Machine Learning Lab Assignment: Predictive Modeling for Insurance Claims

Objective:
Build a predictive model to determine if a building will have an insurance claim during a specific period using building characteristics. In this assignment, you will explore and apply four machine learning algorithms: Support Vector Machine (SVM), Linear Regression, k-nearest Neighbors (KNN), and Naive Bayes. The evaluation metric for this assignment is the Area Under the Curve (AUC).

Tools Required:
- Google Colab
- Python
- Scikit-Learn
- Pandas
- Matplotlib
- Seaborn

Dataset Description:
1. VariableDescription.csv: Contains descriptions of variables.
2. train_data.csv: Training dataset with the target variable.
3. test_data.csv: Testing dataset for predictions.
4. SampleSubmission.csv: Submission format for the competition.

Tasks:

1. Data Exploration and Preprocessing:
   - Load the datasets using Pandas.
   - Explore the data's structure, distributions, and missing values.
   - Handle missing values appropriately (imputation or removal).
   - Encode categorical variables if necessary.

2. Feature Selection:
   - Select relevant features for building the predictive model.
   - Justify the selection of features based on insights from the data exploration.

3. Model Building:
   - Implement SVM, Linear Regression, KNN, and Naive Bayes models using Scikit-Learn.
   - Train the models using the training dataset.
   - Tune hyperparameters for each model using techniques like Grid Search or Random Search.

4. Model Evaluation:
   - Evaluate the models using the Area Under the Curve (AUC) metric.
   - Compare the performance of different models.
   - Visualize the ROC curve for each model.

5. Submission Preparation:
   - Make predictions on the test dataset using the best-performing model.
   - Prepare the submission file following the format of SampleSubmission.csv.

Submission Requirements:

- Submit the Colab notebook containing the code for data exploration, preprocessing, feature selection, model building, evaluation, and submission preparation.
- Include comments and explanations for each step to ensure understanding.
- Submit the final model predictions in the specified format.

Additional Tips:

- Students should experiment with different algorithms and hyperparameters.
- Discuss the importance of feature selection and its impact on model performance.
- Students should be in a position to interpret ROC curves and understand AUC.