AI project report

Done by:

1- Ali Ahmed 20201701725

2-Ahmed Medhat 20201701703

3-Mina George 20201701741

4- Diaaeldeen Amr 20201701720

5-Mohamed Ibrahim 20201701732

6-Mohamed Reda 20201701733

Pre-processing techniques

1-Label encoding: refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

```
from sklearn import preprocessing
label = preprocessing.LabelEncoder()

df['work-class']= label.fit_transform(df['work-class'])
df['marital-status']= label.fit_transform(df['marital-status'])
df['position']= label.fit_transform(df['position'])
df['relationship']= label.fit_transform(df['relationship'])
df['race']= label.fit_transform(df['race'])
df['sex']= label.fit_transform(df['sex'])
df['native-country']= label.fit_transform(df['native-country'])
df['education']= label.fit_transform(df['education'])
df2['workclass']= label.fit_transform(df2['workclass'])
df2['marital-status']= label.fit_transform(df2['marital-status'])
df2['occupation']= label.fit_transform(df2['occupation'])
df2['relationship']= label.fit_transform(df2['relationship'])
df2['race']= label.fit_transform(df2['race'])
df2['sex']= label.fit_transform(df2['sex'])
df2['native-country']= label.fit_transform(df2['native-country'])
df2['education']= label.fit_transform(df2['education'])
```

2-Standard scaler: it's an important technique that is mainly performed as a preprocessing step before many machine learning models, in order to standardize the range of functionality of the input dataset.

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train = pd.DataFrame(scaler.fit_transform(X_train), columns = X.columns)

X_test = pd.DataFrame(scaler.transform(X_test), columns = df2.columns)
```

3- Replace: This method replaces the specified value with another specified value and searches the entire Data Frame and replaces every case of the specified value.

```
df['position'] = df['position'].replace([' ?'], 'Prof-specialty')
df['native-country'] = df['native-country'].replace([' ?'], 'United-States')
df['work-class'] = df['work-class'].replace([' ?'], 'private')
df2['occupation'] = df2['occupation'].replace([' ?'], ' Machine-op-inspct')
df2['native-country'] = df2['native-country'].replace([' ?'], 'United-States')
df2['workclass'] = df2['workclass'].replace([' ?'], 'private')
```

4-Data splitting: is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model. Data splitting is an important aspect of data science, particularly for creating models based on data.

```
X = df.drop("salary", axis = 1).copy()
y = df["salary"].copy()

# X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3,random_state=101)

X_train = X
X_test = df2.copy()
y_train = y
```

Analysis on the dataset

1- **Correlation is the statistical measure of the relationship between two variables. There are different types of correlation coefficients like Pearson coefficient (linear) and Spearman coefficient (non-linear) which capture different degrees of probabilistic dependence but not necessarily causation.**

2- **The drop() function is used to drop specified labels from rows or columns. Remove rows or columns by specifying label names and corresponding axis, or by specifying directly index or column names. When using a multi-index, labels on different levels can be removed by specifying the level.**

- **Drop (`work-fnl`) `column because the` correlation is small**
- **Drop (`fnlwgt`) `column because` of the same preprocessing of train data**

```
sns.clustermap(df.corr(), cmap = "vlag", dendrogram_ratio = (0.1, 0.2), annot
= True, linewidths = .8, figsize = (9,10))
plt.show()
################## Train data ######################
df.drop("work-fnl", axis=1, inplace=True)
#################### Test data ####################
df2.drop("fnlwgt", axis=1, inplace=True)
```

Size of the dataset

- **Before pre-processing**

```
Train dataset(22792, 15)
Test  dataset (9769, 14)
```

- **After pre-processing**

```
Train dataset(22792, 14)
Test  dataset (9769, 13)
```

Accuracy of each model

- **GuassianNB : 80%**
- **Random Forest: 85%**
- **Svm : 80%**
- **Logistic Regression : 83%**
- **Decision Tree: 81%**
- **MLP: 84%**
- **Cat Boost : 87%**

Conclusion

The way we live and work is being changed by AI. We use AI every day, whether it's by asking Siri for directions, getting movie recommendations from Netflix, or having drones deliver products to customers. Using AI subsets like machine learning, deep learning, and natural language processing, businesses are developing novel solutions.

The design and development of algorithms that are able to learn from and make predictions based on data is the subject of the artificial intelligence field known as machine learning. The creation of algorithms that can automatically improve with additional data is the goal of machine learning.

Data preprocessing is a method for transforming raw data into a clean set of data. To put it another way, the raw format in which the data are collected from various sources makes it impossible to conduct analysis. It is used in Machine Learning projects to improve results from the model used. The data must be formatted correctly. The Random Forest algorithm, for instance, does not support null values; consequently, null values must be managed from the initial raw data set in order for the Random Forest algorithm to be executed. Another consideration is that the data set ought to be formatted in such a way that multiple Machine Learning and Deep Learning algorithms can be run on it, and the most effective one will be selected.

In research, correlation analysis is a statistical technique for calculating and measuring the strength of the linear relationship between two variables. In a nutshell, correlation analysis determines the degree to which one variable is affected by another. A high correlation indicates a strong connection between the two variables, whereas a low correlation indicates a weak connection.

A model of machine learning can be thought of as a computer program that has been trained to spot patterns in new data and make predictions. A mathematical function that takes requests in the form of input data, makes predictions based on those predictions, and then produces an output is how these models are represented. After being trained on a set of data, these models are given an algorithm to reason over data, find patterns in feed data, and learn from those patterns. These models can be used to predict the unknown dataset once they have been trained

References

- https://www.geeksforgeeks.org/
- https://thecleverprogrammer.com/
- https://www.w3schools.com/default.asp
- https://c3.ai/
- https://www.w3resource.com/index.php
- https://www.techtarget.com/searchenterpriseai/
- https://middle-east.alibabacloud.com/
- https://www.educanada.ca/index.aspx?lang=eng
- https://www.questionpro.com/