Konzeptvorschlag für SSBI-Projekt

Analyse und Prognose der Covid-19-Fallzahlen ausgewählter Länder anhand von Google-Mobilitätsdaten und weiteren Variablen

Andreas Herb & Sebastian Wölk

Inhaltsverzeichnis

Einleitung und Kurzvorstellung des Themas

Mobilität im Alltag wird oft als Grundbedürfnis von Menschen gesehen. Je mobiler ein Mensch ist, desto unabhängiger wird er von den Beschränkungen seines Lebensraums. In Folge der Corona-Pandemie wurde die Mobilität vieler Menschen in Europa ab der starken Ausbreitung im März 2020 drastisch eingeschränkt, um die Verbreitung des Virus einzudämmen oder gar zu stoppen. So werden von Google, zur Bekämpfung des Coronavirus, täglich aktualisiert und anonymisiert für nahezu alle Länder der Welt sogenannte Mobilitätsberichte bereitgestellt. Einfach gesagt stellen diese Berichte Mobilitätstrends von Menschen in Diagrammform dar, aufgeschlüsselt nach geografischen Regionen und Kategorien von Orten – beispielsweise Einzelhandel und Freizeit, Läden des täglichen Bedarfs, Parks, Bahnhöfe und Haltestellen, Arbeitsstätten und Wohnorte.

Im Rahmen der Veranstaltung “Self-Service-BI” soll in diesem Projekt, anhand der von Google frei verfügbaren Mobilitätsberichte, untersucht werden, inwiefern sich die Mobilität der Menschen in Europa verändert hat und welche Auswirkungen dies auf die Corona-Fallzahlen hatte. Ergänzt werden sollen diese Daten um länderspezifische Angaben, z.B. Feiertage, Schulferien und Wochenenden, um die Auswirkungen konkreter und differenzierter betrachten zu können. Darauf aufbauend wird versucht, mittels Machine-Learning-Methoden ein Regressions- oder Klassifikationsmodell zu erstellen, welches zur Prognose der Corona-Fallzahlen herangezogen werden könnte.

Technologien und Beschreibung des geplanten Tool-Stacks

Um die angedachte Projektidee umsetzen zu können, soll primär das von Microsoft bereitgestellte Self-Service-BI-Tool “Power BI” genutzt werden. Wie in der Abbildung unten zu sehen, werden damit vom Laden der Daten über die Aufbereitung bis hin zur Modellierung und der Visualisierung jegliche Arbeitsschritte durchgeführt. Um die angedachten Machine-Learning-Modelle aufstellen zu können, wird auf die von Microsoft angebotene “Azure Cloud” und dem dort integrierten Machine-Learning-Studio zurückgegriffen, um das finale Modell mit der besten Vorhersageleistung in Power-BI zu integrieren. Dabei liegt der Fokus zunächst auf den von Azure bereitgestellten Möglichkeiten des automatisierten maschinellen Lernens, ggf. soll der ebenfalls zur Verfügung stehende Azure Machine Learning-Designer für das Aufbauen der Modelle genutzt werden. Ziel dabei ist es, die über die Schnittstellen zwischen MS PowerBI und der Azure Cloud (Microsoft Ökosystem) bereitgestellten Modellierungs- und Machine-Learning-Funktionen mit “Low-Code”-Ansatz auf ihre Praxistauglichkeit hin zu verproben und im Rahmen eines praktischen Use-Cases zu testen.

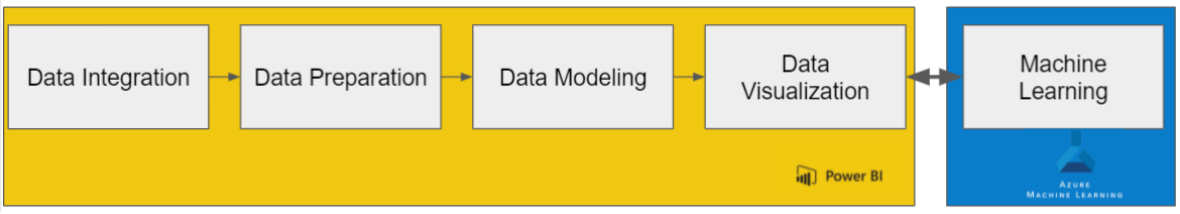

Abb.1: Geplante Technologien und deren Einordnung

Vorgehensweise und Datenmodell

Im Folgenden soll die geplante Vorgehensweise in Bezug auf die oben beschriebene Projektidee unter Einsatz der bereits vorgestellten Technologie-Bausteine kurz vorgestellt, die zur Verwendung geplanten Datensätze erläutert und deren Beziehungen ohne finalen Anspruch auf Vollständigkeit und Endgültigkeit erläutert und dargestellt werden.

Der diesem Projekt zugrunde liegende Hauptdatensatz soll der Mobilitätsbericht von Google (verfügbar unter: https://www.google.com/covid19/mobility/) sein. Dieser Datensatz enthält Informationen über die Veränderung der Bewegungsmuster von Bürgerinnen und Bürgern verschiedener Länder durch die Coronakrise. Der Datensatz ist grundsätzlich öffentlich zugänglich und dient primär der Nutzung durch die Gesundheitsbehörden zum Treffen fundierterer Entscheidungen zur Eindämmung des Coronavirus. Aufgezeichnet wird die Veränderung der Zahl der Besucher an definierten Ortskategorien (z.B. Geschäfte des täglichen Bedarfs, Parks, Arbeitsstätten etc.) im Vergleich zu einem Referenzzeitraum, der durch Google auf den Zeitraum 03. Januar 2020 bis 06. Februar 2020 , kurz vor der Corona-Pandemie, festgelegt wurde. Hierbei wird über einen Zeitraum von 5 Wochen jeweils der Medianwert (Zahl der Besucher einer bestimmten Ortskategorie) für jeden Wochentag als Referenztag festgehalten.

Angereichert werden - durch Zusammenführung oder Join - soll dieser Datensatz durch Corona-Fallzahlen (Anzahl Neuinfektionen, Inzidenzwerte, Verlauf aktiver Infektionen), die ebenfalls an verschiedenen Stellen im Web verfügbar sind.

Geplant ist darüber hinaus die Einrichtung von vier weiteren “Dimensionstabellen”, die sich im Datenmodell um diesen Hauptdatensatz zzgl. der Corona-Fallzahlen ansiedeln. Hierbei soll eine Dimension “Zeit”, die Informationen zu Datum, Wochentag, Wochenende und Jahreszeit enthält, geschaffen werden. Diese in der Dimension “Zeit” enthaltenen Variablen können ggf. Einflüsse auf die Frequentierung bestimmter Ortskategorien des Mobilitätsdatensatzes von Google beschreiben und den verstärkten Einfluss auf die Bewegungsmuster der Bevölkerung an bestimmten Tagen erklären. Außerdem soll eine Dimension “Feiertage” und eine Dimension “Schulferien” mit dem Hauptdatensatz in Beziehung gesetzt werden, da Schulferien und Feiertage vermutlich ebenso einen stärkeren Einfluss auf die Bewegungsmuster einer Bevölkerung haben können. Als vierte Dimension wird an die Dimension “Länder” gedacht, worin eine Zuordnung der europäischen Länder zu Teilkontinenten (Nord-, Ost-, Süd- und Westeuropa) anhand des Ländernamens und/oder des Länder-ISO-Codes zur aggregierten Analyse der vorliegenden Daten vorgenommen werden soll.

Unter Einsatz der bereits vorgestellten Technologie-Bausteine sollen die Datensätze anschließend zusammengeführt bzw. gejoint werden und ein entsprechendes Datenmodell für die weitere Analyse in MS PowerBI erstellt werden. Anschließend werden die in ein Datenmodell überführten Tabellen im Rahmen der Data Preparation bereinigt (Löschen/Hinzufügen relevanter Spalten, ggf. Kategorisierung von Variablen, Formatierung von Datentypen, Erstellen selbst berechneter Spalten und/oder Metriken etc.) und für die anschließende Visualisierung und Modellierung vorbereitet werden.

Im Anschluss an eine kurze Data Exploration, um das Datenmodell und dessen Inhalt ein wenig kennenzulernen, ist geplant, ein statistisches Modell (z.B. Regressionsmodell oder Klassifikationsmodell) zur Vorhersage der Corona-Fallzahlen in Abhängigkeit zu den Mobilitätsmustern verschiedener Gesellschaften innerhalb der MS Azure Cloud zu erstellen - also zu trainieren und zu testen - und das endgültige Modell mit der besten Performance

anschließend in eine Dashboard-Visualisierung einzubinden. Das Deployment erfolgt entsprechend wieder in MS PowerBI.

Ein denkbares Datenmodell - allerdings noch ohne Anspruch auf Vollständigkeit und Endgültigkeit - könnte in etwa wie folgt aussehen:

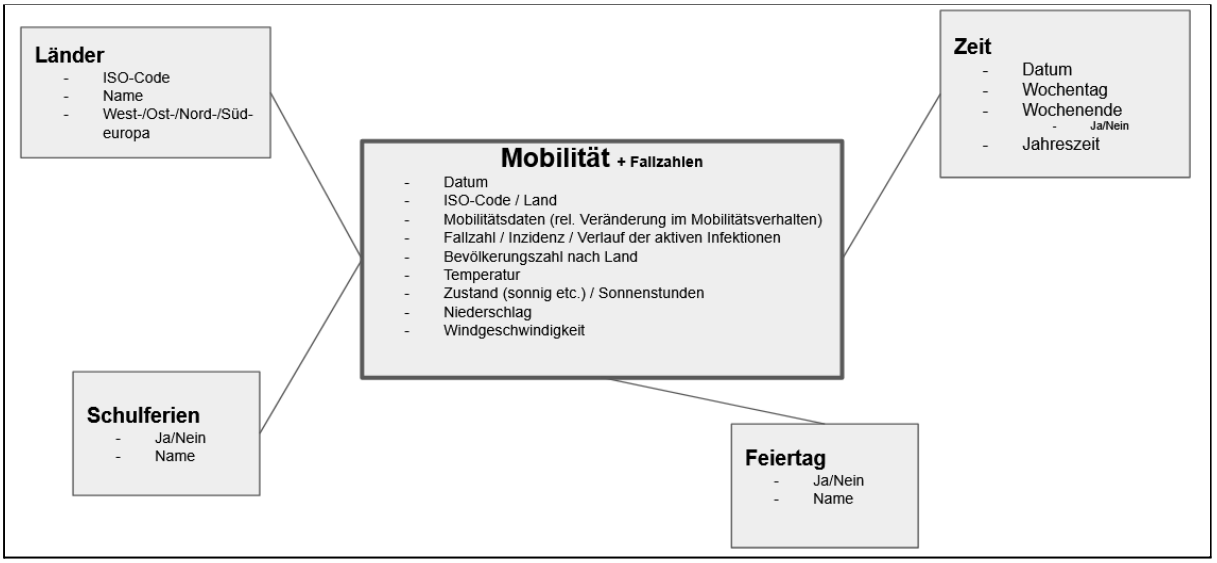

Abb.2: Unter Vorbehalt geplantes Datenmodell

Fazit

Die in den vorangegangenen Abschnitten vorgestellte Projektidee, unter Einbezug der erläuterten Technologie- und Tool-Bausteine und der geplanten Vorgehensweise inkl. Datenmodell, soll schlussendlich zur Beantwortung einiger analytischer Fragestellungen in Bezug zu dem erläuterten Hauptdatensatz von Google verwendet werden.

Zusammenfassend steht demnach im Fokus der Projektarbeit die Frage, inwieweit sich die Corona-Fallzahlen in Abhängigkeit zur Frequentierung (Anzahl Besucher) bestimmter Ortskategorien vorhersagen lassen.

Darüber hinaus soll, im Rahmen der visuellen Analyse von Zusammenhängen im Rahmen einer kurzen Data Exploration-Phase, die grundsätzliche Frage beantwortet werden, ob signifikante Zusammenhänge zwischen den Bewegungsmustern einer Gesellschaft und dem Verlauf der Corona-Fallzahlen erkannt werden können.

Die zentrale Fragestellung, die diese Projektarbeit im analytischen Kontext zu beantworten versucht, ist, ob sich die Bewegungsmuster ausgewählter Gesellschaften signifikant auf den Corona-Fallzahlen-Verlauf auswirken und ob wesentliche Unterschiede in den Bewegungsmustern der Länder, die die Corona-Pandemie bisher relativ erfolgreich bewältigt haben und jenen Ländern, die relativ stark von den Einflüssen der Corona-Pandemie betroffen sind oder waren, erkennbar sind.