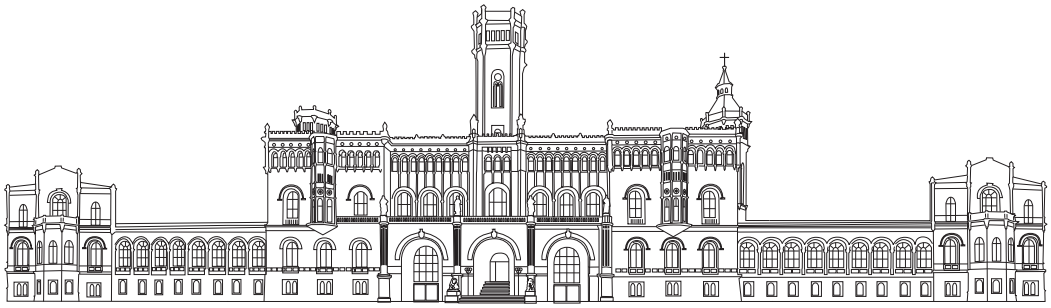

LEIBNIZ UNIVERSITÄT HANNOVER

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

1 BASICS

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

1.1 DATASET

- Give an overview over the dataset(s) used in the thesis
- Probably short, the points are
  - (!) Structure of the dataset in the database and the way it gets converted
  - (!) Size of the dataset
  - (?) Possible contents or their form
  - (?) Where the data came from
- This could be its own chapter.

1.2 MACHINE LEARNING

- Explain machine learning (probably just a short overview, although the target group does not really now anything about it)
- Quick explanation of the used ml-library

1.2.1 *Categories of Machine Learning*

- Supervised
  - Classification
  - Regression
- Unsupervised

1.2.2 *Auto-Sklearn*

1.2.2.1 *Scikit-Learn*

2 EXISTING ALGORITHMS

2.1 NAIVE ALGORITHMS

- This will be very short as the basic way of finding unique columns is known by the target group
- If the naive algorithms are implemented with flipped tables, explain it here

2.2 RELATED WORK

- Maybe here, maybe at the end
- Possibly a small comparison here and a longer one in its own section

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

4.2 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