metin, saat içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Python Foundations**

**Instructor: Atlas Khan**

**Teaching Assistant: Nail Şenbaş**

**Group 3 Members: Tinashe Hafe, M. Yusuf Kansiz**

**Project Title: Predicting Loan Eligibility**

Contents

[Abstract 2](#_Toc95153082)

[1. Introduction 3](#_Toc95153083)

[2. PROJECT 4](#_Toc95153084)

[2.1. Methodology 4](#_Toc95153085)

[2.2. Getting the system ready and loading the data 4](#_Toc95153086)

[2.3. Data Exploration 5](#_Toc95153087)

[2.4. Raw Data visualization 5](#_Toc95153088)

[2.5. Data cleaning 11](#_Toc95153089)

[2.6. Machine learning discussion 11](#_Toc95153090)

[2.6.1. Logistic Regression 11](#_Toc95153091)

[2.6.2. Decision Tree 12](#_Toc95153092)

[2.6.3. Random Forest 13](#_Toc95153093)

[3. Conclusion 14](#_Toc95153094)

[4. References 15](#_Toc95153095)

**List of Figures**

[Figure 1 Loan Prediciton Architecture 3](#_Toc95153120)

[Figure 2 - Graph – Loan Status 6](#_Toc95153121)

[Figure 3- Graph - Gender Analysis 7](#_Toc95153122)

[Figure 4- Graphs Dependents Analysis 7](#_Toc95153123)

[Figure 5- Graph Property Area Analysis 7](#_Toc95153124)

[Figure 6- Graph Loan Status Analysis 8](#_Toc95153125)

[Figure 7 - Logistic Regression Block Diagram [7] 12](#_Toc95153126)

[Figure 8- Decision Tree Block Diagram [8] 12](#_Toc95153127)

[Figure 9- Random Forest Block Diagram [9] 13](#_Toc95153128)

[Figure 10- Results Graph 14](#_Toc95153129)

# Abstract

*It has become common knowledge that loan applications are increasing by the day. Choosing the right client can be a cumbersome process. In this project, we proposed a machine learning model that predicts the acceptance and rejection of an applicant. Machine learning model techniques are going to be implemented with the use of the applicant’s information.*

# Introduction

Loans department is a critical part of business of banks. The majority profit comes from the interest generated by loans. The loan corporations approve a loan after thorough analysis of intensive method of verification and validation. Considering all this, they still don’t have guarantee whether the applicant would be able to repay the loan with no challenges which makes Loan Eligibility prediction a very critical process. The main objective of this paper is established effective, fast, immediate method for the meriting candidates This can help financial institutions or lenders by saving time and minimizing the risk of default. Credit Eligibility Estimation System will estimate whether the candidates are eligible for credit by evaluating their features such as income, credit history, age, property area, etc. of each loan requesting candidate.

Figure 1 Loan Prediction Architecture



# PROJECT

## Methodology

The following steps were done

1. Data Loading
2. Data Exploration
3. Data Cleaning
4. Data Preprocessing
5. Training the Dataset
6. Predicting the output
7. Error/Accuracy Checking

The given problem is a supervised classification problem if the applicant is eligible for a loan or not that is Yes or

No.

Initially we dive into the data by doing some analysis and visualization then later on the following algorithms were implemented:

a. Logistic Regression

b. Random Forest

c. Decision Tree

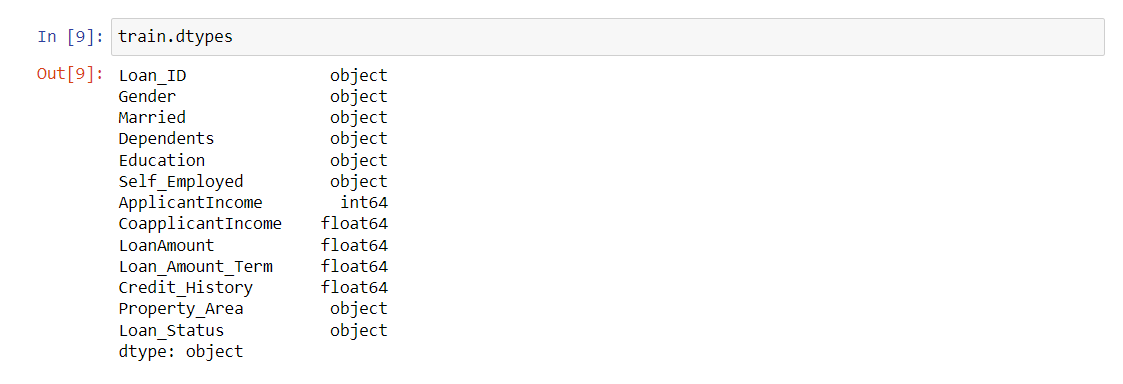
**Specifications:**

Python and the following libraries / packages were used **“Pandas**, **seaborn**, **sklearn, yellowbricks”**

## Getting the system ready and loading the data

In this section, we will explain all the steps taken to get the system ready

* The data is downloaded from GIT HUB
* For this problem, we have three CSV files: train, test and sample submission.
* Train file will be used for training the model, i.e. our model will learn from this file. It contains all the independent variables and the target variable.
* Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data.
* Sample submission file contains the format in which we have to submit out predictions
* Our data set has the following attributes



Picture 1 Code Snippet- Data Types

## Data Exploration

Various libraries were imported for data exploration then we took a quick look on our data contents using train .head() and train. Tail

## Raw Data visualization

We checked the shape of our data

train .shape the output was (614,13) implying that our data has 614 rows and 13 columns

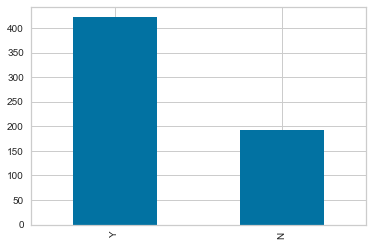
test.shape () the out put was (367,12) also we counted our values:

**Categorical features**: These features have categories (Gender, Married, Self Employed, Credit\_History, Loan\_Status)

**Ordinal features:** Variables in categorical features having some order involved (Dependents, Education, Property\_Area)

**Numerical features:** These features have numerical values (Applicant Income, CoapplicantIncome, Loan Amount, Loan\_Amount\_Term)

Figure 2 - Graph – Loan Status



The bar graph shows that 422 of the loan applicants were accepted which accounts for about 69%

Visualizing the variables separately

Figure 3- Graph - Gender Analyze

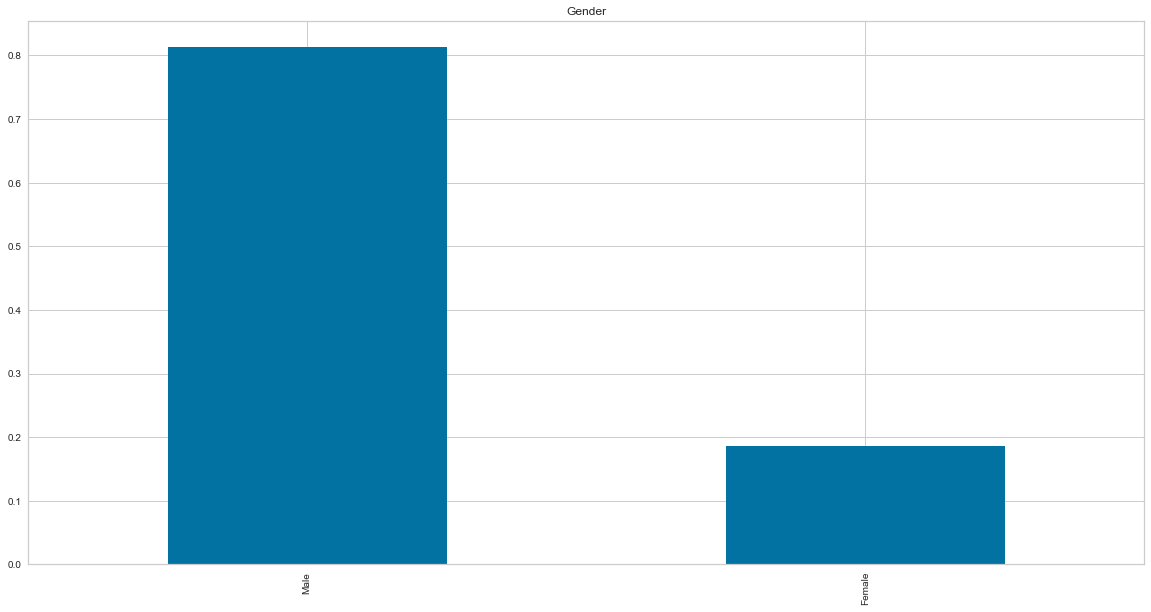


Figure 4- Graphs Dependents Analyze

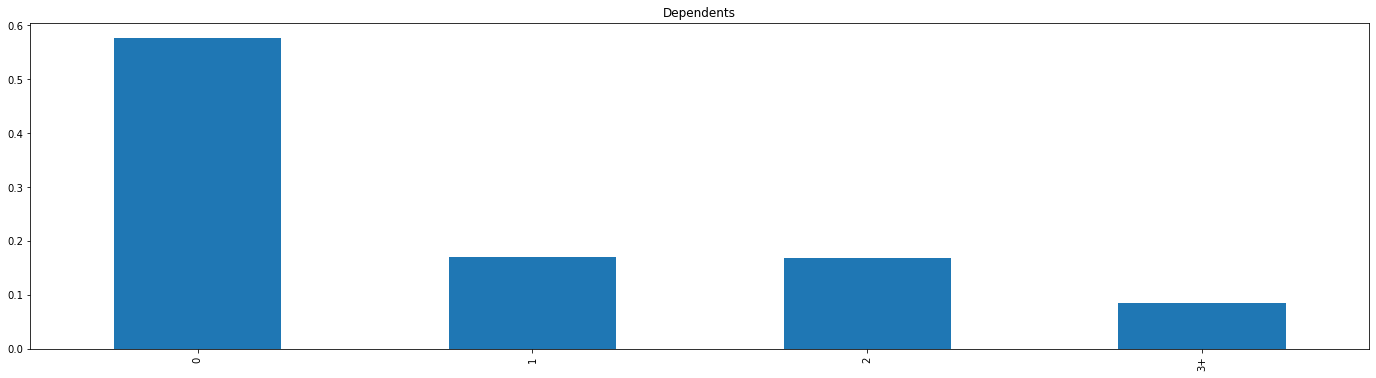


Figure 5- Graph Property Area Analyze

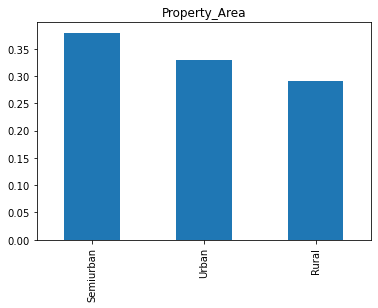
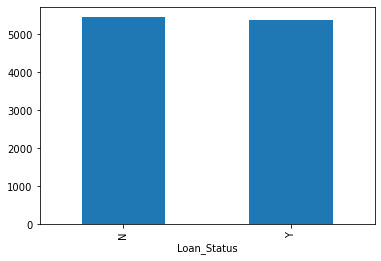
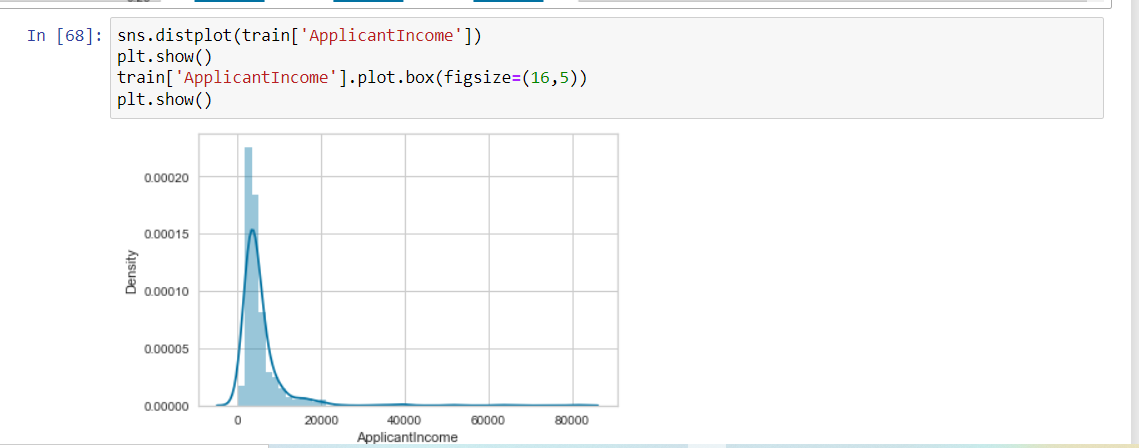


Figure 6- Graph Loan Status Analyze



**Inferences from Visualizations**:

* It can be deduced from the above bar plots that:
* 80% applicants in the dataset are male.
* Approximately 65% of the applicants in the dataset are married.
* Approximately 15% applicants in the dataset are self-employed.
* Approximately 85% applicants have repaid their doubts.

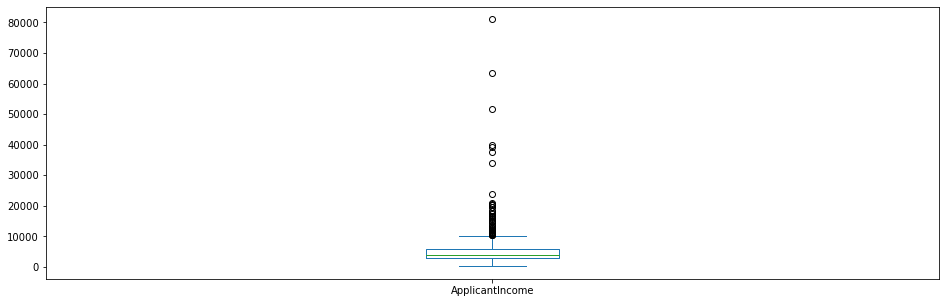


Picture 2 - Code Snippet and Plot

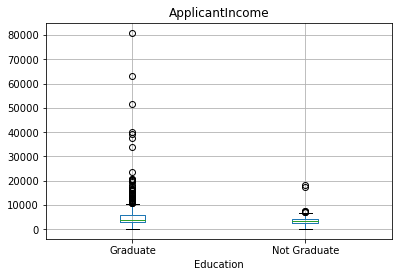
Count function was implemented to have an overview of the applicant income distribution.

It can be observed the distribution of applicant income is towards left which means it is not normally distributed (negatively skewed)

We normalized it later using log transformation



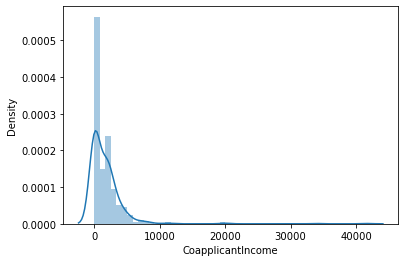
Picture 3- Applicant Income Distribution



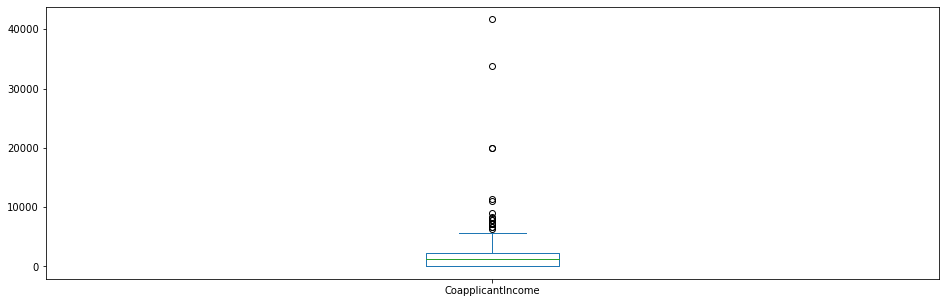
Picture 4- Graph Applicant Income- Education Distribution

The above box plot elaborates how the income is distributed with respective to graduation status

We can deduce that they are a lot of graduates with higher incomes which are the outliers



Picture 5 -Graph Coapplicant Income Distribution



Picture 6- Graph Coapplicant Income Distribution

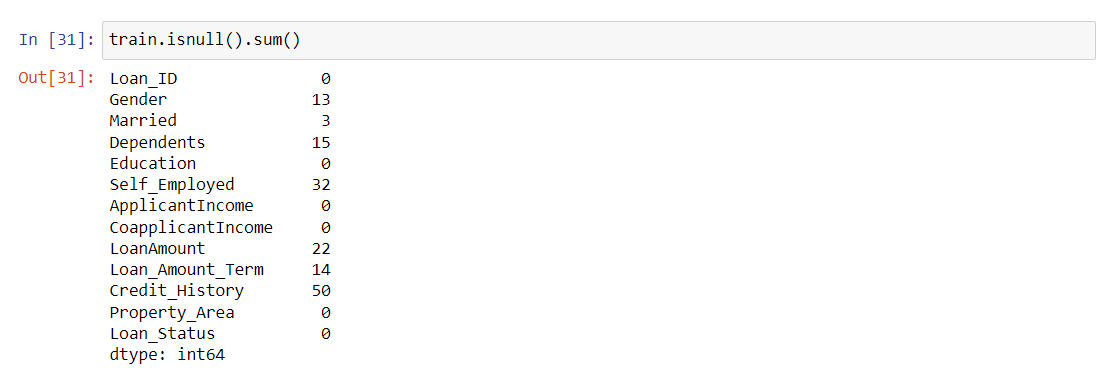
It can be observed that the distribution of co-applicant’s is almost similar to the applicant’s and has some outliers too. Inferences from graphics:

* It can be inferred that the proportion of male and female applicants is more or less the same for both approved and unapproved loans.
* The proportion of married applicants is higher for approved loans.
* Distribution of applicants with 1 or 3+ dependents is similar across both the categories of
* Loan Status.
* There is nothing significant we can infer from Self Employed vs Loan Status plot.

## Data cleaning

We checked if we had missing values from our data using the following function .is.null.sum()

train.is.null.sum ()



Picture 7- Code Snippet - Missing Data

There are missing values in Gender, Married, Dependents, Self\_Employed, Loan Amount, Loan\_Amount\_Term and Credit\_History features. We will treat the missing values in all the features one by one. We can consider these methods to fill the missing values:

* For numerical variables: imputation using mean or median
* In our case we used the mode because we had outliers
* For categorical variables: imputation using mode

## Machine learning discussion

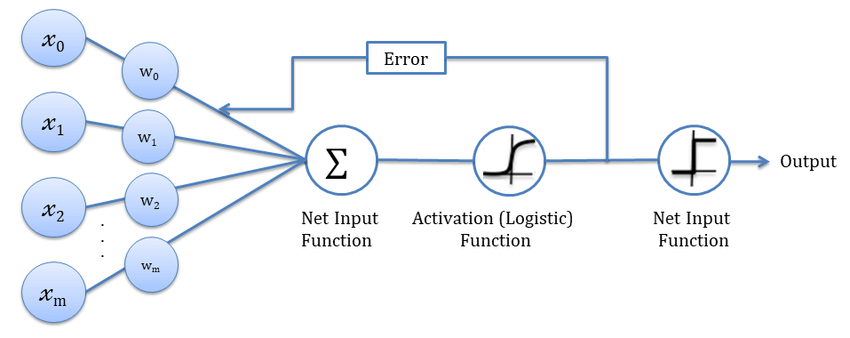
### Logistic Regression

Mean Validation Accuracy 0.775039661554733

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. Logistic regression is an estimation of Logit function. Logit function is simply a log of odds in favor of the event.

This function creates a s-shaped curve with the probability estimate, which is very similar to the required step wise function

Figure 7 - Logistic Regression Block Diagram [7]

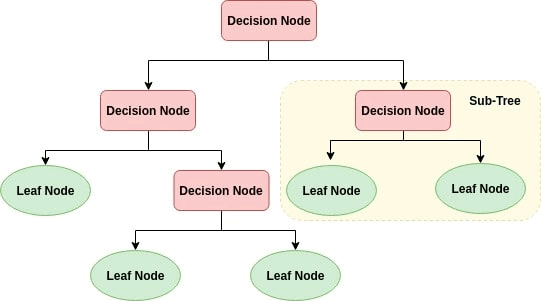


### Decision Tree

Mean Validation Accuracy 0.6941565309360126

Decision tree builds regression or classification models in the form of a tree structure. It **breaks down a dataset into smaller and smaller subsets** while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes.

Figure 8- Decision Tree Block Diagram [8]



As we see in the above picture the node is split into sub-nodes. We can also select the best split point in the decision tree.

### Random Forest

Mean Validation Accuracy 0.8077702252432362

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction

Figure 9- Random Forest Block Diagram [9]

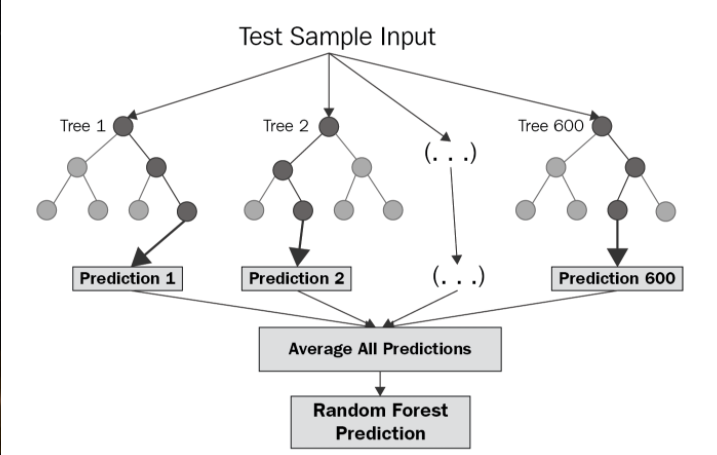
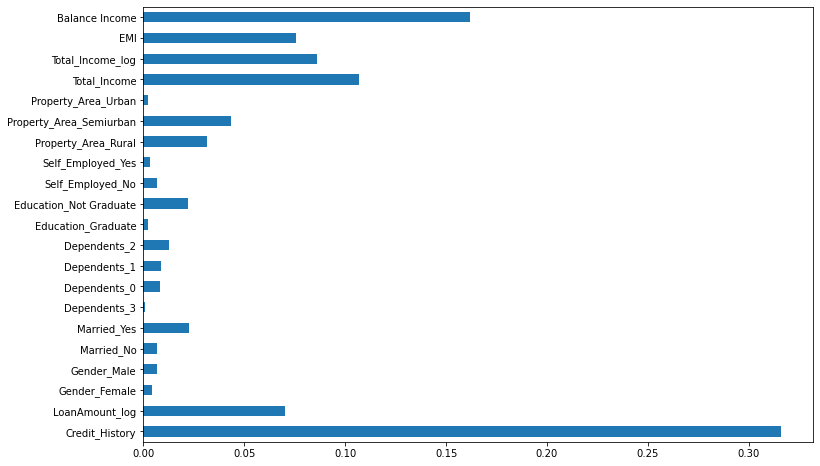


Figure 10- Results Graph



We can observe from the diagram that Credit History is the most important feature followed by Balance Income, Total Income, EMI. So, feature engineering helped us in predicting our target variable.

# Conclusion

Choosing the right loan applicant is and has always been a critical issue for the financial sector.

Now with the new technology that is AI and machine learning models which help in deciding the right or the eligible applicant can be devised so we decided to work on some of them.

In our model by utilizing logistic regression, stratified logistic regression, Random Forest and Decision Tree. Random Forest had a better accuracy model we finally predicted whether or not the loan is approved or not. So, to implement this, numerous input variables were required to get the output.

# References

1. Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. *IOSR J. Comput. Eng.*, *18*(3), 18-21.

2. Goyal, A., & Kaur, R. (2016). A survey on ensemble model for loan prediction. *International Journal of Engineering Trends and Applications (IJETA)*, *3*(1), 32-37.

3. Li, Y. (2019, August). Credit risk prediction based on machine learning methods. In *2019 14th International Conference on Computer Science & Education (ICCSE)* (pp. 1011-1013). IEEE.

4. Abakarim, Y., Lahby, M., & Attioui, A. (2018, November). Towards an efficient real-time approach to loan credit approval using deep learning. In *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)* (pp. 306-313). IEEE.

5. Hamid, A. J., & Ahmed, T. M. (2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal*, *3*(1), 1-9.

6. Bekhet, H. A., & Eletter, S. F. K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, *4*(1), 20-28.

Hossain, M. S., & Arefin, M. S. (2019). Development of an Intelligent Job Recommender System for Freelancers using Client's Feedback Classification and Association Rule Mining Techniques. J. Softw., 14(7), 312-339.

7. Wu, J. (2019). Investigating commuting time patterns of residents living in affordable housing: a case study in Nanjing, China. Promet-Traffic&Transportation, 31(4), 423-433.

8. Çavuşoğlu, Ü. (2019). A new hybrid approach for intrusion detection using machine learning methods. Applied Intelligence, 49(7), 2735-2761.