**Research paper *:*** *MACHINE LEARNING BASED CUSTOMER CHURN PREDICTION IN BANKING Manas Rahman*

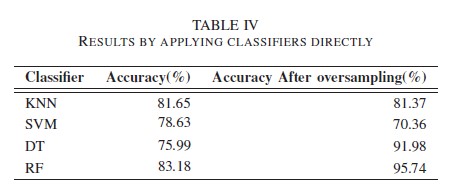
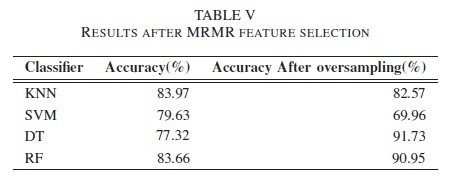
*Department of Computer Science Central University of Kerala Periye, Kasaragod manasrahmanpadiyoor@gmail.com V Kumar Department of Computer Science Central University of Kerala Periye, Kasaragod vkumar@cukerala.ac.in*

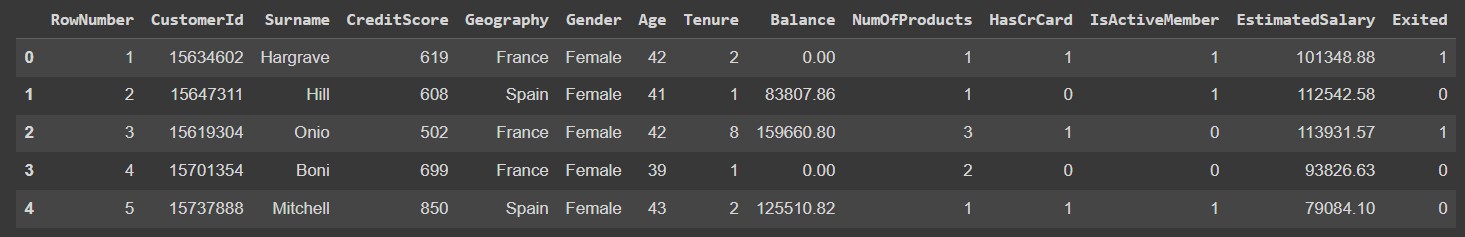
**Introduction:**

This paper focuses on customer churn and engagement in the banking sector. It proposes a machine learning-based method to predict customer churn using techniques such as KNN, SVM, Decision Tree, and Random Forest. The study also explores feature selection methods to identify relevant features and improve system performance. The experimentation is conducted on a churn modelling dataset from Kaggle. The results show that the Random Forest model after oversampling performs better in terms of accuracy compared to other models. The paper concludes that predicting customer churn in the early stages is crucial for banks, and the proposed method can help in achieving this goal.

**Data set:**

The dataset used in the analysis consists of information on 10,000 bank clients. The target parameter is a binary variable indicating whether the customer has left the bank or is still a customer. The dataset includes 13 feature vectors that provide information on customer data and transactions. These features include credit score, gender, age, tenure, balance, number of products, credit card ownership, active membership, and estimated salary. The dataset is pre processed to remove irrelevant features and transform certain attributes. The study focuses on predicting customer churn in the banking sector using machine learning techniques such as KNN, SVM, Decision Tree, and Random Forest. The results of the models are evaluated based on accuracy, and oversampling is used to address the imbalance in the dataset. Feature selection methods, such as MRMR and Relief, are also applied to improve the performance of the models. The study concludes that the Random Forest model after oversampling achieves the highest accuracy in predicting customer churn.

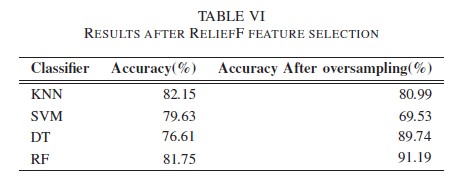




**Observations from the paper.**

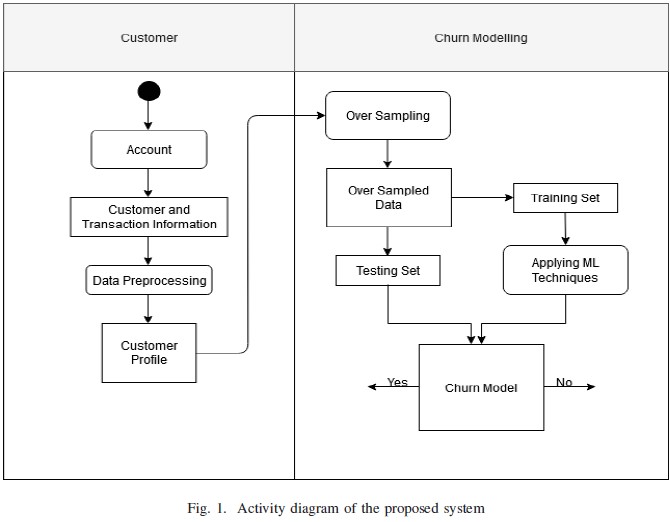
1. The paper focuses on predicting customer churn in the banking sector using machine learning techniques.
2. The dataset used in the analysis consists of information on 10,000 bank clients, with 7,963 positive class samples (maintained) and 2,037 negative class samples (exited).
3. The study explores different machine learning models such as KNN, SVM, Decision Tree, and Random Forest for predicting customer churn.
4. Oversampling is used to address the imbalance in the dataset, and it is found to improve the accuracy of certain classifiers, such as Decision Tree and Random Forest.
5. Feature selection methods, such as MRMR and Relief, are applied to identify relevant features and improve system performance.
6. The study suggests that customers who use more banking services (products) are more likely to be loyal, and the bank should focus on customers who use fewer than three products to retain them.
7. The paper highlights the importance of early identification of customer churn in the banking sector and the need for a generalized system that can predict churn in the early stages.
8. SVM model struggles with the imbalanced dataset.

**Results from the paper:**



1. These are the results given by the paper.
2. We tried to increase the accuracy by at most 1% to 2% for Accuracy

**Methodology used in the paper:**



**We have used the same methodology which is used in the paper and used other selected features for our proposal**

# RESULTS BY DIRECTLY APPLYING CLASSIFIERS DIRECTLY

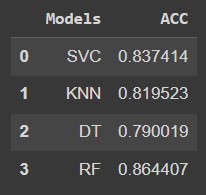
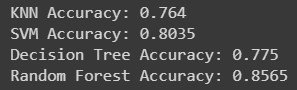
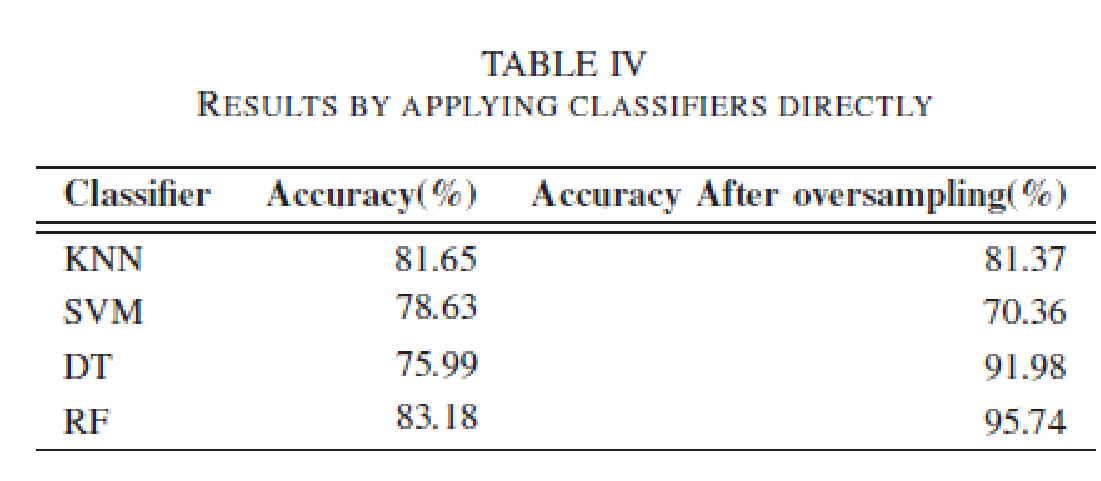
## COMPARING OUR RESULTS WITH PAPER

**Accuracy(%) Accuracy after over sampling(%)**

1. As we can see we have increased Decision tree accuracy by 2% and Random forest accuracy by 2.3%
2. We can also see we have increased SVC by 13% and KNN by 1% after overs sampling

# RESULTS AFTER MRMR SELECTION

## COMPARING OUR RESULTS WITH PAPER

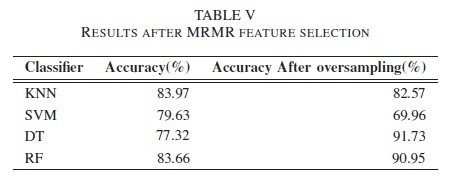
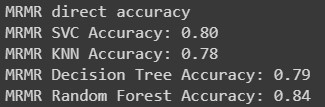


Accuracy

(

%

)



Accuracy after over sampling(%)

1. As we can see we have increased SVC by 1% Decision tree accuracy by 2% and random forest accuracy by 1%
2. We have also increased Random forest by 2% after over sampling

# RESULTS AFTER RELIEF FEATURES

## COMPARING OUR RESULTS WITH PAPER

**Accuracy(**

**%**

**)**

**Accuracy after over**

**sampling(**

**%**

**)**

1.

As we

can see

we h

a

ve

increased

random forest accuracy by 2

%

SV

M

accuracy

by 1

%

2.

W

e

have als

o increased decision

tree

accuracy

after

o

ver sampling

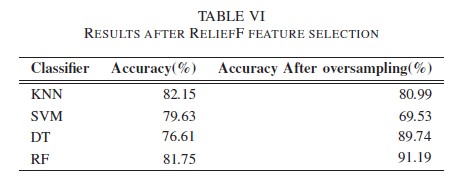
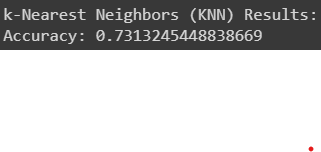
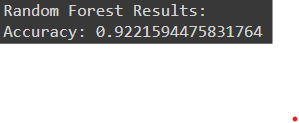
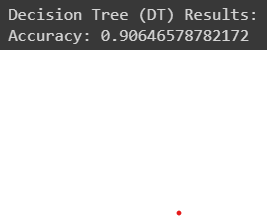
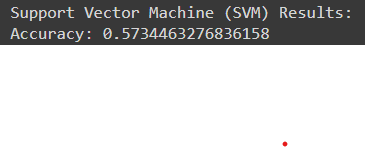
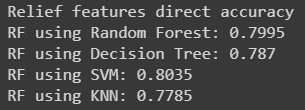
by

1

% and random forest

accuracy by over

sampling by 1%



**Algorithms and techniques used:-**

1. **SMOTE**
2. **Oversampling**
3. **Random forest**
4. **KNN**
5. **Decision tree**
6. **SVM**
7. **MRMR feature selection**
8. **Relief feature selection**