Exploring the Utility of Machine Learning Across Varied Data Formats

Cărămidă Iustina-Andreea - 332CA

Faculty of Automatic Control and Computer Science
University Politehnica of Bucharest
iustina.caramida@stud.acs.upb.ro

Abstract. This study investigates the applicability of machine learning techniques on diverse datasets. We explore the effectiveness of two algorithms, Linear Regression and Multi-Layered Perceptron (MLP), on predicting both health outcomes and financial well-being. Specifically, we utilize a stroke prediction dataset to assess the modelś ability to identify individuals at risk of stroke. Additionally, we employ a salary prediction dataset to evaluate the modelś capacity to classify individuals earning above a specific income threshold (e.g., $50,000 per year). Through comparative analysis, this research aims to elucidate the strengths and limitations of each algorithm when applied to these contrasting data types, offering insights into their suitability for various prediction tasks. Furthermore, we present a framework for data analysis, outlining essential steps for data cleaning, exploration, and preparation, which can be applied to enhance the effectiveness of machine learning models across diverse datasets.

Keywords: *Machine Learning · Heterogeneous Data · Comparative Analysis · Prediction Modeling · Data Analysis Techniques · Stroke Prediction · Salary Prediction · Linear Regression · Multi-Layered Perceptron · DataPreprocessing*

1 Introduction

1.1 Motivation: The Power and Nuance of Machine Learning Data

Machine learning (ML) has become a cornerstone of progress in numerous disciplines. Its ability to extract valuable insights from vast and complex datasets has fueled breakthroughs in healthcare, finance, and social sciences. However, the effectiveness of ML models is not a one-size-fits-all proposition. Different data types possess unique characteristics, and understanding these nuances is essential for selecting the most appropriate ML algorithms. Data can be structured (organized in tables) or unstructured (text, images), numerical or categorical, and may exhibit linear or non-linear relationships between features. Choosing the right algorithm depends heavily on these factors. This research delves into this crucial aspect of ML application by exploring the performance of two distinct algorithms on contrasting datasets.

1.2 Research Focus: Delving into Stroke Prediction and Salary Prediction

This study focuses on the application of ML techniques to two contrasting datasets: stroke prediction and salary prediction. Stroke, a leading cause of disability and death globally, poses a significant public health burden. Stroke prediction models aim to identify individuals at high risk of experiencing a stroke, allowing for preventive measures and early intervention. These models typically analyze factors such as age, blood pressure, cholesterol levels, and smoking history.

Conversely, salary prediction models attempt to classify individuals based on income thresholds. This information can offer valuable insights into economic trends, such as income inequality, and inform policy decisions. Salary prediction models might analyze factors like education level, work experience, and industry sector. By investigating these two distinct datasets, this research aims to gain a broader understanding of how ML algorithms perform on different data types with varying underlying structures and complexities.

1.3 Methodology: Unveiling the Algorithms - Linear Regression and Multi-Layered Perceptron

To investigate the performance on these contrasting datasets, this study employs two prominent ML algorithms: Linear Regression and Multi-Layered Perceptron (MLP).

Linear Regression is a wellestablished technique known for its interpretability and efficiency in uncovering linear relationships between features (data points) and target variables (what we want to predict). This makes it a valuable tool for understanding the underlying factors influencing a particular outcome, such as the relationship between blood pressure and stroke risk. However, its strength lies in capturing linear relationships. If the underlying relationships in the data are more complex and non-linear, then Linear Regression might not be as effective.

On the other hand, Multi-Layered Perceptrons (MLPs) are a type of artificial neural network capable of learning complex, non-linear patterns within data. Unlike Linear Regression, MLPs are not limited by linearity and can potentially capture more intricate relationships between features and target variables. This capability makes them particularly suitable for datasets with complex underlying structures, such as the factors influencing an individual's salary, which might involve a combination of education, experience, industry, and other factors interacting in non-linear ways.

1.4 Research Objectives: Evaluating Algorithms, Unveiling Strengths and Weaknesses

By comparatively analyzing the performance of Linear Regression and MLP on stroke and salary prediction tasks, this research seeks to achieve several key objectives:

Evaluate the Suitability of Algorithms for Diverse Data Types: This involves assessing the effectiveness of each algorithm in capturing the underlying relationships within the stroke prediction and salary prediction datasets. We will determine which algorithm performs better on each dataset, offering insights into their suitability for different data types, such as linear datasets (blood pressure and stroke risk) versus potentially non-linear datasets (factors influencing salary).

Gain Insights into Algorithmic Strengths and Weaknesses: By analyzing the comparative performance, we aim to highlight the scenarios where each algorithm excels and identify areas where one might outperform the other. This will provide valuable guidance for researchers and practitioners in selecting the most appropriate algorithm for their specific prediction tasks. For instance, if interpretability is crucial (e.g., understanding the factors influencing stroke risk), Linear Regression might be preferred. If the data is likely to have complex, non-linear relationships, then an MLP might be a better choice.

Demonstrate Best Practices for Data Analysis in ML Applications: Effective data analysis is crucial for building robust ML models. This research will showcase essential steps for data cleaning, exploration, and preparation, emphasizing their importance in enhancing model performance across diverse datasets. These steps may include handling missing values, identifying outliers, and feature engineering (creating new features from existing data) to improve the model's ability to learn from the data.

1.5 Expected Contribution: Advancing the Application of ML on Heterogeneous Data

Through this exploration, the research aims to contribute valuable knowledge to the field of machine learning, particularly the application of ML on heterogeneous datasets. The findings can guide researchers and practitioners in selecting appropriate algorithms for their specific prediction tasks and data types. Furthermore, by demonstrating best practices for data analysis, this research can contribute to the development of more robust and reliable ML models across diverse application domains. This can lead to advancements in areas like healthcare (improved stroke prediction for preventive measures) and economics (better understanding of factors influencing income inequality). Ultimately, the research aims to contribute to the responsible and effective use of ML for tackling complex problems across various fields.

2 Exploratory Data Analysis

3 Data Preprocessing

4 Algorithms Designs

5 Evaluation

6 Conclusions

References

1. A* search algorithm - Wikipedia
2. Hill Climbing - Wikipedia
3. Moodle - Artifical Intelligence Course
4. Multiple constraint satisfaction problems using the A-star search algorithm
5. Courses timetabling based on hill climbing algorithm