1. Projektbeschreibung „Uni-Abbrecher“

1.1 Einleitung und Ist-Analyse

Universität Giggle Tech Institute

Die Studentenverwaltung & Psychologische Berater des GTI möchten vorbeugend u.a. Kampagnen durchführen, um das abbrechen der Universität unter den Studierenden zu minimieren.

Dafür müssen die Profile der Studenten analysiert werden, um Schlüsse daraus zu ziehen, welche Gegenmaßnahmen bei zukünftigen Studenten fruchten könnten, damit Sie nicht abbrechen.

Hierfür haben Wir einen Datensatz von einer Hochschuleinrichtung bekommen, Ende 2021 das letzte Mal bearbeitet, der aus verschiedenen unabhängigen Quellen zusammengestellt und uns zur Verfügung gestellt wurde.

1.2 Zusammensetzung der Daten

Spaltenname(Original), Übersetzung & mögliche Hypothese jeweils:

Marital status: Familienstand (kategorisch)

Frage: Sind Singles evtl. eher Abbrecher?

Application mode: Bewerbungsmodus: Verwendete Bewerbungsmethode. (kategorisch)

Frage: Wenn Studenten sich persönlich vorgestellt haben, blieben Sie dann eher?

Application order: Die wievielte Einschreibung an einer Universität wurde angenommen. (Numerisch)

Frage: Wenn der Student nicht seine liebste Uni bekommt, ist er schneller ein Abbrecher?

Course: Kurs (kategorisch)

Frage: Gibt es bestimmte Kurse, die zum abbrechen animieren?

Daytime/evening attendance: Student hat tagsüber/ abends Unterricht. (kategorisch)

Frage: Abends Studierende, brechen Sie eher ab?

Previous qualification: Abschluss vor der Einschreibung. (kategorisch)

Frage: Bleibt man eher an der Uni mit besseren Abschluss?

Nationality: Die Nationalität des Studenten. (kategorisch)

Frage: Hat die Abbrechrate was mit der Herkunft zutun?

Mother's qualification: Die Qualifikation der Mutter des Schülers. (kategorisch)
Father's qualification: Die Qualifikation des Vaters des Schülers. (kategorisch)

Fragen jeweils: Hat der familiäre Hintergrund etwas mit Erfolg zutun?

Displaced: Gibt an, ob der Student aus einer andren Stadt kommt. (kategorisch)

Frage: Gibt es eine Verbindung der Abbruchrate zu der Entfernung zum Wohnort

Educational special needs: Hat der Student besondere pädagogische Bedürfnisse (kategorisch)

Frage: Durch mehr Betreuung evtl. weniger Abbrüche?

Debtor: Schuldner: Gibt an, ob der Student einen Kredit aufgenommen hat. (kategorisch)

Frage: Brechen verschuldete Studenten weniger ab?

Tuition fees up to date: Ob die Studiengebühren aktuell bezahlt sind. (kategorisch)

Frage: Sieht man am „bezahlt“ Status, ob Studenten schneller abbrechen?

Gender: Das Geschlecht des Schülers. (kategorisch)

Frage: Sagt das Geschlecht etwas über den Erfolg an der Uni aus?

Scholarship holder: Gibt an, ob der Student ein Stipendium besitzt. (kategorisch)

Frage: Bricht man ohne Stipendium eher ab?

Age at enrollment: Alter des Studierenden zum Zeitpunkt der Einschreibung. (Numerisch)

Frage: Brechen jüngere Studenten eher ab?

International: Gibt an, ob der Student aus dem Ausland kommt. (kategorisch)

Frage: Brechen Studenten von ganz weit weg weniger ab?

SPALTEN mit jeweils Infos zum 1. und 2. Semester

Curricular units (credited): Die Anzahl erhaltener Credis für bestandene Kurse des Semesters

Curricular units (enrolled): Die Gesamtzahl der eingeschriebenen Kurse in jedem Semester.

Curricular units (evaluations): Die Anzahl der bewerteten Kurse pro Semester.

Curricular units (approved): Anzahl der Kurse, die in jedem Semester mit "bestanden" bewertet wurden

Curricular units (grade): Die Durchschnittsnote für jedes Semester.

Curricular units (without evaluations): Anzahl der Kurse ohne Bewertung pro Semester.

Frage: Hat der Verlauf der Kurse und der Noten etwas mit dem Abbruch der Uni zutun?

Spalten mit Infos zum Heimatort

Unemployment rate: Arbeitslosenquote im Heimatort

Frage: Sorgt höhere Arbeitslosigkeit zuhause dafür, dass man lieber nicht abbricht?

Inflation rate: Inflationsrate im Heimatort

Frage: Sorgt eine hohe Inflationsrate zuhause dafür, lieber nicht abzubrechen?

GDP: Bruttoinlandsprodukt im Heimatort

Frage: Kann das BIP einen Einfluss auf das Abbrechen der Uni haben?

Spalten die wir von Anfang an nicht verwendet haben

Mother's occupation: Der Beruf der Mutter des Schülers. (kategorisch)
Father's occupation: Der Beruf des Vaters des Schülers. (kategorisch)

Grund: Scheinbar passten die Daten nicht zu der Beschreibung des Datensatzes und waren nicht logisch kategorisierbar.

Und schließlich unser angepeiltes Ziel

Target: Hat der Student sein Studium abgebrochen, oder später bestanden

1.3 Bewertung der Datenqualität

Die Datenqualität ist grundsätzlich super! Es gibt keine fehlenden Werte, nur wenige Ausreißer die man eigentlich missachten könnte, und es gab nur beim lesen des Datensatzes kleine Unstimmigkeiten, wenn man Ihn mit Excel aufgemacht hat. Allerdings keine Probleme bei KNIME und POWER BI.

Für das KNIME ML Model haben wir in der „Application order" Spalte 4 Aussreißer entfernt.

Für Power BI haben wir aus der gleichen Spalte nur einen Wert mit 0 auf 1 gesetzt.

2. Backlog und Projektziele

2.1 Sammlung von User Storys und Potenzialen

Nr.1 Als Mitglied der Studentenverwaltung möchte ich eine Übersicht über die aktuellen Studienabbruchraten erhalten, um die Dringlichkeit des Problems zu verstehen.
Akzeptanzkriterien:

Ein Dashboard zeigt die aktuellen Studienabbruchraten nach verschiedenen Kriterien (z. B. Kurs, Bewerbungsmodus) an

Nr.2 Als Psychologische Beratung der Studenten möchte ich frühzeitig Informationen über gefährdete Studenten erhalten, um intervenieren zu können.
Akzeptanzkriterien:

Das ML-Modell identifiziert Studierende mit erhöhtem Abbruchrisiko
Ein Dashboard zeigt eine Liste gefährdeter Studierender und ihre relevanten Merkmale an

Nr.3 Als Verantwortlicher für innerschulische Kampagnen möchte ich die wichtigsten Einflussfaktoren auf Studienabbrüche verstehen, um gezielte Maßnahmen zu planen.
Akzeptanzkriterien:

Das Dashboard zeigt die wichtigsten Merkmale und ihre Auswirkungen auf die Studienabbruchvorhersage an. Dabei brauchen wir grafische Darstellungen und statistische Informationen um bestehende Beziehungen zu verdeutlichen.

2.2 Priorisierung

Folgende Spalten haben wir bevorzugt behandelt:

- Alter
- Familienstand
- Kurse
- Qualifikationen der Mutter
- Qualifikation des Vaters

2.3 Auswahl der erreichbaren Projektziele für den ersten Sprint

Erster Sprint(4Std):

-Aufbereitung des Datensatzes und Weitergabe an PowerBI

-Dokumentation bis Punkt 2

-In Power BI explorative Datenanalyse

3. Projektausarbeitung

3.1 Zuordnung von Hypothesen zu User Storys und Potenzialen

Ziel: Personalisierte Beratung und Unterstützung

Funktionalität: Frühwarnsystem für gefährdete Studierende, Evaluierung von Interventionsmaßnahmen

Hypothese: Studierende mit regelmäßiger Teilnahme an speziellen Unterstützungsworkshops haben eine höhere Wahrscheinlichkeit, das Studium erfolgreich abzuschließen.

Hypothese: Studierende, die in ihrer Bewerbung "besondere pädagogische Bedürfnisse" angegeben haben, könnten von gezielter Unterstützung und Anpassungen im Studium profitieren.

Hypothese: Studierende, die von auswärts zugezogen sind, könnten Schwierigkeiten haben, sich an die neue Umgebung anzupassen. Gezielte Beratung könnte ihnen helfen, erfolgreich zu bleiben.

4. Projektmetriken

4.1 Messung des Fortschritts und Erfolgs

Für das Model in KNIME

Genauigkeit (Accuracy): Der Prozentsatz der korrekt vorhergesagten Studienausgänge im Vergleich zu den tatsächlichen Ergebnissen.

Precision und Recall: Diese Metriken zeigen das Verhältnis von richtig positiven Vorhersagen zu allen positiven Fällen (Precision) sowie das Verhältnis von richtig positiven Vorhersagen zu allen tatsächlichen positiven Fällen (Recall). Es wird ausbalanciert, da es mehr Studenten im Datensatz gibt, die bestanden haben.

F-Measure: Die F-Maßnahme ist eine Metrik, die in der Evaluierung von Klassifikationsmodellen verwendet wird, insbesondere wenn ein Ausgleich zwischen Präzision (Precision) und Rückruf (Recall) erforderlich ist. Die F-Maßnahme kombiniert Precision und Recall in einer einzigen Zahl und hilft dabei, die Leistung eines Klassifikators in Bezug auf wahre positive, falsche positive und falsche negative Ergebnisse zu bewerten.

Für das Dashboard in Power BI

Anzahl der personalisierten Unterstützungen: Wenn das Dashboard genutzt wird, um gefährdete Studierende zu identifizieren, verfolgen Sie die Anzahl der angebotenen personalisierten Unterstützungsmaßnahmen.

Veränderung der Abbruchraten: Verfolgen Sie, ob die Abbruchraten nach der Einführung von Interventionen und Unterstützungsmethoden sinken.

Für die Dokumentation

Klarheit und Verständlichkeit: Mitglieder der Studentenverwaltung dürfen gerne ab und an die Dokumentation gegenlesen und Feedback geben, ob alles klar und verständlich ist.

Vollständigkeit & Benutzerfreundlichkeit: Wir sollten immer mal wieder darauf achten, dass alle relevanten Aspekte des Projekts in der Dokumentation abgedeckt sind bzw. es sollte auch mal geschaut werden, ob alles schön strukturiert ist.

Umso einfacher ist es für andere und uns selber zu verstehen, was gemacht wurde.

4.2 Definitions of Done

1. In KNIME sollte der Datensatz aufbereitet werden.

2. Danach kann Power BI den Datensatz laden und explorative Analyse wird durchgeführt.

3. KNIME ML Model zur Vorhersage fertig stellen

4. Power BI Dashboard für das Studentenprofil fertigstellen

5. Dokumentation fertig stellen

6. Power Point Präsentation

5. Planung des ersten Sprints

5.1 Aufgabenpakete und Aufwandsschätzung

Erster Sprint(4Std):

-Aufbereitung des Datensatzes und Weitergabe an PowerBI. Kleiner Aufwand, 2 Std

-Dokumentation bis Punkt 2. Mittlerer Aufwand, 4 Std

-In Power BI explorative Datenanalyse Mittlerer Aufwand, 4 Std

6. Technische Beschreibung des ersten Sprints anhand des Machine Learning Canvas

1. Datenimport

Der Datensatz wurde als Excel-CSV übergeben und wir haben Ihn mit dem CSV Reader Knoten eingelesen

2. Dataillierte Analyse von Dateninhalten und Statistiken

Hier wurden die statistica Knoten benutzt, zusätzlich Linear Correlation und Data Explorer Knoten um ein allgemeines Verständnis für den Datensatz zu bekommen.

3. Aufbereitung der Daten für das Modell

Unter Visualisierungen findet man verschiedene Plot-Knoten(Box Plot, Scatter Plot& Histogram) die Ausreißer erkennen lassen bzw. kategorische Daten analysiert und darstellt

4. ML Modell

Hier haben wir mit 3 verschiedenen Methoden, 3 verschiedene Modelle trainiert und ausgewertet.

Modelle waren Random Forest, logistisches Regressionsmodell(ja, für Klassifizierungsproblem), und das XGBoost-Klassifikator-Modell.

Methoden waren Bayes´sche Methode, Zufalls-Such-Methode, eine weitere optionale Methode.

7. Ergebnisse der EDA und der Hypothesenüberprüfung

Thema Alter: Wir hätten Anfangs gedacht, dass eher die älteren Leute Ihr Studium erfolgreich beenden. (Mehr Durchhaltevermögen und Erfahrung)

ABER: Es war genau anders herum. Die 17-22 Jährigen schaffen viel häufiger das Studium, als alle Altersgruppen darüber. Wenn man „bestanden" & „abgebrochen" in einem Säulendiagram nach Altersgruppen anzeigen lässt.

ALLERDINGS: Schaut man sich die Abbrecher alleine an nach Altersgruppen, stechen die jungen hervor, da es einfach eine Große Masse von Ihnen gibt. Die Sichtweise bestimmt das Ergebnis oder die Erkenntnis daraus.

Thema Familienstand: Wir hätten gedacht, das evtl. verheiratete eher ein Studium bestehen, da Unterstützung besteht.

ABER: Scheinbar kommen die ledigen durchschnittlich besser durch. Wenn man wieder abgebrochen und bestanden zusammen darstellt.

Die Masse liegt aber bei den ledigen Studenten und war sogar bei den weiblichen Studenten noch höher (evtl. Schwangerschaften, Familie gründen).

Thema Kurse: Es lag nahe, dass auch gewisse Kurse dafür verantwortlich sein könnten, dass Menschen sich entschließen, das Studium abzubrechen.

Kurz und knapp können wir sehen, dass wir Recht hatten und Kurs 7 und Kurs 1 die Abbruchrate Durchschnittlich stark erhöhten in der Vergangenheit.

8. Bedienungsanleitungen und Ergänzungen zur Dokumentation

Da wir in **KNIME** unser komplettes Modell ausführlich beschrieben haben, schauen Sie sich gerne das komplette Canvas an.

Sie finden zu jedem Knoten, jeder Komponente bzw. Methanode und zu jedem Abschnitt Erklärungen auch in den gelben Kästen, was genau in welchem Schritt gemacht wird und was wir uns dabei gedacht haben.

Zusätzlich werden über den Komponenten Quick-Info eingeblendet.

Das **Power BI Dashboard** ist sehr übersichtlich, zu erwähnen sind aber 3 wichtige Buttons die in der Browser-Version sich mit Linker Maustaste (bei Desktop zusätzlich mit STRG) aktivieren lassen:

„Zurück-Button“ ist ganz links oben in der Ecke und sorgt dafür, dass alle Filter zurück gesetzt werden.

„Mortarboard-Button“ sind die Beiden Button rechts oben mit dem schwarzen Hut. Der erste zeigt alle Studenten an, die bestanden haben. Der durchgestrichene Hut ist für die Anzeige der Studienabbrecher.