

ITEC613 - Machine Learning with Scikit-Learn, Semester 1, 2025

Assessment Task 3

Submitted By: Bishal Bidari

Student ID: S00391799

Contents

[Introduction 3](#_Toc200491211)

[Dataset Preparation (MNIST): 3](#_Toc200491212)

[Classification Models 3](#_Toc200491213)

[Task 3: Model Evaluation & Comparison 3](#_Toc200491214)

[Task 4: Ethical Considerations 3](#_Toc200491215)

[Task 5: Regression Models (Synthetic Housing Data) 4](#_Toc200491216)

[Task 6: Visualizations 4](#_Toc200491217)

[Task 7: Strategic Insights 4](#_Toc200491218)

[Task 8: Conclusion 4](#_Toc200491219)

# Introduction

This project investigates advanced machine learning ideas and their uses - it applies classification plus regression models. The goal is to build models that are accurate and explainable. It uses datasets that come from the real world. The project evaluates models with suitable metrics - it gathers strategic insights, which fit ethical AI practices.

# Dataset Preparation (MNIST):

The MNIST dataset served as image classification. To prepare the data, pixel values were put to scale with StandardScaler. That normalized the input features. A dimension cut occurred through Principal Component Analysis (PCA). It kept 95 % of the data's variance. This lowered the feature count and improved training speed - it also reduced overfitting.

The dataset split into training plus test sets. Custom names were applied - bishal\_train and bidari\_test. A fixed random seed, random\_state=99, was used. For the regression, a synthetic dataset was created. It copied the California Housing dataset - this dataset had 20,640 samples also eight features - it was also scaled for a consistent measure.

# Classification Models

Three classification models went into place and then a comparison happened. The first model, k-Nearest Neighbors (kNN) with k=10, appeared simple and worked well. But it costs a lot to compute large datasets. The second model, a Support Vector Machine (SVM), underwent tests with linear in addition to RBF kernels. The linear SVM trained quickly yet offered little bend. An RBF SVM adjusted using GridSearchCV with a subset of the training data. This model worked much better, because it could model complex decision boundaries by using the kernel trick. A Deep Neural Network (DNN), constructed with TensorFlow, was the last model. The network held two hidden layers which had dropout besides ReLU activations. A softmax output layer followed - it trained for 15 epochs - this model got the best accuracy among all the models.

# Task 3: Model Evaluation & Comparison

All classification models were evaluated using accuracy, precision, recall, and F1-score. A performance comparison table was created to summarize the results. The neural network showed the best overall performance, followed closely by the RBF SVM. Linear SVM performed better than kNN, but not as well as RBF or the DNN. kNN, while simple, struggled with prediction efficiency on larger datasets. These evaluations provided a clear picture of the strengths and limitations of each classifier in terms of performance and scalability.

# Task 4: Ethical Considerations

During the development of the neural network model, the team thought about several ethical ideas in AI. They handled fairness by using balanced datasets, and a model evaluation looked at many metrics. Accountability meant writing down model decisions and architecture. Transparency then centered on explaining the model’s outputs. The developers mentioned tools like SHAP or LIME as possible additions to interpret complex neural network predictions. Such points matter for putting machine learning models into use in a responsible way.

# Task 5: Regression Models (Synthetic Housing Data)

The regression task used four models. We put in place Linear Regression, Polynomial Regression (degree=2), Ridge Regression along with Lasso Regression. We checked the models with Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R²).

Polynomial regression looked like it fit the data too closely, even with a high R². Ridge regression offered the best mix of model intricacy and its work. Lasso simplified the model, but it worked a little less well. The outcomes show that regularization is important for handling overfitting plus multicollinearity.

# Task 6: Visualizations

‍To show the findings, the analysis contained various pictures. Confusion matrices helped to check classifier predictions. Bar charts showed classification model metrics, which summarized how they did. For regression models, scatter plots drew the connection between actual and predicted numbers. This helped to tell how well the model fits - this visual help helped to explain the results and let us talk about them more clearly.

# Task 7: Strategic Insights

From the analysis, several recommendations appear. For simple classification tasks, models such as linear SVM run fast and are easy to understand. For complex jobs that need more accuracy, neural networks or RBF SVMs are better. In regression, linear models perform poorly in high dimensional spaces without regularization. Ridge regression balances complexity plus generalization well, but Lasso helps when one wants to reduce features. This information can direct model choice for actual applications, depending on data complexity and how one plans to use the model.

# Task 8: Conclusion

The project built practical machine learning skills across classification and regression - it incorporated ethical considerations plus effective visual storytelling. Applied skills included dimensionality reduction through PCA and model tuning also optimization using GridSearchCV. The project also addressed model evaluation, comparison along with ethical awareness in AI deployment. This experience showed the importance of model interpretability. It also stressed the balance between simplicity and performance - these competencies are necessary for responsible data science.