Smart Data Analytics Task

Aufgabe: Epileptische Anfälle

Es wurden zwei Modelle gebildet. Das erste Modell klassifiziert in die Klassen 1 bis 5 und das zweite Modell in Klasse 1 und alle anderen als eine Klasse. Es wurden beim zweiten Modell, die Klassen 2 bis 5 zusammengefasst, da diese als „Nicht Kritisch“ eingestuft wurden und die Aufgabe fordert, dass das Modell epileptische Anfälle („Sehr kritisch“) vorhersagen soll.
Im ersten Schritt wurden die Werte skaliert um sie besser verarbeiten zu können.
Im zweiten Schritt wurde eine Principal Component Analyse (PCA) vorgenommen, um die Dimension zu reduzieren.
Im Anschluss wurden die Daten in ein Trainings- und ein Testdatenset aufgeteilt (60% zu 40%). Danach wurde der SGD (Stochastic Gradient Descent) Klassifikator auf das Trainingsset angewendet. Beim ersten Modell ergibt sich dafür eine Zuverlässigkeit von etwa 40%. Der True-Positive Wert von Klasse 1 („Sehr Kritisch“) liegt bei etwa 80%.
Die der Klassifikation mittels SVC (Support Vector Classifier) ergibt bei Klasse 1 („Sehr Kritisch“) einen True-Positive Wert von etwa 90%.
Beim ersten Modell ist die Vorhersage der Klassen 2 bis 5 generell sehr schlecht. Somit wurden im nächsten Schritt diese Klassen zusammengefasst und bei Klassifikatoren auf diesen Daten angewendet.
Beim zweiten Modell ergibt sich beim SGD Klassifikator für Klasse 1 („Sehr Kritisch“) ein True-Positive Wert von ebenfalls etwa 90%. Beim SVC Klassifikator nur knapp besser mit 91%. Bei beiden Klassifikatoren ist die Vorhersage von Klasse 2-5 auf über 98% gestiegen.
Die beiden Klassifikatoren wurden gewählt, da es sich zum einen natürlich um eine Klassifikation handelt und zum anderen nur verhältnismäßig wenig Beispiele gegeben sind. Es wurde auch der KNeighbors Klassifikator ausprobiert, dieser hat allerdings schlechtere Ergebnisse geliefert als SGD und SVC.

Aufgabe: Waldbrand

Bei dieser Aufgabe konnte ich kein Modell generieren, welches die verbrannte Fläche ansatzweise vorhersagen kann. Es wurde zunächst der Datensatz nach Korrelation etc. untersucht. Danach wurden die Monate in numerische Werte umgewandelt und die X- und Y-Koordinaten zu einem Merkmal zusammengefasst. Ein einfaches SVR (Support Vector Regression) Modell konnte nicht die verbrannte Fläche vorhersagen. Ausgewertet wurden die Ergebnisse mittels „MAE“ (Mean average error), „MSE“ (Mean Squared Error) und „RMSE“ (Rooted Mean Squared Error) ausgewertet. Die grafische Darstellung der vorhergesagten Werte lässt allerdings schon erkennen, dass sich das Modell nicht entsprechend verhält. Auch mit skalierten Werten und PCA ließ sich kein besseres Ergebnis erzeugen. Das Modell wurde gewählt, da nur sehr wenige Beispiele gegeben sind und SVR hier angemessen ist.