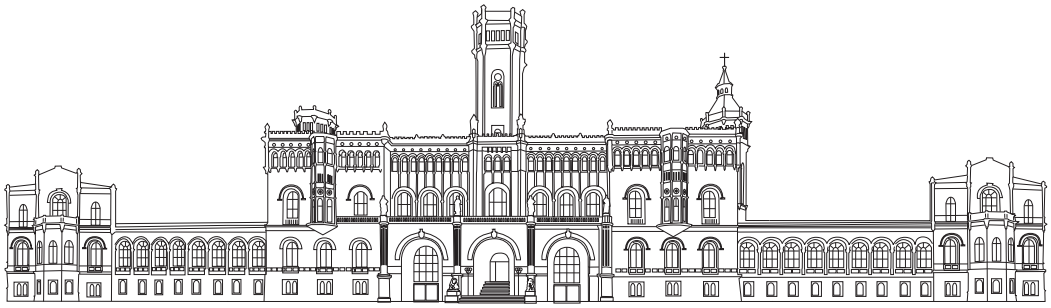

LEIBNIZ UNIVERSITÄT HANNOVER

FAKULTÄT FÜR ELEKTROTECHNIK UND INFORMATIK
INSTITUT FÜR PRAKTISCHE INFORMATIK

Improving Primary Key Detection with Machine Learning

Bachelor Thesis

submitted by

JANEK PRANGE

on 03.07.2022

First Examiner	:	Prof. Dr. Ziawasch Abedjan
Second Examiner	:	Prof. Dr. Sören Auer
Supervisor	:	Prof. Dr. Ziawasch Abedjan

DECLARATION

I hereby affirm that I have completed this work without the help of third parties and only with the sources and aids indicated. All passages that were taken from the sources, either verbatim or in terms of content, have been marked as such. This work has not yet been submitted to any examination authority in the same or a similar form.

Hannover, 03.07.2022

Janek Prange

ABSTRACT

Short summary of the contents in English. . . a great guide by Kent Beck how to write good abstracts can be found here:

https://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

AI Artificial Intelligence

1

FUNDAMENTAL KNOWLEDGE

- Introduce the subjects used in the thesis
- The explanations have to be sufficient for a student after the lecture DBS I

1.1 MACHINE LEARNING

Machine learning is a part of the field of Artificial Intelligence (AI). The focus is on extracting information from large amounts of data, with algorithms gradually improving themselves to mimic human learning [1].

Machine learning algorithms generally require labeled data to extract information. This labelled data consists mostly of a table where each column corresponds to a feature, which is a specific aspect the algorithm can use to infer some information.

TODO: Deep Learning, unlabeled data (images etc)

1.1.1 *Categories of Machine Learning*

- Supervised
  - Classification
  - Regression
- Unsupervised

1.1.2 *Scikit-Learn and Auto-Sklearn*

1.2 USED PACKAGES AND LIBRARIES

1.2.1 *Pandas*

1.2.2 *Scikit-Learn and Auto-Sklearn*

2

PROBLEM STATEMENT

3

PROPOSED METHOD

- explain the idea behind the feature extraction (it has to be understandable by a student)
- explain each feature
- if they exist, explain feature which were not included

4

EXPERIMENTS

4.1 CORRECTNESS

- measure accuracy and precision of the method
  - with 5, 10 and 20 rows as input (do more rows make a difference with large tables?)
  - overall and only non trivial
  - trained on one dataset and used on another
  - trained on multiple datasets
- explore the weaknesses of the method
  - what does a column look like which causes problems

4.1.1 *Experiment Data*

4.1.2 *Comparing models with different input sizes*

4.1.3 *Altering the training time*

4.1.4 *Changing the scoring functions*

4.1.5 *Conclusions*

4.2 EFFICIENCY

- compare the speed to the naive algorithm
  - with 5, 10 and 20 rows as input
  - measure the time with locally saved tables and tables from the database
- is a higher efficiency possible by reading only as many rows/-columns as necessary?

4.2.1 *Experiment Data*

4.2.2 *Comparing models with different input sizes*

4.2.3 *Changing the portion of unique columns*

4.2.4 *Conclusions*

5

CONCLUSION

5.1 POSSIBLE APPLICATIONS

5.2 LIMITATIONS OF THE METHOD

6

RELATED WORK

BIBLIOGRAPHY

[1] IBM Cloud Education. *Machine Learning*. July 15, 2020.
URL: https://www.ibm.com/cloud/learn/machine-learning
(visited on 05/18/2022).