OST
Ostschweizer
Fachhochschule

ML End-to-End Projekt

Mechatronik BSc
HS 2025

Aufgaben

Machine Learning im Modul ML

1 End-to-End Machine-Learning Projekt

Ihr End-to-End Machine-Learning Projekt (ML-Projekt) ist eine Gelegenheit für Sie, ein interessantes maschinelles Lernproblem Ihrer Wahl im Kontext eines realen Datensatzes zu untersuchen. Am Schluss des Dokuments finden Sie einige Projektideen. Aber die beste Idee wäre, maschinelles Lernen mit Problemen in Ihrem eigenen Interessens- und Erfahrungsgebiet zu kombinieren.

Jedes ML-Projekt muss rechtzeitig in Form eines ausführbaren `Jupyter-Notebooks` (Python $>$ 3.5) an die Dozenten abgegeben werden und folgende Abschnitte enthalten:

(a) *Header*:
Projektname, Autoren, Datum, OST-Logo

(b) *Einleitung*:
Beschreibung der Aufgabe, Anwendungsbereich und Bedeutung innerhalb dieses Anwendungsbereichs.

(c) *Zielsetzung und Vorgehensweise*:
Was soll erreicht werden? Welchen Ansatz wählen Sie, um dieses Ziel zu erreichen? Welche Metrik wählen Sie, um die Leistung zu bewerten? Welches Level dieser Metrik muss für eine zufriedenstellende Erfüllung der Aufgabe erreicht werden?

(d) *Explorative Datenanalyse*:
EDA, gilt als Zwischenbericht

(e) *Pipeline*:
Generieren und Einsatz einer Pipeline

(f) *No Free Lunch*:
Auswahl von geeigneten Klassen von Lernern

(g) *Validierung*:
Kreuzvalidierte Performance dieser Lerner und Hyperparameter-Tuniung

(h) *Entscheid*
Auswahl eines geeigneten Lerners und Begründung

(i) *Schlussfolgerung und Ausblick*:
Was könnte man weiterführend untersuchen? Welche Methoden würden sich auch noch eignen?

(j) *Referenzen*:
Quellenangaben von relevanten Papern, Büchern, Internet-Link der verwendeten Quellen und Daten.

Folgende Voraussetzungen gelten für das Projekt:

(a) Projekte können *einzeln* oder *in Teams von zwei Studenten* durchgeführt werden. Bei einer Zweiergruppe sind die Gruppenmitglieder dafür verantwortlich, die Arbeit gleichmässig zu verteilen und sicherzustellen, dass jedes Mitglied seinen Beitrag leistet.

(b) Das End-to-End ML-Projekt (ML-Projekt) muss von ihnen *eigenständig* erstellt worden sein. Wenn Sie Ideen oder Code von anderen Quellen verwenden, müssen diese Quellen klar gekennzeichnet sein. Fügen Sie ihrem Jupyter-Notebook eine *Eigenständigkeitserklärung* bei (siehe am Schluss des Dokuments).

(c) In Ihrem ML-Projekt muss es um neue Dinge gehen, die Sie in diesem Semester (HS2023) gemacht haben. Sie dürfen keine älteren Arbeiten verwenden.

(d) Die rechtzeitige Abgabe des `Jupyter Notebooks` des ML-Projekts ist Voraussetzung für die Teilnahme am Prüfungsgespräch.

Wenn Sie Probleme beim Schreiben eines Vorschlags haben, können Sie sich gerne an den Dozenten wenden. Sobald Sie Ihren Vorschlag eingereicht haben, geben wir Ihnen Feedback. Natürlich liegt die endgültige Verantwortung für die Definition und Ausführung einer interessanten Arbeit bei Ihnen.

2 Termine und Bewertung des Projekts

2.1 Termine

Projektziel	**KW**	**Termin**	**Umfang**
Vorschlag und Gruppeneinteilung	40	29. September 2025	1 Seite
Zwischenbericht (EDA, Jupyter-NB)	46	10. November 2025	2-3 Seiten
Endgültiges Jupyter Notebook	3	12. Januar 2026	-

2.2 Bewertung

Jedes Ergebnis Ihres Projekts wird anhand folgender Faktoren bewertet:

(a) *Vollständigkeit*: wurden die Abschnitte gemäss obiger Vorlage erstellt und ausgeführt?

(b) *Verständlichkeit*: Der Schreibstil und die Klarheit der schriftlichen Arbeit.

(c) *Korrektheit*: Wurden die Methoden korrekt angewendet?

(d) Die *Neuheit* der Projektideen und Anwendungen. Die Gruppen werden ermutigt, originelle Ideen und neue Anwendungen für die Projekte zu entwickeln.

(e) Der *Umfang* der Studie und der Experimente.

(f) *Qualität*: Ein Projekt, das ein intelligenteres System durch die Kombination mehrerer ML-Techniken hervorbringt, oder ein Projekt, das gut konzipierte Experimente und eine gründliche Analyse der experimentellen Ergebnisse beinhaltet, oder ein Projekt, das verschiedene Anwendungen aus der Praxis gut integriert, werden höher bewertet.

3 Projektvorschläge

3.1 Kaggle Contests

(a) IEEE-CIS Fraud Detection
Can you detect fraud from customer transactions?

(b) LANL Earthquake Prediction
Can you predict upcoming laboratory earthquakes?

(c) Mechanisms of Action (MoA) Prediction
Can you improve the algorithm that classifies drugs based on their biological activity?

(d) Cassava Leaf Disease Classification
Identify the type of disease present on a Cassava Leaf image

Predict the popularity of shelter pet photos

(e) PetFinder.my - Pawpularity Contest
Predict the popularity of shelter pet photos

(f) 15.071x - The Analytics Edge (Spring 2015)
Test your analytics skills by predicting which New York Times blog articles will be the most popular

(g) University of Liverpool - Ion Switching
Identify the number of channels open at each time point

(h) Google Brain - Ventilator Pressure Prediction
Simulate a ventilator connected to a sedated patient's lung

(i) Microsoft Malware Prediction
Can you predict if a machine will soon be hit with malware?

(j) Global Wheat Detection
Can you help identify wheat heads using image analysis?

(k) Bengali.AI Handwritten Grapheme Classification
Classify the components of handwritten Bengali

(l) TensorFlow - Help Protect the Great Barrier Reef
Detect crown-of-thorns starfish in underwater image data

(m) Happywhale - Whale and Dolphin Identification
Identify whales and dolphins by unique characteristics

(n) Cornell Birdcall Identification
Build tools for bird population monitoring

(o) Personalized Medicine: Redefining Cancer Treatment
Predict the effect of Genetic Variants to enable Personalized Medicine

(p) Plant Pathology 2020 - FGVC7
Identify the category of foliar diseases in apple trees

(q) Africa Soil Property Prediction Challenge
Predict physical and chemical properties of soil using spectral measurements

(r) Rainforest Connection Species Audio Detection
Automate the detection of bird and frog species in a tropical soundscape

(s) ALASKA2 Image Steganalysis
Detect secret data hidden within digital images

(t) WSDM - KKBox's Music Recommendation Challenge
Can you build the best music recommendation system?

(u) Generative Dog Images
Experiment with creating puppy pics

(v) Freesound Audio Tagging 2019
Automatically recognize sounds and apply tags of varying natures

(w) BirdCLEF 2021 - Birdcall Identification
Identify bird calls in soundscape recordings

3.2 Data Mining

Gehen Sie auf die Webseite `https://pslcdatashop.web.cmu.edu/` und klicken Sie auf «Öffentliche Datensätze». Auf der Webseite stehen etwa 30-40 öffentliche Datensätze zur Verfügung. Wählen Sie nur die Datensätze aus, deren Status als «vollständig»gekennzeichnet ist. Wenn Sie auf die einzelnen Datensätze klicken, finden Sie eine allgemeine Beschreibung und die dazugehörigen Publikationen. Um sich die Datensätze anzusehen, klicken Sie auf den Link «Export».

Projektidee 1: In allen Datensätzen können Sie mehrere maschinelle Lerntechniken (mindestens zwei bis drei verschiedene ML-Methoden) vergleichen, um «Correct First Attempt values»(in der Spalte «Outcome»aufgeführt) vorherzusagen. Die Hypothese hier ist, dass es keinen absoluten Gewinner geben kann. Verschiedene maschinelle Lerntechniken können in unterschiedlichen Aufgabenbereichen und auf unterschiedlichen Datensätzen effektiv sein (no free lunch).

Projektidee 2: In allen Datensätzen können Sie mehrere maschinelle Lerntechniken (mindestens zwei bis drei verschiedene ML-Methoden) vergleichen, um «Correct First Attempt values»(in der Spalte «Outcome»aufgeführt) vorherzusagen. Untersuchen Sie, ob ein Voting Classifier oder Regressor besser performt als jeder der einzelnen Learner alleine.

3.3 Weitere Datensätze

- Google Dataset Search
  `https://toolbox.google.com/datasetsearch`
- UC Irvine hat ein Repository, das für Ihr Projekt nützlich sein könnte:
  `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
- Sam Roweis stellt mehrere Datensätze zur Verfügung:
  `http://www.cs.toronto.edu/~roweis/data.html`.
- Kaggle Datensätze:
  `https://www.kaggle.com/datasets`
- Übersicht aus Wikipedia:
  `https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_`

```
research
```

4 Eigenständigkeitserklärung

Die folgende Eigenständigkeitserklärung sollten Sie unterzeichnen und als Anhang zum abgegebenen `Jupyter Notebook` beilegen.

«*Hiermit bestätige ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen) entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.*»

Lösungen