***Abstract***

**The project deals with predicting values based on classification algorithm which can found based on problem statement and target variable**

**The datasets of the project is linked below**

**[The X Dataset](https://github.com/ukishore33/Analysing-Credit-Card-Dataset/blob/main/Credit_card.csv)**

[**The Y Dataset**](https://github.com/ukishore33/Analysing-Credit-Card-Dataset/blob/main/Credit_card_label.csv)

**The main aim of the Project was to perform Exploratory Data Analysis and Perform cleaning and perform predictions from the Credit Card Dataset.**

**Based on the Problem statement target variable, we can conclude that we should use classification algorithm in this project.**

**Imported various Python libraries such as Pandas, NumPy, Seaborn and Matplotlib into Google Collab and Readed the insurance dataset using read\_csv() function and merged the both datasets. we gone ahead in observing the dataset's unique values in every column separately.**

**While observing every column separately, we observed the values which are wrongly typed or in some values in which the values should be dropped, we cleaned the data accordingly. We checked the data for null values and filled that with appropriate values like mean, median and mode which better fits that column.**

**The Dataset after every cleaning is done is saved as a file and attached below.**

[**The Cleaned Dataset**](https://github.com/ukishore33/Analysing-Credit-Card-Dataset/blob/main/Cleaned_dataset%20(1))

**while performing Exploratory data analysis, we performed univariate analysis that is histogram and pie chart for some columns, Bivariate Analysis that is box plot,Scatter plot, correlation between data using heatmap and listed down the observations accordingly.**

**While performing Feauture engineering, we performed encoding using label encoder, we performed scaling using standard scalar.**

**While splitting the data according the vif factors previously checked, we selected the x and y columns accordingly, performed predictions using logistic regression, knn, decision tree and random forest,SVM.**

**We Finally concluded that Random Forest is the best way to perform Predictions using the Given Data.**

**The Project's Data Cleaning part, EDA and Feauture Engineering,ML is Done in the file Below.**

**[The Project File ](https://github.com/ukishore33/Analysing-Credit-Card-Dataset/blob/main/Credit\_Card\_ML.ipynb)**

**Performed Data Analysis using MySQL Tool and obtained Some Insights from it.**

**The Queries we performed is saved in a file and linked below.**

[**My SQL File**](https://github.com/ukishore33/Analysing-Credit-Card-Dataset/blob/main/Credit_Card.sql)

1. ***Introduction***

Introduction:-

A credit card is a convenient tool that allows you to buy items now and pay for them later.

Credit card is a physical payment card that allows you to get credit from a financial institution.

If you buy something with credit, you are in debt.

This means you owe money to the company that gave you the credit card.

If you don't pay the entire amount at the end of each month, you pay a fee for the credit card called interest.

If managed correctly, credit cards can be great way to build credit and manage your money.

Importance in Today's world:-

Credit score cards are a common risk control method in the financial history.

It uses personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings.

The bank is able to decide whether to issue a credit card to the applicant.

Credit scores can objectively quantify the magnitude of risk. Credit score is a number that depicts a consumer's credit worthiness.

Importance of Predicting a good client:-

Credit risk as the board in banks basically center around deciding the the probability of customer's default or credit decay and how expensice it will end up being assuming it happens.

It is important to consider major factors and predict beforehand the probability of consumers defaulting given their conditions, which is where a machine learning model comes in handy and allows the bank and major financial institutions to predict whether the customer will default or not.

This project builds a machine learning model with the best accuracy possible.

Impact on Banking Sector:-

Banks receive a lot of credit card applications.

Many of the applications do not get approved for a variety of reasons, like increased loan balances or poor-income levels.

Manually analysing these applications can be very time consuming and full of human errors.

Hence we can automate this task with the help of machine learning.

Impact on Future Tools:-

The knowledge of knowing of applicants who are having car, property, mobile phone, work phone, email id and also their basic details helps the banks a lot to analyse whether to approve the credit cards to the applicants or not. It would definitely help the banks in future too.

* 1. ***Initial Hypothesis***

The decision on approval or rejection of applications will be mainly dependent on Income Source, Annual Income, Education, Property owner, Car Owner.

Based on the Problem statement and target variable we can conclude that the machine learning techniques we should use here are related to Classification Algorithms.

* 1. ***Data Analysis Approach***

EDA Approach:-

We Imported Libraries such as pandas, Numpy, Seaborn, Matplotlib, Plotly, etc.

We imported the datasets and we merged the dataset.

After Merging we understood and manipulated the data by checking the unique values and filling the null values with appropriate measures.

And we started with EDA Approach through some Graphs where we observed many useful insights such as

- Histogram and Barchart

63.3% of the applicants are females and 36.7% of the applicants are males.

40.3% of the applicants own car, and 59.7% of the applicants do not own car.

65.2% of the applicants own property and 34.8% of the applicants do not own property.

70.5% of the applicants do not have children.

Approximately 50% of the applicants source of income is through working.

Approximately 10% of the applicants do not own House/apartment.

We can see in annual income histogram, the graph is right skewed, which implies most of the applicants are present towards right of the peak. The peak is pointed at approximately 0.12.

We can see that most of the applicant's education level is secondary/secondary special.

Most of the applicants are Married.

The Birthday count values are normally distributed.

Many of the Applicants have lesser employed Days.

Each and Every applicants has mobile phone.

Most of the Applicants Do not have Work phones.

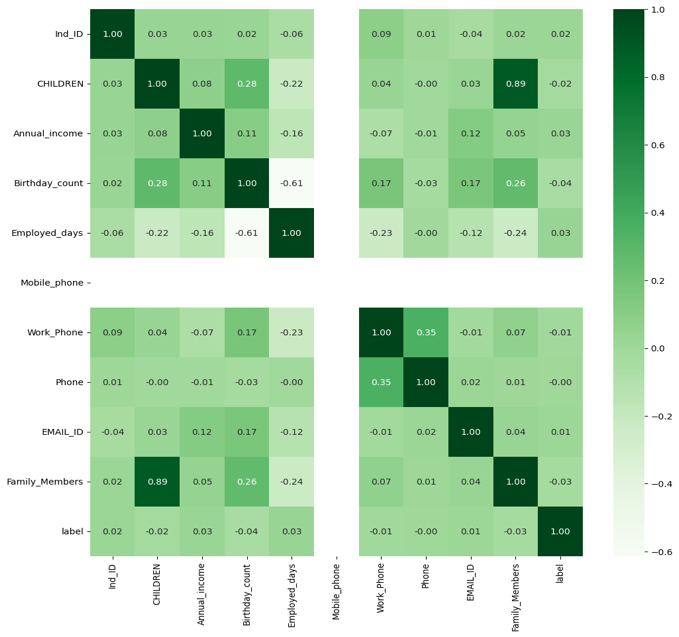
Only few of the applicants have E-Mail ID's.

Most of the applicants are labourers by Occupation.

Most of the applicants have two members in their Family.

Most of the applicant's credit card is approved.

* Correlation with Heat Map and Pairplot



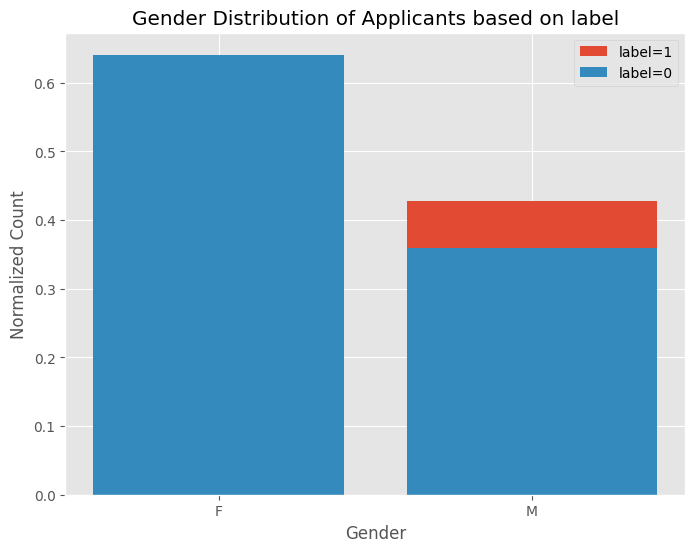
The family members and children column are highly correlated.

The annual income and Employed days are more correlated than any other columns.

A grid of red squares

Description automatically generated

* Bar Chart which shows Gender Distribution of Applicants Based on Label.



Every female applicant's credit card is approved.

Few of the male applicant's credit is rejected

* Bar Chart Showing Education Level

A graph of a graph with red and white bars

Description automatically generated

From the graph above, most of the applicants have secondary/secondary special level of Education, which is followed by Higher Education. The remaining applicants contribution towards Education level is minimal/negligible.

**Feauture Engineering Methods:-**

Here we changed the categorical variables into numerical columns by Label Encoding Techniques

Label Encoding means giving a distinct value starting with 0 value for each unique string of qualitative column.

Here we changed the categorical variables for columns such as

1) Gender

2) Car Owner

3) Property Owner

4) Type of Income

5) Education

6) Marital Status

7) Housing Type

8) Type Occupation

And we moved on Next to Scaling where we used standard scalar to change every column’s values to a single scale except the target variable.

And then we just moved on to splitting the dataset into train and test variable based on 70/30 division randomly and further I divided the train and test into X\_train, y\_train, X\_test, y\_test for the purpose of machine learning model.

* 1. ***Machine Learning Approach***

Based on the problem statement, we just analysed that we should work on the classification algorithms.

The classification algorithms we used here are Logistic Regression, KNN Algorithm, Decision Tree Algorithm, Random Forest Algorithm.

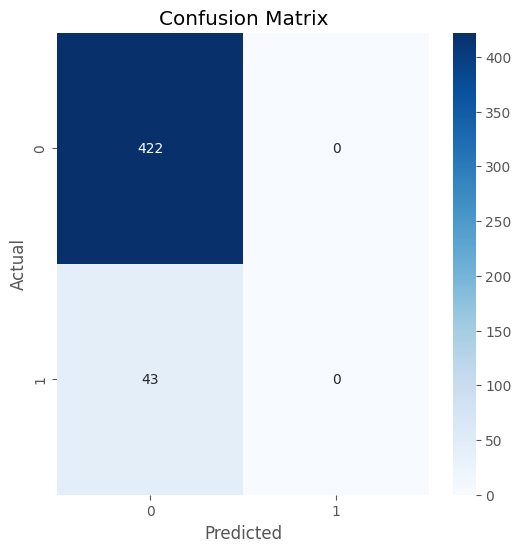
Logistic Regression Algorithm:-

Here First we built the algorithm using X\_train and y\_train and predicted the values using X\_test.

Then we Evaluated the predictions using y\_test

During visualisation of evaluation metrics, we obtained some important measures such as Accuracy, Precision, Confusion Matrix, Classification Report

A graph of a graph showing a red square

Description automatically generated

KNN Algorithm:-

Here First we built the algorithm using X\_train and y\_train and predicted the values using X\_test.

Then we Evaluated the predictions using y\_test.

During visualisation of evaluation metrics, we obtained some important measures such as Accuracy, Precision, Confusion Matrix, Classification Report

A graph of a bar graph

Description automatically generatedA blue squares with white text

Description automatically generated

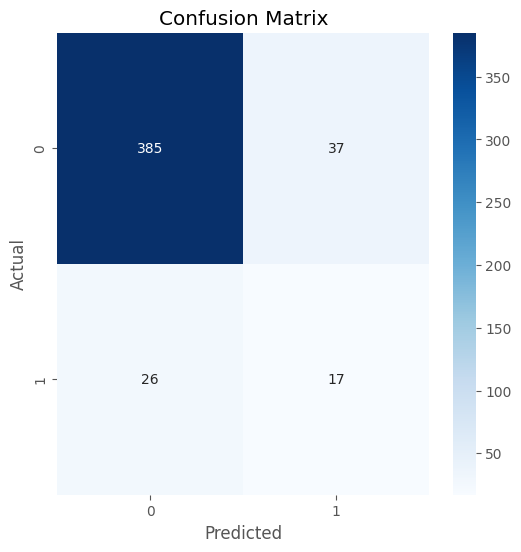
Decision tree Algorithm:-

Here First we built the algorithm using X\_train and y\_train and predicted the values using X\_test.

Then we Evaluated the predictions using y\_test.

During visualisation of evaluation metrics, we obtained some important measures such as Accuracy, Precision, Confusion Matrix, Classification Report.

A graph of a bar graph

Description automatically generated

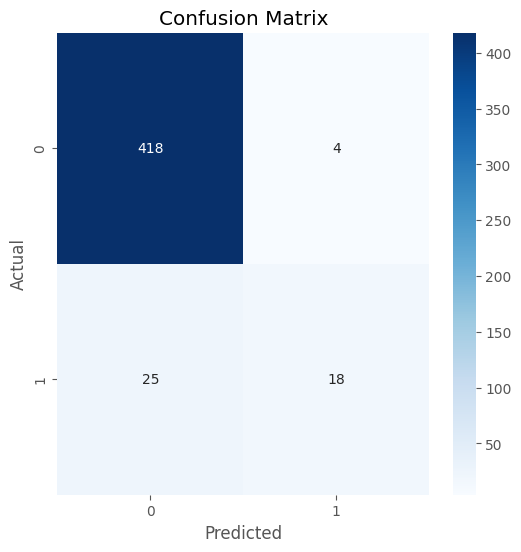
Random Forest Algorithm:-

Here First we built the algorithm using X\_train and y\_train and predicted the values using X\_test.

Then we Evaluated the predictions using y\_test.

During visualisation of evaluation metrics, we obtained some important measures such as Accuracy, Precision, Confusion Matrix, Classification Report.

A graph of a bar graph

Description automatically generated

SVM Algorithm:-

Here First we built the algorithm using X\_train and y\_train and predicted the values using X\_test.

Then we Evaluated the predictions using y\_test.

During visualisation of evaluation metrics, we obtained some important measures such as Accuracy, Precision, Confusion Matrix, Classification Report.

A graph of a graph showing a red square

Description automatically generated A graph of blue squares

Description automatically generated

Conclusion:-

Based on the Evaluation, we can conclude that the Random Forest is the Algorithm we can use for making predictions in the Dataset.

***5)Getting Some Useful Insights Using My SQL Queries***

We had imported the Cleaned Dataset through Import Wizard after Making Data Cleaning approach in Python.

After that we just made some quick queries to get some specific Insights from the Data.

The Specific Queries and Insights we gained from the Data are –

Group the customers based on their income type and find the average of their annual income.

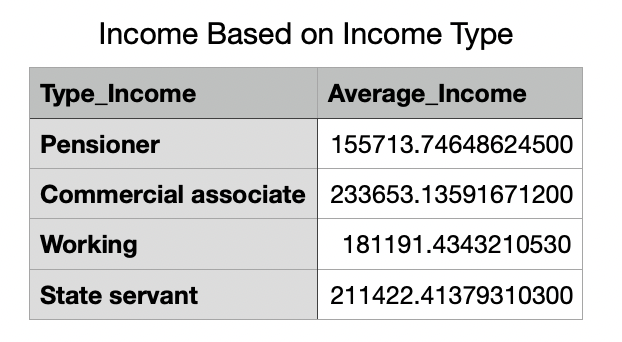
The Query we used here was

SELECT Type\_income as Type\_Income, AVG(Annual\_income) AS Average\_Income

FROM credit\_dataset

GROUP BY Type\_income;

The Result we get here was



**Insights:-**

Based on what we get above we see that the Commercial Associate gets High Average Salary than any other Applicants.

Find the female owners of cars and property

The Query we used here was

SELECT count(\*) FROM credit\_dataset

WHERE GENDER = 'F' AND (Propert\_Owner = 'Y' and Car\_Owner = 'Y');

**Insights:-**

We could get that the 177 Females Own Cars And Properties.

Find the female customers who are staying with the family.

The Query we used here was

SELECT \* FROM credit\_dataset

WHERE GENDER = 'F' AND Marital\_status= 'Married' AND Family\_Members > 1;

**Insights:-**

We can get that in overall 739 female customers are staying with the family

We can get that 625 married female customers are staying with the family

Please list the top five people having the highest income.

The Query we used here was

SELECT \* FROM credit\_dataset

ORDER BY Annual\_income DESC

LIMIT 5;

**Insights:-**

The list of top five Applicants who have highest incomes are having their own cars, house, property, and they work as commercial associate and they completed their Higher education and they all have atleast 1 children and their family consists of more than 2 members.

How many married couple are having bad credit.

The Query we used here was

SELECT COUNT(\*) AS num\_married\_couples\_bad\_credit FROM credit\_dataset WHERE label = '1' AND marital\_status = "Married";

**Insights:-**

There are 114 Married Couples who have Bad Credit

what is the highest education and what is the total count.

The Query we used here was

Select EDUCATION, count(\*) as Total\_Members from credit\_dataset

Group by EDUCATION order by count(\*) desc limit 1;

**Insights:-**

The Education Most of the Applicants had finished was Secondary/Secondary Special Education Level. Total Number of Count of that specific Education level of applicants are 1031.

Between Married males and females who is having more bad credit.

The Query we used here was

SELECT gender, COUNT(\*) AS Total\_Count FROM credit\_dataset

WHERE label= '1' And Marital\_status='Married' GROUP BY gender ORDER BY Total\_Count DESC;

**Insights:-**

The Applicants who have more bad credit is 110 in which 59 is Females and 51 is male.