2024

Analysis of Housing Ownership and their pred

Tileshwar Narayan

Abstract:

The research aims to understand the pattern of dwellings ownership using the support vector models with the range of the linear, radial, and polynomial kernels techniques thanks to the provided data obtained from the US Census using IPUMS USA. The prescreening was designed to mark out the most senior person in each home as the most important individual and to order others under the senior's name. The incorporation of SVC analytical method which takes into consideration the testing accuracies as the measure of model performance was the focus of the analysis. Linear Kernel SVC proved its effectiveness presenting almost the same train and test accuracy 82.61 % and 82.81% for both normal and cross-validated models. On the other hand, in the next model, the test accuracy of RBF algorithm was 83.07% which gradually decreased later after implantation of cross-validation that was 82.07%. As regards training accuracy RBF – this value increased from 82.91% to 86.52%. A polynomial kernel scored 82.71% in testing, and after cross-validation, it also got a slight lift of 83.01% in test accuracy, while the training data we have the accuracy of 81.90% and after cross validation we have 83.01%. The investigated robust kernel support vector classification (RBF kernel SVC) proved to be the most effective method for housing occupancy recognition, considering the accuracy test it has achieved. This outcome testifies to the fact that kernel type choice is of substantial importance in support vector models and is also indicative of the gradient effect of cross-validation on model's performance. The research will provide useful information of the application of machine learning techniques for authenticating the dwellers of premises which will be a great notion when we come to housing policy and urban planning.

Introduction:

In this analysis a data-driven model has been created based on SVMs which is a classification method rooted in census data for Washington State. Knowing the factors of household occupancy is of value for the urban planning and making policies. The recognition of different drivers of occupancy which include the demographic data and housing statistics, helps policymakers to develop interventions and improve the housing financial affordability and the economic gap. The research aim is to use SVM algorithms for assessing and distinguishing structures whether they are owner occupied or a renter occupied and to identify relationships between variables and occupancy status. The goal of research is to examine the application of SVMs for an occupancy classification problem and to identify the factors that influence building occupancy ownership.

Overview:

The research utilizes support vector classifiers (SVCs) as basis for dwelling occupancy (owned or rented) prediction by means of Washington State Census data. The study will accomplish this by analyzing SVM performance using linear, radial, and polynomial kernels along with cross validation on different parameters which will help to provide a picture of connections between predictor variables including age, income, education level, marital status, number of family members, housing features like number of bedrooms and number of rooms and occupancy status. The dataset provides an excellent opportunity for exploring individual demographics and housing characteristics in depth. It will thus enable a comprehensive survey of the causes of ownership of the dwelling. The study adopts a robust classification modeling technique and identifies significant predictors which provide informative direction for policy makers, urban planners, and other actors in the housing sector.

Description of the Dataset:

The original data set provided has 24 different variables with one target variable. The data consists of many factors like personal data, demographic and dwelling details. The variable 'OWNERSHP' is our target variable. Other important variables are "DENSITY," representing the population density of the surrounding area, "HHINCOME" for household income, "ROOMS" denoting the number of rooms in a dwelling, and "BUILTYR2" specifying the year of construction for the dwelling. Additionally, demographic attributes such as "AGE," "MARST" (marital status), and "EDUC" (educational attainment) contribute to a comprehensive understanding of the dataset. We have a very comprehensive data set in hand which will be used for the analysis of the ownership of the dwellings.

Theoretical Background:

Support Vector Classifier (SVC) **(ref: 3)** is one of classifiers that belong to supervised learning. This is used when we calculate accuracy through classification and classify the factors affecting the study. It divides data with classes to achieve the highest possible accuracy and finally minimizes the error. The hyperplane that was formed from the model with the best fit is the most robust. Through the variation of parameters and conditions, SVC can also take different forms. This classifier is not only separating from linear pattern but non-linear classifier as well because of employing many kernel techniques. SVC by these kernels makes it possible to image object by effecting the 2D data into higher-dimensional spaces where the classes become the linear separation lines and shapes that lead to the complex relation modelling and selecting the decision boundaries. Also, SVC is a classifier which is driven mainly on high margin maximization to define its optimization process in which margin is increased by moving the decision boundary away from the support vectors which are data points so close to the boundary. Regularizing parameters such as C (cost), gamma and degree help to regulate the model performance between the highest margin increase and the lowest error decrease, they are good to control the overfitting problem itself. Even though SVC could possess either its applicability and achievements or it can be a complex technique in case of dealing with many data points, the computational problem might arise in SVC.

There are three kernels that we have used in our analysis i.e., Linear, RBF (Radial Basis Function), and Polynomial. Every single one of the kernels affords a unique angle for SVC to perceive and categorize data samples. The Linear kernel assumes a simple approach dealing with data points on a plane divided by lines or hyperplanes. This makes it easy to use for datasets that have easily distinguishable splits. The linear kernel uses C or cost for as hyper tunning parameter which can be done through any number. While, the RBF kernel takes data through a many-dimensional space, to capture detailed connections between features. RBF indeed allows us to handle more complicated, nonlinear boundaries and can be adaptive in terms of processing of the datasets of various levels of complexity. The RBF Model uses Cost and gamma points to do the hypertunning the model. On the other hand, the Polynomial kernel feature has mathematical expertise by generating the curves in the data for SVM to carry out the tasks that a simple boundary can't. This type of kernel can be used to classify datasets of all degrees of polynomial functions, as it accounts for those datasets with curved decision boundaries. The polynomial kernel uses C the cost and degree values which has minimum point of 2 to hypertune the model.

Methodologies:

The data provided has been cleaned and we have used different subsets for 3 models. The first step in updating the data as per serial number and Age, as we have the serial number which are repeating with the number of people living in that dwelling. If we have 4 people in the house the serial will be repeated 4 times. So, to clear our data I have considered the person who is eldest in the house and considered the person as the ownership/renter. So, we have removed all other rows which were reporting. Now since the first level is done, I have used some variables which have been used for our model building. I have used the following variables like DENSITY, OWNERSHIP, HHINCOME, ROOMS, BEDROOMS, VEHICLES, NFAMS, NCOUPLES, AGE, MARST, and EDUC. The DENSITY has categorized under 3 categories, all values which are under 2000 will be set to 1, values from 2000 to 4000 will be set to 2 and values 4000 and more will be categorized as 3.

Similarly, I have categorized the values of HHINCOME into 3 categories too. In this case, any values which is less than 100000 has been categorized as 1, values between 100001 and 250000 have been categorized as 2 and anything above 250001 has been categorized as 3. In the VEHICLES column we have observed that those who do not have any vehicle have been categorized as 9 which now has been changed to 0 for better understanding. For the variable MARST or marital status has been categorized as 3 separate variables i.e., isMarried, isDivorced and isNeverMarried. The values 1 and 2 have been combined to make it one and similarly values 3, 4 and 5 i.e., Separated, Divorced and Windowed has been combined as one and NeverMarried has been left alone. Similarly, I have created separate variables for Education level. The first variable contains NoSchooling who have NA or no education, PrimarySchooling for those who have studied till class 4. The person who are between 5 to 12 have been classified under Schoolings and all above data has been separated and considered under College.

We have been doing the train-test split by a ratio of 70:30 i.e., 70% train and 30% test with a random state of 42. We have created 2 models each for Linear, RBF and Polynomial. The first model has been created taking Cost C as 0.1 directly and the next step we have used cross validation with different C values. The rbf linear model was created over normal model at first and then cross validation has been provided over that. It has used Cost and gamma values to do the model. At last, the svc with polynomial model where again we have created a single normal model with C cost and degrees and after that we have done the parameter tuning under cross validation.

For linear model we have created a set of variables that has been used to create the model like 'DENSITY', 'OWNERSHP', 'HHINCOME','BEDROOMS', 'VEHICLES', 'NFAMS', 'NCOUPLES', 'AGE', 'isMarried', 'NoSchooling', 'PrimarySchooling'. For the next rbf model has been creating a dataframe with 'OWNERSHP', 'HHINCOME', 'BEDROOMS', 'VEHICLES', 'AGE', 'isNeverMarried', 'College', and the polynomial kernel model is being created by considering 'OWNERSHP', 'HHINCOME', 'BEDROOMS', 'VEHICLES', 'NFAMS', 'AGE', 'isDivorced'.

Computational Results:

Linear Kernel:

Support Vector Classifier with Linear Kernel:

```
Training accuracy: 0.8261212374194147
Testing accuracy: 0.8281571258521805
Training error rate: 0.17387876258058532
Testing error rate: 0.1718428741478195
```

Support Vector Classifier with Linear Kernel with Cross Validation:

```
Grid Search Result:
GridSearchCV (cv=KFold (n_splits=5, random_state=1, shuffle=True),
              estimator=SVC (),
              param_grid= {'C': [0.1, 1, 5], 'kernel': ['linear']}

Best parameters: {'C': 1, 'kernel': 'linear'}
```

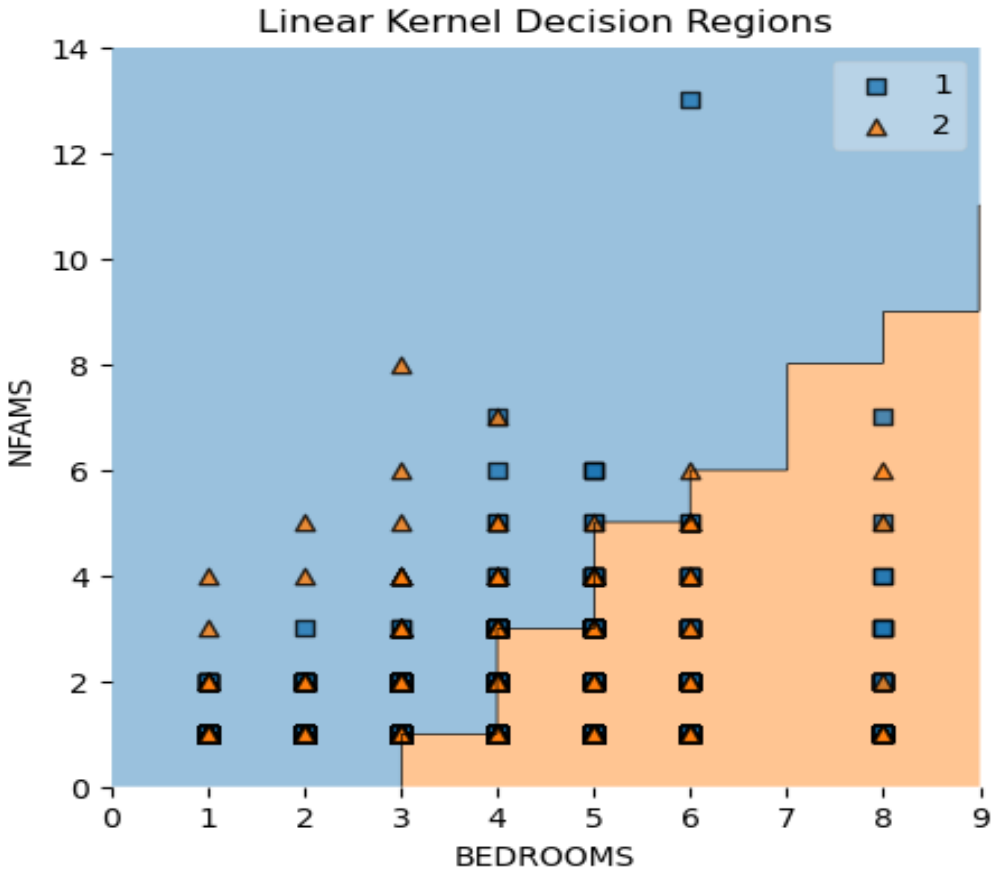
Accuracy Rates:

```
Training accuracy (CV): 0.8261212374194147
Testing accuracy (CV): 0.8280489124553619
Training error rate (CV): 0.17387876258058532
Testing error rate (CV): 0.17195108754463806
```

Variables Importance:

	Feature	Importance Score
2	BEDROOMS	0.735266
9	PrimarySchooling	0.710123
4	NFAMS	0.559557
7	isMarried	0.468555
8	NoSchooling	0.413048
1	HHINCOME	0.411556
3	VEHICLES	0.236372
5	NCOUPLES	0.203581
0	DENSITY	0.118222
6	AGE	0.029405

Decision boundary plot with best features:



Radial Basis Function Kernel:

SVC RBF without Cross Validation:

```
SVC (C=1, gamma=0.1)

Number of support vectors: 9786
```

Accuracy Rate:

```
Training accuracy: 0.8291823199295023
Testing accuracy: 0.8307542473758252
Training error rate: 0.1708176800704977
Testing error rate: 0.16924575262417485
```

SVC RBF with Cross Validation:

```
GridSearchCV (cv=KFold (n_splits=5, random_state=1, shuffle=True),
              estimator=SVC (),
              param_grid= {'C': [1, 5, 10], 'gamma': [1, 2, 5],
                           'kernel': ['rbf']})

Best parameters (RBF): {'C': 1, 'gamma': 1, 'kernel': 'rbf'}
```

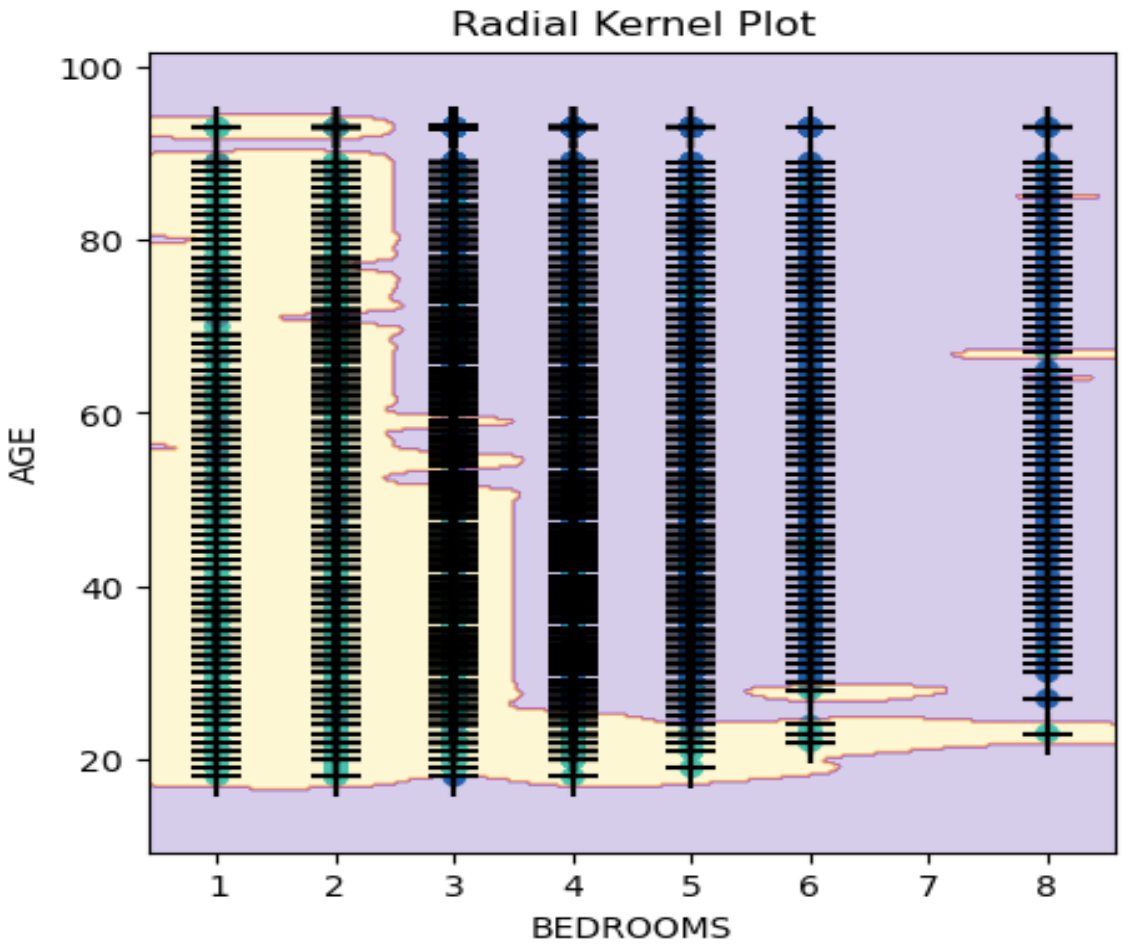
Accuracy Rates:

```
Training accuracy (RBF, CV): 0.8652659895181114
Testing accuracy (RBF, CV): 0.8207986148685207
Training error rate (RBF, CV): 0.1347340104818886
Testing error rate (RBF, CV): 0.17920138513147932
```

Important Features:

```
Feature: BEDROOMS, Importance Score: 0.09338816145438807
Feature: AGE, Importance Score: 0.05772102586300183
Feature: VEHICLES, Importance Score: 0.03878368141975974
Feature: HHINCOME, Importance Score: 0.012747538145222382
Feature: isNeverMarried, Importance Score: 0.003484471377556542
Feature: NFAMS, Importance Score: 0.0020560545395519767
Feature: College, Importance Score: 0.00043285358727409575
```

Decision Boundary Graph:



Polynomial Kernel:

SVC Poly without Cross Validation:

```
SVC (C=1, degree=2, kernel='poly')

Number of support vectors: 10692
```

Accuracy Scores:

```
Training accuracy: 0.8190714716386067
Testing accuracy: 0.8271832052808138
Training error rate: 0.18092852836139328
Testing error rate: 0.1728167947191862
```

SVC Poly with Cross Validation:

```
Best parameters (Poly): {'C': 10, 'degree': 2}
```

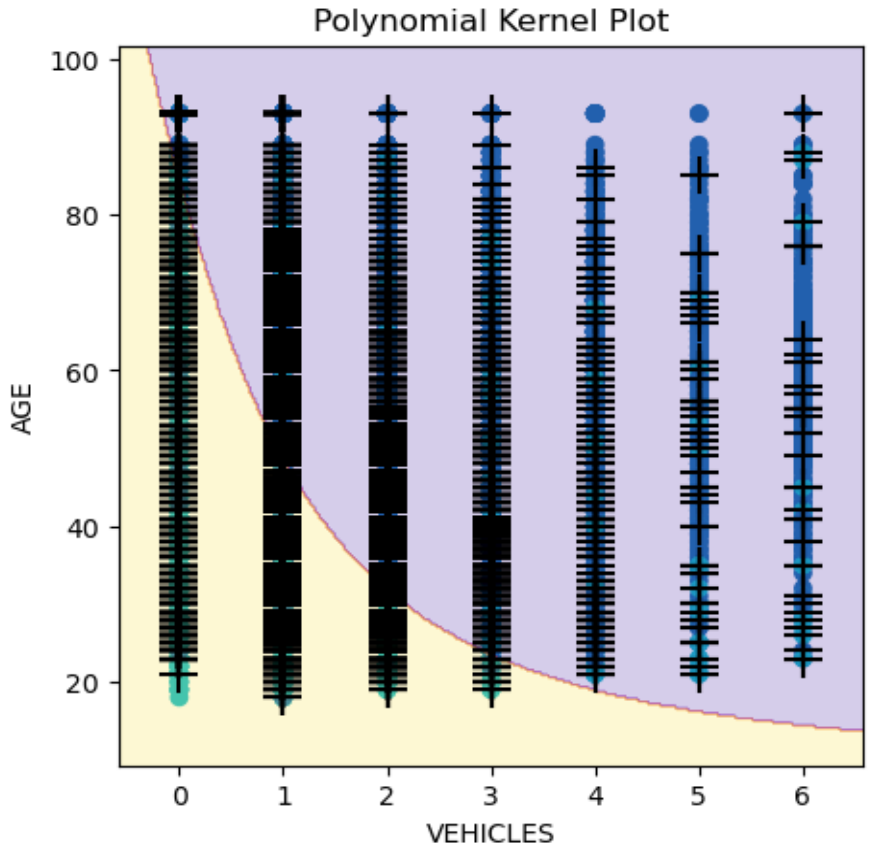
Accuracy Scores:

```
Training accuracy (Poly, CV): 0.8214832336162515
Testing accuracy (Poly, CV): 0.830104966994914
Training error rate (Poly, CV): 0.17851676638374847
Testing error rate (Poly, CV): 0.16989503300508602
```

Important Features:

```
Feature: BEDROOMS, Importance Score: 0.10877610648198248
Feature: AGE, Importance Score: 0.07224326371604806
Feature: VEHICLES, Importance Score: 0.01352667460231578
Feature: HHINCOME, Importance Score: 0.009933989827940715
Feature: NFAMS, Importance Score: 0.005670381993290774
Feature: isDivorced, Importance Score: 0.0019045557840060567
```

Decision Boundary Graph:



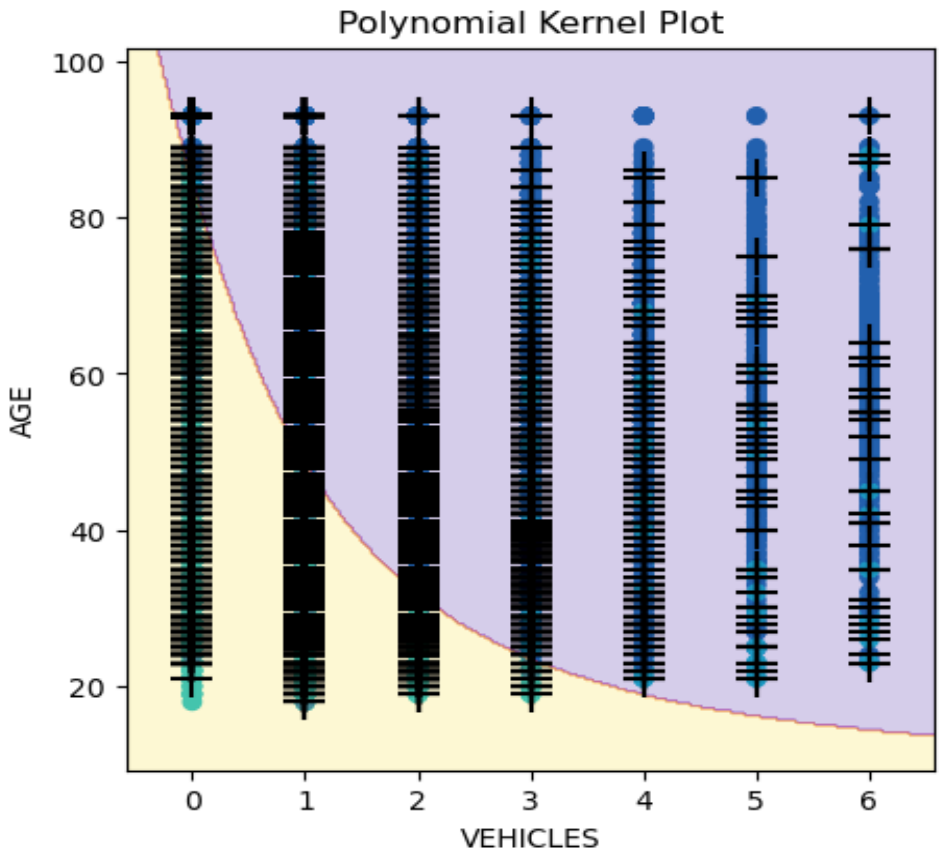
Discussion:

The analysis has gathered more information that was required to come to any conclusion. The first model with Linear kernel without any cross validation has yielded an accuracy of 82.61% for training data and an accuracy of 82.81% for test data. The cross validation on this model has provided approximately the same results as 82.61% for training data and 82.80% for test data. While the important features concluded after the analysis were bedrooms in the house, primary education, number of family living in the house and is married status. This model indicates a pattern where if the number of bedrooms is more and the number of family living together indicates a higher household income which is one the factor which can influence the purchase of the dwellings. The analysis further uses the radial basis function kernel with SVC where the results without cross validation the accuracy scores are 83.16% for training data and 83.17% for test data. The accuracy of the model after cross-validation is 87.26% for training data and 81.98% for test data, which is better than the rbf model without cross validation. Some of the important features that came out during this analysis are the number of bedrooms, age, number of vehicles, and household income. This indicates that this model provides the first important feature as number of bedrooms along with age, number of vehicles and the household income, higher these numbers are the higher the chances that the person living in the dwelling is an owner. Now for our final model SVC with Polynomial kernel without cross validation, we have got accuracy scores of 81.90% for training data and 82.71% for test data. Then after cross validation the model's accuracy scores are 82.14% for training data and 83.01% for test data. The model performance has been improved slightly after cross validation and some of the important features classified by this model are number of bedrooms, age, vehicles, and household income. Indicating similar results from earlier models.

The analysis suggests that there are 4 main variables which are getting the highest importance namely the number of bedrooms in the dwelling which means the higher the number of bedrooms in the dwelling the chances are the residence is an owner. Similarly, Age is also one the factor which suggest that the person with higher age can have accumulated more amount of money to buy a dwelling which is also one of the factors influencing the analysis. Number of vehicles is also a similar indicator which includes the higher the number of vehicles may correspond to higher income level or higher worth which also provides a significant position to own a dwelling. One of the last important factors is the household income, which works like more the household income the higher the chance that the resident is an owner as they will be financially stronger than the other people which has single income source making their way to take some risk and own a property. Some other factors instead of these 4 can be number of family members and education level.

Model	Training Accuracy	Testing Accuracy
Linear Kernel	0.826121	0.828157
Linear Kernel (CV)	0.826121	0.828049
Radial Basis	0.829182	0.830754
Radial Basis (CV)	0.865266	0.820799
Polynomial Kernel	0.819071	0.827183
Polynomial Kernel (CV)	0.821483	0.830105

Here is one of the Decision boundary graphs from polynomial model for our further analysis.



The following graph is between Age and Vehicles. The pattern over here provides a pattern in the ownership. The yellow side is for the renters (2) and the blue part is for the owners (1). The model here indicates that as the number of vehicles increases the chances of being an owner are more powerful as we can observe from the graph where the number of veh icles is more than 4. Having a low age of 20 but having a minimum of 4 cars indicates that the person's background is ver y strong, and he is financially not at risk which makes him counted as an owner at this age. As age increases along with t he number of vehicles owned, the person will be classified as an owner. The group which does not own a car are most lik ely to be a renter which is very clearly classified under this model. Persons with one vehicle but more than the age of 55 can be under owners. People who have 2 cars and are above 35 are also classified under owner. '+' signs in the indicates the concentration of the classification. A '+' sign on the yellow side indicates that less vehicles with higher age are likely t o be owners and the same on blue side indicates a pattern where the vehicles are more, dweller can be an owner even a t the age is less indicating a financial independence and strong background. Any point near the decision boundary line wi ll show a possible error or overlapping of the results or we can say its misclassified. Overall, the 4 main factors i.e., numb er of bedrooms, number of vehicles, Age and household income are important factors which affect our overall model by 10% to 12% in our accuracy indicating the importance of these factors.

Interpretation:

The analysis of dwelling occupancy using the support vector models with linear, radial, and polynomial kernels provides almost consistent results across the three models. Higher accuracies are generally associated with variables such as num ber of bedrooms, age, number of vehicles, and household income, indicating their importance in determining dwelling o wnership. Specifically, larger households and greater financial stability, as indicated by factors like more bedrooms, vehic les, and higher income, are strongly correlated with dwelling ownership. The influence of age suggests the accumulation of wealth over time as a contributing factor. While model accuracies vary slightly, with improvements observed after cro ss validation in some cases, the fundamental relationship between these key variables and dwelling ownership remains c onsistent throughout the analysis, emphasizing their significance in housing ownership variations.

Conclusion:

The analysis done on this dataset with support vector models provides different training and testing accuracies across dif ferent kernels. For the linear kernel model, both with and without cross-validation, training accuracies are around 82.61 % with slightly higher testing accuracies, approximately 82.81%. The radial basis function (RBF) kernel model generated a higher training accuracy, around 82.91%, and slightly higher testing accuracy of about 83.07%, but its performance dec reases after cross-validation, particularly in testing accuracy, which drops to approximately 82.07% but the training accu racy has been improved to 86.52%. In the next model, the polynomial kernel model displays slightly lower training accur acies, approximately 81.90% without cross-validation and 82.14% after cross-validation, with testing accuracies around 8 2.71% and 83.01.0%, respectively. Important variables identified across models include the number of bedrooms, age, n umber of vehicles, and household income, highlighting their significance in determining dwelling occupancy patterns. Th ese findings emphasize the importance of considering both training and testing accuracies, as well as the impact of cross -validation, in assessing the effectiveness of support vector models for analyzing dwelling occupancy.

References:

1. Feature Importance for Any Model using Permutation by Taylor Jensen (Codes) https://medium.com/@T_Jen/feature-importance-for-any-model-using-permutation-7997b7287aa

2. Permutation feature importance (Codes for Feature Importance) https://scikit-learn.org/stable/modules/permutation_importance.html#:%7E:text=The%20permutation%20feature%20importance%20is,model%20depends%20on%20the%20feature

3. Implementing SVM and Kernel SVM with Python's Scikit-Learn (Theory)(Page 3) https://www.geeksforgeeks.org/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/