UNIMINUTO
Corporación Universitaria Minuto de Dios
Educación de Calidad al alcance de todos
Vigilada MinEducación

NRC 407 Machine Learning
Week 2

José Ignacio Gutiérrez Guzmán

Professor: Claudia Mosquera

Selecting the correct Dataset.

For the dataset selection, some possible topics were considered, including health, sports, and the pandemic. However, given the personal focus on software development, an interesting survey conducted by a well-known platform was interesting. The survey collected data from developers worldwide, providing insights into their current conditions, most commonly used programming languages, years of professional experience, salaries, and more. Some fields contain multiple data points, making it even more interesting to analyze.

Dataset given:

https://www.kaggle.com/datasets/dheemanthbhat/stack-overflow-annual-developer-survey-2022

You can see a copy of this procedure in next GitHub repo URL and run as many times as you want:

https://github.com/joseiguti/machinelearning/blob/master/semana2/dataset1.ipynb

Next, we will see selected columns and explain them.

Column	Description	Will be processed
Employment	Current employment status	
Age	Range age	*
Gender	Gender	
Ethnicity	Ethnicity	
RemoteWork	Work situation	
CodingActivities	Coding activities	
EdLevel	Education level	
YearsCode	Years of coding	*
YearsCodePro	Years of coding in professional way	*
DevType	Developer type	
OrgSize	Organization size where developer works	*
Country	Country	
ConvertedCompYearly	Annual salary	*
LanguageHaveWorkedWith	Programming languages worked with	
DatabaseHaveWorkedWith	Databases have worked with	
PlatformHaveWorkedWith	Platform have worked with	
WebframeHaveWorkedWith	Frameworks have worked with	
OpSysProfessional use	Operative system have worked with	

How we process data and convert into results?

```
# Import enough libraries for initialize de dataset.
import pandas as pd
import re

# Load the data from the csv file.
df = pd.read_csv('https://joseiguti.com/machinelearning/survey_results_public.csv’)

# How many rows and columns we have
print("Rows:", df.shape[0])
print("Columns:", df.shape[1])
```

```
Rows: 73268
Columns: 79
```

In some cases was necessary to clean and prepare the data in columns, for example, because data stored in column **Age** is a String like “25-34 years old”, we made a function that expects a string and returns the average between the first and second one number. In this case the result returned would be (30) rounding to up. So due is not possible to know the current developer age due survey limits, we assumed that like an average.

```
def extract_age(text):
    pattern = r'(\d+)-(\d+) years old'
    match = re.match(pattern, str(text))
    if match:
        start_age = int(match.group(1))
        end_age = int(match.group(2))
        return (start_age + end_age) // 2
    return None
```

```
# Son we apply the function to extract the age
df['Age'] = df['Age'].apply(extract_age)

# We convert the age to int value
df['Age'] = df['Age'].astype(int)

# We ensure about the changes and print it
datos_agrupados = df['Age'].unique()
```

Now, finally we expose the final results of the operations. But before that, we ensure only numeric data will be processed. Let's see.

```
data_numeric = df.select_dtypes(include='number')
```

```
# Finally we print the media data
media =
data_numeric.mean(numeric_o
nly=True)
print(media)
```

```
Age 30
YearsCode 12
YearsCodePro 7
OrgSize 1488
ConvertedCompYearly 170761
```

```
media = data_numeric.median()
print(media)
```

```
Age 29
YearsCode 9
YearsCodePro 4
OrgSize 59
ConvertedCompYearly 67845
```

```
media = data_numeric.mode()
print(media)
```

```
Age 29
YearsCode 10
YearsCodePro  0
OrgSize  0
ConvertedCompYearly 150000
```

Conclusions:

- Technically is possible to manipulate String data and set the behavior like a number, then get the expected results.
- Deleting records with missing data is not always necessary. Another strategy can be employed to replace them to avoid delete important data related in the record.
- Before to decide to make operations in a specific column where a simple sight their data is numeric, is better check the unique values to eval if is there any uncommon value and set it in other compatible value.

"OpenAI. (2021). ChatGPT [Modelo de lenguaje generativo]. Recuperado de https://openai.com"