Siva Sushmitha Meduri
Written Homework-2

Dwelling Ownership Prediction in Washington State 2023

Support Vector Machine Analysis

ABSTRACT

This report explores the prediction of homeownership in Washington State using Support Vector Machines (SVMs) based on data collected from the US Census and accessed through IPUMS USA Version 13.0. Through SVM models, the goal is to predict whether a dwelling is occupied by owners or renters based on economic, demographic and housing factors. Three SVM kernels - linear, radial, and polynomial - are employed and compared to identify the most effective model for predicting homeownership. Findings highlight "Number of Rooms", "Household Income" and "Age" as strong predictors of ownership. Model results and their implications for dwelling ownership are discussed.

INTRODUCTION

Homeownership is a key aspect of housing stability and wealth accumulation for individuals and families. Predicting homeownership can provide valuable insights for policymakers, housing practitioners, and researchers. By leveraging machine learning techniques, aim is to uncover the underlying factors that influence whether a dwelling is occupied by owners or renters. The dataset utilized for this analysis originates from IPUMS USA, a comprehensive repository of census and American Community Survey (ACS) data spanning from 1790 to the present. The dataset encompasses a wide range of variables related to demographics, housing characteristics, and economic factors. Variables include age, household income, education level, marital status, household size, housing type, and more.

This report employs Support Vector Machine(SVM), a powerful machine learning algorithm, to classify dwellings as either owner-occupied or renter-occupied based on a set of predictor variables. SVM models are particularly well-suited for binary classification tasks and have been widely used in various fields, including housing research. Three different SVM kernels - linear, radial, and polynomial are explored - to identify the most effective model for predicting homeownership. Also, the strongest factors influencing the homeownership using feature importance, we aim to understand the factors driving homeownership associated with housing tenure. By identifying key predictors of homeownership, policymakers and stakeholders can develop targeted interventions to promote housing affordability and equity.

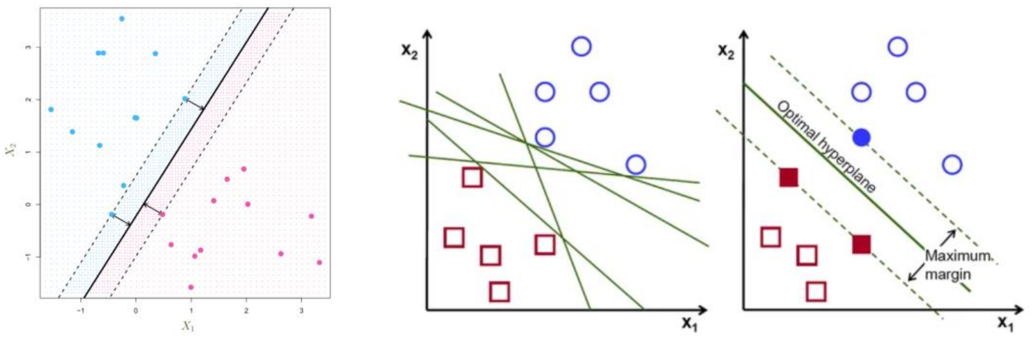
THEORITICAL BACKGROUND

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs are best suited for classification problems than the regression ones. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection.

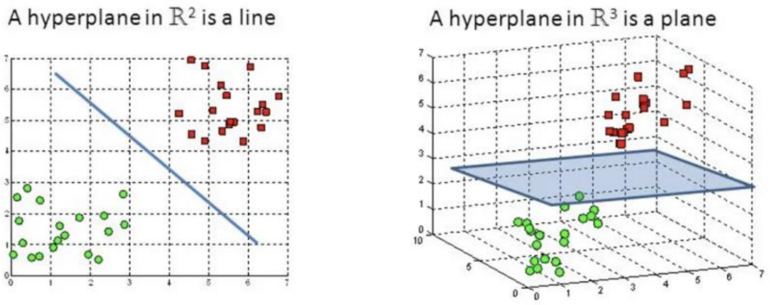
SVM can be of two types:

- **Linear SVM**: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM**: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Linear SVM: The main objective of the SVM algorithm is to find the "**optimal hyperplane**" in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible.

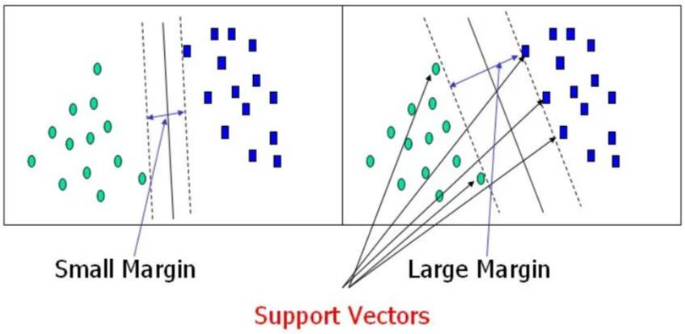


Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.



Hyperplanes in 2D and 3D feature space

A natural choice is the maximal margin hyperplane (also known as the maximal optimal separating hyperplane), which is the separating hyperplane that is farthest from the training observations. That is, we can compute the (perpendicular) distance from each training observation to a given separating hyperplane; the smallest such distance is the minimal distance from the observations to the hyperplane and is known as the "**margin**". The wider margin indicates better classification performance.

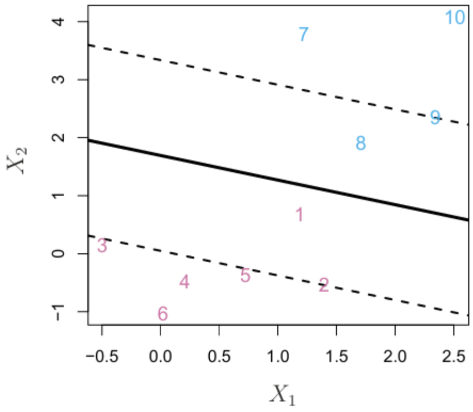


Large Margin Intuition: In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the sigmoid function. If the squashed value is greater than a threshold value(0.5) we assign it a label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values([-1,1]) which acts as margin.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane.

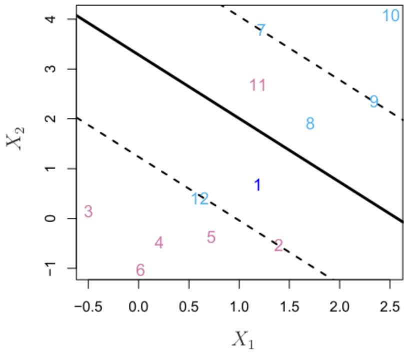
The "maximum-margin hyperplane" or the "**hard margin**" hyperplane is a hyperplane that properly separates the data points of different categories without any misclassifications. SVMs has greater robustness to individual observations, and better classification of most of the training observations.

The support vector classifier, sometimes called a "**soft margin classifier**", support does exactly this. Rather than seeking the largest possible margin so that every observation is not only on the correct side of the hyperplane but also on the correct side of the margin, we instead allow some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane. The margin is soft because it can be violated by some of the training observations.



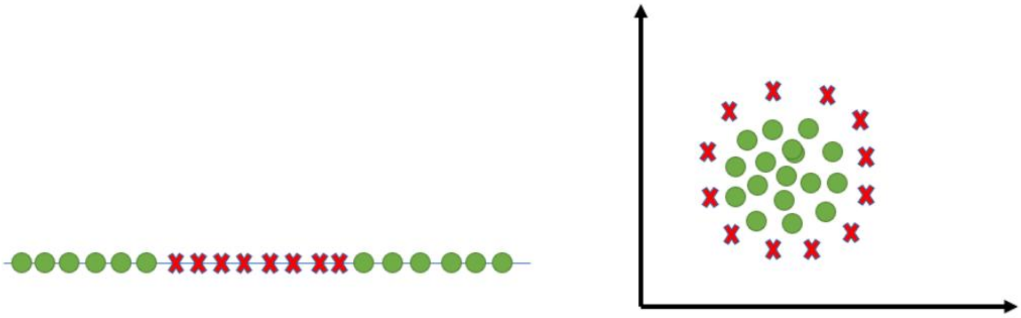
In the above picture, most of the observations are on the correct side of the margin. However, a small subset of the observations are on the wrong side of the margin.

An observation can be not only on the wrong side of the margin, but also on the wrong side of the hyperplane. In fact, when there is no separating hyperplane, such a situation is inevitable. Observations on the wrong side of the hyperplane correspond to training observations that are misclassified by the support vector classifier. This case is shown in the picture below.

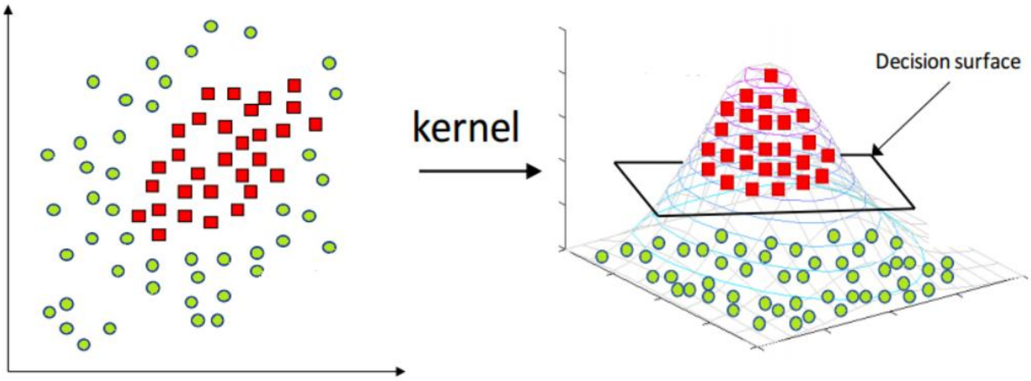


Tuning Parameters and Optimisation Techniques: To achieve optimal performance, SVMs require careful tuning of several parameters.

Kernel is the mathematical function, which is used in SVM to map the original input data points into high-dimensional feature spaces, so, that the hyperplane can be easily found out even if the data points are not linearly separable in the original input space.

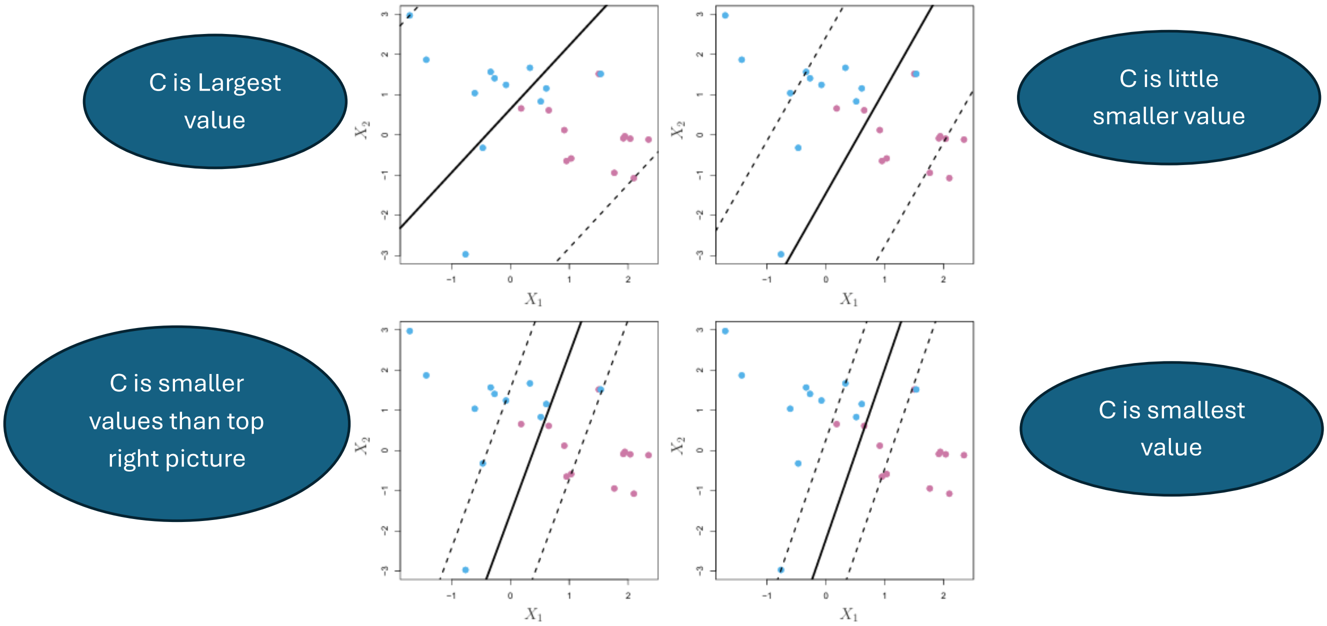


Here we see we cannot draw a single line or say hyperplane which can classify the points correctly. So what we do is try converting this lower dimension space to a higher dimension space using some quadratic functions which will allow us to find a decision boundary that clearly divides the data points. These functions which help us do this are called Kernels and which kernel to use is purely determined by hyperparameter tuning.



Some of the common kernel functions are:
Linear Kernal- Suitable for linearly separable data

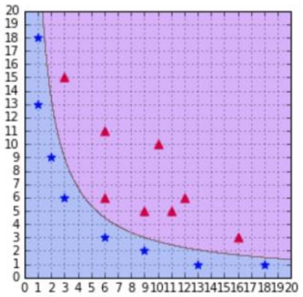
Regularization Parameter (C): The regularization parameter, often denoted as C, controls the trade-off between maximizing the margin and minimizing classification errors. A smaller C value leads to a larger margin but may result in misclassification of some training examples (soft margin). Conversely, a larger C value allows for fewer misclassifications but may lead to overfitting (hard margin).



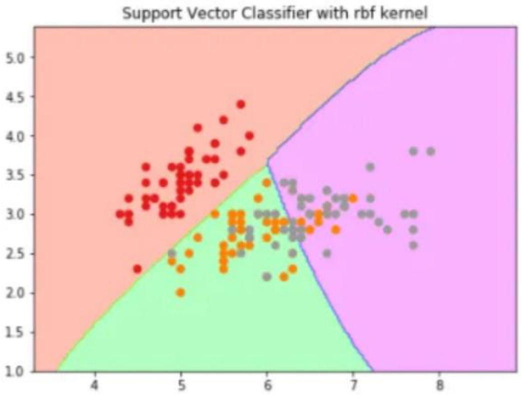
Polynomial Kernal - Suitable for data with polynomial decision boundaries. Requires tuning of the degree parameter.

Degree Parameter (degree): The degree parameter is specific to polynomial kernels and controls the degree of the polynomial function used to separate classes. Higher degree values allow for more complex decision boundaries but may increase the risk of overfitting.

Radial Basis Function(RBF)- Effective for non-linearly separable data; requires tuning of the gamma parameter.



A SVM using a polynomial kernel is able to separate the data (degree=2)



*Gamma Parameter (gamma):*Gamma is a parameter specific to the RBF kernel and determines the influence of individual training samples on the decision boundary. A smaller gamma value implies a smoother decision boundary, whereas a larger gamma value results in a more complex boundary, potentially leading to overfitting.

For calculating the optimal parameters mentioned above, k-fold cross validation is implemented. Cross-validation is a crucial technique for tuning SVM parameters and assessing model performance. Techniques such as k-fold cross-validation split the dataset into multiple subsets (folds), allowing for rigorous evaluation of model performance across different parameter configurations. Grid search and randomized search are commonly used strategies for tuning hyperparameters by systematically exploring a range of values and selecting the combination that yields the best performance on validation data.

Advantages of SVMs:

SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships. Robust to overfitting. Versatile kernel functions allowing for flexibility in modeling complex relationships in the data. Tuneable parameters that allow users to customize the model's behavior according to the specific requirements of the problem. SVMs use only a subset of training points (support vectors) to define the decision boundary, making them memory efficient, especially for large datasets.

Disadvantages of SVMs:

Computationally intensive especially for large datasets, non-linear kernels, or high-dimensional feature spaces. SVMs inherently support binary classification and extending them to handle multiclass problems requires additional techniques such as one-vs-one or one-vs-all strategies. The performance of SVMs is sensitive to the choice of kernel function and its parameters, requiring careful selection and tuning for optimal results.

METHODOLOGY

Data Exploration:

The methodology employed in this analysis began with data preparation, utilizing the data collected from the US Census and accessed through IPUMS USA Version 13.0. The dataset named “Housing_revised.csv” is loaded. Few column names in this revised dataset are changed to human readable names. To check the range of values and their structure, summary statistics is calculated and investigated the data structure. Based on the codebook, the N/A values are checked for the selected features which are utilised in this analysis. If existed, they are removed in the data cleaning process. Out of the full dataset, only five predictors are chosen for analysis. The are Number of Rooms, Household Income, Age, Marital Status, Education. The reason for picking these predictors are: **Rooms:** The number of rooms in a dwelling is often indicative of its size and value. Larger homes tend to be associated with higher incomes, which are often necessary for homeownership.

Age: Age can be a proxy for financial stability and life stage. Younger individuals might be less likely to own homes due to financial constraints, while older individuals might have had more time to accumulate wealth and transition to homeownership. **Household Income:** This is a direct and intuitive predictor of homeownership. Higher-income households are generally more likely to be able to afford a home. **Education:** Education level is often linked to higher-paying jobs and increased financial stability, making homeownership more attainable. **Marital Status:** Marital status can influence household finances and living arrangements. Married couples, for example, might be more likely to pool resources and prioritize homeownership. Out of these five, Marital Status and Education are categorical variables. As, SVMs are fundamentally based on finding the optimal hyperplane that separates different classes in your data. This relies on calculating distances between data points. Categorical variables (e.g., "red", "blue", "green") don't have meaningful numerical distances that can be directly used. So, these two categorical variables are encoded into continuous numerical variables by implementing ordinal encoding.

For Marital Status column, it is divided into three columns 'MARRIED', 'DIVORCED', and 'SINGLE'. It does this by applying conditions to the original 'MARITAL_STATUS' column and converting the results into binary (0 or 1) values. Similarly, the Education column is divided into three new 'EDU_PRIMARY_SCHOOL', ‘EDU_HIGH_SCHOOL', 'EDU_COLLEGE’. Later, the target variable Ownership is also checked for N/A values and removed if existed. Then, this target variable is also converted to into numerical variables using binary encoding, with categories like 'Rented' and 'Owner'. It creates a new column: ‘OWNERSHIP_BINARY': Gets a value of 1 if the 'OWNERSHIP' value is 'Owner'. Gets a value of 0 if the 'OWNERSHIP' value is 'Rented' (or any other value). This new variable OWNERSHIP_BINARY is considered as target variable for our analysis.

These are the predictors: 'ROOMS', AGE','HOUSEHOLD_INCOME','EDU_PRIMARY_SCHOOL', 'EDU_HIGH_SCHOOL', 'EDU_COLLEGE', 'MARRIED', 'DIVORCED', 'SINGLE'. Then dataset is divided into training and testing sets using an 80-20 split ratio to facilitate model evaluation. The training set is utilized to train SVM models, while the testing set is reserved for evaluating model performance. Once the features are set, preprocessing step such as feature scaling is also implemented.

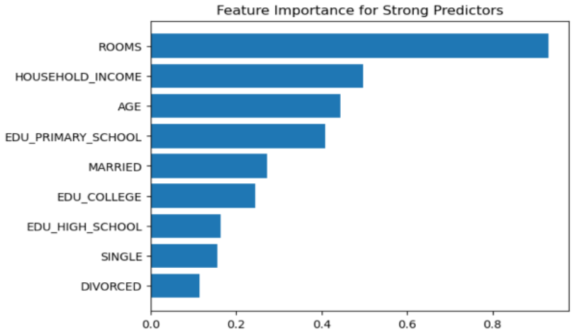
Modelling:

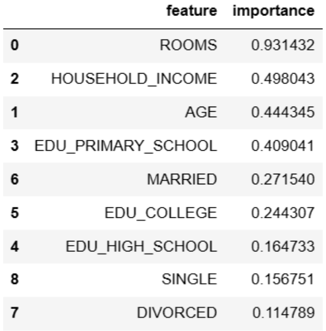
Three SVM models were trained using different kernels: linear, RBF, and polynomial. First, a SVM model with some random parameters of gamma, C and degree was trained on a subset of the data. The performance on test set was evaluated with accuracy.

Tuning Parameters : The linear SVM model was trained with various values of the regularization parameter (C) to find the optimal balance between model complexity and generalization. The RBF SVM model was trained with different values of the regularization parameter (C) and the kernel coefficient (gamma) to optimize the decision boundary. The polynomial SVM model was trained with different degrees of the polynomial kernel to capture non-linear relationships between features. To obtain the optimal cost value gamma and degree, k-fold cross validation is performed. New model with optimal parameters are fit and the performance of each SVM model was evaluated using accuracy metrics on a test set. Feature Selection: Additionally, the strongest predictors of the Homeownership through feature importance are calculated. It’s known that Rooms and Household Income are top two strongest predictors in predicting the Homeownership analysis. The models' decision boundaries were visualized using plots to interpret their performance and understand the relationships between predictors and homeownership.

COMPUTATIONAL RESULTS AND DISCUSSIONS:

SVM Decision Boundary- Linear kernel: This is the plot generated for finding out the top two strongest predictors obtained through feature selection: “ROOMS” and “HOUSEHOLD_INCOME”.



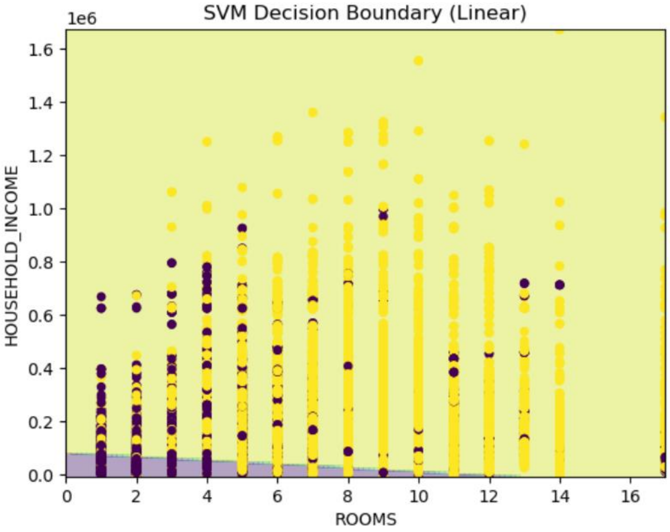


	feature	importance
0	ROOMS	0.931432
2	HOUSEHOLD_INCOME	0.498043
1	AGE	0.444345
3	EDU_PRIMARY_SCHOOL	0.409041
6	MARRIED	0.271540
5	EDU_COLLEGE	0.244307
4	EDU_HIGH_SCHOOL	0.164733
8	SINGLE	0.156751
7	DIVORCED	0.114789

It is evident from the graph and table above that, Rooms and Household_Income are strongest predictors.

PLOT:

Below is a plot demonstrating an SVM decision boundary on two strong predictor variables, namely ROOMS and HOUSEHOLD_INCOME. This plot illustrates how the SVM classifier divides the feature space into regions corresponding to different classes (e.g., owners vs. renters).



DISCUSSION:

The X-axis represents the number of rooms in a house. The y-axis represents the household income. Each data point likely represents a house with a certain number of rooms and a corresponding price. In an SVM, the decision boundary is the hyperplane that best divides the data points between the target classes. In this case, it's likely separating "rented" from "owned" housing units based on the number of rooms and household income. It can be interpreted that based on the areas above and below the decision boundary: **Upper Right:** This area likely represents houses with more rooms and higher household income. The SVM have classified these as more likely to be “Owners”. **Lower Left:** This area likely represents houses with fewer rooms and lower household income. The SVM have classified these as more likely to be “Rented”. In some cases, there may be data points that are misclassified by the model. These are points that lie on the wrong side of the decision boundary. For example, if a blue data point lies in the yellow region, or vice versa is indicated as a misclassification.

From the interpretation of the SVM plots, it seems that individuals or households with higher income levels and larger dwelling sizes (measured by the number of rooms) are more likely to own their homes. This suggests that financial stability and access to larger housing spaces play crucial roles in homeownership.

These findings have important implications for addressing societal challenges surrounding access to housing. Policies aimed at promoting homeownership should focus on improving income equality, providing affordable housing options, and facilitating access to financial resources for prospective homebuyers.

RESULTS:

After experimenting with three different kernel functions (linear, radial basis function (RBF), and polynomial) and different optimal parameters, the optimal Regularization parameter C attained through cross-validation for linear kernel is C=5, and its resultant SVM model obtained a test accuracy of 80.5%. The optimal kernel parameters gamma = 1 and C = 1 is attained through cross-validation for Radial Basis Function Kernel, and its resultant SVM model obtained a test accuracy of 81.3%. The optimal degree of polynomial for the polynomial kernel attained through cross-validation is degree=5 and its resultant SVM model obtained a test accuracy of 81%. Key Takeaways: From our analysis, we identified several strong predictors of homeownership, including: HOUSEHOLD_INCOME: Higher household income levels were associated with a higher likelihood of homeownership. ROOMS: Dwellings with a greater number of rooms tended to be owned rather than rented.

CONCLUSION:

In this study, Support vector Machine is employed to conduct analysis and to investigate the “Dwelling Ownership” from the data from the US Census and accessed through IPUMS USA. Among the models tested, the RBF and polynomial SVM models achieved the highest accuracy of 81%(small variation compared with linear SVM model). This suggests that the non-linear decision boundaries captured by the RBF kernel were better suited for separating the classes in the feature space compared to the linear and polynomial kernels. Based on our findings, we can make the following hypothetical recommendations to policymakers: Implement policies aimed at increasing access to affordable housing options, particularly for low- and moderate-income households, to promote homeownership. Develop programs to provide financial assistance and incentives for first-time homebuyers, such as down payment assistance and mortgage loan subsidies.

REFERENCES

1. Steven Ruggles, Sarah Flood, Matthew Sobek, Danika Brockman, Grace Cooper, Stephanie Richards, and Megan Schouweiler. IPUMS USA: Version 13.0 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D010.V13.0
2. Plot scikit-learn (sklearn) SVM decision boundary / surface. https://stackoverflow.com/questions/51297423/plot-scikit-learn-sklearn-svm-decision-boundary-surface
3. Saini, A. (2024, January 23). Guide on Support Vector Machine (SVM) Algorithm. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/
4. Gandhi, R. (2018, June 7). Support Vector Machine — Introduction to Machine Learning Algorithms. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
5. Support Vector Machine (SVM) Algorithm. (2023, June 10). GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm/
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009): Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. ISBN: 978-0-387 84857-0