Compulsory 2 – ML

Report: Predicting Employee Salaries and Job Satisfaction Using Regression Models

Introduction:

This report analyzes the "Salary&JobSatisfaction Dataset.txt" dataset to predict employee salaries using linear regression and classify job satisfaction using logistic regression. The dataset contains 1000 entries with features such as Years_of_Experience, Education_Level, Job_Complexity, Work_Hours_Per_Week, Company_Size, Salary (Euros), Buyer_Income (Euros), and Job_Satisfaction. In this task I preprocessed and analyzed the data and then trained and evaluated two machine learning models.

Data Preprocessing and Analysis

Initial Steps

The dataset was loaded into a Pandas DataFrame using pd.read_table() with a comma delimiter, as the .txt file structure was comma-separated. The first few rows showed there was a mix of numerical (Years_of_Experience, Work_Hours_Per_Week, Salary (Euros), Buyer_Income (Euros)) and categorical (Education_Level, Job_Complexity, Company_Size, Job_Satisfaction) variables.

```
# Reads the .txt file and separates by commas
sjs = pd.read_table('Salary&JobSatisfaction Dataset.txt', delimiter=',')
# Makes a dataframe of the data
sjs_df = pd.DataFrame(sjs)
```

Missing Values

A check for missing values using sjs_df.isnull().sum() showed no missing entries across all columns, eliminating the need for imputation or row deletion. This simplifies preprocessing and ensures all data points contribute to model training.

```
# Checks for missing values in the dataset
print("Missing Values:\n", sjs.isnull().sum())
```

Feature Exploration and Encoding

To assess linear relationships with Salary (Euros) for linear regression, scatter plots were used for continuous variables like Years_of_Experience and Work_Hours_Per_Week, while boxplots were used for discrete variables (Education_Level, Job_Complexity, Company_Size). These visualizations help

determine feature relevance. For instance, a positive linear trend between Years_of_Experience and salary is expected, while categorical features show salary variations across different levels.

Encoding and Scaling:

Features were standardized using StandardScaler to ensure that all variables have a mean of 0 and a standard deviation of 1, this places all the features on a comparable scale.

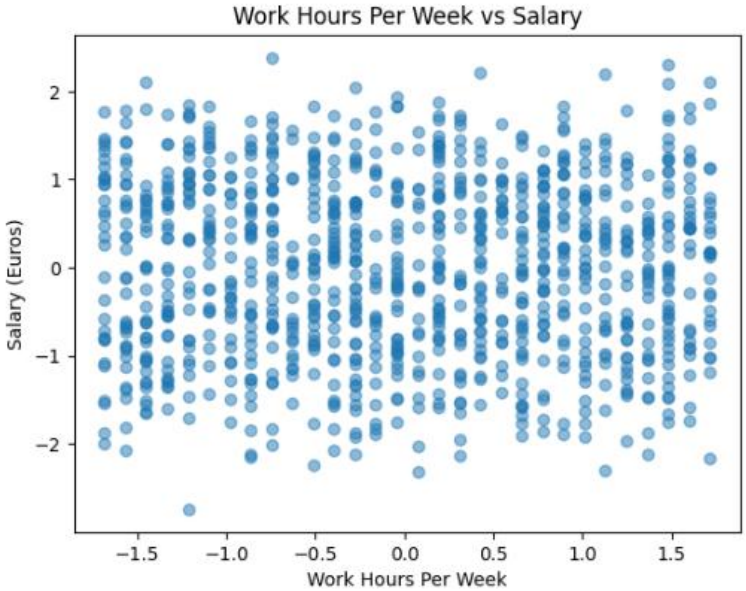
Categorical variables (Company_Size, Job_Satisfaction) were encoded using the map() function from Pandas to convert them into numerical format, so that they can be better used by the models.

```
# Define the numerical features
num_features = ['Years_of_Experience', 'Work_Hours_Per_Week', 'Buyer_Income (Euros)', 'Salary (Euros)']
# Scale the numerical features
scaler = StandardScaler()
sjs_df[num_features] = scaler.fit_transform(sjs_df[num_features])
# Encode the categorical features using map()
company_size_mapping = {'Small': 0, 'Medium': 1, 'Large': 2}
sjs_df['Company_Size'] = sjs_df['Company_Size'].map(company_size_mapping)
job_satisfaction_mapping = {'No': 0, 'Yes': 1}
sjs_df['Job_Satisfaction'] = sjs_df['Job_Satisfaction'].map(job_satisfaction_mapping)
```

Model Training and Evaluation

1 - Linear Regression: Salary Prediction

Features Selected: Years_of_Experience, Education_Level, Job_Complexity, Buyer_Income (Euros), Company_Size and Job_Satisfaction. Work_Hours_Per_Week was excluded as I didn't find a linear relationship to Salary (Euros).



Process:

Data was split into 90% training and 10% testing sets using train_test_split (random_state=42), I chose 9:1 ratio because the dataset is quite small and this ratio produced the best results, I found.

A LinearRegression model was trained on the scaled training data.

Predictions were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Performance:

- MAE: 0.11 (indicating an average prediction error of 0.11 units in standardized salary).
- MSE: 0.02.
- RMSE: 0.14.
- R^2: 0.98 (suggesting 98% of salary variance explained).

2 - Logistic Regression: Job Satisfaction Classification

Features Selected: Years_of_Experience, Education_Level, Job_Complexity, Buyer_Income (Euros), Company_Size and Salary (Euros).

Process:

Data was split and scaled similarly.

A LogisticRegression model was trained on the scaled training data.

Performance:

- Training Accuracy: 87.11%.
- Testing Accuracy: 86.00%.
- Cross-Validation Accuracy: 0.87 ± 0.02.
- Classification Report:
- Class 0 (No): Precision 0.93, Recall 0.79, F1-score 0.85.
- Class 1 (Yes): Precision 0.80, Recall 0.94, F1-score 0.87.
- Overall Accuracy: 86%.
- Confusion Matrix (Testing): [[41, 11], [3, 45]] (indicating 11 false positives and 3 false negatives).

Feature Importance:

- Salary (Euros): 5.40 (strong positive predictor).
- Years_of_Experience: 1.04 (positive).
- Work_Hours_Per_Week: -0.06 (negative).
- Buyer_Income (Euros): -6.86 (strong negative predictor).

Discussion of Model Performance

Linear Regression

The R^2 of 0.98 indicates an excellent fit, capturing almost all salary variance. The low MAE (0.11) and RMSE (0.14) suggest high prediction accuracy, with errors being minimal in the standardized scale. This performance implies that the features used Level are highly predictive, though the linear assumption may still miss minor nonlinear effects.

Logistic Regression

The 86% testing accuracy is decent (without the random_state set to 42 it varied from 86% to 91%),it also got (0.87 ± 0.02) on cross-validation indicating consistency. The classification report shows a higher recall for "Yes" (0.94) but lower precision (0.80), suggesting the model is better at identifying satisfied employees but occasionally misclassifies dissatisfied ones (11 false positives). The feature importance shows Salary (Euros) as a strong driver of satisfaction.

Challenges

Linear models may miss complex interactions between features.

The models' ability to generalize to broader populations may be limited due to the dataset containing only 1000 entries. Furthermore, some features weren't very relevant to part 2 of this task and restrict deeper insights (no data on work-life balance for example).

Findings

Salary Prediction: The linear regression model excels ($R^2 = 0.98$), with minimal errors (MAE = 0.11), indicating strong predictive power of features like experience and education.

Job Satisfaction: The logistic regression model achieves 86% accuracy, with salary positively influencing satisfaction.

Insights

Salary is a key driver of job satisfaction, suggesting employers should focus on compensation.

The high R^2 for salary prediction underscores the importance of experience and education in determining pay.

Future work could use nonlinear models (decision trees for example) to capture interactions and improve satisfaction prediction.