FREIE UNIVERSITÄT BERLIN

FACHBEREICH WIRTSCHAFTSWISSENSCHAFT

Abschlussbericht
Business Intelligence Semester Projekt

des Fachbereichs Wirtschaftswissenschaft
der Freien Universität Berlin

West Nile Virus Prediction in Chicago

Berlin, 16.08.2023

1 Project Understanding

Die Ausbreitung von Infektionskrankheiten durch Vektoren, insbesondere von durch Mücken übertragenen Krankheitserregern, stellt eine erhebliche Gefahr für die öffentliche Gesundheit dar. Ein besonders bedrohlicher Erreger in diesem Kontext ist das West-Nil-Virus (WNV), das durch infizierte Stechmücken auf den Menschen übertragen wird. Die Viruserkrankung kann von asymptomatischen Fällen bis hin zu schweren neurologischen Erkrankungen reichen und in einigen Fällen sogar tödlich verlaufen. Die hohe Relevanz des Themas wurde deutlich, als im Jahr 2002 die ersten menschlichen Fälle des West-Nil-Virus in Chicago gemeldet wurden. Infolgedessen etablierte die Stadt Chicago bis 2004 ein umfassendes Überwachungs- und Bekämpfungsprogramm, das bis heute genutzt wird.

Dieses Projekt verfolgt das Ziel, mithilfe des CRISP-DM (Cross Industry Standard Process for Data Mining) - Vorgehens Vorhersagen über die Verbreitung des West-Nil-Virus in Mückenpopulationen zu treffen. Dabei werden verschiedene Datenquellen, welche durch die Plattform Kaggle zur Verfügung standen, einschließlich Wetterdaten, Standortinformationen, Testergebnisse und Sprühdaten, genutzt, um ein Vorhersagemodell zu entwickeln. Dieses Modell soll dabei unterstützen, die Verbreitung des West-Nil-Virus frühzeitig zu erkennen und geeignete Maßnahmen zur Prävention und Bekämpfung zu ergreifen.

Die Hypothese, dass bestimmte klimatische Bedingungen die Verbreitung des Virus begünstigen könnten, bildet die Grundlage für diese Untersuchung. Hierbei wird die Annahme getroffen, dass warmes und trockenes Wetter die Wahrscheinlichkeit für positive Testergebnisse bei verschiedenen Moskitoarten erhöht. Durch die präzise Vorhersage von infizierten Moskito-Populationen kann die Stadt Chicago Ressourcen gezielter und effizienter einsetzen, um die Übertragung des West-Nil-Virus auf Mensch und Tier zu minimieren. Dieses Projekt verdeutlicht somit den Wert datengetriebener Ansätze im Gesundheitssektor.

2 Data Understanding

Um die Herausforderung der Vorhersage des West-Nil-Virus in Mückenpopulationen effektiv anzugehen, ist ein umfassendes Verständnis der bereitgestellten Datensätze von großer Bedeutung. Im Mittelpunkt steht die Analyse von Wetterdaten, Daten aus geografischen Informationssystemen und historischen WNV-Testergebnissen mit dem Ziel, vorherzusagen,

ob WNV zu einem bestimmten Zeitpunkt, an einem bestimmten Ort und bei einer bestimmten Mückenart auftreten wird.

Der Hauptdatensatz umfasst die WNV-Testergebnisse. Jedes Jahr werden in der Zeit von Ende Mai bis Anfang Oktober in ganz Chicago strategisch platzierte Mückenfallen aufgestellt. Diese Fallen, die von Montag bis Mittwoch aktiv sind, sammeln Mücken ein, die dann am Ende der Woche auf das Virus getestet werden. Jeder Datensatz enthält die Anzahl der Stechmücken, die Art der Stechmücken und die Angabe, ob in der Kohorte WNV nachgewiesen wurde oder nicht. Wenn die Anzahl der gesammelten Mücken 50 übersteigt, werden die Daten in zusätzliche Datensätze aufgeteilt, wobei eine Obergrenze von 50 Mücken pro Datensatz eingehalten wird.

Die Fallen werden anhand der Blocknummer und des Straßennamens identifiziert, wobei diese Attribute für die Analyse in GPS-Koordinaten umgewandelt werden. Bei einigen Fallen handelt es sich um "Satellitenfallen", die zur besseren Überwachung in der Nähe von bereits bestehenden Fallen aufgestellt wurden. Diese Satellitenfallen sind mit Buchstabenpostfixen gekennzeichnet. Es ist wichtig zu wissen, dass nicht alle Standorte einheitlich getestet werden und Aufzeichnungen nur existieren, wenn eine bestimmte Mückenart in einer bestimmten Falle und zu einem bestimmten Zeitpunkt identifiziert wurde. Der Testzeitraum erstreckt sich über mehrere Jahre, wobei Trainingsdaten aus den Jahren 2007, 2009, 2011 und 2013 verfügbar sind. Die Herausforderung erfordert Vorhersagen für die Jahre 2008, 2010, 2012 und 2014.

Kommentiert [1]: was genau bedeutet das?

Es werden auch Sprühdaten “Spray” bereitgestellt, die die Mückenbekämpfungsmaßnahmen der Stadt Chicago in den Jahren 2011 und 2013 widerspiegeln. Sprühaktivitäten können die Mückenpopulationen beeinflussen und sich möglicherweise auf die WNV-Prävalenz auswirken. Die Wetterdaten, die von der National Oceanic and Atmospheric Administration (NOAA) stammen, sind eine wichtige Komponente. Der Datensatz umfasst die Jahre 2007 bis 2014 und deckt sich mit den Testmonaten. Diese Informationen sind wertvoll, da davon ausgegangen wird, dass warme und trockene Bedingungen die WNV-Übertragung begünstigen. Darüber hinaus helfen Kartendaten von OpenStreetMap bei der Visualisierung, indem sie die räumlichen Aspekte der Daten in einen Kontext stellen.

Kommentiert [2]: Sollten wir das nicht lieber rausnehmen? Wir verwenden die Daten ja nicht

Kommentiert [3]: Ja, würde ich auch eher rausnehmen

Der Hauptdatensatz enthält verschiedene Attribute: Datum des WNV-Tests, ungefähre Adresse, Stechmückenart, Blocknummer, Straßenname, Fallenbezeichnung, Koordinaten, Anzahl der Stechmücken und WNV-Präsenz. Der Sprühdatensatz enthält Einzelheiten über Ort und Zeitpunkt des Sprühens zur Mückenbekämpfung. Der Wetterdatensatz umfasst die

klimatischen Bedingungen in den Testjahren. Eine kurze Übersicht der einzelnen Wetterattribute kann der Tabelle 1 entnommen werden.

Zusammenfassend lässt sich sagen, dass die Phase des Data Understandings die Vielschichtigkeit der bereitgestellten Datensätze verdeutlicht, die Mückentests, Wetterbedingungen, räumliche Informationen und Maßnahmen zur Mückenbekämpfung umfassen. Ein gründliches Verständnis dieser Elemente ist für die nachfolgenden Phasen der Data Preparation, Modeling und Evaluation unerlässlich.

Spaltenname	Beschreibung
Station	Wetterstation 1 oder 2
Date	Datum
Tmax	Maximaltemperatur in °F
Tmin	Minimaltemperatur in °F
Tavg	Durchschnittstemperatur in °F
Depart	Abweichung von Normaltemperatur in °F
DewPoint	Durchschnittlicher Taupunkt in °F
WetBulb	Durchschnittliche Feuchtkugel in °F
Heat	Saison beginnt im Juli
SeaLevel	Durchschnittlicher Meereshöhe Druck in Zoll Quecksilber

ResultSpeed	Resultierende Windgeschwindigkeit in Meilen pro Stunde
ResultDir	Resultierende Richtung, ganze Grad in Meilen pro Stunde
AvgSpeed	Durchschnittliche Geschwindigkeit in Meilen pro Stunde

Tabelle 1: Übersicht über die Attribute des Wetterdatensatzes.

3 Data Preparation

Im dritten Kapitel dieser Arbeit werden die Trainings- und Wetterdaten genauer betrachtet. Während des ersten Schritts erfolgt die "Data Selection", in der nur die für die Problemstellung relevanten Daten ausgewählt werden. Anschließend werden die Daten im Rahmen der "Data Transformation" so transformiert, dass sie für die Modellierung am geeignetsten sind und die besten Ergebnisse erzielt werden. Das Augenmerk liegt hierbei auf der Erhöhung der Datenqualität. Der letzte Schritt dieses Kapitels beinhaltet die "Data Integration", wobei die Trainings- und Wetter-Data Frames zu einem Data Frame vereint werden.

3.1 Data Selection

Im ersten Teil der "Data Selection" wird das Trainings-Data Frame betrachtet. Folgende Attribute werden für die Problemstellung als relevant eingestuft: "Date", "Species", "Latitude", "Longitude", "WnvPresent". Das Datum ist wichtig, um Auskunft über die Jahreszeiten zu erhalten. In Wintermonaten ist die Temperatur niedrig, wodurch davon ausgegangen werden kann, dass das WNV weniger verbreitet ist als im Sommer bei heißen Temperaturen. Außerdem ist das Datum in beiden Data Frames vorhanden, sodass "Date" als Mergevariable genutzt werden kann. Wie der Abbildung XY zu entnehmen ist, breitet sich das WNV nicht bei allen Mückenarten gleichermaßen aus. Aus diesem Grund wird das Attribut "Species" in die Modellierung eingeschlossen. Um den Ort, an dem das WNV nachgewiesen worden ist, zu lokalisieren, werden die Attribute "Latitude" und "Longitude" anstelle der Adressangabe verwendet. Gründe dafür sind die höhere Genauigkeit der GPS-Koordinaten sowie das Format. Bei den Attributen "Latitude" und "Longitude" handelt es sich um numerische Attribute, mit denen bei der Modellierung einfacher zu arbeiten ist. In der nachfolgenden Modellierung wird mit Modellen gearbeitet, die nur numerische Attribute verarbeiten können. "Address",

“Block”, “Street” und “AddressNumberAndStreet” werden dementsprechend aus dem Data Frame gelöscht. Die Attribute “Trap” und “NumMosquitos” werden ebenfalls aus dem Data Frame gelöscht, da sie im Hinblick auf die Problemstellung als irrelevant angesehen werden. “WnvPresent” stellt das Zielattribut der Modellierung dar. Der Spraydatensatz wird für die Modellierung nicht verwendet, da die Daten keine Aussagekraft für die Vorhersage des West-Nil-Virus haben.

Im zweiten Teil dieses Kapitels werden die im Hinblick auf die Problemstellung relevanten Attribute der Wetterdaten ausgewählt. In Chicago gibt es zwei Wetterstationen mit zwei verschiedenen Standorten. Eine Station befindet sich eher im Norden von Chicago und die andere Station eher weiter im Süden. Um zu lokalisieren, wo welche Wetterdaten gemessen wurden, wird das Attribut “Station” zunächst im Datensatz belassen. Das Attribut “Date” wird aus den gleichen Gründen wie bei den Trainingsdaten als relevant angesehen. Gemäß der Annahme, dass heißes und trockenes Wetter günstiger für die Ausbreitung des WNV sind als kaltes und nasses Wetter, werden die Attribute “Tmax”, “Tmin” und “WetBulb” in der Modellierung berücksichtigt. “Tavg” wird nicht genutzt, da es auch Tage geben kann, an denen die Minimal- und Maximaltemperatur weit auseinanderliegen. Diese Gegebenheit könnte ebenfalls einen Einfluss auf die Ausbreitung des West-Nil-Virus haben und wird mit den Attributen “Tmax” und “Tmin” in der Modellierung berücksichtigt. Die Attribute “Water1”, “SnowFall” und “PrecipTotal” werden trotz ihrer Aussagekraft bezüglich der Feuchtigkeit gelöscht. In der Modellierung wird hierfür lediglich das Attribut “WetBulb” inkludiert, da diese Daten Feuchtigkeit besser widerspiegeln als nur die Niederschlags- und Schneemenge. Eine hohe Luftfeuchtigkeit kann bestehen ohne Regen- oder Schneefall. Die Feuchtkugeltemperatur verbindet die Temperatur der trockenen Luft mit der Luftfeuchtigkeit. Das Attribut hat somit dieselbe Einheit wie “Tmax” und “Tmin” (Fahrenheit). Neben der Temperatur und Feuchtigkeit werden auch die Attribute “ResultSpeed” und “ResultDir” für die Vorhersage als relevant eingestuft. Die Windgeschwindigkeit und -richtung liefern Auskunft darüber, in welche Richtung und wie schnell sich das WNV ausbreitet. Folgende Attribute werden als irrelevant angesehen, da sie in keiner Beziehung zu der Ausbreitung des WNV stehen oder in hohem Maße unvollständig sind: “SeaLevel”, “BewPoint”, “StnPressure”, “Cool”, “Heat”, “Depth”, “CodeSum”, “AvgSpeed” und “Depart”. Die Attribute “Sunset” und “Sunrise” könnten auf den ersten Blick wichtige Informationen über die Ausbreitung des West-Nil-Virus liefern, da bei Dämmerung Mücken besonders aktiv sind. In der Modellierung werden die

Daten jedoch nur tageweise und nicht stündlich betrachtet, weshalb die beiden Attribute ebenfalls gelöscht werden.

3.2 Data Cleaning & Transformation

Wie bereits erwähnt, verarbeiten die in der Modellierung verwendeten Modelle lediglich numerische Attribute, weshalb das Attribut "Species" aus dem Trainings-Data Frame transformiert wird. Für jedes der acht Mückenarten wird ein Attribut mit den Werten 0 und 1 (Dummies) erstellt, wobei 1 für das Vorhandensein der jeweiligen Mückenart steht. Der Datensatz wird demnach um die folgenden binären Attribute erweitert: "Culex Pipiens/Restuans", "Culex Restuans", "Culex Pipiens", "Culex Erraticus", "Culex Salinarius", "Culex Tarsalis", "Culex Territans" und "Unspecified Culex". Das ursprüngliche Attribut "Species" wird gelöscht.

Im gesamten Wetter-Dataframe wird nach Missing Values gesucht und diese werden durch "0" ersetzt. Außerdem wird das Attribut "WetBulb" ebenfalls in ein numerisches Attribut umgewandelt, um in der Modellierung damit arbeiten zu können. Wie der Abbildung XY2 zu entnehmen ist, kommen in dem Feuchtigskeitsdatensatz zwei Ausreißer vor, die vermutlich vor der Umwandlung von Missing Values in "0" fehlende Messwerte darstellen. Da es sich hierbei um vollständig zufällig fehlende Daten (MCAR) handelt, können die fehlenden Werte mithilfe der einfachen Imputation geschätzt werden. Hierfür wird der Mittelwert des Attributs berechnet und die fehlenden Werte mit dem Mittelwert ersetzt.

Der nachfolgende Schritt wird für den bereits vereinten Data Frame aus Trainings- und Wetterdaten durchgeführt. Der Merge wird jedoch erst im nachfolgenden Kapitel genauer erläutert. Für die Vorhersage des WNV hat das jeweilige Jahr keine Aussagekraft. Lediglich die Monate sind entscheidend. Aus diesem Grund wird in dem vereinten Data Frame das Attribut "Date" zunächst in einen Datumstyp umgewandelt, um im nächsten Schritt die Monate als neues Attribut separat darzustellen. Das ursprüngliche Attribut "Date" wird gelöscht, so dass in der Modellierung nur noch die Monate Einfluss nehmen.

3.3 Data Integration

Um die Trainings- und Wetterdaten in einem Data Frame zu vereinen, wird im ersten Schritt eine neue Spalte in den Trainings-Data Frame hinzugefügt. Das neue Attribut heißt "Station" und unterteilt Chicago in zwei Gebiete (1 und 2) gemäß der Wetterstationen aus dem Wetter-

Data Frame. Die Wetterstation 1 hat die folgenden Koordinaten: Latitude: 41.995, Longitude: -87.933. Die Wetterstation 2 hat die folgenden Koordinaten: Latitude: 41.786 und Longitude: -87.752. Um Chicago in zwei Gebiete einzuteilen, wird der Mittelwert der Latitude der beiden Wetterstationen gebildet (41.8905). So werden alle Datensätze in dem Trainings-Data Frame mit einer Latitude größer als 41.8905 der Station 1 und alle Datensätze mit einer Latitude kleiner/gleich 41.8905 der Station 2 zugeordnet. Der Merge der Trainings- und Wetter-Data Frames erfolgt nun mit den Attributen "Date" und "Station" als Schlüssel. Das bedeutet, dass alle Zeilen aus beiden Data Frames basierend auf den übereinstimmenden Werten dieser beiden Attribute kombiniert werden.

4 Modeling

Im folgenden Kapitel werden die Modelle vorgestellt, die zur Vorhersage, ob das WNV für eine bestimmte Zeit und einen bestimmten Ort vorhanden ist, herangezogen wurden.

4.1 Modellauswahl

Bei dem in dieser Arbeit vorliegenden Klassifikationsproblem liegt das Ziel in der Vorhersage, ob die Mücken an einer bestimmten Teststation positiv oder negativ auf das WNV getestet werden. Mithilfe einer möglichst genauen Vorhersage kann die Stadt Chicago zielgerichtet an den richtigen Orten Pestizide sprühen, um eine Ausbreitung des WNV zu verhindern. Aufgrund des konkreten Ziels werden folgende Modelle aus dem Supervised Learning genutzt: K-nearest-neighbor Classifier (k-NN-Algorithmus), Support Vector Machine (SVM), Logistische Regression, Random Forest Classifier.

Der k-NN-Algorithmus wird verwendet, da dieser auf lokalen Ähnlichkeiten basiert und daher in Hinblick auf die Berücksichtigung räumlicher Beziehungen zwischen Mückenfallenstandorten und Wetterdaten sowie der Präsenz des WNV geeignet sein kann. SVM wird als Modell herangezogen, da das Modell effizient mit Daten umgehen kann, die in einem höherdimensionalen Merkmalsraum liegen. Zum Beispiel wäre es denkbar, dass die Mücken- und Wetterdaten verschiedene Merkmale haben, die nicht nur die räumliche Verteilung, sondern auch die Wetterbedingungen umfasst. Des Weiteren ist SVM aufgrund der Margin-Maximierung robust gegenüber Overfitting. Außerdem wird die logistische Regression herangezogen, welche eine häufig verwendete Methode für binäre Klassifikationsprobleme wie dieses darstellt. In der Praxis kommt diese Methode vor allem aufgrund der Recheneffizienz

oft zum Einsatz. Der Random Forest Classifier kann gut mit komplexen Daten umgehen und ist im Vergleich zu einem einzelnen Baum robuster gegenüber Overfitting. Daher eignet sich diese Methode zur Untersuchung des vorliegenden Klassifikationsproblems.

Es würden auch weitere Modelle, wie z.B. neuronale Netzwerke, für dieses Klassifikationsproblem in Frage kommen, allerdings beschränken wir uns aufgrund des Umfangs dieser Arbeit ausschließlich auf die oben genannten Modelle.

4.2 k-NN-Algorithmus

Der k-NN-Algorithmus betrachtet für die Klassifikation die k nächsten Nachbarn, wobei in diesem Fall k = 3 gewählt wurde. Dabei wird ein Datenobjekt so klassifiziert, wie die Zielvariable bei der Mehrzahl der drei nächsten Nachbarn ausgeprägt ist. Bei diesem Modell ist es wichtig, dass nur Features in die Suche nach den nächsten Nachbarn einfließen, die relevant sind, wie z.B. die Wettereinflüsse, da es ansonsten zu Verzerrungen kommen kann („Fluch der Dimensionalität“). Aus diesem Grund wird die Variable „Station“ nicht mit einbezogen und es werden nur die Monate („Month“) und nicht jedes einzelne Datum (Variable „Date“) betrachtet. Zur Vermeidung von Overfitting wird eine Kreuzvalidierung (Cross-Validation) mit cv = 5 durchgeführt, d.h. der Datensatz wird in fünf gleich große Teile aufgeteilt und in jeder Iteration wird ein anderer Teil als Testdaten ausgewählt. Dadurch ist erkennbar, wie die Leistung zwischen den Datensätzen variiert und es wird eine bessere Verallgemeinerungsfähigkeit erzielt.

4.3 SVM

SVM ist ein maschinelles Lernverfahren, bei dem eine lineare Diskriminante so zwischen den Werten in den Trainingsdaten gesetzt wird, dass der Abstand zu den nächsten Datenpunkten maximiert wird. Um die Gefahr von Overfitting zu berücksichtigen, wird eine Kreuzvalidierung mit cv = 5 durchgeführt. Allerdings klassifiziert das Modell in diesem Fall alle Datensätze als negativ, d.h. die Methode eignet sich nicht zur Vorhersage basierend auf dem vorliegenden Datensatz.

4.4 Logistische Regression

Um Overfitting zu vermeiden, wird eine Kreuzvalidierung mit cv=5 durchgeführt. Bei der logistischen Regression wird dann für einen Datensatz berechnet, mit welcher Wahrscheinlichkeit dieser zu einer bestimmten Klasse gehört. Die Zuordnung erfolgt in die Klasse, bei der die Wahrscheinlichkeit am höchsten ist. Damit das Modell funktioniert und die Features in einem vergleichbaren Bereich sind, wurden die Eingangsdaten skaliert und die maximale Anzahl an Iterationen erhöht. Allerdings kann das Modell durch Ausreißer schnell verzerrt werden, wodurch sich die Genauigkeit verschlechtert, was bei der Evaluation des Modells beachtet werden sollte.

4.5 Random Forest Classifier

Der Random Forest Classifier trifft eine Klassifikationsentscheidung, indem er die Vorhersagen mehrerer Entscheidungsbäume aggregiert und die Klasse auswählt, die von den meisten Bäumen vorhergesagt wird. Da der Datensatz viele Attribute aufweist, anhand dessen Splits vorgenommen werden können, eignet sich diese Methode im vorliegenden Fall sehr gut. Außerdem wird eine Kreuzvalidierung (cv=5) durchgeführt, um die Leistung mit verschiedenen Anzahlen von Entscheidungsbäumen zu bewerten und die optimale Anzahl zu ermitteln, welche in diesem Fall zehn beträgt. Abbildung 1 stellt grafisch die Vorhersagen auf Basis des Random Forest Classifiers dar:

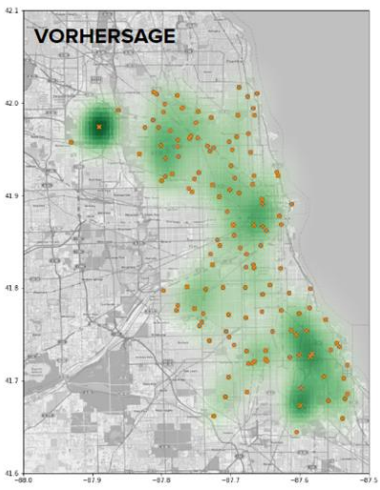

Abbildung 1: Vorhergesagte Koordinaten auf Basis des Random Forest Classifiers.

5 Evaluation

Die Konfusionsmatrix ist in unserem Projekt eine aussagekräftige Bewertungsmethode, da sie direkt aus den tatsächlichen Vorhersagen und den wahren Werten abgeleitet wird. In dieser Matrix werden die vier grundlegenden Szenarien dargestellt, die bei der Klassifikation auftreten können:

- True Positives (TP): Diese Fälle repräsentieren tatsächlich positive Werte, die korrekt vom Algorithmus erkannt wurden. In diesem Fall sind das die Stationen, bei denen das West-Nil-Virus korrekt vorhergesagt wurde.
- False Positives (FP): Hierbei handelt es sich um Fälle, in denen der Algorithmus fälschlicherweise positive Werte vorhergesagt hat, obwohl sie in Wirklichkeit negativ sind. In diesem Beispiel wären das die Stationen, bei denen der Algorithmus falsch das West-Nil-Virus vorhergesagt hat.
- True Negatives (TN): Diese Fälle repräsentieren tatsächlich negative Werte, die vom Algorithmus korrekt erkannt wurden. Das sind die Stationen, bei denen das Fehlen des West-Nil-Virus korrekt vorhergesagt wurde.
- False Negatives (FN): Hierbei handelt es sich um Fälle, in denen der Algorithmus fälschlicherweise negative Werte vorhergesagt hat, obwohl sie in Wirklichkeit positiv sind. Das sind die Stationen, bei denen der Algorithmus das West-Nil-Virus fälschlicherweise nicht erkannt hat.

Kommentiert [4]: falls wir noch etwas kürzen müssen, könnte man die Erklärungen zur Konfusionsmatrix vlt. etwas kürzer fassen

Die Konfusionsmatrix erlaubt es, aus diesen Werten wichtige Metriken zur Leistungsbewertung abzuleiten, darunter:

- Genauigkeit: Anteil der korrekt vorhergesagten Werte (TP und TN) an der Gesamtzahl der Vorhersagen.
- Präzision: Anteil der korrekt vorhergesagten positiven Werte (TP) an allen vorhergesagten positiven Werten (TP und FP).
- Recall: Anteil der korrekt vorhergesagten positiven Werte (TP) an allen tatsächlich positiven Werten (TP und FN).
- F1-Score: Ein gewichteter Durchschnitt von Präzision und Recall, der eine ausgewogene Leistungsbewertung ermöglicht.

Da die Konfusionsmatrix direkt die Anzahl der korrekten und falschen Vorhersagen für jede Klasse zeigt, bildet sie die Grundlage für diese wichtigen Leistungsmetriken. Daher ist die

Konfusionsmatrix in diesem Fall eine unverzichtbare Methode zur Bewertung der Algorithmus-Leistung.

Die k-NN Classifier-Konfusionsmatrix zeigt eine vergleichsweise hohe Anzahl von TN, was darauf hindeutet, dass der Algorithmus gut darin ist, negative Werte korrekt zu erkennen. Allerdings ist die Anzahl der FN recht hoch, was auf eine suboptimale Fähigkeit hinweist, positive Werte zu erkennen. Die Präzision (TP / (TP + FP)) ist relativ niedrig, während der Recall (TP / (TP + FN)) ebenfalls niedrig ist. Der F1-Score (2 * (Präzision * Recall) / (Präzision + Recall)) zeigt einen Kompromiss zwischen Präzision und Recall.

Die Random Forest-Konfusionsmatrix zeigt eine sehr niedrige Anzahl von TP und eine relativ hohe Anzahl von FN. Dies resultiert in einer niedrigen Präzision und einem niedrigen Recall. Auch dieser Algorithmus hat Schwierigkeiten, sowohl positive als auch negative Werte korrekt zu klassifizieren.

Die Konfusionsmatrix für die SVM mit k-Folds zeigt, dass der Algorithmus überhaupt keine positiven Werte erkannt hat. Die Präzision, der Recall und der F1-Score sind daher alle auf null gesetzt. Dies deutet darauf hin, dass der Algorithmus entweder schlecht angepasst ist oder es Probleme bei der Datenverarbeitung gibt. Ähnlich wie bei der SVM hat auch die logistische Regression mit k-Folds keine positiven Werte erkannt, was zu einer Präzision, einem Recall und einem F1-Score von null führt. Der Majority Dummy Classifier zeigt keine hohe Anzahl von FP und viele FN. Die Präzision, der Recall und der F1-Score sind ebenfalls null.

Modell	Genauigkeit	Präzision	Recall	F1-Score	Konfusions-matrix
k-NN Classifier	0.8919562113279391	0.7307692307692307	0.0794979079497908	0.14339622641509434	[19 7 220 1855]
Support Vector Machine	0,886244645406949	0,0	0,0	0,0	[0 0 239 1862]
Logistische Re-gression	0,886244645406949	0,0	0,0	0,0	[0 0 239 1862]
Random Forest Classifier (10 verwendete Bäume)	0.9876308277830638	0.9649122807017544	0.6111111111111112	0.7482993197278912	55 2 35 2010]
Dummy Classifier	0.4902427415516421	0.11552680221811461	0.5230125523012552	0.18925056775170326	[125 957 114 905]
Majority (Dummy) Classifier	0.886244645406949	0.0	0.0	0.0	[0 0 239 1862]

Tabelle 2: Evaluation der Performance aller Modelle.

Die Evaluation der verschiedenen Modelle basierend auf den gegebenen Konfusionsmatrizen im Rahmen des CRISP-DM Prozesses ermöglicht eine tiefere Einsicht in die

Kommentiert [5]: In dem vorliegenden Datensatz wird versucht, eine kleine Anzahl an positiven Fällen zu finden, d.h. die Klassenverteilung ist unausgewogen. Daher funktioniert die Bewertung auf Grundlage der Genauigkeit und der anderen Gütemaße nicht. Die Präzision, der Recall und der F1-Score sind bei der logistischen Regression, SVM und dem Majority Classifier Null, weil keine Vorhersagen für die positive Klasse gemacht werden, was dazu führt, dass der Nenner in der Präzisionsformel Null ist und sich daher die Metrikwerte nicht berechnen lassen. Stattdessen wird zur Evaluation der Performance die Konfusionsmatrix der Modelle betrachtet. Aus der Matrix kann abgelesen werden, wie viele Datenobjekte richtig und wie viele fälschlicherweise als positiv oder negativ klassifiziert wurden.
Um festzustellen, ob eines der aufgestellten Modelle einen Mehrwert bietet, werden sie mit zwei Baseline-Ansätzen verglichen. Zum einen mit einem Dummy Classifier, welcher rein zufällig Vorhersagen trifft. Zum anderen mit einem Majority Classifier, welcher immer die Mehrheitsklasse des Trainingsdatensatzes wählt (in diesem Fall: negativ). Zweiterer ist besser geeignet als Vergleich, da eine unausgewogene Klassenverteilung vorliegt, wodurch die Genauigkeit von diesem sehr hoch ist.

Klassifikationsleistung für die Erkennung des WNV. Fast alle Modelle zeigen eine hohe Genauigkeit von über 90 %. Die Genauigkeit ist in diesem Fall jedoch keine geeignete Leistungsmetrik und könnte irreführend sein: Lediglich 5,2 % der getesteten Mücken waren tatsächlich mit dem WNV infiziert. Infolgedessen neigt die Genauigkeit dazu, hoch zu sein, wenn das Modell hauptsächlich negative Vorhersagen macht. Da die Daten ein deutliches Klassenungleichgewicht aufweisen, ist die Genauigkeit allein nicht ausreichend aussagekräftig. Diese Tatsache lässt sich bei SVM, bei der logistischen Regression und dem Majority Classifier erkennen, da keine positiven Vorhersagen für die Testdaten gemacht wurden, dennoch zeigt sich eine Genauigkeit von etwa 89 %. Dies unterstreicht, wie die Genauigkeit bei ungleichmäßig verteilten Klassen an Aussagekraft verlieren kann. Die Konfusionsmatrix bietet einen tieferen Einblick in die spezifischen Fehlervarianten und ermöglicht eine fundierte Entscheidung bei der Wahl des geeignetsten Modells.

Zusammenfassend zeigt die Analyse der Konfusionsmatrizen, dass der Random Forest die beste Leistung im Hinblick auf die Problemstellung hat. Für die Stadt Chicago ist es gefährlicher, wenn FP-Werte hoch sind, denn so kann sich das Virus unentdeckt verbreiten. Der Random Forest hat im Vergleich zu den anderen Modellen eine hohe TP Rate und die niedrigste FN Rate und trifft bessere Vorhersagen als ein Majority-Classifier.

6 Konkrete Vorhersage

Basierend auf den Algorithmen lassen sich Vorhersagen über betroffene Koordinaten in der Stadt Chicago treffen. Diese werden nun veranschaulicht:

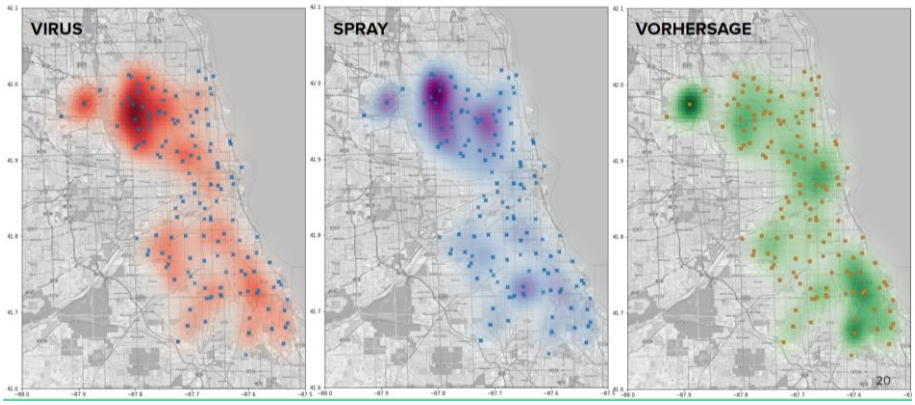

Abbildung 3: Visualisierung der GPS Koordinaten der Trainings-, Spray- und Vorhersagedaten.

Kommentiert [6]: das gehört doch eigt. auch zur Evaluation oder? Würde vorschlagen, dass wir dafür kein extra Kapitel aufmachen

Kommentiert [7]: Naja habe dem Kapitel ein treffenderen Namen gegeben. Ich denke das sollte ein eigenes Kapitel sein, da es um die Anwendung des Algorithmus geht und nicht mehr um Evaluation - oder?

Kommentiert [8]: Ich glaube, ich würde es auch unter Evaluation fassen

Links im Bild sind alte Messdaten des Virus zu sehen - die Teststationen sind mit einem blauen x dargestellt. Es wurde davon ausgegangen, dass das Virus sich radial zum bekannten Messpunkt verhält. Daher sind diese Ausbreitungen mit einer roten Farbe gezeichnet. Dort, wo viele radiale Kreise sich kreuzen, wird die Farbe dunkler. Daneben ist dargestellt, wo die Stadt Chicago das Pestizid versprüht hat - auf dem Hintergrund der roten Virus-Daten. Man kann gut erkennen, dass das rot eingezeichnete Virus von der blauen Farbe der Pestizide überdeckt ist. Die Pestizide wurden demnach gut verteilt.

Anhand des besten Algorithmus, dem Random Forest, wurden in der rechten Grafik die vorhergesagten Koordinaten grün eingefärbt. Wenn man demgegenüber die Koordinaten der Spray-Aktionen der Stadt stellt, sieht man eindrucksvoll, dass nicht an den richtigen Stellen gesprüht wurde bzw. sich das Virus vor allem südlich und auch hinter dem Fluss am Ufer ausbreiten wird.

Kommentiert [9]: Wollen wir am Ende vlt. noch ein kurzes Unterkapitel "Ausblick" einfügen (das fand ich in der Bsp.doku ganz gut)?

Kommentiert [10]: Ja das hatte er doch auch als Kritikpunkt, dass wir kein Ausblick haben