**Loan Prediction**

**Report**

**Ahmad Shamayl**

Contents

[ABSTRACT 2](#_Toc124896129)

[INTRODUCTION 2](#_Toc124896130)

[Data mining: 2](#_Toc124896131)

[Project: 2](#_Toc124896132)

[Models that we are using: 3](#_Toc124896133)

[Reasons for using this model: 4](#_Toc124896134)

[Versions of Model: 4](#_Toc124896135)

[Related work to our model: 5](#_Toc124896136)

[Results when model is applied: 5](#_Toc124896137)

[PREPROCESSING 6](#_Toc124896138)

[Literature Review: 8](#_Toc124896139)

[Methodology: 9](#_Toc124896140)

[Flow Chart: 11](#_Toc124896141)

[Refrences: 13](#_Toc124896142)

# **ABSTRACT**

Data Mining also known as knowledge discovery in data (KDD) is a process of discovering patterns and valuable information from huge amount of data sets. In this project we are making a classification Machine Learning model, to predict whether the loan should be given to some person or not based on previous data. The model we have chosen is Support Vector Machine (SVM) and Logistic Regression using jupyter notebook with python language. The dataset we have used is Loan Prediction. We are also doing some preprocessing in dataset to clean it and normalize it with some python built-in modules and libraries.

# **INTRODUCTION**

## **Data mining:**

DataMiningalso known as knowledge discovery in data (KDD) is a process of discovering patterns and valuable information from huge amount of data sets. In data mining we find the anomalies and correlations within huge amounts of data to make predictions or predict outcomes.

## **Project:**

We are creating a classification Machine Learning model, so that we can classify who should be given loan and who should not. The model we have chosen is Support Vector Machine (SVM) and Logistic Regression. The dataset we have used is Loan Prediction.

## **Models that we are using:**

* Support Vector Machine
* Logistic Regression

1. **Support Vector Machine (SVM):**

Support Vector Machine (SVM) is a quick and reliable classification algorithm accomplishes very well and easily with finite amount of data to analyze. The main agenda behind SVM algorithm is simple and applying on Natural Processing Language (NLP) doesn’t require the complicated stuff. It has a decision line or also known hyper plane divides two classes which is also called decision boundaries.

We use kernel for transforming our data from 1 dimension to another.

1. **Logistic Regression:**

Logistic Regression is an analytical algorithm to predict binary outcomes such as, (0,1) or (Yes, No) based on the previous observations of the data. There is a predefined threshold in logistic regression which is usually 0.5 that if the value is below 0.5, the model will predict it as 0 or No class and if the value is greater than or equal to 0.5 then the model will predict it as 1 or Yes

## **Reasons for using this model:**

It is a classification model and Support Vector Machine is one of the best models for it, but we are also using Logistic Regression model to show the difference of why Support Vector Machine (SVM) is one of the best models for both classification and regression.

## **Versions of Model:**

There are two types of **SVM model** which is used for basics classification and regression whereas the other version of SVM is used with kernel, it is the mathematical function which transforms the original data to a greater number of dimensions, it is more tensile for non-linear data.

In our project we are using SVM with kernel.

There are three types of **Logistic Regression** model, which are binary, nominal and ordinal. In binary classification we have to predict only two classes either 1(YES/TRUE) or 0(NO/FALSE), in nominal we have two or more than two states it is the generalization of binary attributes. Ordinal classification is used to predict the ordinal data, a data whose attribute’s values depends on arbitrary scale.

In our project, we are using Binary Logistic Regression as we have to predict only two classes whether the loan is given or not.

# **Related work to our model:**

Support vector machine has been used in the following things:

1. Face Detection:

It classifies parts of the face images and non-face images to create a boundary around the face.

1. Bioinformatics:

In this SVM is used to identify the classes of genes of patients by using genes and biological problems. This is also used for cancer detection.

1. Handwriting recognition:

In thus SVM is used to predict the hand writing by giving it previous data of widely used handwriting.

# **Results when model is applied:**

**DATASET**

Insurance companies uses the largest amount of data handling/analytics or data science methods. This dataset provides a part of the working of the insurance companies or organizations. It defines what kind of backlashes they face and the different strategies used by them to overcome them. Also, it shows which external or internal factors affect the outcomes, they can be both positive and negative. It is a classification problem, containing 615 rows and 13 columns.

In this dataset using customer details, we want to automate the loan approval criteria. There are different factors like Education, Loan amount, Gender, etc. In this dataset we are given a query to identify the customers that are eligible for the requested loan amount to automate the process.

# **PREPROCESSING**

Data preprocessing is a process in which data manipulation is done through various tasks, these tasks or steps may include:

* Data cleaning
* Data reduction
* Data integration
* Data transformation (Standardization or Normalization)

**Data Cleaning:**

In this we have to get rid of missing values, smoothing the noisy data and to make the data consistent

* data.isnull().sum()
* new\_data = data.dropna()
* new\_data.shape
* new\_data.isnull().sum()

**Data Reduction:**

In data reduction we can delete those attributes which are not important for our model for prediction, this process is called Dimensionality Reduction and it can be done through Stepwise algorithm

There is also one more thing numerosity reduction in which we can add alternatives of data to make it more correlated

* again = new\_data.replace({'Loan\_Status': {'Y':1 , 'N':0}})
* Loan\_predicting\_df = again.replace(to\_replace='3+' , value = 4)
* mydata = replaced\_dataset.drop(columns = 'Loan\_ID' , axis =1)

**Data integration:**

In integration, we get data from multiple sources and store in a database.

This dataset is collected by different sources it is already integrated.

**Data transformation:**

Subtracting mean and then dividing it by its standard deviation is called Standardization.

Dividing a vector by its length is called normalization, this also transforms your data between 1(YES/TRUE) or 0(NO/FALSE).

* **Manually Normalization:**

replaced\_dataset = Loan\_predicting\_df.replace({'Gender' : {'Male' : 1 , 'Female' : 0} , 'Married': {'Yes' :1 , 'No':0 'Self\_Employed' : {'Yes' : 1 , 'No' : 0} , 'Education' : {'Graduate' : 1 , "Not Graduate" : 0} , 'Property\_Area' : {'Rural' : 0 , "Semiurban" : 1 , 'Urban' : 2} })

* **PCA Standardization:**

pca = PCA()

pca\_transformed\_data = pca.fit\_transform(X)

* **MinMax Normalization:**

minmaxscalar= MinMaxScaler()

min\_max\_scalar\_transformation\_data = minmaxscalar.fit\_transform(X)

* **Unit Vector Scaling:**

unit\_vector = scale(mydata)

# **Literature Review:**

The linked work analysis provides results on numerous healthcare datasets, where analysis and prediction were performed using various approaches and procedures. Various prediction models have been devised and applied by researchers utilizing data mining techniques, machine learning algorithms, or a combination of these approaches.

Dr Saravana Kumar N M, Eswari, Sampath P, and Lavanya S (2015) used Hadoop and the Map Reduce approach to analyze diabetic data. This approach forecasts the kind of diabetes as well as the risks connected with it. The Hadoop-based solution is cost-effective for any healthcare institution. [1]

Aiswarya Iyer (2015) investigated hidden patterns in a diabetes dataset using a classification algorithm. This model made use of Nave Bayes and Decision Trees. The performance of both algorithms was compared, and the usefulness of both methods was demonstrated as a consequence. [2]

Another study used supervised learning to estimate the rank of scientific research papers by Mohamed El Mohadab, Belaid Bouikhalene [4], and Said Safi. Kumar Arun, Garg Ishan, and Kaur Sanmeet [3] are working on predicting how a bank would accept a loan. They presented a model using machine learning technologies such as SVM and neural networks. This evaluation of the literature assists us in carrying out our job and developing a trustworthy bank loan forecast model.

# **Methodology:**

**SVM**

In different machine learning applications, Support Vector Machines (SVM) is a strong classification and regression tool. Its goal as an unsupervised machine learning approach is to build a decision boundary between two classes that allows label prediction using one or more feature vectors.

* Accuracy in manual transformed data:

The average cross-validation score is: **0.7875**

No. of K-folds: **5**

Testing accuracy: **1.0**

Training Accuracy: **0.7916666666666666**

**Logistic Regression**

Logistic regression is a statistical analytic approach that uses past observations of a data set to predict a binary result, such as yes or no. A logistic regression model forecasts a dependent variable by examining the connection between one or more existing independent variables.

* Accuracy in manual transformed data:

The average cross-validation score is: **0.7875**

No. of K-folds: **5**

Testing accuracy: **1.0**

Training Accuracy: **1.0**

# **Flow Chart:**

First, we load our loan dataset. Then we preprocess our data using PCA Transformation and MINMAX Normalization. After that we manually transform it. After that by using Train Test Split, we split out data into training and testing. Then the Training and Testing data is passed through the models (Support Vector Machine and Logistic regression). After that the model is evaluated using K-folds cross validation method. If the model does not acquire the desired output than the data is again sent to the models. If it passes the validation criteria than the prediction is calculated which can be YES or NO in this case.

MODEL EVALUATION/VALIDATION

SVM/LOGISTC REGRESSION

TRAINING DATA

TESTING DATA

TRAIN TEST SPLIT

MANUALLY TRANSFORMED

MINMAX NORMALIZATION

PCA TRANSFORMATION

PRE-PROCESSING

LOAN DATASET

START

PREDICTION

END

Yes

NO

**Performance Assessment Table:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classes** | **Accuracy** | **Precision** | **Recall** | **Sensitivity** | **Specificity** | **F0.5** | **F1** | **F2** |
| **Class-1** | 0.6875 | 0.9677 | 0.6818 | 0.75 | 0.6818 | 0.66 | 0.8 | 0.66 |
| **Class-2** | 0.5833 | 0.7419 | 0.657 | 0.384 | 0.657 | 0.66 | 0.69 | 0.66 |
| **Class-3** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.66 | 1.0 | 0.66 |
| **Average** | 0.7569 | 0.9032 | 0.7796 | 0.711 | 0.7796 | 0.66 | 0.83 | 0.66 |

**Accuracy:** the proportion of right classifications achieved by a trained machine learning model, defined as the number of correct predictions divided by the total number of predictions across all classes.

**Precision:** It is one measure of a machine learning model's performance since it measures the accuracy of a positive prediction provided by the model. Precision is calculated by dividing the number of true positives by the total number of positive predictions.

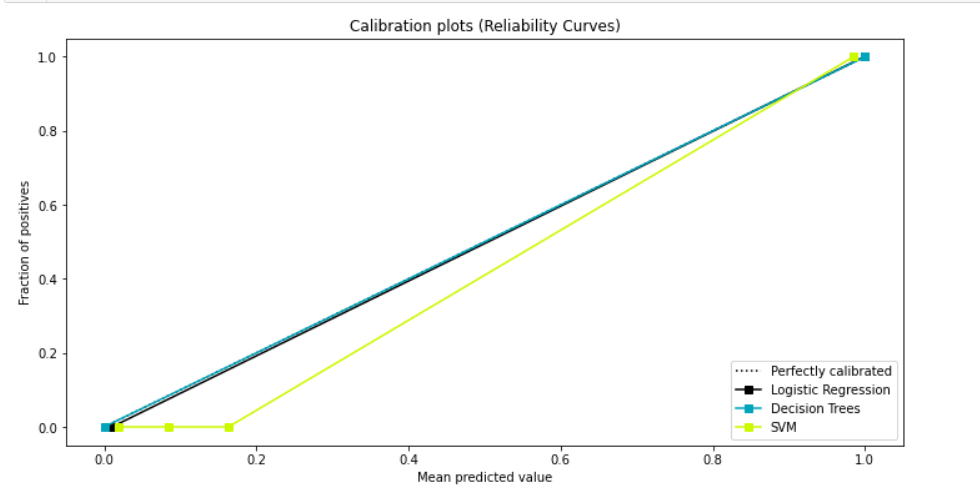
**Recall:** The fraction of real positive labels properly detected by the model is measured by recall.

**Sensitivity:** It quantifies how successfully a machine learning model detects positive examples. It's also referred to as the true positive rate (TPR) or recall. Sensitivity is used to assess model performance since it shows how many positive cases the model accurately identified.

**Specificity:** Specificity is defined as the algorithm's or model's ability to predict a genuine negative of each accessible category. It is also known as the actual negative rate in the literature.

**F-Score:** The F-score (also known as the F1 score or F-measure) is a performance indicator for Machine Learning models. Precision and recall are combined into a single score.

**Line Graph:** A line graph, also known as a line plot or a line chart, is a graph that connects individual data points with lines. A line graph depicts quantitative data over a certain time period.

****

# **References:**

**[1]** [Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S,” Predictive Methodology for Diabetic Data Analysis in Big Data”, 2nd International Symposium on Big Data and Cloud Computing,2015.](https://www.sciencedirect.com/science/article/pii/S1877050915005700)

**[2]** [Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly,” Diagnosis of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.](https://arxiv.org/ftp/arxiv/papers/1502/1502.03774.pdf)

**[3]** [Kumar Arun, Garg Ishan, Kaur Sanmeer, Loan Approval Prediction based on Machine Learning Approach](https://www.iosrjournals.org/iosr-jce/papers/Vol18-issue3/Version-1/O1803017981.pdf)

**[4]** [Mohamed El Mohadab, Belaid Bouikhalene, Said Safi, ‘Predicting rank for scientific research papers using supervised learning’Applied Computing and Informatics 15 (2019) 182–190.](https://www.sciencedirect.com/science/article/pii/S2210832717302703)