**236501**

**מבוא לבינה**

**תרגיל בית 3**

לק ב'

3ד .2. חשוב לשמור על עקביות זו כוון שנרצה להשוות בין מסווגים. לפיכך נרצה להיות עקביים עם המבחנים איתם אנו בוחנים את המסווגים, כדי למנוע מצב בו מסווג מקבל ציון נמוך בגלל חלוקה קשה יותר של הנתונים הגולמיים ולא בגלל איכות סיווג נמוכה יותר.

5. 2.

5.3. כפי שניתן לראות ערכי השגיאה המינימאליים והדיוק המקסימלי התקבלו עבור .

5.4. ניתן להבחין כי המסווגים באופן כללי הניבו סיווג טוב של הדוגמאות. הופתענו לראות שהמסווג שקיבל את הציון הגבוה ביותר הוא דווקא זה עם הפרמטר , כלומר המסווג שמסווג את הדגימה על סמך אובייקט יחיד שהוא הקרוב ביותר ב- DataSet לדגימה. ציפינו שדוגמאות רועשות ב-DataSet יובילו ליותר סווגים שגויים במקרה של החלטה על סמך שכן יחיד קרוב ביותר. נראה שבחלוקה של מסד הנתונים אותה ביצענו הדוגמא הסמוכה ביותר לכל דגימה באופן מובהק יותר דוגמא טובה, ואילו רק בהרחבת החיפוש לדוגמאות רחוקות יותר מסתננות דוגמאות רועשות אשר משבשות את ההחלטה.

7.4. התוצאות הטובות ביותר התקבלו עבור מסווג KNN עם פרמטר .

**חלק ג:**

נתאר את התהליך שעשינו בבניית המסווג.

בחרנו להשתמש בסיווג שפועל לפי אלגוריתם עבור שכן יחיד, שכן במהלך פתרון התרגיל מסווג זה הגיע לתוצאה בסיסית הכי גבוה, ולכן בחרנו להשתמש בו בתור בסיס.

על מנת לשפר אותו רצינו לתת משקלים לתכונות שונות שלפיהן מחושב המרחק. בשביל להבין באיזה תכונות מדובר רצינו לראות כיצד ניראה עץ החלטות בשיטת ID3, מה שיתן לנו הערכה כללית מיהן התכונות החשובות. לאחר בניית העץ השתמשנו בספרייה (graphiz ) שמצאנו באינטרנט שיודעת להציג את העץ ולייצא אותו לקובץ PDF. להלן התוצאה שקיבלנו:

ניתן לראות כי בעץ יש צמתי מפתח בעץ (מסומנים באדום). צמתי מפתח אלו גורמים לחיתוך של הסיווגים בצורה ברורה לסיווג מסוים – ועל כן היינו רוצים להגדיל את המשקל של המרחק מצמתים אלו. ולכן הכפלנו כל צומת כזה במשקלים גדולים (10-30) בהתאם לכמה הסיווג שהם יוצרים נחרץ.

כמו כן בעבור דגימה שנדרש לסווג, בעבור דוגמא שקיימת במערכת ומקיימת שהיא נמצאת בצד השני ביחס לצמתים אלו ניתן עונש נוסף למרחק – ולכן דגימות שמסווגות באופן דומה יהיו קרובות לדגימה שנדרש לסווג.

בנוסף, ניתן לראות כי בעץ ישנם צמתים חשובים אחרים (מסומנים בכחול), צמתים אלו גורמים להחלטה לא נחרצת ומבלבלת, לכן בעבור תכונות אלו איפסנו את פאקטור המרחק, שכן נחשיב תכונות אלו כלא מוסיפות מידע רלוונטי כלל אלא מבלבלות את המערכת.

סה"כ לאחר שיפורים הלו, הצלחנו להעלות את רמת הדיוק של הסיווג של אלגוריתם KNN עם שכן אחד באחוז שלם – שזה שיפור של יותר מ 20% ביחס לאלגוריתם המקורי.