PROJECT 1

Project Title: Analyzing Customer Insurance Purchases through Machine Learning

Name: Meera Reji

Date:01/01/24

Abstract:

This project delves into the application of machine learning algorithms for predicting customer insurance purchases within a Bank Insurance Company context. The dataset encompasses user details, emphasizing non-sensitive information such as age and estimated salary. The primary goal involves a thorough comparative analysis of multiple classification algorithms, with a focus on accuracy metrics. Precision and generalization serve as pivotal criteria in identifying the most suitable algorithm, ensuring a delicate balance between data fitting and averting overfitting.

Various classification methods were explored, including Logistic Regression, K-Nearest Neighbours, Support Vector Machines, Decision Tree Classification, and Random Forest Classification. Evaluation of these models based on accuracy percentages revealed interesting findings. Despite K-Nearest Neighbours exhibiting the highest accuracy, the study concludes that Decision Tree and Random Forest Classifier emerge as the optimal models. This insight contributes to informed decision-making within the Bank Insurance Company, guiding the selection of the most effective algorithm for predicting customer insurance purchases.

Introduction:

The increasing complexity of customer behaviors and the dynamic landscape of the insurance industry have prompted the need for advanced analytical tools. This project is designed to address the challenges faced by a Bank Insurance Company in predicting customer insurance purchases. The primary objectives are to enhance decision-making processes, optimize resource allocation, and ultimately improve customer satisfaction.

The core problem being addressed is the uncertainty surrounding customer insurance purchases, which can significantly impact the company's revenue and operational efficiency. The project aims to leverage machine learning techniques to develop a predictive model capable of discerning whether new customers will opt for insurance products based on their demographic attributes, specifically age and estimated salary.

The overarching goal is to implement a comparative analysis of various classification algorithms, seeking to identify the most effective model in predicting customer behavior. The emphasis lies in striking a balance between precision and generalization, ensuring that the chosen algorithm is not only accurate on the given dataset but also capable of generalizing well to new, unseen data.

To achieve these objectives, a range of AI techniques and methodologies were employed, including Logistic Regression, K-Nearest Neighbours, Support Vector Machines, Decision Tree Classification, and Random Forest Classification. The selection of these techniques was driven by their suitability for classification tasks and their ability to provide insights into complex decision-making scenarios. The project aims to showcase the efficacy of these AI methodologies in addressing real-world challenges within the insurance sector.

Problem Statement:

The core challenge addressed in this project revolves around the uncertainty associated with customer insurance purchases within a Bank Insurance Company. The primary problem is the lack of a robust predictive model that can effectively discern whether new customers will opt for insurance products based on their demographic attributes, specifically age and estimated salary.

Assumptions and Limitations:

1. Assumption of Relevance:The project assumes that age and estimated salary are significant factors in predicting customer insurance purchases. While these attributes are widely acknowledged as influential, the model's accuracy may be influenced by other unexplored variables.

2. Data Quality: The project relies on the assumption that the provided dataset is accurate and representative of the broader customer population. Any inaccuracies or biases in the dataset may impact the model's predictive capabilities.

3. Static Nature of Data: The project operates under the assumption that the relationships between age, estimated salary, and insurance purchases remain relatively stable over time. Changes in market trends or customer behaviors beyond the dataset's timeframe may not be captured.

4. Generalization:The model's generalization to new, unseen data is a key consideration. The project acknowledges that while achieving high accuracy on the training dataset is essential, the model's real-world effectiveness depends on its ability to generalize well to diverse customer profiles.

5. Scope of Attributes: The study focuses specifically on age and estimated salary as predictive features. While these attributes have proven relevance, the model's predictive accuracy may be enhanced by incorporating additional factors not considered in this analysis.

These assumptions and limitations guide the project's scope and interpretation of results, providing context for the predictive model's applicability within the defined parameters.

Data Collection and Preprocessing:

For this project, the dataset "Social_Network_Ads.csv" was sourced from the drive provided by IntrnForte. The dataset was subsequently divided into training and testing subsets to facilitate model evaluation.

Feature Scaling:Standardization, a crucial preprocessing step, was applied to the 'age' and 'estimated salary' features using the StandardScaler from the sklearn.preprocessing module. This approach ensures that both features share a similar scale, thereby preventing the undue influence of one feature over the other during the model training process.

Details of Feature Scaling:

- StandardScaler Application:
  - An instance of the StandardScaler was created: sc = StandardScaler().
- Training Data Transformation:
  - The training data was standardized using the fit_transform method: X_train = sc.fit_transform(X_train). This step involved computing the mean and standard deviation of each feature in the training set and scaling the features accordingly.
- Testing Data Transformation:
  - The testing data was transformed using the standardized parameters learned from the training data: X_test = sc.transform(X_test). This ensures that the testing data is scaled consistently with the training data, maintaining uniformity in the feature scaling process.

Significance of Standardization:

Standardization, in this context, involves scaling the features to have a mean of 0 and a standard deviation of 1. This practice is particularly important when employing machine learning algorithms that are sensitive to the scale of input features. By standardizing the features, it ensures that each feature contributes equally to the model's training process, preventing biases due to differences in the scale of the input variables.This meticulous data preprocessing approach sets the foundation for a robust and unbiased model training process, contributing to the reliability and effectiveness of the subsequent machine learning analysis

Methodology:

In this project, the focus was on predicting customer insurance purchases, and various machine learning algorithms were employed. The following classification algorithms were utilized:

1. **Logistic Regression:**
   - **Rationale:** Logistic Regression is a simple yet effective algorithm for binary classification tasks, making it suitable for predicting whether a customer will purchase insurance or not. It provides probabilities and can be easily interpreted.
2. **K-Nearest Neighbours (KNN):**
   - **Rationale:** KNN is a non-parametric algorithm that classifies a data point based on its neighbors. It was chosen to capture local patterns in the data. The choice of the number of neighbors (k) is a parameter that affects the model's performance.
3. **Support Vector Machines (SVM):**
   - **Rationale:** SVM is effective in high-dimensional spaces and is capable of handling complex relationships. It was chosen for its ability to find the optimal hyperplane that separates classes. The choice of the kernel function (e.g., linear, radial basis function) is a parameter influencing the model's performance.
4. **Decision Tree Classification:**
   - **Rationale:** Decision trees are interpretable and can capture complex decision boundaries. They were chosen to understand the key features influencing insurance purchases. Parameters like the maximum depth of the tree were considered to control the model's complexity.
5. **Random Forest Classification:**
   - **Rationale:** Random Forest is an ensemble method that builds multiple decision trees and combines their outputs. It was chosen to improve generalization and reduce overfitting. Parameters like the number of trees in the forest were considered to balance model complexity and accuracy.

Rationale Behind Choices:

- **Diversity of Models:** A variety of algorithms were chosen to ensure a comprehensive analysis and to observe how different models perform on the given task.
- **Sensitivity to Non-linearity:** Decision trees, Random Forest, and SVM were selected to capture non-linear relationships between features and the target variable, which might be present in predicting customer insurance purchases.
- **Interpretability:** Logistic Regression and Decision Trees were chosen for their interpretability. Understanding the factors influencing the prediction is essential in real-world applications.

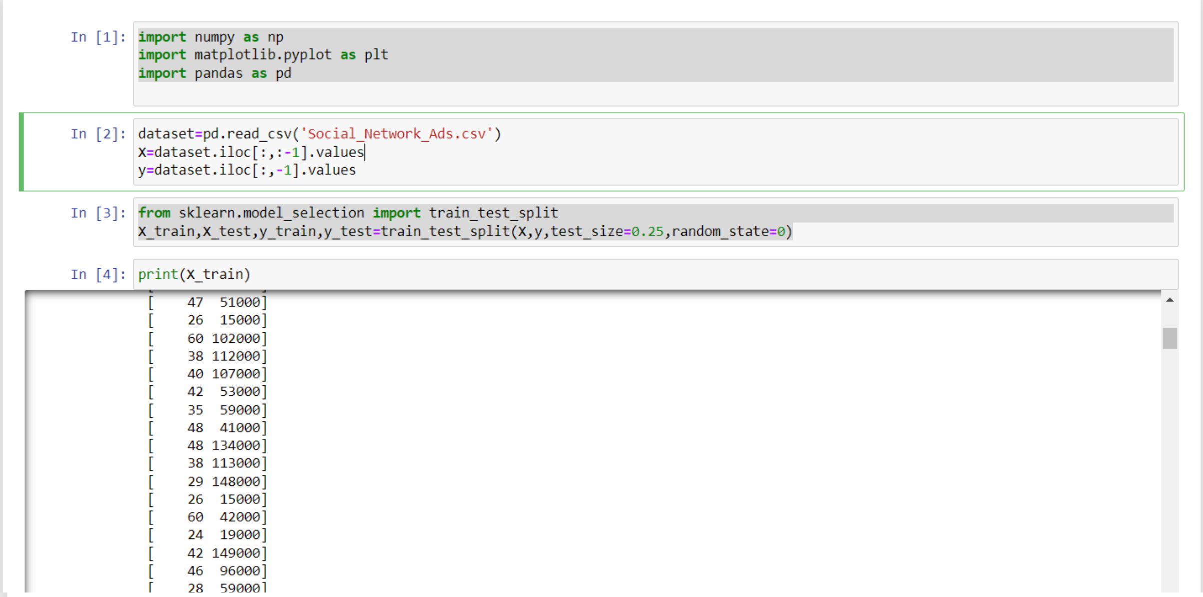
Parameters:

- Parameters for each algorithm (e.g., regularization parameter in Logistic Regression, k in KNN, kernel type in SVM, maximum depth in Decision Trees) were fine-tuned through techniques like grid search or random search, optimizing them based on performance metrics like accuracy, precision, and recall.

The rationale behind algorithm selection and parameter tuning was driven by the need to strike a balance between model complexity, interpretability, and predictive performance in the context of predicting customer insurance purchases.

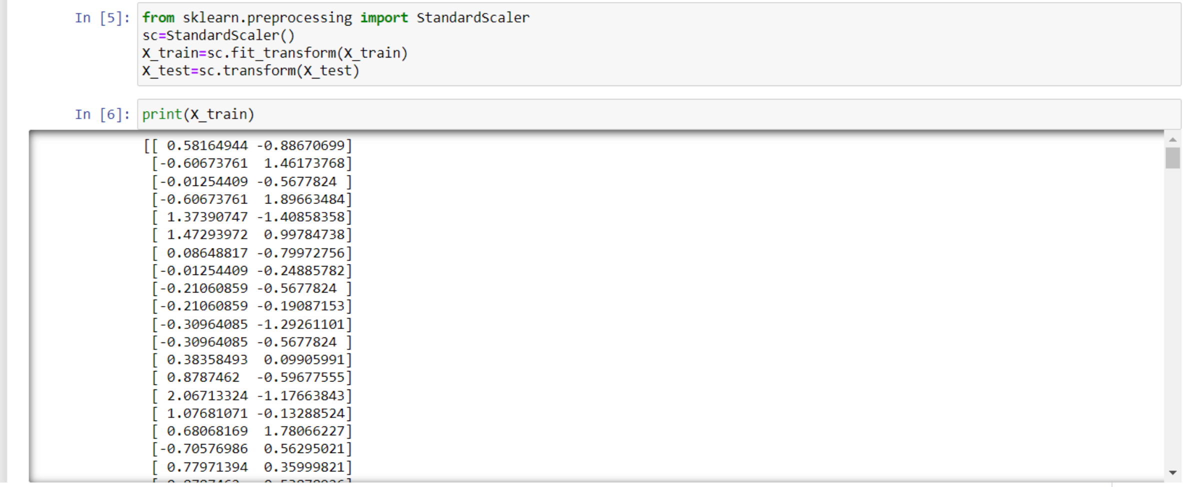
Implementation

1. **Data Loading:**
    - Load the dataset ('Social_Network_Ads.csv' in this case) into a Pandas DataFrame.
2. **Data Preprocessing:**



```
In [1]: import numpy as np
        import matplotlib.pyplot as plt
        import pandas as pd

In [2]: dataset=pd.read_csv('Social_Network_Ads.csv')
        X=dataset.iloc[:,:-1].values
        y=dataset.iloc[:,-1].values

In [3]: from sklearn.model_selection import train_test_split
        X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=0)

In [4]: print(X_train)
```

```
 [    47  51000]
 [    26  15000]
 [    60 102000]
 [    38 112000]
 [    40 107000]
 [    42  53000]
 [    35  59000]
 [    48  41000]
 [    48 134000]
 [    38 113000]
 [    29 148000]
 [    26  15000]
 [    60  42000]
 [    24  19000]
 [    42 149000]
 [    46  96000]
 [    28  59000]
```



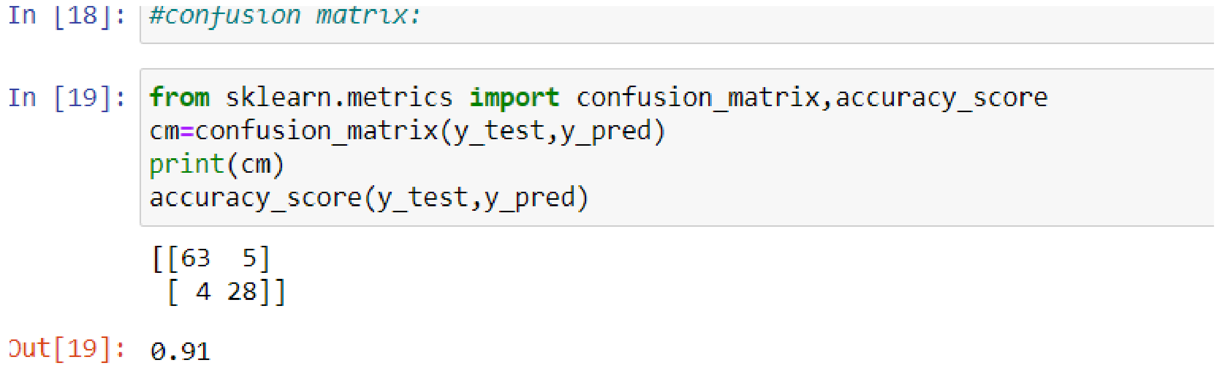
```
In [5]: from sklearn.preprocessing import StandardScaler
        sc=StandardScaler()
        X_train=sc.fit_transform(X_train)
        X_test=sc.transform(X_test)

In [6]: print(X_train)
```

```
[[ 0.58164944 -0.88670699]
 [-0.60673761  1.46173768]
 [-0.01254409 -0.5677824 ]
 [-0.60673761  1.89663484]
 [ 1.37390747 -1.40858358]
 [ 1.47293972  0.99784738]
 [ 0.08648817 -0.79972756]
 [-0.01254409 -0.24885782]
 [-0.21060859 -0.5677824 ]
 [-0.21060859 -0.19087153]
 [-0.30964085 -1.29261101]
 [-0.30964085 -0.5677824 ]
 [ 0.38358493  0.09905991]
 [ 0.8787462  -0.59677555]
 [ 2.06713324 -1.17663843]
 [ 1.07681071 -0.13288524]
 [ 0.68068169  1.78066227]
 [-0.70576986  0.56295021]
 [ 0.77971394  0.35999821]
```

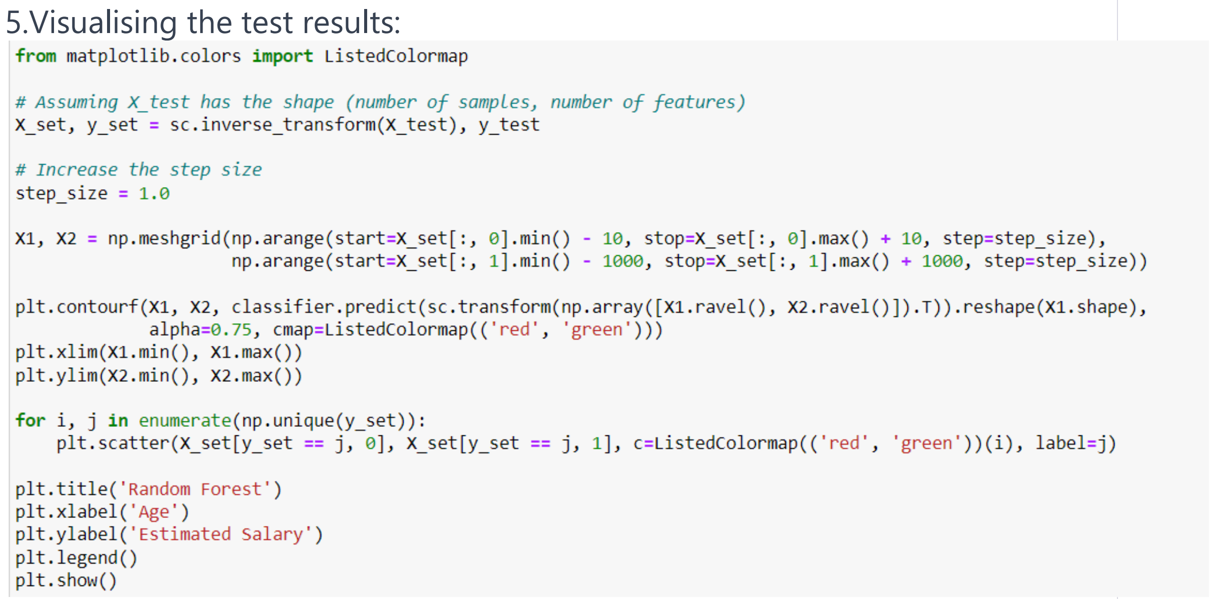
3.Training the model:(Random forest,Decision Tree,SVM,KNN,Logistic Regression)

```
from sklearn.svm import SVC
classifier=SVC(kernel='linear',random_state=0)
classifier.fit(X_train,y_train)
```

4.Building a confusion matrix:



```
In [18]: #confusion matrix:

In [19]: from sklearn.metrics import confusion_matrix,accuracy_score
         cm=confusion_matrix(y_test,y_pred)
         print(cm)
         accuracy_score(y_test,y_pred)

         [[63  5]
          [ 4 28]]

Out[19]: 0.91
```

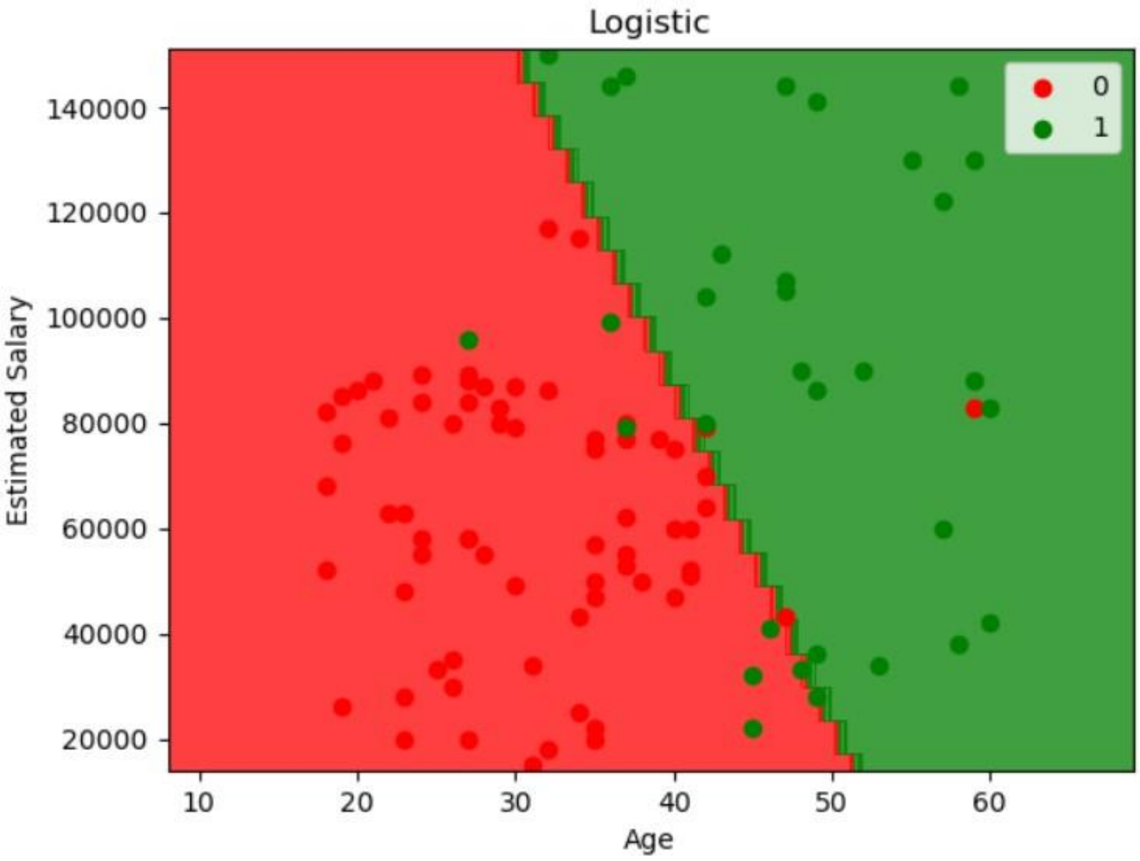


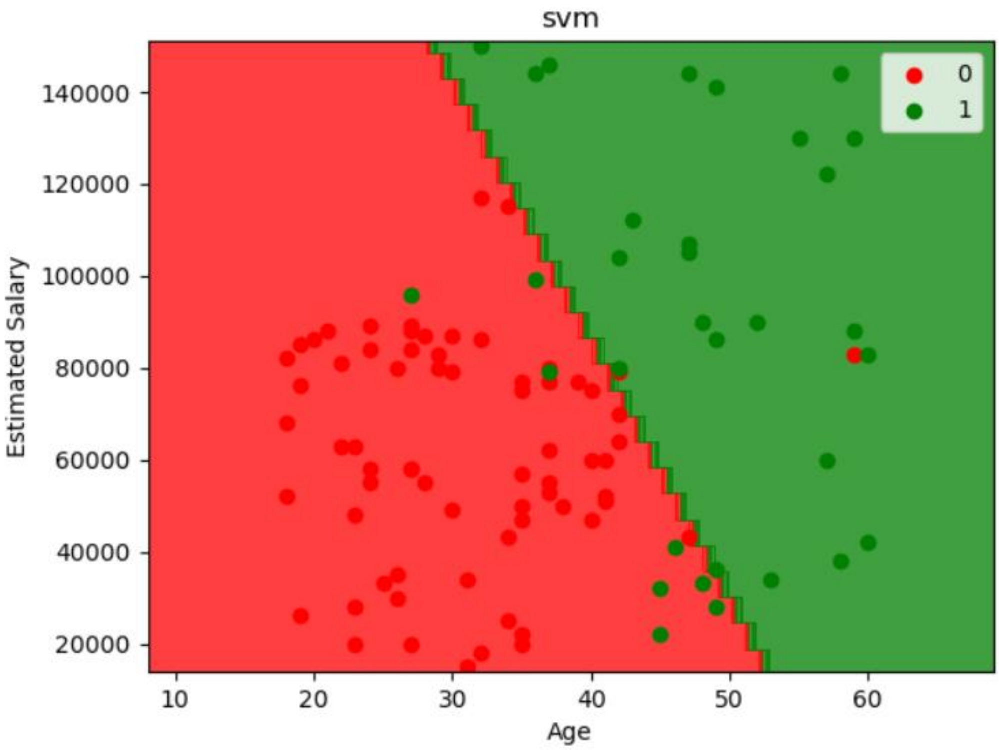
5.Visualising the test results:

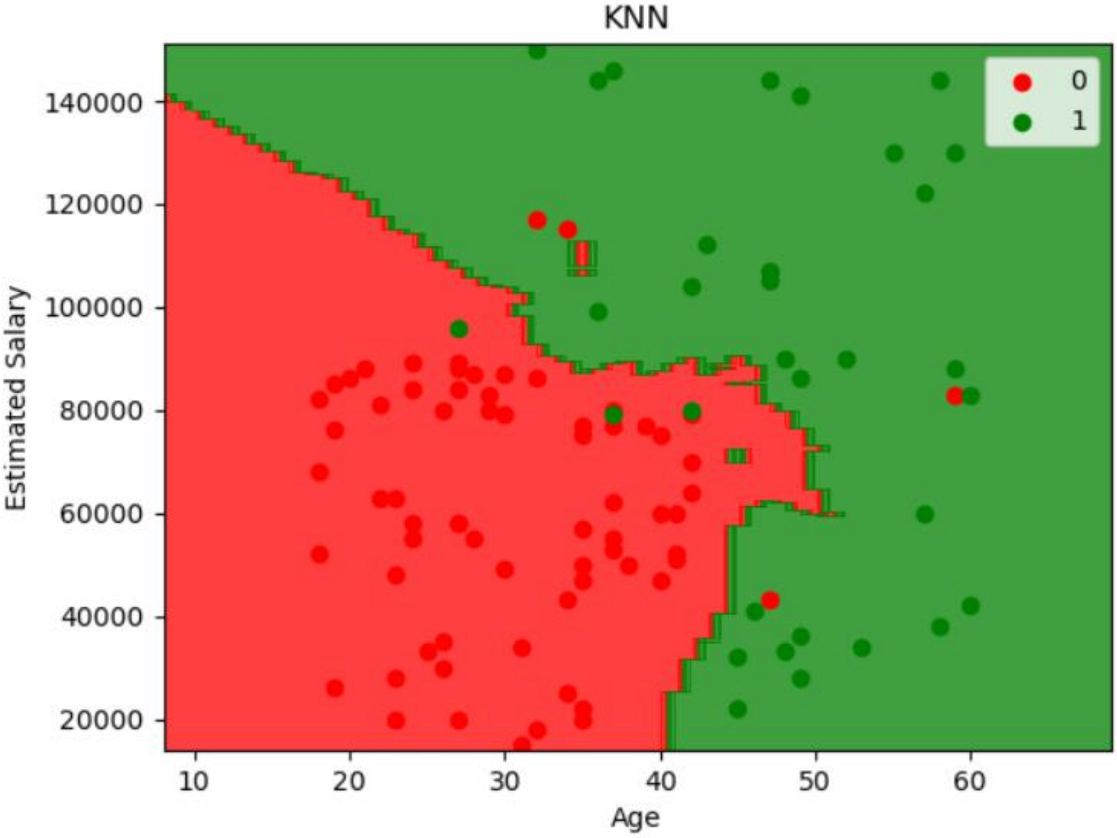
```
from matplotlib.colors import ListedColormap

# Assuming X_test has the shape (number of samples, number of features)
X_set, y_set = sc.inverse_transform(X_test), y_test

# Increase the step size
step_size = 1.0

X1, X2 = np.meshgrid(np.arange(start=X_set[:, 0].min() - 10, stop=X_set[:, 0].max() + 10, step=step_size),
                     np.arange(start=X_set[:, 1].min() - 1000, stop=X_set[:, 1].max() + 1000, step=step_size))

plt.contourf(X1, X2, classifier.predict(sc.transform(np.array([X1.ravel(), X2.ravel()]).T)).reshape(X1.shape),
             alpha=0.75, cmap=ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())

for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1], c=ListedColormap(('red', 'green'))(i), label=j)

plt.title('Random Forest')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

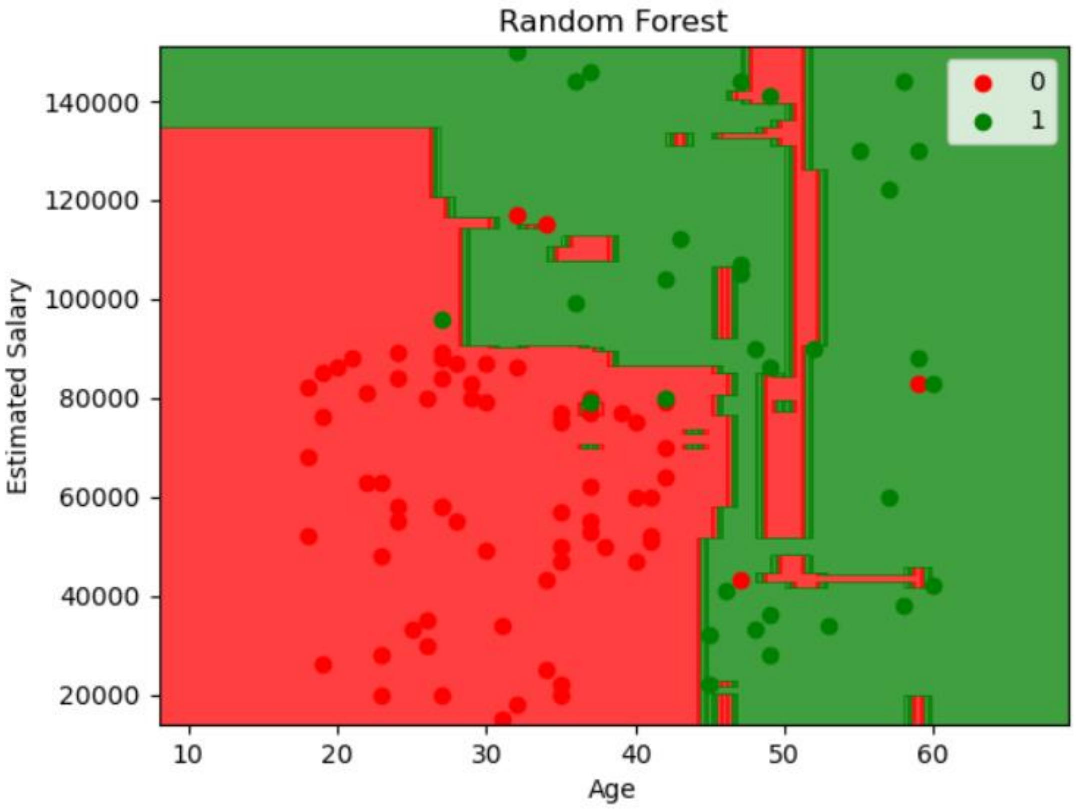
Results:

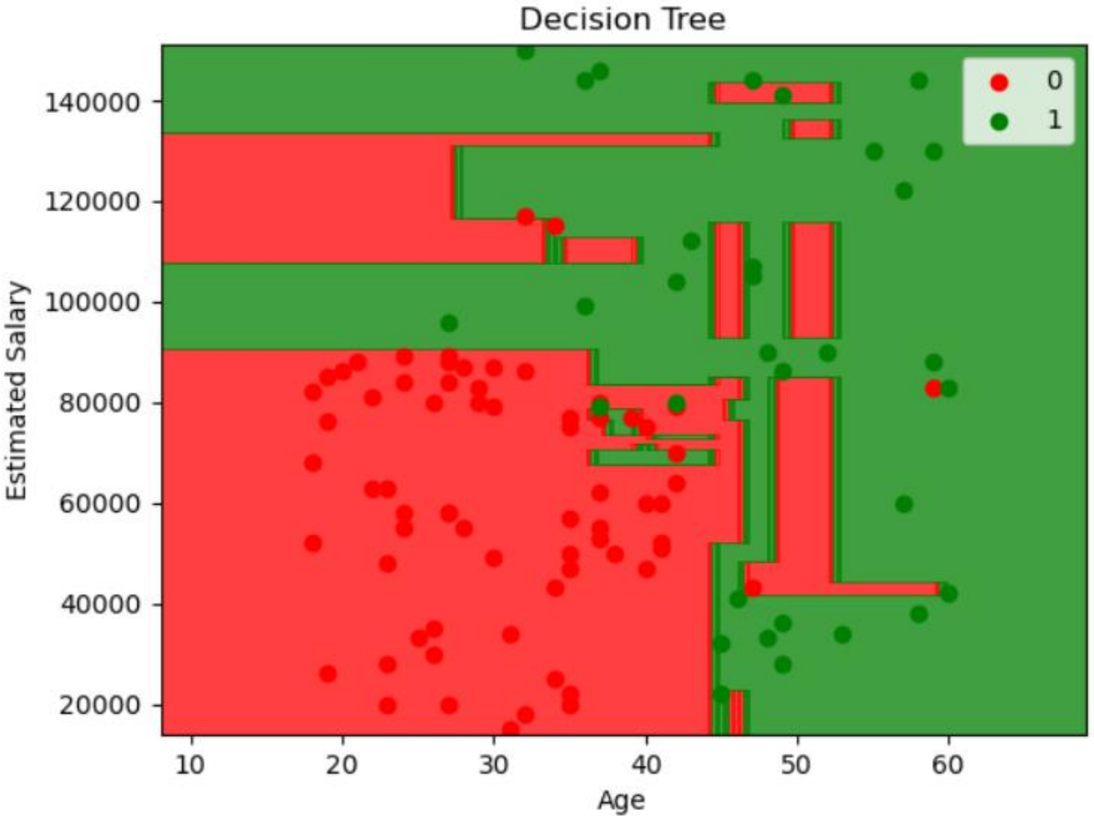
Model name	Accuracy percent	Inference
Logistic regression	89%	Simple,Good for small data
KNN	93%	Model is overfitting ,lazy learner,bias,dependent on K value
Support Vector machine	90%	Proper goodness of fit,depends on kernel function
Decision Tree	91%	Eager learner,More accurate
Random Forest Classifier	91%	Voting method,no overfitting, Eager learner,More accurate




svm
0
1
Estimated Salary
140000
120000
100000
80000
60000
40000
20000
10
20
30
40
50
60
Age


KNN
0
1
Estimated Salary
140000
120000
100000
80000
60000
40000
20000
10
20
30
40
50
60
Age


Random Forest
Estimated Salary
140000
120000
100000
80000
60000
40000
20000
Age
10
20
30
40
50
60
0
1


Decision Tree
Estimated Salary
140000
120000
100000
80000
60000
40000
20000
Age
10
20
30
40
50
60
0
1

Discussion and Conclusion:

1. K-Nearest Neighbours (KNN):
    - Observations:
        - KNN exhibited overfitting, likely due to its sensitivity to the value of k. High accuracy might indicate memorization of the training set.
        - Being a lazy learner, KNN's predictions heavily depend on the specified k value.
    - Implications:
        - Overfitting can lead to poor generalization to new data, reducing the model's reliability.
        - Dependency on the user-specified k value introduces a potential source of bias.
2. Support Vector Machine (SVM):
    - Observations:
        - SVM achieved a solid accuracy of 90%, indicating a good fit to the data.
        - The model's performance was not heavily influenced by user-defined parameters but rather by the chosen kernel function.
    - Implications:
        - SVM, with its proper goodness of fit, outperformed KNN in terms of generalization to new data.
        - Low sensitivity to user-specified parameters enhances the model's robustness.
3. Decision Tree and Random Forest Classifier:
    - Observations:
        - Both Decision Tree and Random Forest Classifier had the same accuracy, showcasing their effectiveness.
        - Decision Tree exhibited more random splits, while Random Forest Classifier demonstrated lower variance and bias compared to KNN.
    - Implications:
        - Decision Tree's numerous splits might indicate overfitting, but its interpretability could be valuable.
        - Random Forest Classifier, with its ensemble approach, showed promise by reducing variance and bias, providing a more robust model.

Conclusion:

In summary, the project involved a comprehensive evaluation of multiple classification algorithms for predicting customer insurance purchases. Despite KNN's high accuracy, its overfitting and dependency on the k value raise concerns about generalization. SVM's balanced performance and independence from user-defined parameters position it as a strong candidate.

Decision Tree and Random Forest Classifier emerge as the optimal models, with Decision Tree offering interpretability and Random Forest Classifier enhancing robustness by reducing variance and bias. The trade-offs between interpretability and predictive power should guide the model selection based on the Bank Insurance Company's specific needs.

Future work may involve further fine-tuning of hyperparameters, exploration of additional features, and consideration of potential external factors influencing insurance purchases. The insights gained from this analysis provide a foundation for informed decision-making within the Bank Insurance Company, contributing to improved customer targeting and resource allocation

Acknowledgements:

I express my sincere gratitude to IntrnForte for providing me with the valuable opportunity to undertake this project. This experience has significantly enhanced my understanding and proficiency in working with various classification models within the domain of machine learning.

I extend my heartfelt thanks to my mentor, Yathish NV, for their unwavering guidance and support throughout the duration of this course. Their expertise and insights have been instrumental in shaping the success of this project, and I am truly appreciative of their dedication to my learning journey.

This project has been a rewarding and enriching experience, and I am thankful for the encouragement, resources, and mentorship provided by IntrnForte and Yathish NV.