www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(1), Jan-April 2025, 75-89. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i1.500

Predicting Urban Expansion in Zacatecas-Guadalupe cities: A Support Vector Machine Approach Enhanced by SHAP Values

A. Carmina Llamas-Valenzuela[,1], José I. de la Rosa[*,1], G. Moreno-Chávez[1], Efrén Gonzales-Ramírez[1], Jesús Villa[1], and José M. Celaya-Padilla[1]*

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, México
allamas@uaz.edu.mx, ismaelrv@ieee.org

Abstract. Cities are centers that generate employment, innovation, and improve the quality of life, basic services, and housing. Rapid and unplanned expansion brings with it undesirable consequences for social and economic development. In Mexico, three out of four people live in a city. For this reason. the aim in this paper is to model and predict urban expansion in Zacatecas-Guadalupe cities (Mexico) using support vector machines, and thus carry out a better planning. In order to achieve this objective, land use and land cover maps corresponding to the period 2000–2020 were used, as well as the inclusion of socioeconomic, topographic, and cultural attribute variables. A soft SVM model was developed with a training accuracy of 92.4%, a validation accuracy of 93% and an F1-Score of 86.3%. In the obtained results, it can be observed that the proximity to already urbanized areas and the type of land use have a high influence on urbanization

Keywords: Urban expansion, Support vector machine, Urban expansion prediction, SHAP values, Satellite images,

Article Info
Received July 18, 2024
Accepted Feb 10, 2025

1 Introduction

Cities are centers that generate employment, innovation, and improve the quality of life, basic services, and housing (United et al., 2018). These opportunities generate a urban expansion, this means that this expansion changes the landscape around the city (Roy, 2021). This is called land use land cover change, where the natural cover adapts or changes to something useful for humans. Nonetheless, this transformation may possess either a favorable or unfavorable impact on the environment or society (Amri et al., 2020). Environmental damage (Deng et al., 2018), insufficient housing for demand, deficient transportation systems, increased social segregation (Rangel et al., 2022) and increased traffic (Kim & Kim, 2022) are some of the negative aspects.

In order to address these unfavorable impacts, it is important to have long-term planning supported by various instruments (González-Madrigal et al., 2020). Numerous approaches to modeling expansion have been devised, some of which rely on data obtained from remote sensing, utilizing satellite imagery to ascertain land use and land cover. Moreover, these models are extensively supported by Geographic Information Systems (GIS) (Nugroho & Al-Sanjary, 2018). Several mathematical models have been utilized for urban expansion, including Cellular Automata (CA) (Kara & Doratlı, 2021; Liang et al., 2018; Rangel et al., 2022; S. W. Wang et al., 2021), a model commonly employed to simulate urban expansion (H. Wang et al., 2021). The CA is a simplification of reality, constructed with key elements (Rangel et al., 2022). However, it has been assumed that urban areas are spatially homogeneous, making it challenging to accurately represent individual decisions and socioeconomic interactions.

Thus, in order to address this limitation, researchers have employed not only statistical methods, such as different types of regressions (Hyandye, 2015; Rangel et al., 2022), but also various machine learning techniques, such as XGBoost-SHAP (Kim et al., 2023; Kim & Kim, 2022), support vector machines (SVM) (Karimi et al., 2019; Mirbagheri & Alimohammadi, 2018), decision trees (Karimi et al., 2021), random forests (Frimpong & Molkenthin, 2021) and others. However, these studies were conducted outside of Mexico.

It is important to note that by 2020, nearly one hundred million Mexicans lived in unplanned urban settlements (Rangel et al., 2022). Moreover, water resources in Mexican cities are at risk due to the lack of urban expansion plans and preventive measures (Garrocho et al., 2020). This situation prompted the creation of the study proposed in this paper. The present study aims to model and predict urban expansion in Zacatecas-Guadalupe cities (Mexico) using support vector machines. In order to achieve this objective, land use and land cover maps corresponding to the period 2000–2020 were used, as well as the inclusion of socioeconomic, topographic, and cultural attribute variables. The obtained results indicate that the proposed model makes reasonable predictions for urban expansion for the year 2030.

The remaining of the paper is organized as follows, section 2 summarizes the materials and methodologies employed in the investigation. Section 3 provides an overview of the experimental setup. Section 4 presents results, an analysis, and discussion of the findings. Finally, Section 5 concludes with some recommendations for future research.

2 Materials and Methods

The Fig. 1 presents the general proposed methodology, which is divided into two phases: (1) developing the urban expansion model using data from 2000 and 2010 and (2) computing the SHAP values and predicting urban expansion for the Zacatecas-Guadalupe (Mexico) area. In the first phase, data collection and preprocessing were carried out. Having the data preprocessed, modeling was performed using the Soft Support Vector Machines (SVM) algorithm. The selection of the previous algorithm was based on its robust behavior and also it has been used in related studies with good results (Karimi et al., 2019). Moreover, the evaluation metrics were calculated with the aim of selecting the best model. In the final phase (2), the calculation of the SHAP values and the prediction of the Zacatecas-Guadalupe area to the year 2030 were carried out, using the best obtained model.

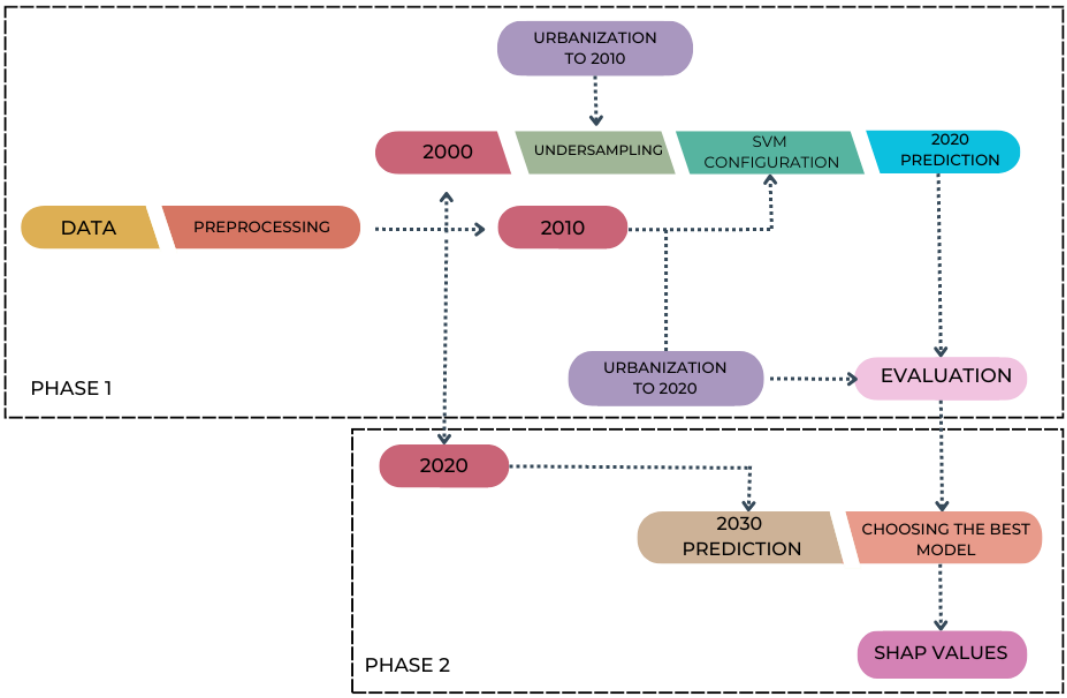

Fig. 1 Methodology proposed and followed in this study.

2.1 Area under study and period

In this study, the cities of Zacatecas-Guadalupe are selected as the study area. These cities belong to the state of Zacatecas, Mexico. The Zacatecas state is in the north-center zone of the country. Turning it into a state of high connectivity by land transportation. The reason for this situation lies in the proximity of the cities of Aguascalientes, San Luis Potosí and Durango (Pérez et al., 2020). Furthermore, this state represents the 3.8% of the surface of the country (INEGI, 2021). The above contributed to the creation of

the Zacatecas-Guadalupe metropolitan area. These metropolitan areas are among the 48th in Mexico. This region is characterized by rapid urban expansion, cultural diversity, and a substantial demand for natural resources and services (Garbutt, 2024).

The study covers the period of 2000 to 2020 and has a duration of 20 years. In this period, the zone has experiment different changes. According to INEGI (INEGI, 2021), there exists an increment of the particular homes from 299,249 in 2000 versus to 442,263 in 2020. And the average of occupants per home changed from 4.5 in the year of 2000 to 3.7 in 2020. Additionally, SEDATU (SEDATU, 2018) reports an average annual growth rate of 2.4% in the period of 2000 – 2015, and (Garbutt, 2024) reports a 2.03% in the period 2010 – 2020 in the metropolitan area of the cities Zacatecas-Guadalupe.

2.2 Data

In order to evaluate and execute the prediction of the Zacatecas-Guadalupe (Mexico) cities area, a data set was prepared with twenty variables. The variables shown in Table 1 characterize the topographic, socioeconomic, environmental, and cultural status of the area in the years 2000 – 2020. Additionally, a neighborhood variable was included. The variables were selected considering related studies (Garrocho et al., 2020, 2021; Karimi et al., 2019, 2021; Kim et al, 2022; Rangel et al., 2022), state religious statistics (Diversidad Zacatecas, n.d.) and other population characteristics. Due to the number of variables, these were grouped into the following categories: i) land use, ii) topography, iii) environmental, iv) socioeconomic, v) cultural and vi) neighborhood.

Table 1. Variables used in this study.

Category	Variable
Topographic	(1) DEM, (2) Aspect, (3) Slope,
Land Use	(4) Land Use Land Cover Map, (5) Distance to build areas, (6) streets and highways, (7) railways and (8) city centers,
Environmental	(9) Distance to green areas/parks,
Socioeconomic	(10) Population density, (11) employed population, (12) female head and (13) male head, (14) quarterly income per home, and (15) number of homes per cell,
Cultural	(16) Distance to temples, (17) cemeteries, and (18) schools, (19) religious population density,
Neighborhood	(20) Pixel change probability.

The analyzed data was created using census results from the years 2000, 2010, and 2020 collected from the National Institute of Statistics, Geography, and Informatics (INEGI). Additionally, vector topographic maps were used. Furthermore, the Landsat Geomedian raster images and Digital Elevation Model (DEM) were collected in the years 2000, 2010, and 2020. All the data used in this paper was projected into Universal Transverse Mercator (UTM) projection zone 13 N. The variables were preprocessed with the open-source QGIS software or Python.

The Landsat Geomedian images were used to create the land use and land cover maps. These multispectral images have a spatial resolution of 30 meters by 30 meters. For this reason, all the variables were adjusted to this resolution. The images were carefully cropped to a width and height of 838 x 429 pixels. Following this adjustment, a segmentation process was conducted to separate the constructed pixels from the rest. In order to achieve this segmentation an unsupervised clustering algorithm was used, specifically the Fast and Robust FCM (Fuzzy C-Means) algorithm (Lei et al., 2018), has been chosen due to its effectiveness given the spatial resolution of the analyzed images.

Also, in order to ensure the accuracy and quality of the segmentation, the Jaccard index was employed and calculated as an evaluation metric. Remarkably, the segmentation achieved a consistent Jaccard index score of 0.7 across all processed images, underscoring the robustness and reliability of the applied algorithm. Once the maps were created, the variable representing the distance to urban areas was derived from them. This new variable represented the Euclidean distance to urban areas. Additionally, this segmentation facilitated the creation of density variables within the dataset, enriching the data and enhancing its analytical potential.

In the topography category, there are variables that describe the terrain. The DEM obtained by the ASTER sensor was used for the years 2000 and 2010, with a resolution of 30 meters per cell (Abrams Michael & Crippen Robert, 2019), while for 2020 the continuous model built by INEGI (INEGI, 2013) was used. The slope and aspect were calculated from the DEM using QGIS.

Data preprocessing was carried out mainly on the variables of density and distance. Relevant data were selected and cleaned by removing outliers, and interpolation was performed on the missing values. The distance variables are Euclidean distances calculated in meters from the vector files published by INEGI. Finally, the variables corresponding to population density, housing, and income were calculated from the censuses of 2000, 2010, and 2020, conducted by INEGI. Afterward, they were transformed into images for further analysis. The previous depicted variables can be seen in the Fig. 2.

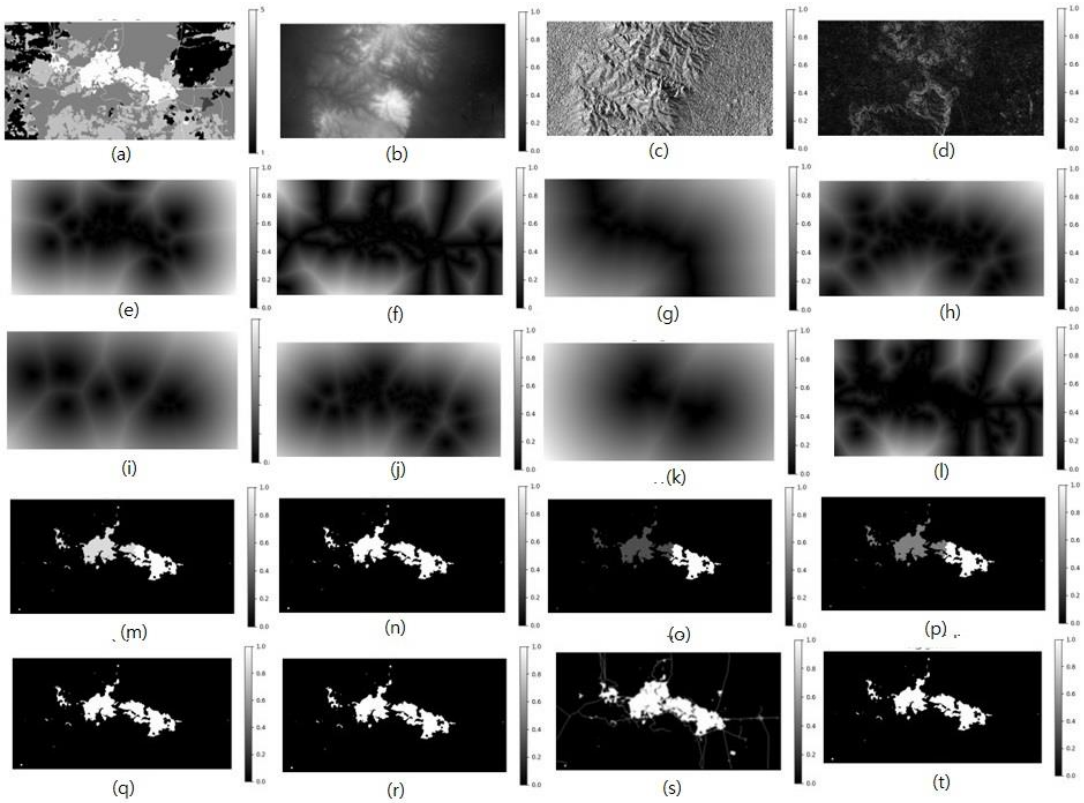

Fig. 2. Training data from the year 2000 used in the study. (a) Land Use Land Cover Map, (b) Digital Elevation Model, (c) Aspect, (d) Slope. (e)-(l) distance variables and (m)-(t) density variables.

Finally, the neighborhood (3×3) variable is the probability of change for each pixel. It is calculated as the fraction of pixels in the neighborhood that are urbanized, normalized by the area of the neighborhood. This procedure generates an image where each element represents the probability that the corresponding pixel will change to urbanized based on the density of neighboring pixels with that land use.

2.3 Soft Support Vector Machine

Support Vector Machines (SVM) were proposed by Vapnik and Lerner in 1963. They were initially introduced for supervised binary classification (González et al., 2017). This technique is based on the construction of an optimal separation hyperplane and the construction of support vectors.

Suppose we have a training set $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_N, y_N)\}$, where $(\vec{x}_i, y_i)$ correspond to the samples for i=1... N , consisting of an n-dimensional feature vector and a label indicating the classes $\{\pm 1\}$ to which each of the samples belong (González et al., 2017).

Since the classes are +1 and -1, the hyperplane is defined by the maximum margin of separation between the classes. Based on this, the support vectors are defined as: $\vec{w} \cdot \vec{x} + b = +1$ and $\vec{w} \cdot \vec{x} + b = -1$, which are parallel to the hyperplane, expressed by $\vec{w} \cdot \vec{x} + b = 0$. The maximum margin of the support vectors is expressed by $2/\|\vec{w}\|$.

However, SVMs were initially proposed with a hard margin, meaning that the data are completely separable. Nonetheless, there are cases where the data is not easy separable. In these cases, a slack variable ξ_i is introduced, allowing for classification errors, but these errors are penalized. With this new parameter, the SVM is softened. Now the problem minimizes the following expression:

$$J(w,\xi)=\frac{1}{2}\|w\|^2+c\sum_{i=1}^{n}\xi_i \quad (1)$$

Subject to:

$$y_i[w_i\cdot x_i+b]+\xi_i-1\geq 0,\ \xi_i\geq 1,\ i=1,\ldots,n \quad (2)$$

Where ξ_i is the slack parameter, and c is the penalty parameter (Karimi et al., 2019). In this new approach, the goal is to find the hyperplane that minimizes the classification errors and maximizes the separation between the support vectors. As the value of c increases, a narrow margin is obtained, which minimizes the number of misclassifications. On the other hand, if c decreases, more incorrect classifications are allowed (Awad Mariette & Khanna Ragul, 2015).

Additionally, the SVM transforms the data into a high-dimensional space. This is done by calculating a data projection using a kernel. The most used is the lineal, polynomial, or radial basis function (RBF) kernel (Karimi et al., 2019).

2.4 Evaluation Metrics utilized

Considering the context of the problem and the proposed methodology, some appropriate evaluation metrics are necessary. During this study, the classification evaluation metrics were selected. A confusion matrix (shown in Table 2) is used to calculate these metrics, introducing us to the concepts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) (Karimi et al., 2019).

The TP refers to the pixels that are built (corresponding to an edification), and the model refers to those classified as built. The FP indicates the unconstructed pixels, while the model places them as constructed. TN pixels are classified as non-constructed, and the actual label is non-constructed. Finally, FN are pixels indicated by the model as not built, but that are actually built (Panesar, 2020).

In this study, the following metrics were obtained: accuracy, precision, and the F1-Score. Accuracy indicates the proportion of correct classifications made by the model regarding the total samples; it was calculated using (3). Equation (4) is used to calculate precision, which focuses on the proportion of TP to the total samples classified as positive. Finally, the F1-Score was calculated according to equation (5) which is the harmonic mean of precision and recall (6). The recall refers to the proportion of TP regarding the instances that are positive (Pagano et al., 2023).

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision=\frac{TP}{TP+FP} \quad (4)$$

$$F1-Score=2\cdot\frac{Precision\cdot Recall}{Precision+Recall} \quad (5)$$

Where recall is defined as:

$$Recall=\frac{TP}{TP+FN} \quad (6)$$

Table 2. Built-unconstructed pixel confusion matrix.

Observed prediction	Built	Unbuilt
Built	True Positive (TP)	False Negative (FN)
Unbuilt	False Positive (FP)	True Negative (TN)

2.5 Shapley Additive exPlanations (SHAP)

Shapley values quantify the marginal contribution of each feature (Kim et al., 2022). This method was proposed by Lundberg and Lee (Lundberg et al., 2017), it is a method that assigns an importance value to each feature for a given prediction. This method is grounded in the theory of Shapley values from cooperative game theory. It is calculated by using equation (7).

$$\Phi_i = \sum_{S \in N} \frac{|S|!\,(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{7}$$

3 Experimental setup

As mentioned previously, twenty variables were used in this study, described in different categories (see Table 1). In the context of the problem, the data set was divided as shown in Table 3.

Due to the disparity of the classes, a random under-sampling was performed, leveling the classes. In this experiment, eight seeds were randomly chosen for under-sampling. Given that our dataset consists of images, we enhanced this approach by incorporating a buffer zone. This buffer ensures that only pixels that are closest to the constructed regions are selected during the under-sampling procedure. Specifically, we constructed a supplementary image layer that covers the buffered pixels, thereby safeguarding the spatial continuity and context of the selected samples. This method mitigates the effects of class imbalance and maintains the structural integrity of the images, facilitating more accurate and reliable analysis.

Table 3. Data division in training and testing.

	Variable	Label
Training Dataset	2000 data layers	binary label of built/not built pixels in 2010
Testing Dataset	2010 data layers	binary label of built/not built in 2020

Once the data preparation was made, the hyperparameters for tuning the SVM model were selected (Table 4). The hyperparameters were selected according to those values used in related work (Karimi et al., 2019). After training and testing the models, the best one was selected. Later, the SHAP values were calculated, and the prediction of urban expansion to the year 2030 was also obtained.

Table 4. Values for every hyperparameter in the grid search for soft SVM.

Hyperparameter	Grid search values
C	0.1, 1, 10, 100
Kernel	Polynomial, RBF
Gamma	1, 2, 3
degree	1, 2, 3, 4, 5

4 Results and discussion

Before developing the predictive model, an analysis of the pixels corresponding to the built and not built labels was carried out. With this, the rate of change of the pixels between the years studied was calculated (as can be seen in Table 5). The growth rate remained constant in the two periods analyzed.

Table 5. Rate of change in the period 2010 to 2010 and 2010 to 2020.

	2000 – 2010	2010 – 2020
Number of pixels changed	16369	14578
Number of pixels unchanged	343133	344924
Growth rate	4.55%	4.06%

Additionally, the area under study was divided into four quadrants: northwest (NW), northeast (NE), southwest (SW), and southeast (SE). The division is shown in Fig. 3. The main purpose of dividing the image into quadrants and analyzing the pixels individually is to identify urban expansion patterns in each section of the image. This allows a deeper and more detailed understanding of the composition of the image.

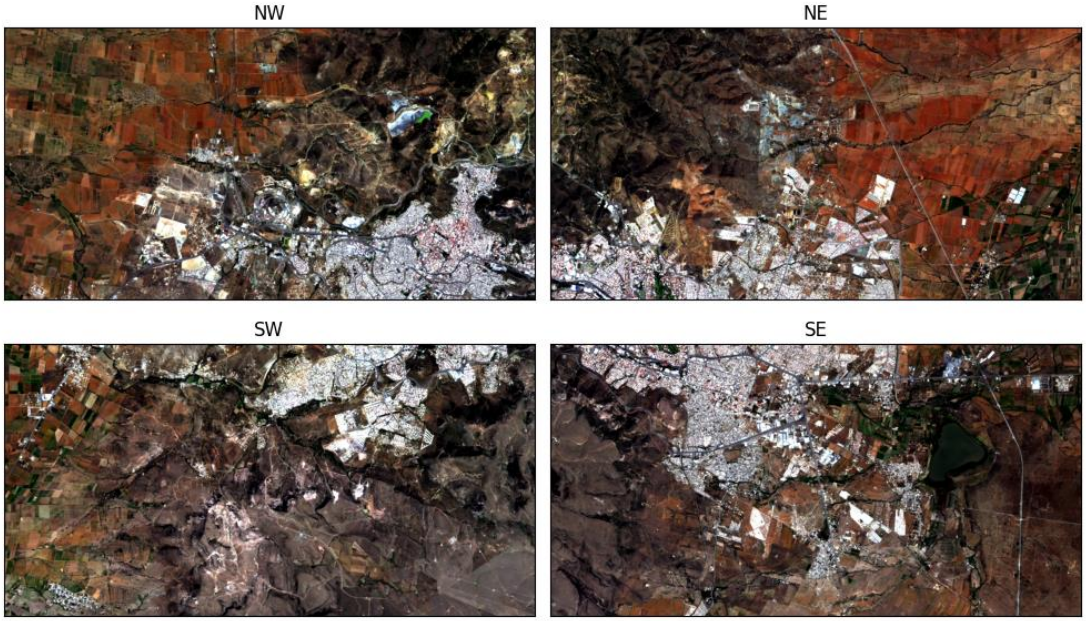

Fig. 3. Division of the study area into four quadrants (NW, NE, SW and SE).

With the divided image, the built-up pixels were counted. The area in km^2 was calculated for each quadrant, this can be seen in Fig. 4., Fig. 4a shows the accumulated built-up pixels in each quadrant throughout the years analyzed. Historically, the largest construction area has been found in the northwest (NW), followed by the southeast (SE). The NW zone corresponds to the area of Zacatecas, and the SE to Guadalupe.

On the contrary, Fig. 4b illustrates the urban expansion in the quadrants between the years 2000 and 2010 and 2010 and 2020, with the aim of identifying which section experienced the highest urban expansion during these periods. During the period spanning from 2000 to 2010, the most significant expansion was observed in the Southeast (SE) region, followed by the Southwest (SW). This indicates that the expansion is concentrated in the Guadalupe city. Through the other time frame, spanning from 2010 to 2020, the growth is observed in the Northeastern (NE) region, followed by the Southeast (SE).

4.1 Soft SVM Model

An experiment was carried out by modifying different values of the parameters of the penalized SVM, using the values indicated in the Table 4. The parameters that took different values in this experiment were: parameter c, kernel, gamma, and polynomial degree. The values for the parameter c were 0.1, 1, 10, 100. Likewise, the radial kernel (RBF) was selected with the parameter

gamma of 1, 2, 3. Finally, for the polynomial kernel, the degree values of polynomial were 1, 2, 3, 4 and 5. The combinations were carried out with the purpose of obtaining the most appropriate predictive model.

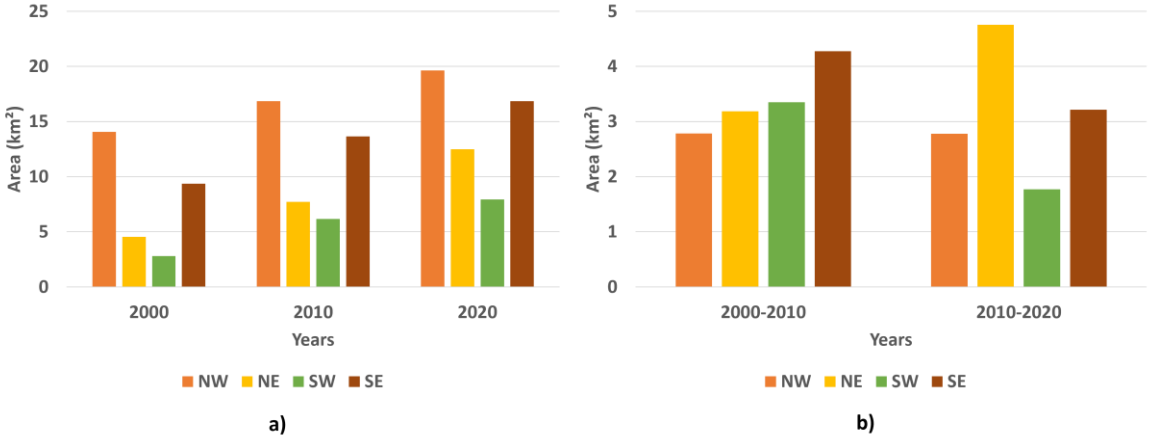

Fig. 4. a) Areas in km^2 of the built zones in the NW, NE, SW and SE quadrants. b) Differences in the areas built in the periods 2000 to 2010 and 2010 to 2020 in km^2 of the NW, NE, SW and SE quadrants.

Model selection was carried out using the accuracy (training and validation), F1-Score and Precision metrics. These metrics were calculated using the confusion matrix of the constructed and unconstructed pixels. The resulting models have a high training and validation accuracy (greater than 90%). These obtained results can be seen in Fig. 5. The highest values in training accuracy are obtained with models with an RBF kernel. On the other hand, the best values for test accuracy are found in models with a polynomial kernel with degree of five.

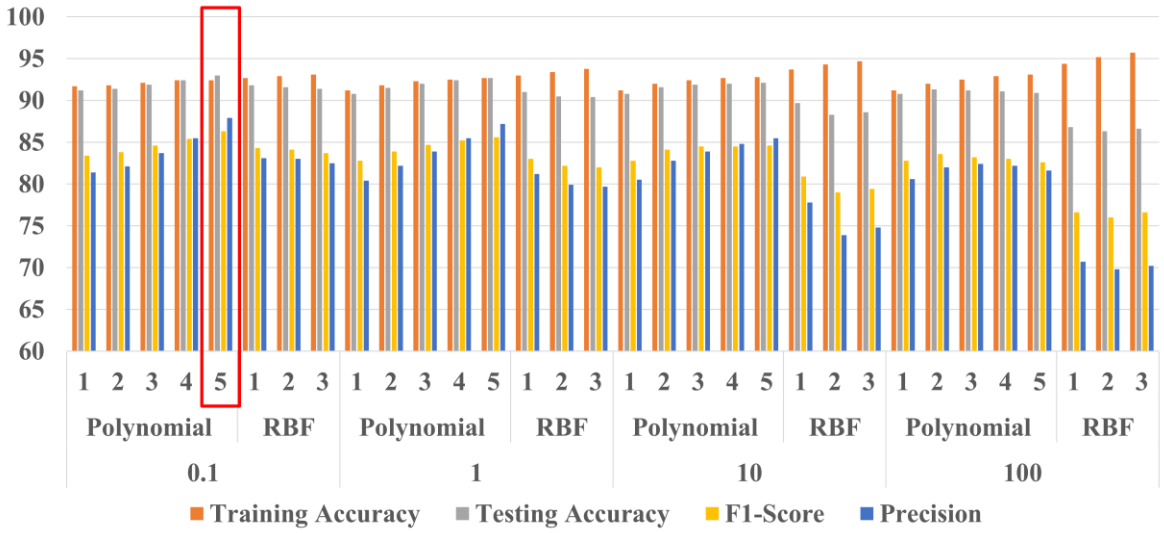

Fig. 5. Result of the experimentation, in the red box is the best model.

However, due to the imbalance between the classes, it was decided to focus on the F1-score. The best results concerning the F1-score are found in the polynomial kernel with degree five, similar to the validation accuracy. These results are around 80%. The precision results are similar to the F1-score, the best is found with the polynomial kernel above values of 80%.

In the case of the values of c, the best results are found with the values of 100 and 0.1. In terms of accuracy (training) and precision, the best values are with a c value of 100. Furthermore, the validation accuracy and F1-score are found in the value of 0.1.

The model that was chosen as the best one, is due that it has one of the best F1-scores and the best balance with the other metrics. This model is the polynomial kernel, with a c value of 0.1 and a polynomial degree of five, the obtained training accuracy was

92.4%, and validation accuracy was 93% and F1-score of 86.3%. In Fig. 5, the red box indicates the selected model. On the other hand, Fig. 6 compares the model results versus to the measured reality of 2020 year.

With the aim of knowing the distribution of built and not built pixels in the prediction result for the year 2020, the km^2 of the previously defined quadrants were calculated. In the Fig. 7 it is shown the difference between the real state of the cities and the validation image according to the SVM model.

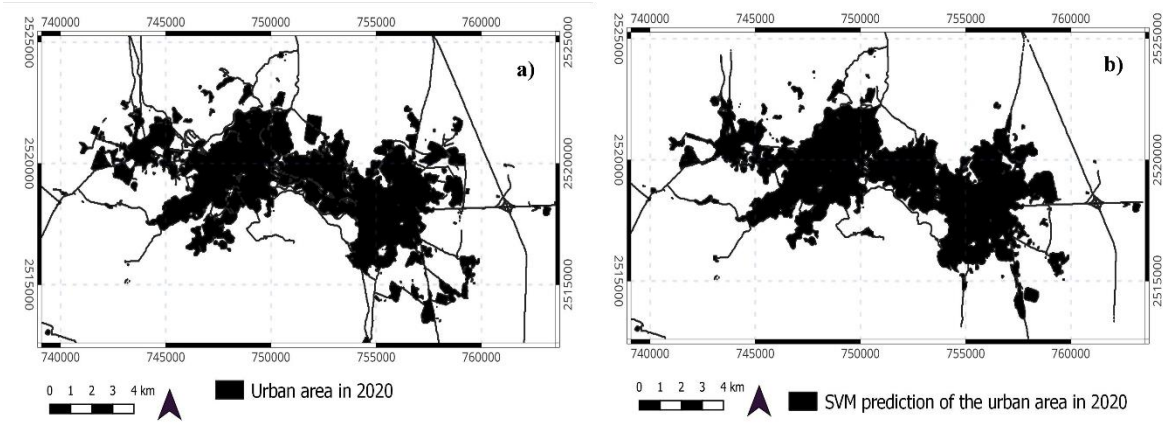

Fig. 6. Comparison of the prediction made by the SVM model and the real urban area for the year 2020.

The SVM results show that in the validation image in the NW, NE and SW quadrants, in the model result the area is smaller. Meanwhile, in the SE quadrant, the validation was greater in the areas. The higher difference is presented in the NE quadrant. The lowest is presented in the SE quadrant.

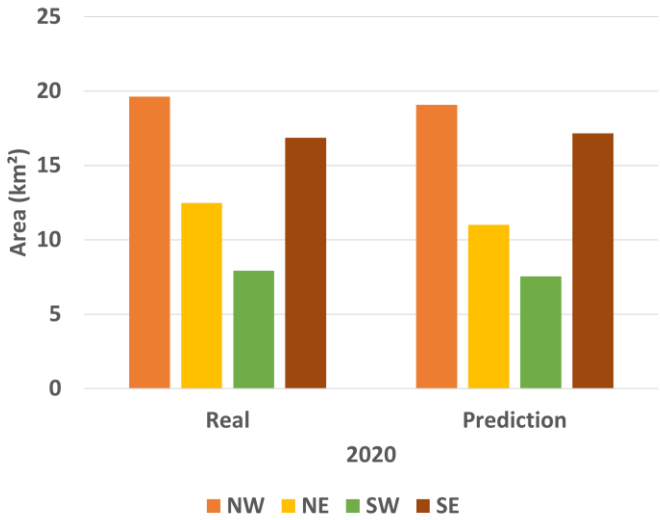

Fig. 7. Comparison of the predicted areas in km^2 made by the SVM model and the real urban area for the year 2020.

4.2 Importance Factor

Due to the computational demand of the methodology, SHAP values were calculated by selecting a subgroup of the data. The selection was carried out by applying the K-means algorithm to the data. Subsequently, the nearest points to the centroids were selected. Fig. 8 provides the SHAP values, which help us to understand the behavior of the independent variables that determine whether the pixels are urbanized.

The variable that most impacts the model is the type of land use land cover maps. The higher the value, the more likely it is to be urbanized. This means that if it is already urbanized it remains urbanized and if it is bare soil or vegetation close to an urban area,

the classifier places it as urbanized. The second variable that has the greatest influence is the probability of urbanization. This indicates that the pixel is urbanized depending on how many pixels within its 3×3 neighborhood it is already urbanized.

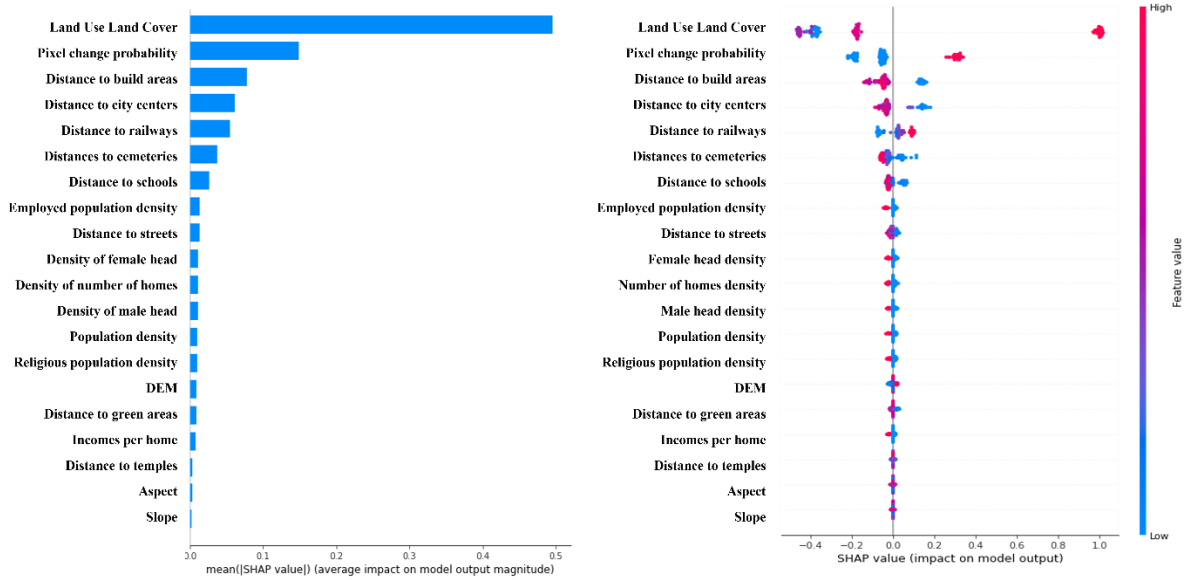

Fig. 8. SHAP values of the independent variables in the soft SVM model.

This implies that when a pixel has a probability close to one, it is likely to have numerous pixels urbanized surrounding it. If the value of this probability is high, the model indicates it as urbanized. This is in agreement with the third variable, which is the distance from an urbanized area. If the distance is small, the model tends to consider it to be urban. On the contrary, the distance from the urban centers is the fourth significant variable. The closer you get to the center, the more attractive to be urbanized.

In contrast, the variables that did not have an impact were the terrain elevation, quarterly income, slope, and aspect. A possible explanation for the low impact of topographic variables on the model could be the spatial resolution of the data used. All variables are set to a spatial resolution of 30 m × 30 m, which reduces the level of detail. The last one can be crucial when carrying out a topographic analysis. In this case, there is a possibility that the resolution is insufficient and cannot accurately capture relevant terrain variations and features. This may result in an underestimation of the influence of topographic variables on the predictive model.

In the case of income, a possible reason for the low impact of the variable is its homogeneity. In the data analyzed, the pixels corresponding to each city have the same amount of income, which may limit their ability to explain variations in urban expansion. This uniformity implies that income does not provide distinguishing information for the model, thus reducing its impact. Finally, even though the city is very religious, the distance to temples is not something that people consider when urbanizing.

4.3 Urban expansion scenario to 2030

The prediction of urban expansion was made for the year 2030, using the soft SVM model selected and data layers from 2020. The prediction result can be seen in the Fig. 9. As a result, it can be seen that the urban expansion of the metropolitan area is greater in the Guadalupe area. This trend has been consistent over the years. The pixel change rate in this prediction is 8.11%. Comparing this to the previous period's analysis, we found the change rate to be double.

Additionally, the change in area in km^2 can be seen in the Fig. 10. It can be observed that the quadrant with the highest increase in built area was the SE quadrant, located in Guadalupe City, followed by the NW quadrant in Zacatecas City. This is contrary to the findings from the period 2010 to 2020. On the other hand, in the Fig 10b, it can be observed that the quadrant with higher urban expansion was the SE, followed by NE, NW and lastly SW.

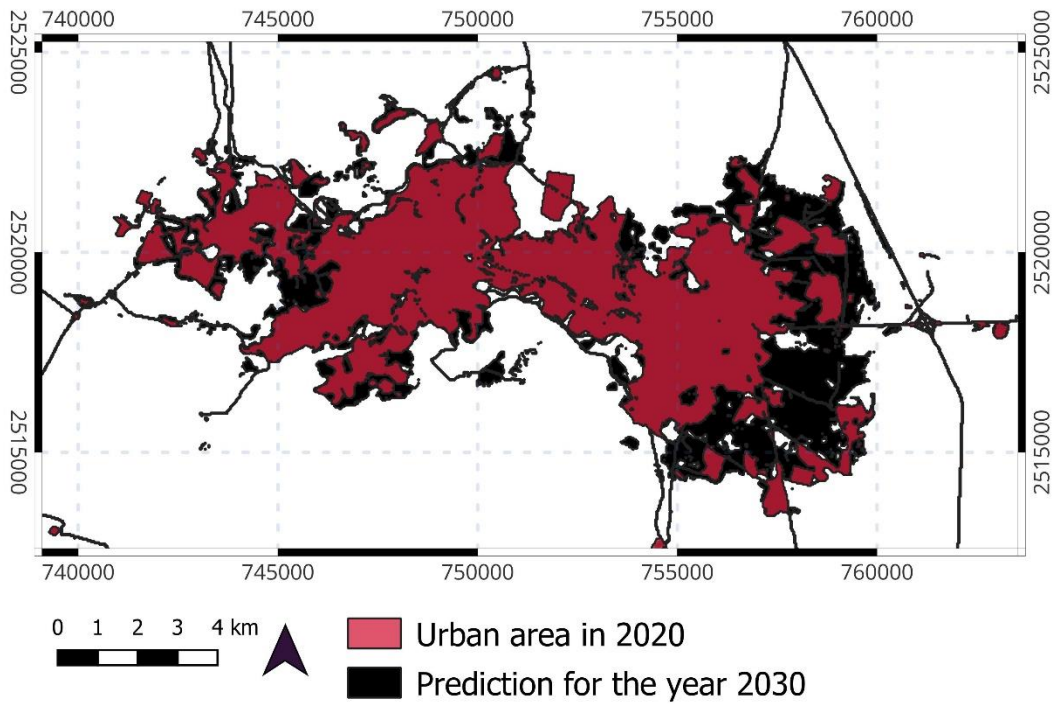

Fig. 9. Urbanized area in the year 2020 and the prediction for the year 2030 as a result of the predictive model.

The disparities in territorial expansion suggest regional differences that necessitate consideration in urban and environmental planning. The expansion of the SE region implies the necessity of proficient land use management to mitigate environmental repercussions. Regions with limited expansion may need approaches focused on optimizing existing land use.

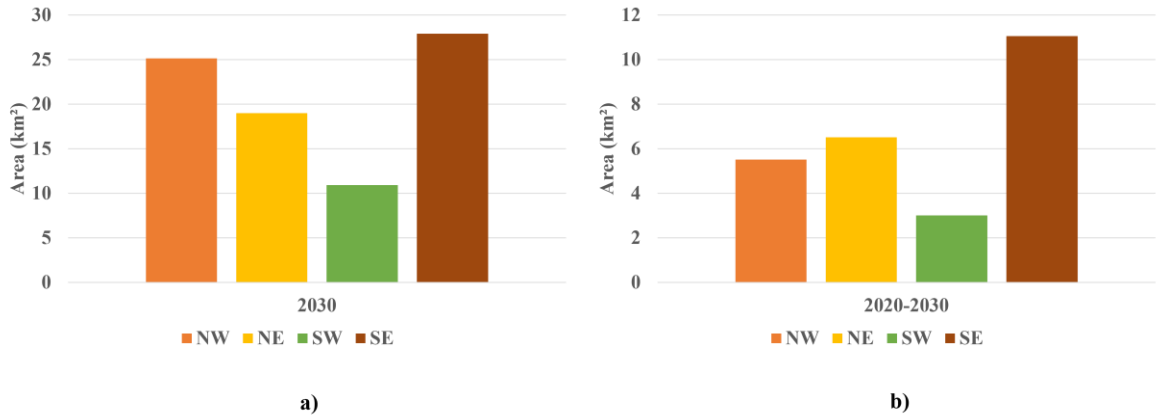

Fig. 10. a) Projected distribution of areas in km^2 for the NW, NE, SW, and SE regions in the year 2030. b) Difference in projected areas in km^2 for the same regions between the years 2020 and 2030.

The predictions made in this study used data up to 2020, allowing for a comparative analysis of urban expansion observed from 2020 to 2024. The expansion observed during this period was examined within the soft SVM model prediction, aiming to validate and understand urbanization decisions, as well as to explain the reasons behind the model's predictions. In the analysis, six areas were selected. On Fig. 11, the chosen areas can be seen, which contains four regions where the model made a correct prediction (blue circle) and two regions where the model did not make a correct prediction (red circle).

Once the regions were selected, they were searched on (Google Earth, 2024). Zones A, B, C, and E were selected due to the accurate predictions made by the model in these regions. The urban expansion presented in the period 2020–2024 can be seen in Fig. 12. Zones C and E were selected because they have a higher density of houses, while Zones A and B were partially developed areas. Finally, Zone A is considered to have the most urban expansion.

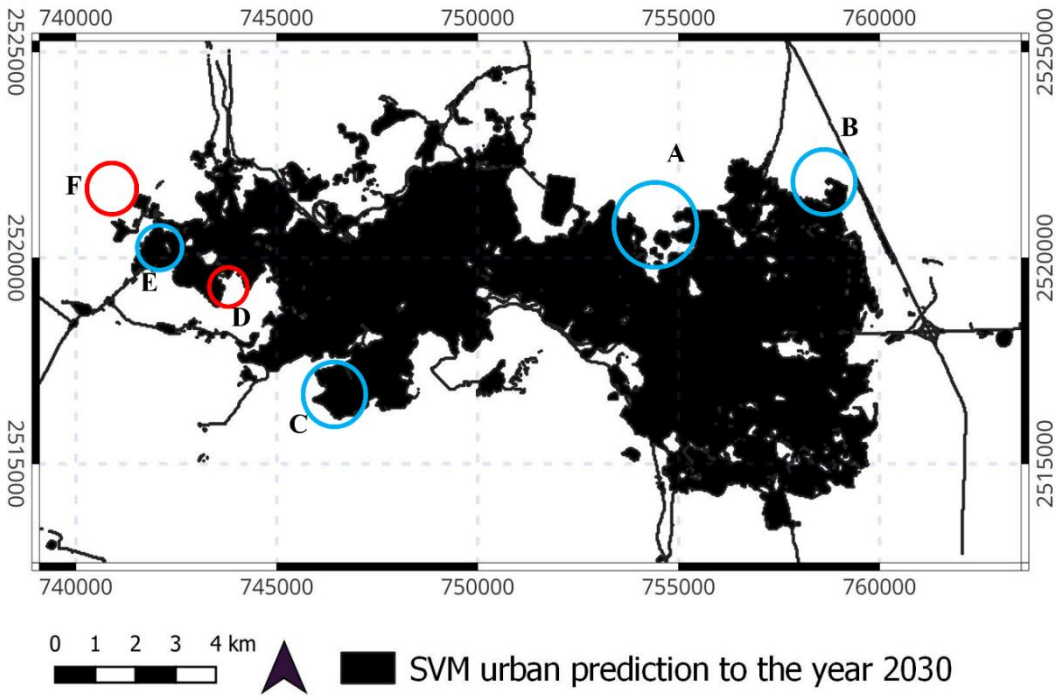

Fig. 11. Analyzed areas to evaluate the accuracy of the urban expansion prediction model. Zones on the blue circles represent areas where the model made correct predictions, while the red circles show differences between the model and reality.

On the other hand, zones D and F were selected because they presented significant discrepancies between the model predictions and the observed reality. Zone F showed much more partially developed areas and more home density not seen in the prediction of the model, possibly due to factors not considered, such as recent changes in urban policies or unexpected investments in infrastructure. Zone D indicated minimal expansion, but houses may have been affected by obstacles not anticipated in the model, such as environmental or socioeconomic constraints that limited urban development. Additionally, this zone can be impacted by traffic and the lack of services, such as schools, supermarkets, and others. These consequences can be driven by the long distances from these zones to the urban areas. Moreover, urban expansion into these zones can cause more traffic on some parts of the existing streets due to the lack of paved roads and main roads.

Furthermore, Zone E is a private residential area, meaning that its main streets are within a perimeter. This can negatively affect Zone F because the street that connects to the road does not reach this area. This creates a secluded area far from the main road, which can cause traffic in different areas and automobile accidents that are currently present in the area. Additionally, this area lacks kindergartens and elementary schools, making it difficult for families to access these services. Moreover, the public transportation routes currently available in the area are deficient due to long waiting times, which encourages people to rely on private vehicles.

4.4 Discussion

Rapid urbanization is the main cause of the change in land use and land cover from vegetation areas, agriculture, and bare soil to urbanization (Rana & Sarkar, 2021). In this study, we have explored the patterns and trends of urban expansion in the Zacatecas-Guadalupe cities (Mexico) during the period from 2000 to 2020, using soft SVM, which is a robust algorithm, and integrating the SHAP values.

The SHAP values revealed that proximity to already urban areas and if the pixel is bare ground indicates a higher probability of urbanization, which is consistent with other studies such as (Karimi et al., 2019, 2021). However, although some authors consider DEM and slope important (Kim & Kim, 2022), the SHAP values calculated in this study indicate that these variables do not impact the urbanization decision, as documented by other authors (Frimpong & Molkenthin, 2021; Rana & Sarkar, 2021).

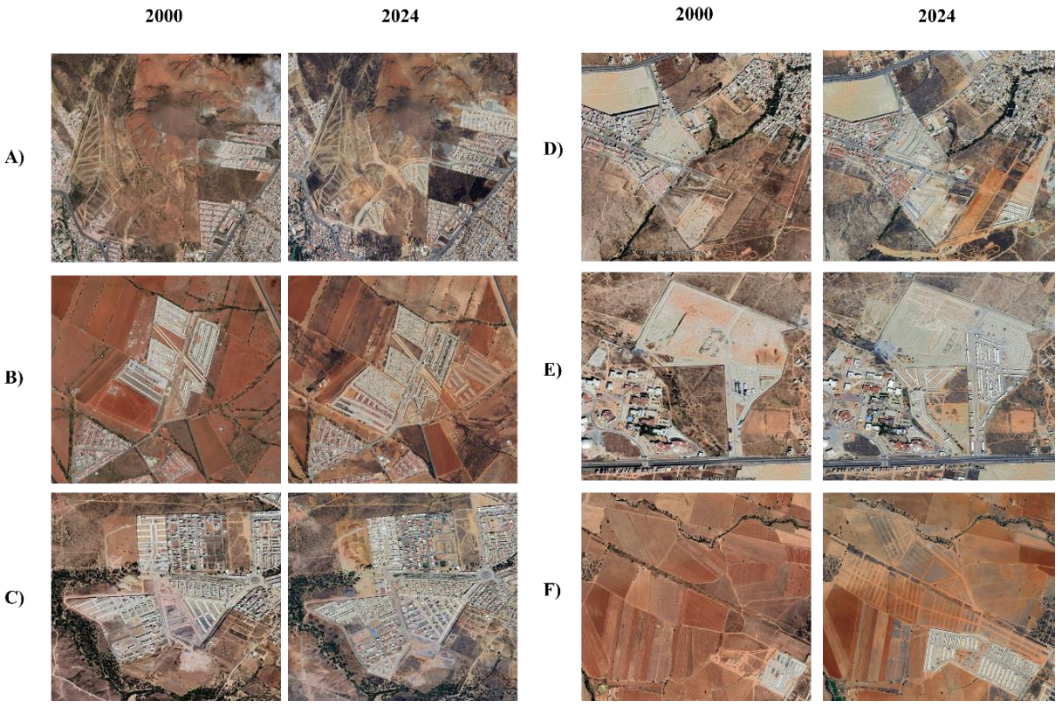

Fig. 12. Comparison of urban expansion between the years 2000 and 2024 in the six regions analyzed.

Distances to city centers, schools, and streets also impact the final prediction, as it has been seen in other articles (Karimi et al., 2019). In the 2030 prediction, it was observed that the distance to streets has a high influence. These conclusions are driven by the concentration in areas with more first and second-order main roads, located in the Guadalupe area. In contrast, the Zacatecas area has a lower density of this type of road, which causes less expansion in that direction. Additionally, this is shown in zones D and F, which are not considered in the prediction because they do not have a nearby street or schools.

The cultural variables, such as distance to temples and religious population density, did not have a significant impact according to the results. Therefore, it would be appropriate to consider other variables that are more influential in determining the attractiveness of a place for habitation

Finally, we suggest that city planners should consider the importance of street construction, as this increases the probability of new constructions and the proximity of new settlements to the city center. Given the significance of streets in the prediction of the model, a robust system of main streets is recommended, along with enhancements to public transportation to reduce the number of private vehicles. Additionally, consideration should be given to creating a city that is much more friendly to pedestrians, bicycles, and motorcycles.

5 Conclusions

The prediction of urban expansion in the Zacatecas-Guadalupe cities was analyzed using a soft SVM algorithm, followed by the calculation of SHAP values. The results show that land use and land cover maps have a high impact, with vegetation areas near urban zones and bare soils being more likely to be urbanized. On the other hand, the slope of the terrain and the aspect do not have much influence on the obtained result.

We consider that prediction can aid city planning, mainly by avoiding uncontrolled traffic, due to the impact that streets have within the model and shape of the city. Furthermore, by utilizing this tool, city planners can reduce isolated areas of the city that lack services

Although the model reflects the interaction of the variables and how the city expands, we believe that there are variables not yet explored that could increase the accuracy of the models. As future work, it would be beneficial to include population projection

variables, property values, and other cultural variables that demonstrate a stronger relationship. Additionally, the performance of a comparative analysis with other algorithms such as random forests and categorical boosting should be also considered.

Acknowledgment

The main author appreciates the support received by the Consejo Nacional de Humanidades, Ciencias y Tecnologías (Spanish for National Council of Humanities, Sciences and Technologies; abbreviated CONAHCYT) through the Scholarship Program for Postgraduate Studies in Mexico.

References

Abrams Michael, & Crippen Robert. (2019). *ASTER GDEM V3 (ASTER Global DEM) User Guide*.

Amri, I., Prabaswara, I. W., & Giyarsih, S. R. (2020). Spatio-Temporal Analysis of Post-Disaster Built-up Expansion in Banda Aceh City, Indonesia. *Proceedings - 2020 6th International Conference on Science and Technology, ICST 2020*. https://doi.org/10.1109/ICST50505.2020.9732823

Awad Mariette, & Khanna Ragul. (2015). *Efficient Learning Machines*.

Deng, Y., Fu, B., & Sun, C. (2018). Effects of urban planning in guiding urban growth: Evidence from Shenzhen, China. *Cities*, *83*, 118–128. https://doi.org/10.1016/j.cities.2018.06.014

Diversidad Zacatecas. (n.d.). Retrieved May 10, 2024, from https://cuentame.inegi.org.mx/monografias/informacion/zac/poblacion/diversidad.aspx

Frimpong, B. F., & Molkenthin, F. (2021). Tracking urban expansion using random forests for the classification of landsat imagery (1986–2015) and predicting urban/built-up areas for 2025: A study of the kumasi metropolis, ghana. *Land*, *10*(1), 1–21. https://doi.org/10.3390/land10010044

Garbutt, O. P. (2024). *Metrópolis de México 2020 Primera edición*. https://www.gob.mx/sedatuhttps://www.inegi.org.mx

Garrocho, C., Jiménez, E., & Chávez-Soto, T. (2020). *Expansión de la ciudad: un instrumento de simulación de escenarios para los sectores público y privado*. https://medium.com/espanol/

Garrocho, C., Soto, T. C., & Jiménez López, E. (2021). Autómata Celular Metro-NASZ: laboratorio experimental de expansión urbana. In *La Situación Demográfica en México* (pp. 149–175). CONAPO. https://medium.

González, R., Barrientos, A., Toapanta, M., & Del Cerro, J. (2017). Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial. *RIAI - Revista Iberoamericana de Automatica e Informatica Industrial*, *14*(4), 394–405. https://doi.org/10.1016/j.riai.2017.07.005

González-Madrigal, J., Solano-Lamphar, H., & Ramírez, M. (2020). *La contaminación lumínica como aproximación a la planeación urbana de ciudades mexicanas*.

Google Earth. (2024). *Vista aérea de Zacatecas, Zacatecas, México [Imagen de satélite]*. https://earth.google.com/

Hyandye, C. (2015). GIS and Logit Regression Model Applications in Land Use/Land Cover Change and Distribution in Usangu Catchment. *American Journal of Remote Sensing*, *3*(1), 6. https://doi.org/10.11648/j.ajrs.20150301.12

INEGI. (2021). *Comunicado de prensa num 57/21*.

Instituto Nacional de Estadistica y Geografia. (2013). *Continuo de Elevaciones Mexicano (CEM) Ayuda general*.

Instituto Nacional de Estadística y Geografía (INEGI). (2021). *Aspectos Geográficos: Zacatecas*. www.inegi.org.mx

Kara, C., & Doratlı, N. (2021). Predict and simulate sustainable urban growth by using gis and mce based ca. Case of famagusta in northern cyprus. *Sustainability (Switzerland)*, *13*(8). https://doi.org/10.3390/su13084446

Karimi, F., Sultana, S., Babakan, A. S., & Suthaharan, S. (2021). Urban expansion modeling using an enhanced decision tree algorithm. *GeoInformatica*, *25*(4), 715–731. https://doi.org/10.1007/s10707-019-00377-8

Karimi, F., Sultana, S., Shirzadi Babakan, A., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems*, *75*, 61–75. https://doi.org/10.1016/j.compenvurbsys.2019.01.001

Kim, M., Kim, D., Jin, D., & Kim, G. (2023). Application of Explainable Artificial Intelligence (XAI) in Urban Growth Modeling: A Case Study of Seoul Metropolitan Area, Korea. *Land*, *12*(2). https://doi.org/10.3390/land12020420

Kim, M., & Kim, G. (2022). Modeling and Predicting Urban Expansion in South Korea Using Explainable Artificial Intelligence (XAI) Model. *Applied Sciences (Switzerland)*, *12*(18). https://doi.org/10.3390/app12189169

Lei, T., Jia, X., Zhang, Y., He, L., Meng, H., & Nandi, A. K. (2018). Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering. *IEEE Transactions on Fuzzy Systems*, *26*(5), 3027–3041. https://doi.org/10.1109/TFUZZ.2018.2796074

Liang, X., Liu, X., Li, X., Chen, Y., Tian, H., & Yao, Y. (2018). Delineating multi-scenario urban growth boundaries with a CA-based FLUS model and morphological method. *Landscape and Urban Planning*, *177*, 47–63. https://doi.org/10.1016/j.landurbplan.2018.04.016

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://github.com/slundberg/shap

Mirbagheri, B., & Alimohammadi, A. (2018). Integration of local and global support vector machines to improve urban growth modelling. *Canadian Historical Review*, *7*(9). https://doi.org/10.3390/ijgi7090347

Nugroho, F., & Al-Sanjary, O. I. (2018). A review of simulation urban growth model. *International Journal of Engineering and Technology (UAE)*, *7*(4), 17–23. https://doi.org/10.14419/ijet.v7i4.11.20681

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Cruz, G. O. R., Peixoto, R. M., Guimarães, G. A. de S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Context-Based Patterns in Machine Learning Bias and Fairness Metrics: A Sensitive Attributes-Based Approach. *Big Data and Cognitive Computing*, *7*(1). https://doi.org/10.3390/bdcc7010027

Panesar, A. (2020). Machine learning and AI for healthcare: Big data for improved health outcomes. In *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-6537-6

Pérez, O. R., Sánchez, M. A. L., Camacho, X. C., & Robledo, M. A. (2020). Methodology for mining economic assimilation in Zacatecas, Mexico. *Economia, Sociedad y Territorio*, *20*(62), 241–272. https://doi.org/10.22136/est20201415

Rana, M. S., & Sarkar, S. (2021). Prediction of urban expansion by using land cover change detection approach. *Heliyon*, *7*(11). https://doi.org/10.1016/j.heliyon.2021.e08437

Rangel, C. G., Soto, T. C., Mata, V., & Jiménez López, E. (2022). *Un modelo de expansión urbana no estacionario en el espacio: Autómatas Celulares y Regresión Geográficamente Ponderada*.

Roy, B. (2021). A machine learning approach to monitoring and forecasting spatio-temporal dynamics of land cover in Cox's Bazar district, Bangladesh from 2001 to 2019. *Environmental Challenges*, *5*. https://doi.org/10.1016/j.envc.2021.100237

Secretaría de Desarrollo Agrario, T. y U. (2018). *Delimitación de las zonas metropolitanas de México 2015*. https://www.gob.mx/sedatu

United, N., of Economic, D., Affairs, S., & Division, P. (2018). *World Urbanization Prospects The 2018 Revision*.

Wang, H., Guo, J., Zhang, B., & Zeng, H. (2021). Simulating urban land growth by incorporating historical information into a cellular automata model. *Landscape and Urban Planning*, *214*. https://doi.org/10.1016/j.landurbplan.2021.104168

Wang, S. W., Munkhnasan, L., & Lee, W. K. (2021). Land use and land cover change detection and prediction in Bhutan's high altitude city of Thimphu, using cellular automata and Markov chain. *Environmental Challenges*, *2*. https://doi.org/10.1016/j.envc.2020.100017