A2EB1 Ermittlung des nächsten Wartungszeitpunkts

Hintergrund:

Neben dem Handel mit Leuchtmitteln per se, bietet die Lichtgestalten GmbH auch Halter für ebendiese an. Dabei werden Sie als Drehteile gefertigt und später in der Montage mit den entsprechenden elektronischen Bauteilen versehen. Die Drehteile werden in einer Werkstatt mit Drehmaschinen von drei Herstellern gefertigt. In Zukunft sollen für diese Maschinen die Wartungszeitpunkte mit Hilfe eines Entscheidungsbaums automatisiert ermittelt werden.

Dabei soll eine Einteilung in „Heute“ für dringende Fälle, „In drei Werktagen“ für demnächst notwendige Wartungen, „In einer Woche“ für baldige Wartungen und zu guter Letzt „In einem Monat“, für wenig dringende Fälle. Dafür erhalten Sie historische Daten bezüglich Kennzahlen wie Betriebstemperatur, Vibrationen, Produktionsauslastung und Betriebsstunden. Weiterhin ist in diesen Daten vermerkt, wie damals zu dem jeweiligen Zeitpunkt entschieden/eingeordnet wurde.

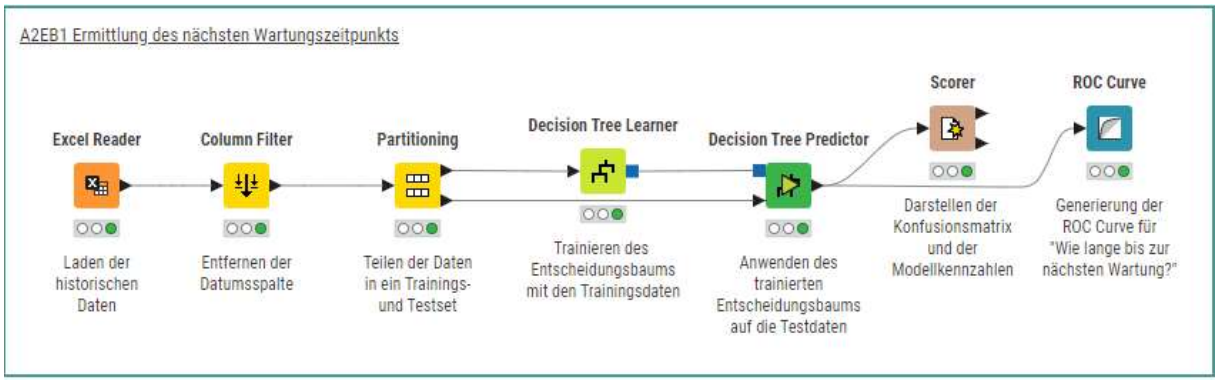
Durch ihren Einsatz zum Training eines Entscheidungsbaums, soll in Zukunft mehr Zeit für andere Tätigkeiten in der Instandhaltung geschaffen werden, indem der Entscheidungsprozess beschleunigt wird.

Aufgabenstellung:

1. Laden sie zunächst mit dem „Excel Reader“-Knoten die historischen Daten in den Workflow („Festlegung_Wartung_historisch.xlsx“).
2. Entfernen Sie mit dem „Column Filter“-Knoten die Spalte für das Datum der letzten Instandhaltung.
3. Trennen Sie die Daten in einen Trainings- und Testdatensatz. Verwenden Sie hierfür den „Partitioning“-Knoten.
4. Mit Hilfe des „Decision Tree Learner“-Knoten trainieren Sie den Entscheidungsbaum. Variieren Sie die Einstellung „Min number records per node“. Was fällt Ihnen auf, wenn Sie den Entscheidungsbaum betrachten?
5. Laden Sie die Testdaten und den trainierten Entscheidungsbaum in den „Decision Tree Predictor“-Knoten. Lassen Sie eine Vorhersage für die Testdaten erstellen.
6. Bewerten Sie mit dem „Scorer“-Knoten die vorhergesagten und tatsächlichen Daten vergleichen. Schauen Sie sich die Konfusionsmatrix an. Was fällt Ihnen auf?
7. Erstellen Sie mit dem „ROC Curve“-Knoten eine Receiver Operation Characteristic Kurve.

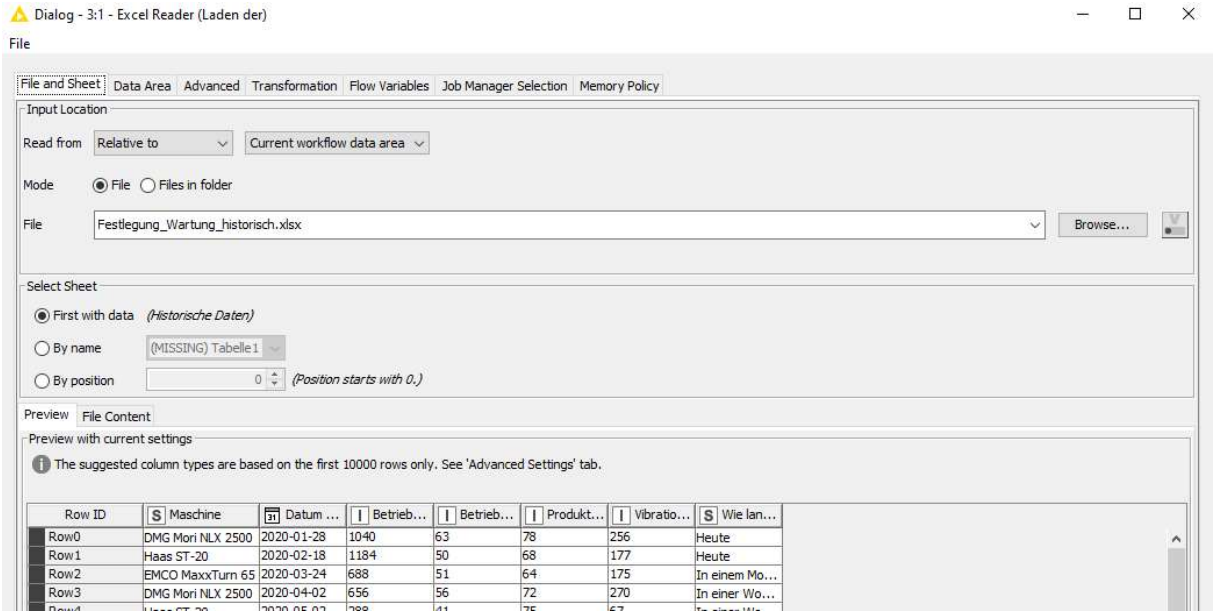
Verwendete Knoten:

Dieser Workflow sollte ihr Ziel sein:



Schritt-für-Schritt-Anleitung

Nachdem Sie den Workflow erstellt haben, laden Sie zunächst mit dem „Excel Reader“-Knoten die Eingangsdaten „Festlegung_Wartung_historisch.xlsx“ aus dem Datenbereich des Workflows. Stellen Sie den Knoten dafür wie auf der nachfolgenden Abbildung zu sehen, ein. (Vorher nicht vergessen, die Eingangsdaten in den „Data“-Ordner abzulegen!)



Nachdem wir nun alle benötigten Daten in unserem Workflow haben, geht es nun daran, sie für das Training des Entscheidungsbaums vorzubereiten.

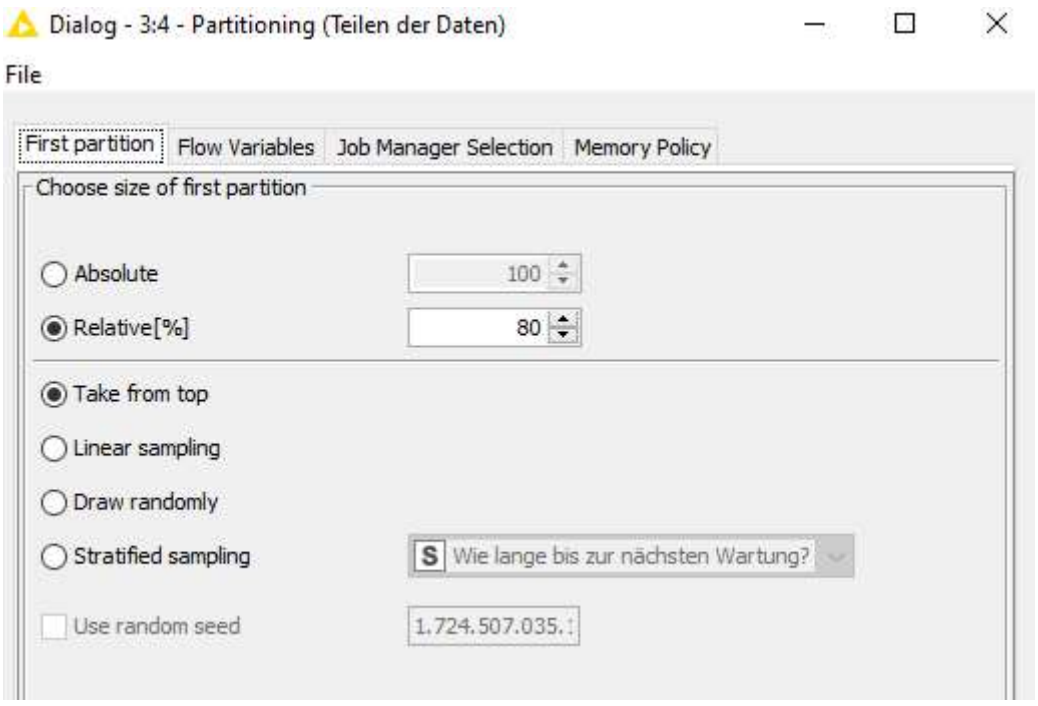
Dafür entfernen wir zunächst mit dem „Column Filter“-Knoten die nicht mehr benötigte Datumsspalte.

Anschließend gilt es noch die Daten in ein Trainings- und ein Testset einzuteilen. Dafür verwenden wir den „Partitioning“-Knoten.

80% der vorhandenen Daten sollen als Trainingsdaten verwendet werden, 20% als Testdaten. Das realisieren wir mit der Einstellung für „Relative[%]“ von 80.

Zunächst teilen wir die Daten sehr einfach auf, indem wir die ersten 80% der Daten zum Training nehmen und die restlichen 20% für den Test. Das wird mit der Strategie „Take from top“ realisiert.

Die Trainingsdaten stehen am oberen Output-Port des Knotens zur Verfügung, die Testdaten am unteren.



Nun sollten wir als Ergebnis zwei Tabellen bekommen:

► 1: First partition (as defined in dialog) ► 2: Second partition (remaining rows) Flow Variables

Rows: 79 | Columns: 6 Table Statistics

#	RowID	Maschine String	Betriebsstunden seit let... Number (integer)	Betriebstemperatur des ... Number (integer)	Produktionsauslastung [... Number (integer)	Vibrationen [Hz] Number (integer)	Wie lange bis zur nächst... String
1	Row0	DMG Mori NLX 2500	1040	63	78	256	Heute
2	Row1	Haas ST-20	1184	50	68	177	Heute
3	Row2	EMCO MaxxTurn 65	688	51	64	175	In einem Monat

Das hier ist der Trainingssatz.

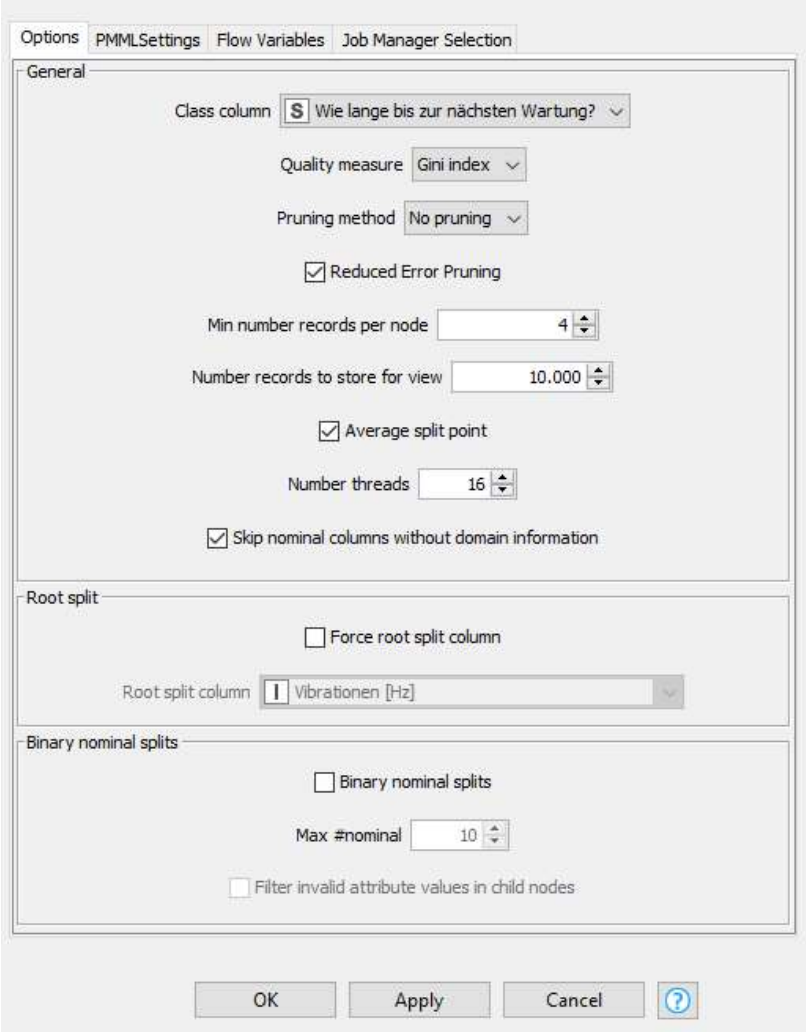
► 1: First partition (as defined in dialog) ► 2: Second partition (remaining rows) Flow Variables

Rows: 20 | Columns: 6 Table Statistics

#	RowID	Maschine String	Betriebsstunden seit let... Number (integer)	Betriebstemperatur des ... Number (integer)	Produktionsauslastung [... Number (integer)	Vibrationen [Hz] Number (integer)	Wie lange bis zur nächst... String
1	Row79	Haas ST-20	160	54	71	114	In einem Monat
2	Row80	EMCO MaxxTurn 65	400	50	86	284	In einer Woche
3	Row81	DMG Mori NLX 2500	512	46	65	61	In einer Woche

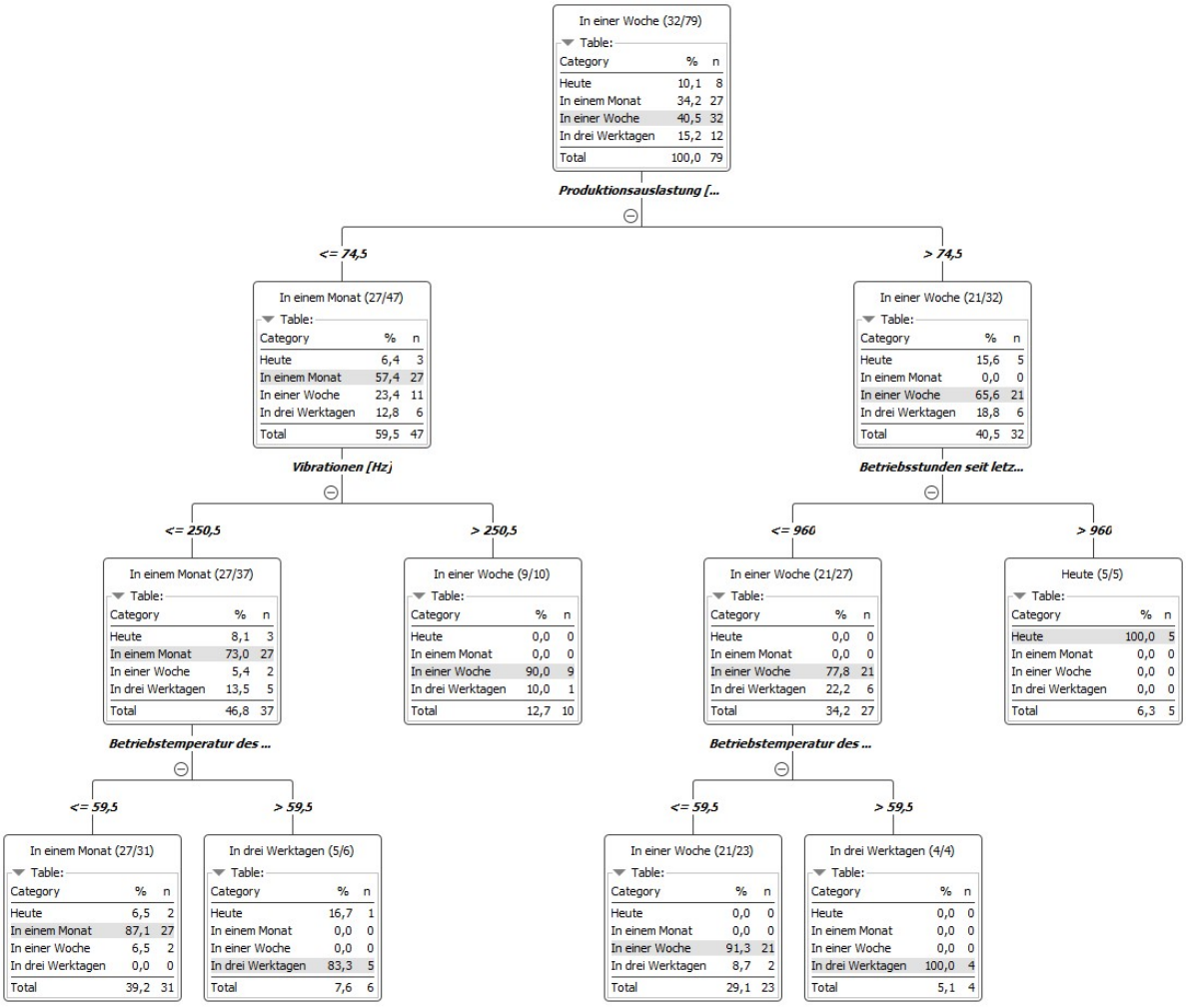
Mit diesem Testsatz überprüfen wir später unseren Entscheidungsbaum.

Im nächsten Schritt werden wir nun beginnen, den Entscheidungsbaum zu trainieren. Verbinden Sie dafür die Trainingsdaten mit dem „Decision Tree Learner“-Knoten. Hier können Sie mit der Einstellung „Min number records per node“ experimentieren. Damit wird die Mindestanzahl von Datensätzen festgelegt, die in einem Endkonten des Entscheidungsbaum vorhanden sein müssen. Je nach eingestelltem Wert verändert sich die Tiefe des trainierten Entscheidungsbaum.



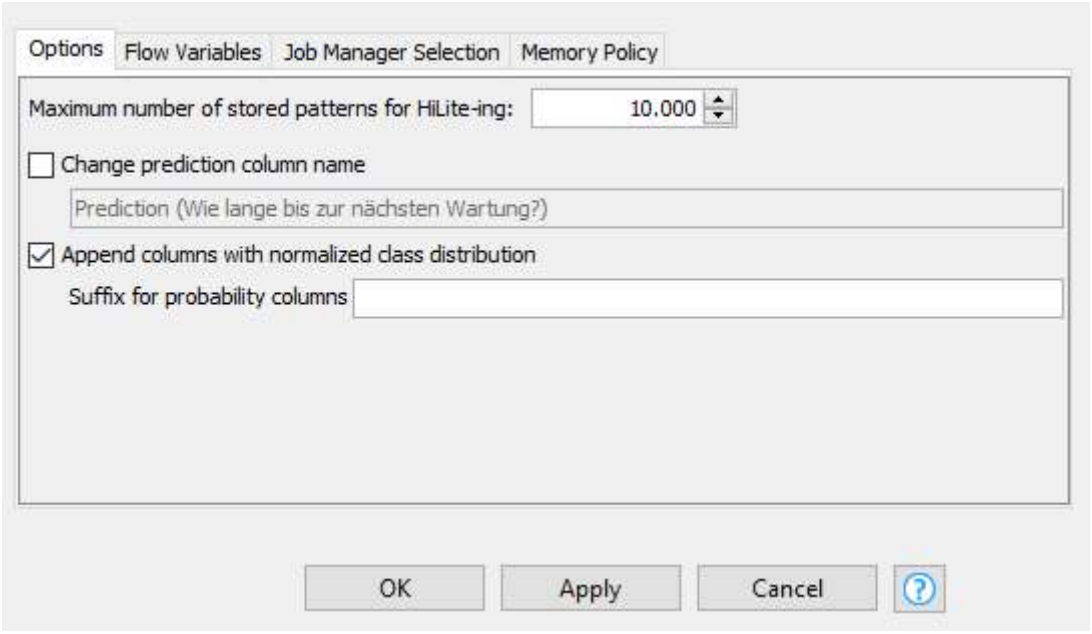
Betrachten können Sie den Entscheidungsbaum, indem Sie die kleine Lupe oberhalb des Knotens öffnen. Um den Entscheidungsbaum schnell zu entfalten, wählen Sie den obersten Knoten an und wählen in dem Dropdown-Menu links oben „Expand selected branch“.

Für die oben gezeigten Einstellungen ergibt sich so bspw. folgender Entscheidungsbaum:



Versuchen Sie nun einmal diesen Baum in eigenen Worten zu erklären.

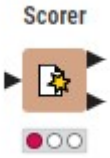
Jetzt müssen wir selbstverständlich noch die Testdaten zum Einsatz bringen, damit wir den Entscheidungsbaum bewerten können. Dafür verwenden wir den „Decision Tree Predictor“-Knoten. Stellen Sie ihn bitte folgendermaßen ein:



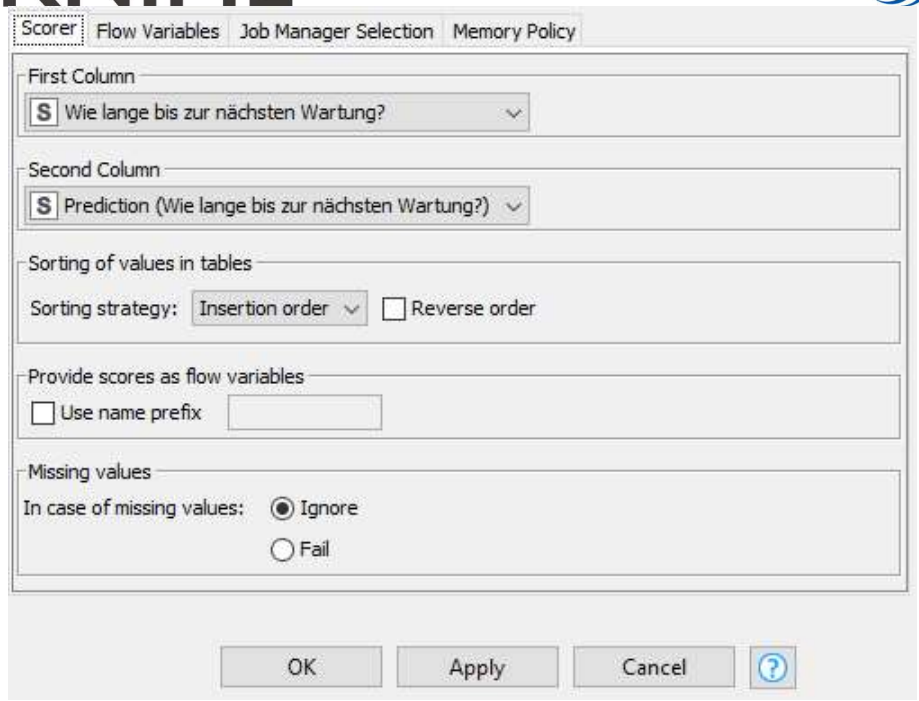
Die Tabelle sollte nun folgendermaßen aussehen:

Rows: 20 | Columns: 11 — Table | Statistics

#	RowID	Maschine (String)	Betriebss... (Number (inte...)	Betriebst... (Number (inte...)	Produktio... (Number (inte...)	Vibration... (Number (inte...)	Wie lange... (String)	P (Wie lan... (Number (dou...)	P (Wie lan... (Number (dou...)	P (Wie lan... (Number (dou...)	P (Wie lan... (Number (dou...)	Predictio... (String)
1	Row79	Haas ST-20	160	54	71	114	In einem Monat	0.065	0.871	0.065	0	In einem Monat
2	Row80	EMCO MaxxT...	400	50	86	284	In einer Woche	0	0	0.913	0.087	In einer Woche
3	Row81	DMG Mori NL...	512	46	65	61	In einer Woche	0.065	0.871	0.065	0	In einem Monat
4	Row82	Haas ST-20	544	58	68	209	In einem Monat	0.065	0.871	0.065	0	In einem Monat
5	Row83	EMCO MaxxT...	256	42	77	94	In einer Woche	0	0	0.913	0.087	In einer Woche



Wie gut unser Modell nun ist, können wir mit Hilfe des „Scorer“-Knoten ermitteln. Dort finden wir die Konfusionsmatrix, sowie eine Angabe der Genauigkeit und der Fehlerrate.



Diese Matrix gibt Aufschluss darüber, wie gut die Vorhersagen oder Klassifikationen mit den tatsächlichen Daten übereinstimmen. Der Accuracy-Wert liefert eine Messung der Übereinstimmung zwischen den beobachteten und erwarteten Klassifikationen. Das sollte nun folgendermaßen aussehen:

File Hilite

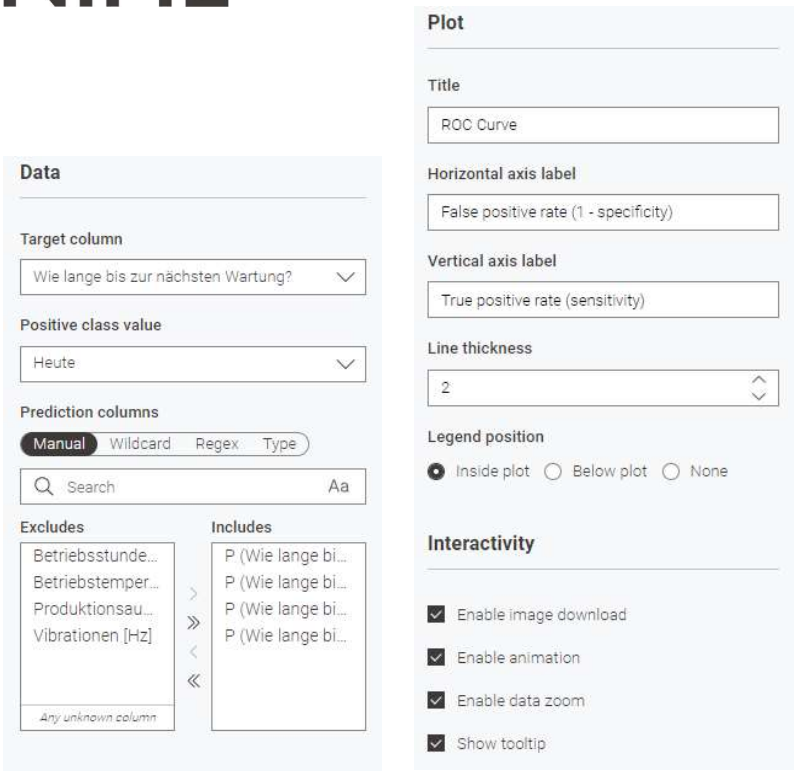
Wie lange ...	In einem M...	In einer W...	In drei We...	Heute
In einem Mo...	4	0	0	0
In einer Wo...	1	9	0	0
In drei Werk...	0	0	3	0
Heute	0	2	1	0

Correct classified: 16 Wrong classified: 4

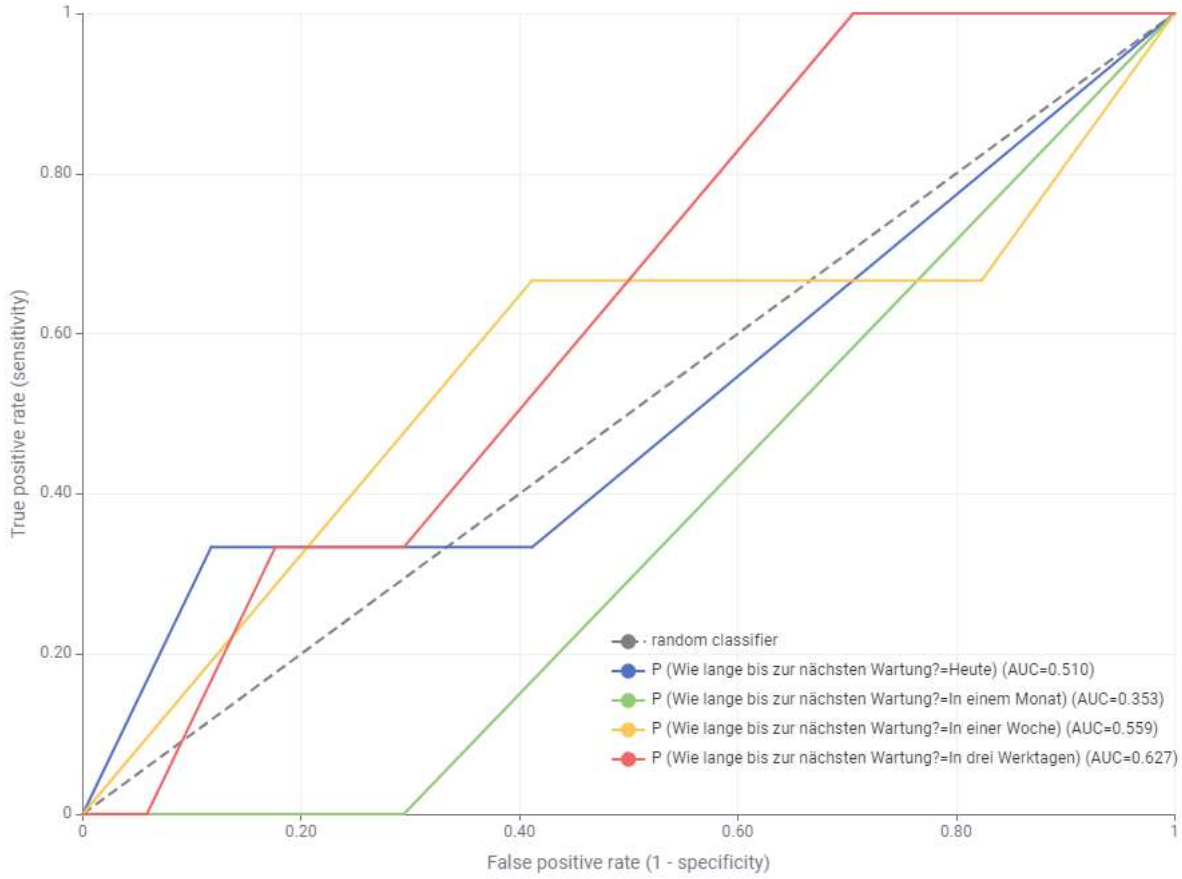
Accuracy: 80% Error: 20%

Cohen's kappa (κ): 0,69%

Darüber hinaus wollen wir uns noch mit dem „ROC Curve“-Knoten befassen. ROC steht für „Receiver Operation Characteristic”. Die generierte Kurve zeigt das Verhältnis zwischen der Sensitivität (True Positive Rate) und der Spezifität (True Negative Rate) bei unterschiedlichen Schwellenwerten für die Klassifikationsentscheidung.



Die erhaltene Kurve sollte nun so aussehen:



In dem Diagramm sind folgende Informationen zu sehen:

- **True Positive Rate (Sensitivität) (Y-Achse):** Die Sensitivität misst, wie gut das Modell in der Lage ist, tatsächlich zu erkennen, welcher Wartungszeitpunkt nun der richtige ist. Es ist das Verhältnis der richtig identifizierten Fälle zur Gesamtanzahl der tatsächlichen Fälle. Je höher die Sensitivität, desto besser ist das Modell darin, die richtigen Wartungszeitpunkte passend zum gesuchten Wert zu finden.
- **False Positive Rate (1 - Spezifität) (X-Achse):** Diese misst, wie viele Fälle falsch eingestuft wurden. Es ist das Verhältnis der falsch identifizierten Fälle zur Gesamtanzahl der tatsächlichen Fälle. Je niedriger die False Positive Rate, desto besser ist das Modell darin, Wartungszeitpunkte richtig zu klassifizieren.

Die graue Kurve zeigt einen Klassifikator an, der reinzufällig Klassen zuweist. In der Praxis wird die ROC-Kurve jedoch oft als Kurve dargestellt. Sie bewertet die Leistung eines Modells, indem Sie den Bereich unter der ROC-Kurve (AUC, hier in der Klammer zu finden) betrachtet. Umso größer die Fläche unter der Kurve ist, desto besser ist die Klassifizierung. Ein AUC von 1 würde auf ein perfektes Modell hinweisen.

In Ihrem Studium haben Sie gelernt, sich nicht mit den „ersten besten“ Ergebnissen von maschinellen Lernverfahren zufriedenzugeben, sondern diese kritisch zu bewerten und auf Verbesserungspotenzial zu überprüfen. Die Accuracy von 80%, die Ihnen der Bewertungsknoten „Numeric Scorer“ für den Test Ihres Entscheidungsbaums anhand der Testdaten ausweist, stellen Sie noch nicht zufrieden. Die Möglichkeiten, die Einstellungen des Decision Tree Lerner-Knotens zu variieren, haben Sie ausgeschöpft. Sie haben aber noch eine weitere Möglichkeit, Ihr Ergebnis zu verbessern:

Erinnern Sie sich an das Verfahren, mit dem Sie die vorhandenen Daten in Trainings- und Testdaten aufgeteilt haben. War das simple Aufteilen dieser Daten der Reihenfolge nach wirklich eine gute Wahl? Was, wenn diese Daten nach dem Ergebnis der Klassifikation sortiert waren? Dann hätte der Lerner-Knoten keine repräsentative Auswahl der vorhandenen Daten erhalten, und entsprechend verfälscht wäre sein Ergebnis!

Eine klügere Strategie zum Aufteilen der Daten ist das „strategic sampling“: mit dieser Einstellung teilt der Partitionierungs-Knoten die Daten im Verhältnis des Auftretens der verschiedenen Klassifikationsergebnisse auf – für das Lernen einer Klassifikation ein weitaus besserer Ansatz!

Gehen Sie zurück zum Partitioning-Knoten und ändern Sie dort die Strategie. Führen Sie die nachfolgenden Knoten alle erneut aus (deren Einstellungen brauchen Sie nicht zu verändern) und vergleichen Sie die Ergebnisse – sowohl den Entscheidungsbaum, als auch die Kennzahlen.