**1. What Changes Were Made for Each Algorithm?**

For every machine learning algorithm, multiple modifications were implemented to improve performance, interpretability, and efficiency. Below are the changes for each algorithm.

**Linear Regression**

* **Feature Standardization**: Standardized the dataset using StandardScaler to ensure that all features contribute equally to model training.
* **Cross-Validation**: Implemented k-fold cross-validation to validate model performance across multiple data subsets.
* **Residual Plots**: Plotted residual errors to check if the model’s assumptions hold.
* **Learning Rate Tuning**: Adjusted the learning rate to improve convergence speed in gradient descent.
* **Additional Metrics**: Added MAE and RMSE as evaluation metrics alongside R-squared.

**Logistic Regression**

* **Feature Scaling**: Used Min-Max Scaling and Standard Scaling to improve optimization performance.
* **Regularization (L1 & L2)**: Added Lasso (L1) and Ridge (L2) regression to prevent overfitting.
* **One-Hot Encoding**: Converted categorical variables into numerical values for better model performance.
* **Precision, Recall, and F1-score**: Included additional evaluation metrics for better performance measurement.
* **Threshold Optimization**: Adjusted decision thresholds to improve classification results.

**K-Nearest Neighbors (KNN)**

* **Optimized K Selection**: Used the elbow method to determine the best number of neighbors.
* **Feature Normalization**: Scaled features to prevent distance-based biases.
* **Weighted Voting**: Implemented distance-weighted voting to improve accuracy.
* **KD-Tree Implementation**: Used KD-Trees to speed up neighbor searches.
* **Decision Boundary Visualization**: Plotted decision boundaries for better interpretability.

**Support Vector Machine (SVM)**

* **Kernel Tuning**: Experimented with linear, polynomial, and RBF kernels.
* **Hyperparameter Tuning**: Used grid search and random search to optimize C and gamma.
* **Soft-Margin SVM**: Implemented soft-margin classification to handle noisy data.
* **Visualization**: Added margin and support vector plots for a better understanding of decision boundaries.

**Decision Tree**

* **Pruning Techniques**: Applied pruning to prevent overfitting.
* **Optimal Depth Selection**: Limited tree depth to improve generalization.
* **Entropy & Gini Index**: Used different criteria to measure node purity.
* **Tree Visualization**: Plotted decision trees for clarity.

**Random Forest**

* **Increased Trees**: Increased the number of trees to stabilize predictions.
* **Hyperparameter Tuning**: Tuned max depth, min sample split, and bootstrap settings.
* **Feature Importance Analysis**: Used SHAP values to analyze feature importance.
* **OOB Error Visualization**: Used out-of-bag error estimates to validate performance.

**K-Means Clustering**

* **Elbow Method**: Used the elbow method to determine the optimal cluster count.
* **Centroid Initialization**: Applied k-means++ for better initialization.
* **Feature Normalization**: Standardized features before clustering.
* **Distance Metrics**: Experimented with Euclidean and Manhattan distance.
* **Cluster Visualization**: Plotted clusters in 2D and 3D.

**Principal Component Analysis (PCA)**

* **Feature Standardization**: Standardized features to remove bias.
* **Dimensionality Reduction**: Reduced features while retaining variance.
* **Explained Variance Ratio**: Computed variance retained by each principal component.
* **Visualization**: Plotted the first two components.

**2. Why Those Changes Were Made?**

Each change implemented in the machine learning models serves a specific purpose, either to improve model performance, prevent overfitting, enhance interpretability, or optimize efficiency. Below is a detailed justification for each modification.

**Linear Regression**

**Feature Standardization** - Ensures all features contribute equally and prevents certain features from dominating.  
**Cross-Validation** - Tests the model on different data splits, reducing overfitting.  
**Residual Plots** - Helps verify that residuals are normally distributed with constant variance.  
**Learning Rate Tuning** - Optimizes gradient descent convergence speed.  
**Additional Metrics (MAE & RMSE)** - Provides better insight into prediction errors.

**Logistic Regression**

**Feature Scaling (Min-Max & Standard Scaling)** - Improves optimization in gradient descent.  
**Regularization (L1 & L2)** - Prevents overfitting by penalizing large coefficients.  
**One-Hot Encoding for Categorical Variables** - Allows the model to correctly interpret categorical data.  
**Precision, Recall, and F1-score** - Provides deeper insights into classification performance.  
**Optimized Decision Threshold** - Improves classification accuracy, especially for imbalanced datasets.

**K-Nearest Neighbors (KNN)**

**Optimized K Value Selection (Elbow Method)** - Determines the best number of neighbors to balance bias and variance.  
**Feature Normalization** - Prevents biased distance calculations.  
**Weighted Voting** - Gives more importance to closer neighbors.  
**KD-Tree Implementation** - Speeds up nearest-neighbor searches for large datasets.  
**Decision Boundary Visualization** - Improves model interpretability.

**Support Vector Machine (SVM)**

**Tuning C and Gamma Parameters** - Enhances classification accuracy by adjusting the margin width and influence of support vectors.  
**Different Kernel Functions (Linear, Polynomial, RBF)** - Ensures the best decision boundary for the dataset.  
**Grid Search and Random Search for Hyperparameter Tuning** - Optimizes parameter selection.  
**Visualization of Support Vectors and Margins** - Helps understand decision boundaries.  
**Soft-Margin SVM Implementation** - Improves generalization in noisy datasets.

**Decision Tree**

**Optimized Tree Depth** - Prevents overfitting while maintaining accuracy.  
**Pruning Techniques** - Reduces complexity by eliminating unnecessary branches.  
**Gini Impurity and Entropy for Best Split** - Determines the best feature splits.  
**Decision Tree Visualization** - Makes decision-making interpretable.  
**Feature Importance Analysis** - Identifies the most influential features.

**Random Forest**

**Increased the Number of Trees** - Improves stability and reduces variance.  
**Tuned Max Depth, Min Samples Split, and Bootstrap Parameters** - Prevents overfitting and optimizes efficiency.  
**Improved Feature Importance Analysis using SHAP Values** - Provides a more accurate breakdown of feature contributions.  
**OOB (Out-of-Bag) Error Estimation** - Estimates generalization error without a separate validation set.  
**Grid Search for Hyperparameter Tuning** - Enhances performance by finding the best hyperparameters.

**K-Means Clustering**

**Elbow Method and Silhouette Score** - Determines the optimal number of clusters.  
**K-Means++ for Better Initialization** - Improves centroid selection for better clustering.  
**Feature Normalization** - Ensures fair clustering by preventing scale bias.  
**Different Distance Metrics (Euclidean, Manhattan)** - Helps find the most suitable distance measure.  
**2D and 3D Cluster Visualization** - Aids in validating clustering performance.

**Principal Component Analysis (PCA)**

**Standardizing Features Before PCA** - Ensures fair contribution from all features.  
**Dimensionality Reduction While Retaining Most Variance** - Speeds up training while preserving information.  
**Explained Variance Ratio** - Helps decide the number of components to retain.  
**Visualization of Principal Components** - Provides insights into transformed data.  
**Applying PCA Before Classifiers** - Improves efficiency and reduces model complexity.

**3. Time (in Hours) to Complete Each Task**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Task** | **Time Taken (hrs)** |
| Linear Regression | Standardization & Residual Plots | 4 |
| Logistic Regression | Regularization & Encoding | 5 |
| KNN | Distance Metric Optimization | 4.5 |
| SVM | Kernel Tuning & Decision Boundaries | 5.5 |
| Decision Tree | Pruning & Depth Optimization | 5 |
| Random Forest | Feature Importance & Tuning | 6 |
| K-Means | Cluster Selection & Initialization | 5 |
| PCA | Dimensionality Reduction & Visualization | 4.5 |

**4. Rating Out of 10 Based on Difficulty**

|  |  |
| --- | --- |
| **Algorithm** | **Difficulty Rating (1-10)** |
| Linear Regression | 5 |
| Logistic Regression | 6 |
| KNN | 7 |
| SVM | 8 |
| Decision Tree | 7 |
| Random Forest | 8 |
| K-Means | 7.5 |
| PCA | 6 |

**5. Importance of Well-Structured Logs**

* **Clearly formatted** for readability.
* **Detailed** to justify all changes.
* **Consistently updated** rather than rushed at the end.
* **Concise yet informative**, avoiding unnecessary clutter.
* **Comprehensive** to showcase the depth of work done.

**6. Notebook Reference - Same Notebook or New Notebook?**

"We have worked on the same notebook and made necessary modifications by taking reference from the existing ones."