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 models ability to identify individuals at risk of stroke. Additionally, we employ a salary prediction dataset to evaluate the models 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

2.1 Datasets attributes description

The initial and crucial step in developing any machine learning algorithm involves a thorough understanding of the data it will be trained on. This understanding is achieved through a comprehensive analysis of the datasets' characteristics. In this vein, the following sub sections will delve into the specific attributes of the two datasets employed in this study: stroke prediction and salary prediction.

A detailed description of each salary prediction attribute is provided in Table 1 and of each stroke prediction attribute in Table 2.

List of all attributes in the Salary Prediction dataset		
Attribute name	Type	Details
fnl	numeric	Socio-economic characteristic of the population from which the individual comes
hpw	numeric	Number of work hours per week
relation	categorical	The type of relationship in which the individual is involved
gain	numeric	Capital gain
country	categorical	Country of origin
job	categorical	The individual's job
edu_int	numeric	Number of years of study
years	numeric	Age of the individual
loss	numeric	Loss of capital
work_type	categorical	The job's type
partner	categorical	The type of partner the individual has
edu	categorical	The individual's type of education
gender	categorical	Individual's gender
race	categorical	Individual's race
prod	numeric	Capital production
gtype	categorical	Type of employment contract

Table 1: Salary Prediction Attributes

List of all attributes in the Stroke Prediction dataset			
Attribute name	Type	Details	Possible values
mean_blood_sugar_level	numeric	The average value of blood glucose throughout the duration observation of the subject	
cardiovascular_issues	categorical	Whether or not the subject has a medical history cardiovascular	0, 1

job_category	categorical	The field in which the person works	child, entrepreneurial, N_work_history, private_sector, public_sector
body_mass_indicator	numeric	Body mass index, which indicates if the person is underweight, within limits normal, overweight or obese	
sex	categorical	The gender of the person	F, M
tobacco_usage	categorical	Current or past smoker indicator	ex-smoker, smoker, non-smoker
high_blood_pressure	categorical	Binary attribute indicating whether a person suffer from high blood pressure or not	0, 1
married	categorical	Binary attribute indicating whether the person a ever been married	Y, N
living_area	categorical	The type of area where he lived most of his life	City, Countryside
years_old	numeric	The person's age in years	
chaotic_sleep	categorical	Binary attribute for a sleep program irregular	0, 1
analysis_results	numeric	The results of medical analyzes of the person, which may include various measurements and indicators relevant to her health	

biological_age_index	numeric	An index that estimates the biological age of a person based on different factors such as lifestyle, health status, measured in an unknown unit	
cerebrovascular_ accident	categorical	Binary indicator indicating whether the person a had a stroke or not	0, 1

Table 2: Stroke Prediction Attributes

2.2 Exploration of Attribute Types and Value Ranges

Prior to applyinga machine learning model to a dataset, a crucial step involves in identifying the types of attributes (features) present and their corresponding values ranges. This anaysis is essential for selecting appropriate algorithms and ensuring optimalmodel performance. In the following pharagraphs we will describe three primary attribute types.

- *Continuous Numeric Attributes:* These attributes possess numerical values that can theoretically take on any value within a specific range. Examples might include: age, weight, temperature etc.
- *Discrete Nominal Attributes:* These attributes represent categorical data with distinct, non-ordered values. Examples include days of the week (Monday, Tuesday, etc.) or types of diseases (cancer, diabetes, etc.).
- *Ordinal Attributes:* These attributes represent categorical data with values that exhibit an inherent order. However, the difference between consecutive values may not be interpretable in terms of a consistent unit. Examples include customer satisfaction ratings (1-star, 2-star, etc.) or movie ratings (G, PG, PG-13, etc.). In ordinal attributes, the numerical value itself might not be as important as the relative order it represents.

Using the *analysys_attributes.py* script, we can identify the Continuous Numeric Attributes and Discrete Nominal Attributes in a specific dataset. The script will output statistics that can be showed in Tables 3 and 5 for numeric attributes and Table 6 and 7 for nominal attributes.

Moreover, the total number of items in the full dataset is 9999 for the Salary Prediction dataset and 5110 for the Stroke Prediction dataset.

List of all Continuous Numeric Attributes in the Salary Prediction dataset							
	fnl	hpw	gain	edu_int	years	loss	prod
count	9.999000e +03	9199.00000	9999.00000	9999.00000	9999.00000	9999.00000	9999.00000
mean	1.903529e +05	40.416241	979.853385	14.262026	38.646865	84.111411	2014.9275 93
std	1.060709e +05	12.517356	7003.7953 82	24.770835	13.745101	3394.0354 84	14007.6044 96
min	1.921400e +04	1.000000	0.000000	1.000000	17.000000	0.000000	-28.000000
25%	1.182825e +05	40.000000	0.000000	9.000000	28.000000	0.000000	42.000000
50%	1.784720e +05	40.000000	0.000000	10.000000	37.000000	0.000000	57.000000
75%	2.373110e +05	45.000000	0.000000	13.000000	48.000000	0.000000	77.000000
max	1.455435e +06	99.00000	99999.0000	206.000000	90.000000	3770.00000	200125.000

Table 3: Continuous Numeric Attributes in Salary Prediction Dataset

List of all Continuous Numeric Attributes in the Stroke Prediction dataset					
	mean_blood_ sugar_level	cardiovascular_ issues	body_mass_ indicator	high_ blood_pressure	years_old
count	5110.000000	5110.000000	4909.000000	5110.000000	5110.000000
mean	106.147677	0.054012	28.893237	0.097456	46.568665
std	45.283560	0.226063	7.854067	0.296607	26.593912
min	55.120000	0.000000	10.300000	0.000000	0.080000
25%	77.245000	0.000000	23.500000	0.000000	26.000000
50%	91.885000	0.000000	28.100000	0.000000	47.000000
75%	114.090000	0.000000	33.100000	0.000000	63.750000
max	271.740000	1.000000	97.600000	1.000000	134.000000

List of all Continuous Numeric Attributes in the Stroke Prediction dataset				
	chaotic_sleep	analysis_ results	biological_ age_index	cerebrovascular_ accident
count	5110.000000	4599.000000	5110.000000	5110.000000
mean	0.054012	323.523446	134.784256	0.048728

std	0.226063	101.577442	50.399352	0.215320
min	0.000000	104.829714	-15.109456	0.000000
25%	0.000000	254.646209	96.710581	0.000000
50%	0.000000	301.031628	136.374631	0.000000
75%	0.000000	362.822769	172.507322	0.000000
max	1.000000	756.807975	266.986321	1.000000

Table 5: Continuous Numeric Attributes in Stroke Prediction Dataset

List of all Discrete Nominal Attributes in the Salary Prediction dataset		
	Non-missing count	Unique values count
country	9999	41
job	9999	14
work_type	9999	9
partner	9999	7
edu	9999	16
gender	9199	2
race	9999	5
gtype	9999	2
money	9999	2

Table 6: Discrete Nominal Attributes in Salary Prediction Dataset

List of all Discrete Nominal Attributes in the Stroke Prediction dataset		
	Non-missing count	Unique values count
job_category	5110	5
sex	5110	2
tobacco_usage	5110	4
married	4599	2
living_area	5110	2

Table 7: Discrete Nominal Attributes in Stroke Prediction Dataset

2.3 Investigation of Class Distribution

TODO

2.4 Analysis of Feature Correlations

TODO

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