Machine Learning – Assignment 3

Due Date: 29.06.2023

Data

You have been provided with data regarding human resources analytics.
The data provides insights into employee-related factors and their impact on organizational dynamics, with features such as employee satisfaction levels, performance evaluations, turnover rates, and more.
By exploring this dataset, we can gain an understanding of the relationships between various employee attributes and uncover patterns that can shape strategies for enhancing employee satisfaction, reducing turnover, and optimizing organizational performance.

The variables description can be found in the **data-variables-description**.pdf file.

Section A (Data Exploration and Pre-processing) *15 pts*

1. Explore the data using tables, visualizations, and other relevant methods (use at least 3 different types of graphs).
   Your graphs should answer the following questions:
   - What is the overall diversity profile of the organization?
   - Is there any relationship between who a person works for and their performance score?
   - Are there areas of the company where pay is not equitable?

   Except for these questions add at **least 2 more** visualizations of your own that shows interesting insights on the data.
2. Apply different methods of pre-processing to the data in order to prepare it for the models you wish to apply in the next sections.

- Provide an explanation to each method you apply. Your choice should reflect an understanding of the method and why it's needed.

Section B (Dimensionality Reduction) *35 pts*

1. Apply PCA algorithm on the data.
2. Create a scatter plot for the new data and color each observation according to the employment status of an employee. Describe your findings (which employees are most similar to one another).
3. Using the PCA principal components that explain the majority of the variance in the data, Identify the features that are most strongly represented in each component (in absolute values). Show it visually.
   - Which features are the **most effective** to separate the employees by their employment status?
     - Explain your findings.
     - Present a resulting clusters visually.
   - Which features are the **least effective** to separate the employees?
     - Explain your findings.
     - Present a resulting clusters visually.
   - There is a specific feature that has the most effect on the separation of the different employment status groups. Can you identify it? Explain your findings, and apply again PCA, but this time without using this feature in your dataset.

     - Explain what changed in your results, and how the separation was affected.

     - Explore and show visually which features now are the most effective to separation.
4. Create a biplot using the first two principal components (PC1 and PC2).

Interpret the biplot: examine the position of the data points in the biplot and their relationships to the variables (features).

Interpret the biplot by considering the following aspects: proximity of data points, angle and direction of vectors, variables' contribution, etc.

5. **Bonus** *5 pts* - Using the PCA, find outliers in your dataset. Print a list of the outliers and explain how you found them and if they have something in common.

Section C (Classification) *25 pts*

Predicting the employment status of the employees (Active / Voluntarily Terminated / Terminated for Cause) - In this section, your goal is to build a predictive model to determine which employees are likely to terminate their employment and which employees are likely to stay.

- Apply **SVM** and at least 2 more machine learning algorithms
  - Provide feature importance for each model (if possible) and explain if the important features make sense.
  - The implementation should include parameter tuning.
  - Compute accuracy for each model and provide sensitivity and specificity measurements for every class in each model. Is there a class in which one of the measurements is relatively low? If so, explain why, and implement a suitable method to improve the results and explain it.
  - Compare the performance results between the models.

Section D (Regression) *25 pts*

Predicting time until employee termination from last satisfaction survey - In this section, your goal is to build a regression model to predict the time (in days) that will pass for an employee to terminate from their last survey of satisfaction.

- Apply at least 3 different machine learning algorithms.

- Provide feature importance for each model and explain if the important features make sense.
- The implementation should include parameter tuning.
- Report suitable measurements to evaluate the performance of each model and compare the results.

- Post process the results of your predictions and find out what will be the exact termination date for each employee.

Section E (Bonus) *15 pts*

- Calculate the Employee Retention Rate for each year from 2008 until 2017 for every recruitment source and display it on a suitable graph.
- Find a suitable measurement to calculate the diversity index for each department based on race, gender and age and display the department in descending order of their diversity index.
- Create a map visualization that shows the number of employees that currently work in the company from each state. The map should display the state ID and the number of employees, color the state by the intensity of the number of employees.

Section F (Performance - Bonus) *5 pts*

Machine learning models that outperformed other students' models for either the classification or regression tasks may get additional points as long as the non-standard methodology to obtain superior results is also explained.

• In order to get the bonus points you may want to apply multiple performance measures to ensure that we can compare your performance on an equal basis to other

projects, and that you did not sacrifice performance in a specific measure to outperform in another.

Submission

- The assignment should be submitted in pairs (only one submission).
- You are required to submit two files including sections A-F. One in **.ipynb** format and one in **.html**. Both files should also include the program's outputs.
- The files' names should be of the form: **ML_HW3_#ID1_#ID2**.
- Assignments submitted late will receive a penalty of **3 points** for each day, up to 5 days. Later submissions will not be accepted.