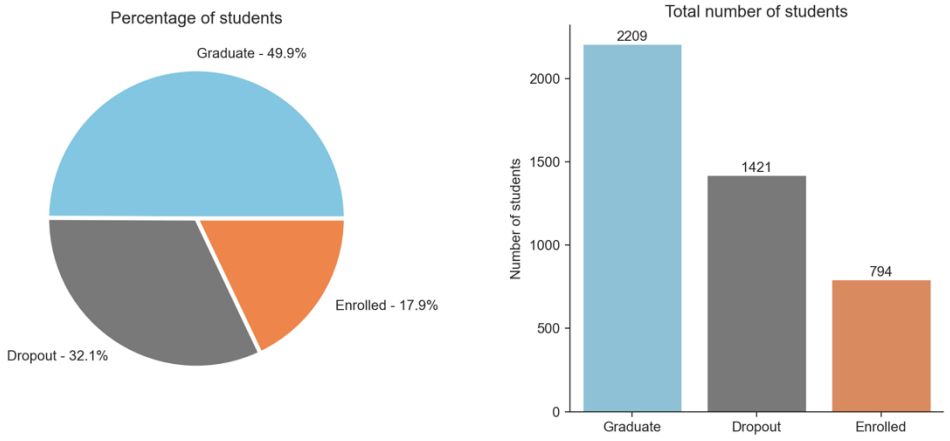
Der Datensatz, den wir hier verwenden, hat schon einige Schritte der Vorverarbeitung durchlaufen. Aus dem Datensatz wurden z. B. schon alle Fehler, Ausreißer und auch fehlende Werte entfernt, bzw. entsprechend aufbereitet.

Quelle [1]

Zudem wissen wir ja schon, dass die meisten Verfahren am besten mit Zahlen umgehen können. Hier sehen wir, dass alle Inputs bereits numerisch sind, die Kategorien also durch Zahlen repräsentiert werden. Nur der Output, hier „Target" genannt, noch nicht. Daher werden wir die Kategorien dann ebenfalls in Zahlen umwandeln.

Visualisierung des Outputs

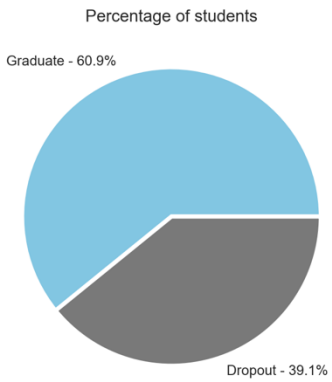
Jetzt wollen wir uns genauer den Output, verschiedene Inputs und schonmal einige mögliche Zusammenhänge untereinander ansehen. Am besten eignen sich dafür Visualisierungen. Beginnen wir mit unserem Output und stellen die absoluten Häufigkeiten der drei Klassen in einem Säulendiagramm und die relativen Häufigkeiten in % anhand eines Kuchendiagramms dar.

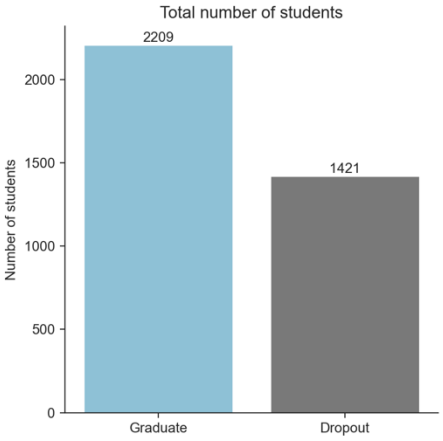


Wir haben hier die Klassen „Graduate", „Dropout" und „Enrolled", also „Studienerfolg", „Studienabbruch" und „noch studierend". Die größte Klasse ist hierbei „Graduate" mit 2209 Studierenden bzw. 49,9 %, gefolgt von „Dropout" mit 1421 Studierenden, bzw. 32,1 %. Die kleinste Gruppe stellen die noch eingeschrieben Studierenden mit 794 Studierenden, bzw. 17,9 %. Insgesamt haben wir 4424 Studierende.

Da wir uns mit einem Klassifikationsproblem mit nur zwei Klassen beschäftigen wollen, entfernen wir alle Beobachtungen, die der Klasse „Enrolled" angehören. Damit verlieren wir natürlich Informationen, aber der Einfachheit halben belassen wir es hier so.

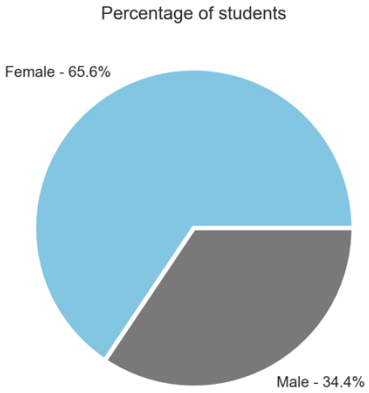
Es bleiben damit 3630 Studierende übrig, von denen 60,9 % der Klasse „Graduate" und 39,1 % der Klasse „Dropout" angehören. Wir haben hier also eine etwas ungleiche Klassenverteilung.





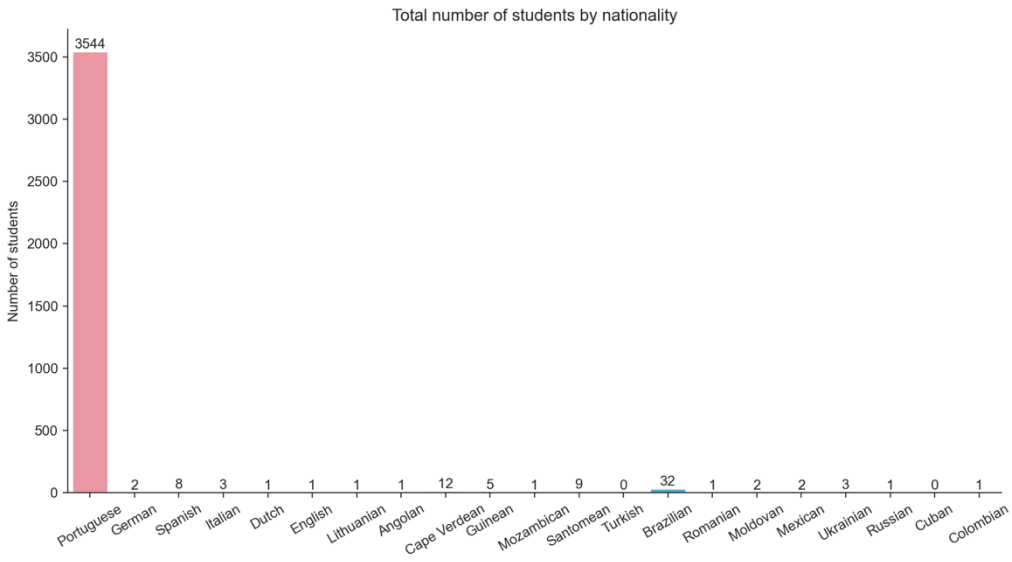
Visualisierung der Inputs

Jetzt schauen wir uns mal beispielhaft zwei der Inputs an. Beginnen wir mit dem „Geschlecht", bzw. „Gender".



Wir sehen hier im Kuchendiagramm, dass 65,6 % der Befragten Frauen und 34,4 % Männer sind. Weitere Kategorien wurden hier leider nicht abgefragt. Männer sind also unterrepräsentiert, wenn wir grundsätzlich von einer Gleichverteilung der beiden Geschlechter in der Grundgesamtheit ausgehen.

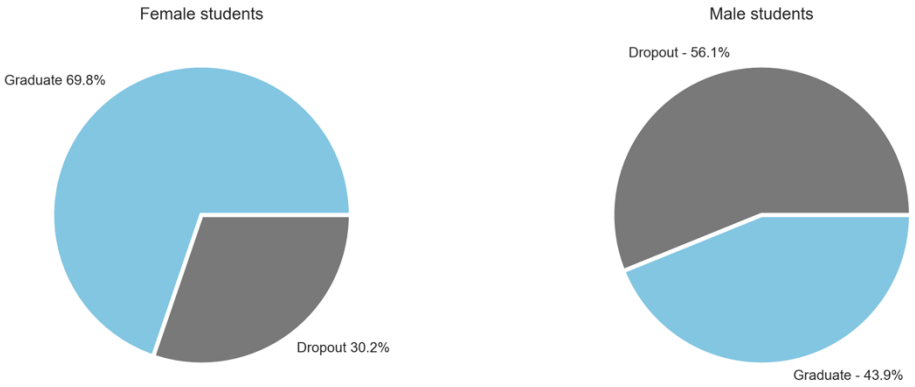
Ein weiterer Input ist die Nationalität der Studierenden. Hier sehen wir, dass die allermeisten, nämlich fast 98 % der Studierenden Portugiesen sind und die anderen Nationalitäten nur sehr wenig besetzt sind.



Diese sehr kleinen absoluten Fallzahlen einzelner Kategorien müssen auch bei der Interpretation der Ergebnisse immer mitberücksichtigen werden. Ggf. ist es sinnvoll, Kategorien zusammenzufassen oder den Input nicht mit in die Analysen einzubeziehen.

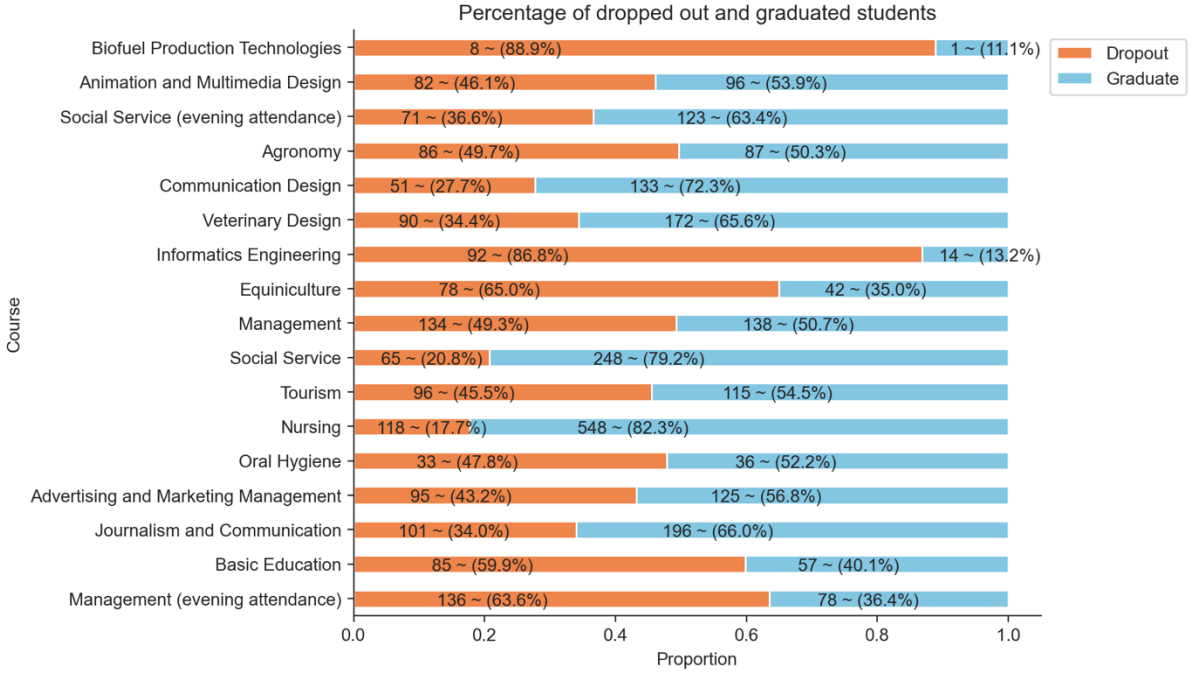
Visualisierung der Zusammenhänge von Output und Inputs

Als nächstes untersuchen wir, ob und wie Output und Inputs zusammenhängen. Um dafür einen ersten Eindruck zu bekommen, eignet sich zunächst auch wieder eine Visualisierung. Z. B. könnten wir uns ansehen, wie viel Prozent der Frauen im Vergleich zu den Männern „Graduates", oder „Dropouts" sind. Das sehen wir hier im Diagramm.

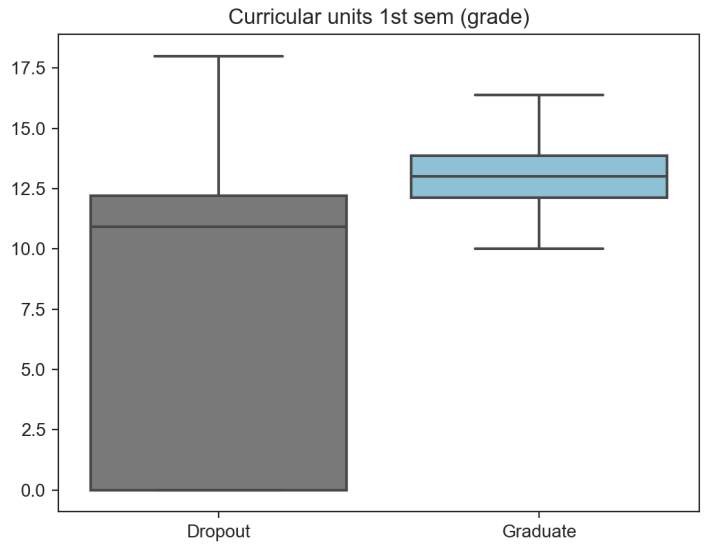


Es fällt auf, dass weibliche Studierende mit 30,2 % eine deutlich niedrigere Studienabbruchquote haben als männliche Studierende mit 56,1 %. Das ist natürlich nur ein erster Hinweis auf einen möglichen Zusammenhang zwischen Geschlecht und Studienabbruch und spiegelt auch keine Kausalität wider.

Ähnlich können wir uns einmal die Erfolgs- und Abbruchquoten in den einzelnen Kursen ansehen. Dieser Barplot zeigt, dass die Abbruchquoten zwischen den einzelnen Kursen stark variieren. So haben z. B. „Nursing" mit 17,7 % und „Social Service" mit 20,8 % die niedrigsten Abbruchquoten. Kurse wie z. B. „Biofuel Production Technologies" mit 88,9 % und „Informatics Engineering" mit 86,8 % haben sehr hohe Abbruchquoten. Auch das ist nur ein erster Hinweis auf einen möglichen Zusammenhang zwischen der Fachrichtung und einem Studienabbruch.



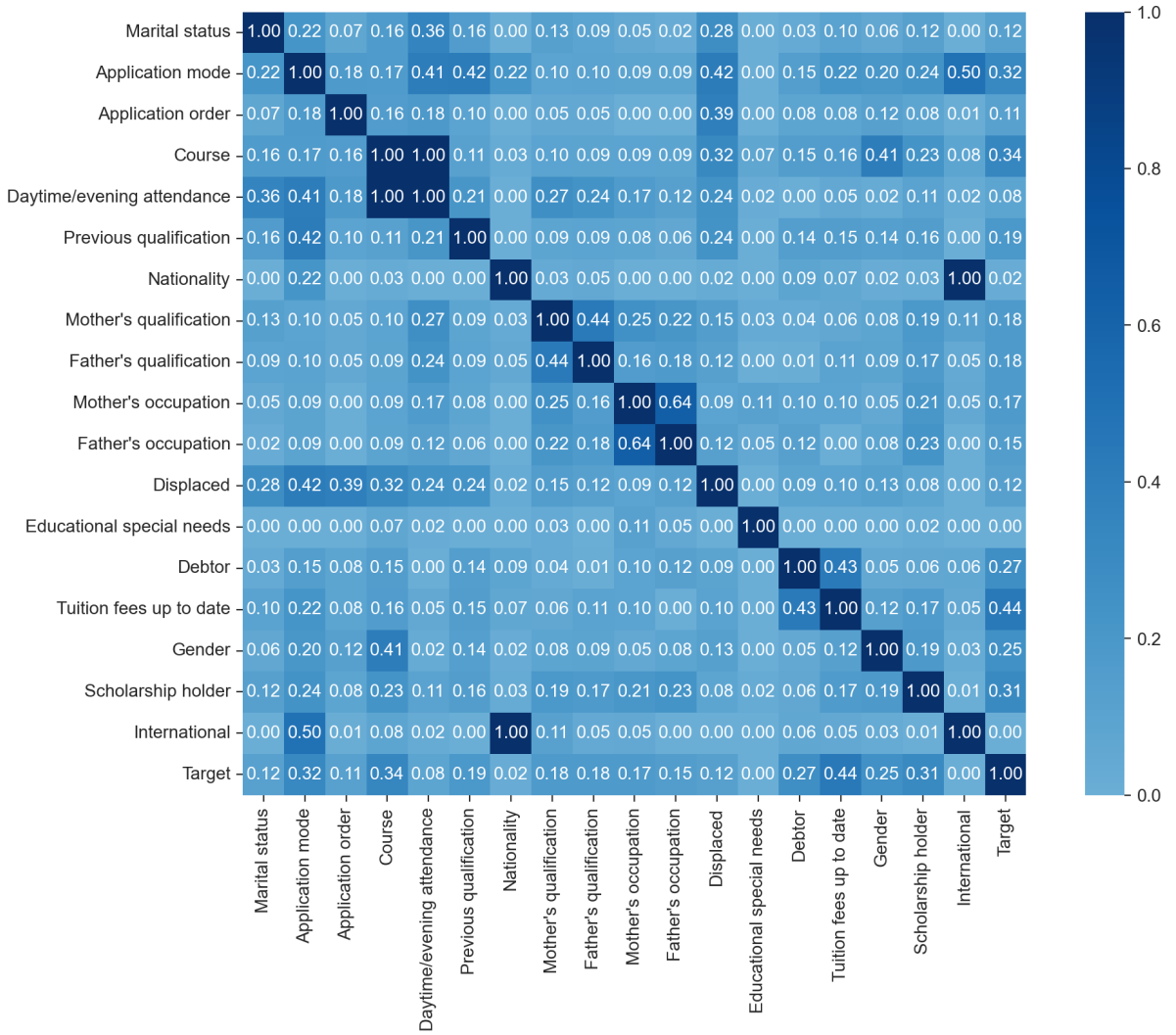
Um uns auch mal einen metrischen Input anzuschauen, sehen wir hier einen Boxplot der Noten am Ende des ersten Semesters für Dropouts und Graduates. Man erkennt deutlich, dass die Noten bei Dropouts niedriger, also hier schlechter sind und stark streuen.



Zusammenhänge und Auswahl von Inputs

Um zu sehen, ob und welche Zusammenhänge es zwischen den einzelnen Inputs und zwischen Inputs und Output gibt, können wir uns eine sogenannte Korrelationsmatrix ansehen, in welcher die einzelnen Korrelationskoeffizienten ausgegeben werden. Wichtig ist hier erstmal nur: Die Farbintensität gibt die Stärke des Zusammenhangs wieder – je dunkler, desto stärker also der Zusammenhang.

Wir sehen hier zur Veranschaulichung nur eine Auswahl der Inputs. Wir können erkennen, dass einige Inputs stark miteinander korrelieren, z. B. der Berufsstatus bzw. die Qualifikation der Mutter und des Vaters. Das bedeutet, dass diese Inputs ähnliche Informationen liefern und davon einige mit dem Ziel entfernt werden können, die Anzahl an Inputs und damit die Dimension zu reduzieren. Man behält meistens den Input mit dem stärkeren Zusammenhang zum Output. Das wäre hier z. B. die Qualifikation der Mutter. Ähnlich gehen wir auch bei den anderen Inputs vor. Zudem entfernen wir Inputs, die nur sehr wenig mit dem Output korrelieren.



Allerdings gibt es hier kein fest vorgeschriebenes Vorgehen. Wir behalten z. B. häufig auch Inputs, die bei unseren Analysen im Fokus stehen, auch wenn sie auf den ersten Blick nicht relevant erscheinen.

Abschluss

Wir haben uns nun einige Details zu unserem Output und verschiedene Inputs angesehen und einen ersten Eindruck der Zusammenhänge in den Daten bekommen. Insbesondere Visualisierungen und Kennzahlen wie Korrelationen liefern hier wichtige Informationen. Diese helfen uns, die Daten passend für weitere Analysen aufzubereiten.

Quellen

Quelle [1] Realinho, V., Machado, J., Baptista, L., Martins, M.V. (2022). Predicting Student Dropout and Academic Success. *Data,* 7(11), 146. https://doi.org/10.3390/data7110146

Grafiken: Dr. Ludmila Himmelspach

Disclaimer